Identification of a Novel Signature Predicting Overall Survival in Head and Neck Squamous Cell Carcinoma

Haige Zheng [1], Huixian Liu [1], Yumin Lu [2] and Hengguo Li [1]*

[1] Medical Imaging Center, The First Affiliated Hospital of Jinan University, Guangzhou, China, [2] The First Affiliated Hospital of Guangxi University of Chinese Medicine, Nanning, China

Background: Head and neck squamous cell carcinoma (HNSCC) is a highly heterogeneous tumor with a high incidence and poor prognosis. Therefore, effective predictive models are needed to evaluate patient outcomes and optimize treatment.

Methods: Robust Rank Aggregation (RRA) method was used to identify highly robust differentially-expressed genes (DEGs) between HNSCC and normal tissue in 9 GEO and TCGA datasets. Univariate Cox regression analysis and Lasso Cox regression analysis were performed to identify DEGs related to the Overall survival (OS) and to construct a prognostic gene signature (HNSCCSig). External validation was performed using GSE65858 dataset. Moreover, comprehensive bioinformatics analyses were used to identify the association between HNSCCSig and tumor immune environment.

Results: A total of 257 reliable DEGs were identified by differentially analysis result of TCGA and GSE65858 datasets. The HNSCCSig including 7 mRNAs (SLURP1, SCARA5, CLDN10, MYH11, CXCL13, HLF, and ITGA3) were developed and validated to identify high-risk group who had a worse OS than low-risk group in TCGA and GSE65858 datasets. Cox regression analysis showed that the HNSCCSig could independently predict OS in both the TCGA and the GSE65858 datasets. Further research demonstrated that the infiltration bundance of CD8 + T cells, B cells, neutrophils, and NK cells were significantly lower in the high-risk group. A nomogram was also constructed by combining the HNSCCSig and clinical characters.

Conclusion: We established and validated the HNSCCSig consisting of SLURP1, SCARA5, CLDN10, MYH11, CXCL13, HLF, and ITGA3. A nomogram combining HNSCCSig and some clinical parameters was constructed to identify high-risk HNSCC-patients with poor prognosis.

Keywords: robust rank aggregation method, signature, nomogram, overall survival (OS), head and neck squamous cell carcinoma (HNSCC)

INTRODUCTION

Head and neck squamous cell carcinoma (HNSCC) ranks sixth among cancer-related deaths worldwide, and includes tumors originating from the mouth, oropharynx, nasopharynx, hypopharynx, larynx, and neck (1). Head and neck cancer is one of the main tumors that threatens human health, with over 600,000 new patients and over 350,000 deaths worldwide every year.

Despite the substantial improvement in surgery, radiation therapy and chemotherapy in the past three decades, early diagnosed patients may achieve good results with surgery or radiation therapy, but survival rates for patients with advanced cancer is only 34.9%, and the median survival time of individuals has been reported to be 6 to 9 months (2, 3). Thus, there is an urgent need to identify effective HNSCC predictive biomarkers for an accurate and effective evaluation of a patient's disease status to improve prognosis and reduce mortality.

As we know, the immune system is an important factor in tumorigenesis. As an emerging effective anti-tumor therapy approach, immunotherapy represented mainly by the programmed cell death 1 (PD-1)/programmed cell death ligand 1 (PD-L1) pathway has shown great therapeutic potential in many tumor types (4, 5). Thus, it is important to study the tumor-immunity correlation and provide additional treatment options for HNSCC patients.

With the development of genome-sequencing technology, there is growing evidence that prognostic gene markers can predict head and neck cancer overall survival (OS). For example, Shen et al. (2017) have identified a 7-gene signature associated with the survival of patients with head and neck carcinoma (6). Liu et al. (2018) selected 5-lncRNA and constructed a prognostic score model for the prediction of OS (7). She et al. (2020) predicted the prognosis of HNSCC using immune-related genes (8). However, the small sample size of previous studies and the differences in sequencing platforms may make the research results unreliable. The Robust Rank Aggregation (RRA) method we use can perfectly avoid this issue. It directly integrates the differentially expressed gene lists analyzed by different data sets and identifies a more robust differentially-expressed gene set.

In this study, we aimed to: (I) identify highly robust differentially-expressed genes (DEGs) between HNSCC and normal tissue using RRA method; (II) identify DEGs associated with OS of HNSCC by Cox regression analysis; (III) construct and validate a prognostic gene signature (HNSCCSig) to identify patients with poor OS; and (IV) analyse association between HNSCCSig and tumor immune environment (TME).

MATERIALS AND METHODS

Source of Data

The 10 gene microarray datasets of HNSCC were derived from the GEO database. Nine gene microarray data sets (GSE6631, GSE13601, GSE30784, GSE31056, GSE33205, GSE37991, GSE51985, GSE59102, and GSE138206) were used for DEGs analysis. GSE65858 with 270 tumor tissues and corresponding clinical information were used to verify the HNSCCSig. We directly downloaded the standardized matrix data from the GEO platform and match the probe to the genetic symbol using the manufacturer's annotation file. When a single gene symbol corresponded to several probes, the method of median sequence value was adopted.

HNSCC patient's normalized HTSeq-counts and HTSeq-fpkm data along with corresponding clinical data were obtained from TCGA databases, which contained data from 502 cancer samples and 44 normal samples. The ensemble ID was converted to a gene symbol through the annotation file. HTSeq-counts data were used to screen for DEGs. Finally, the HTSeq-fpkm data were applied to establish a prognostic model, after removing 10 samples with survival time less than one month, Ultimately, 491 samples were incorporated into the modeling analysis.

Comprehensive Analysis of Data Sets and DEGs Screening

Acquisition of DEGs between HNSCC and adjacent normal tissue were performed using the nine datasets from GEO using "limma" packages from R software (version 3.6.2) (9). The edgeR package was used for DEG analysis of the TCGA queue. The RRA method was used to analyze the DEGs identified from the nine GEO datasets using R package "RobustRankAggreg". All genes with $|\text{LogFC}| > 1.0$, false discovery rate (FDR) < 0.05 and $p < 0.05$ were selected as DEGs. More reliable HNSCC-specific DEGs were obtained by the intersection of nine GEO and TCGA dataset results. The "sva" package was run to eliminate or reduce batch effects between different data sets.

Functional Analysis of Genes

GO and KEGG enrichment analyses were applied to identify potential biological processes (BPs), cellular components (CCs), molecular functions (MFs), as well as KEGG pathway terms. Important signal pathways related to DEGs were identified using the "clusterprofiler" R package.

Identification of the Survival-Related Prognostic Model

Univariate, Lasso Cox regression analyses were conducted to obtain prognostic genes significantly correlated with OS in 491 HNSCC patients with survival time >30 days (10). Firstly, Univariate analysis was used to identify genes associated with prognosis. Secondly, Lasso Cox regression analysis was performed to further select candidate genes significantly related to hnscc and the 10-fold cross-validation was used to determine the optimal lambda value, thereby limiting the error to a minimum of 1 standard error. Ultimately, a seven-gene HNSCCSig was established based on LASSO coefficients (β) derived from LASSO model multiplied by the mRNA expression levels.

Using the median value as the cut-off, 491 HNSCC patients were separated into high- and low-risk groups. Kaplan–Meier analysis and receiver operating characteristic curve (ROC) curve analyses were performed to evaluate the accuracy and sensitivity of the model using the R package "survival" and "survivalROC", respectively (11). The concordance index(C-index) was also used to measure discrimination between the predicted value and the real value of the Cox model.

GSE65858 Cohort for External Validation

The GSE65858 dataset from the GEO database was used for external validation. Risk scores for each included patient were calculated using the same prognostic gene signature. Next, the predictive capability of HNSCCSig was also tested based on the Kaplan–Meier curve, the ROC curve, and the C-index.

Independence of Prognostic Genes

Only 378 patients with complete clinicopathological information including age, sex, survival time, survival status, grade, American Joint Committee on Cancer (AJCC) staging, T staging, N staging and M staging were included in our subsequent analysis. Univariate and multivariate analyses were performed to test whether the prognostic risk score could be independent of other clinical variables (such as age, gender, AJCC-stage, T-stage, N-stage, M-stage and grade).

Establish and Verify a Prognostic Nomogram

A nomogram can be combination with multiple indicators to diagnose or predict disease onset or progression. We constructed the nomogram using T-stage, N-stage, M-stage and risk score to predict the 1-, 3-, and 5-year OS rate of head and neck cancer. Nomogram is generally verified by two indicators, namely, discrimination and calibration. Thus, the nomogram's calibration curves were drawn to compare the actual survival rate and the predicted survival rate, with the ordinate indicating the actual survival rate, and the abscissa indicating the predicted survival rate. While, the C-index was used to determine the discrimination of the nomogram using a bootstrap approach, repeated 1000 times. Kaplan–Meier analysis, Decision Curve Analysis (DCA) curve were also applied to assess the predictability of the prognostic nomogram. DCA was used to predict clinical outcome variables and was performed to quantify the clinical utility of the nomogram and to determine its clinical usefulness (12). The nomogram was used to calculate the total score and patients were separated into two groups. The Kaplan-Meier method was used to plot the survival curves for different risk assessment groups.

Enrichment and Tumor Immunity Analyses

To clarify the potential pathobiological mechanism underlying the gene signature, GSEA analysis was conducted to identify rich GO terms and KEGG pathways between high-risk and low-risk
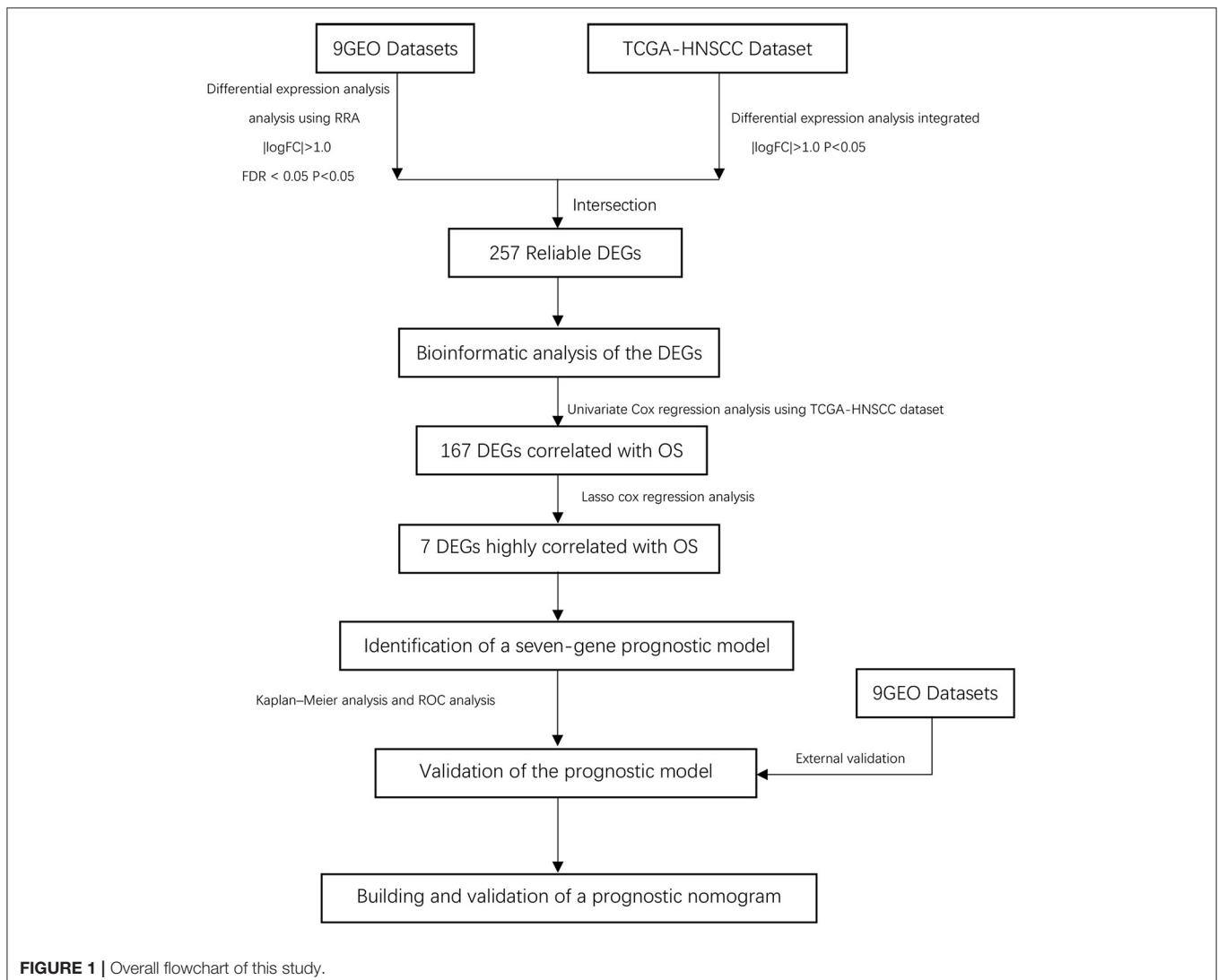
FIGURE 1 | Overall flowchart of this study.

TABLE 1 | Details of the GEO datasets included in this study.

Datasets	Platform	Sample size (tumor/control)	Application
GSE6631	[HG_U95Av2] Affymetrix Human Genome U95 Version 2 Array	44 (22/22)	Identification of DEGs
GSE13601	[HG_U95Av2] Affymetrix Human Genome U95 Version 2 Array	58 (31/27)	Identification of DEGs
GSE30784	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	212 (167/45)	Identification of DEGs
GSE31056	[HG-U133_Plus_2]Affymetrix Gene Chip HumanGenomeHG-U133Plus2 Array	47 (23/24)	Identification of DEGs
GSE33205	[HuEx-1_0-st]Affymetrix Human Exon 1.0 ST Array [transcript (gene) version]	69 (44/25)	Identification of DEGs
GSE37991	Illumina HumanRef-8 v3.0 expression bead chip	80 (40/40)	Identification of DEGs
GSE51985	Illumina HumanHT-12 V4.0 expression bead chip	20 (10/10)	Identification of DEGs
GSE59102	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	42 (29/13)	Identification of DEGs
GSE138206	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	18 (12/6)	Identification of DEGs
GSE65858	Illumina HumanHT-12 V4.0 expression bead chip	270	Validation of prognostic gene

groups, including the KEGG pathway in C2, GO term in C5, and oncogenic signatures of gene sets in C6.

The stromal, immunity, and estimate scores of each head and neck cancer sample were calculated using the ESTIMATE algorithm. Single-sample Gene set enrichment analysis (ssGSEA) approaches can be used to quantify tumor-infiltrating immune cells using the R package "GSVA" (13). The infiltration abundance of 29 immune cells (including B cells, CD4+ T cells, CD8+ T cells, macrophages, neutrophils, dendritic cells, etc.) in each head and neck cancer sample were calculated. Finally, the correlation between stroma, immunity, estimated score and risk score, and the difference in immune cell infiltration between the high-risk and low-risk groups was analyzed.

Statistical Analysis

R software (version 3.6.2; http://www.Rproject.org) and GraphPad Prism (v. 8.0) were used for statistical analysis and to prepare figures. The unpaired t-test was used to estimate the statistical significance of two groups of normally distributed variables. Kaplan-Meier curve analysis was used for survival analysis, the logarithmic rank test was conducted to assess the survival risk in high and low risk groups. Unless otherwise specified, $p < 0.05$ was considered statistically significant.

RESULTS

Screening for Differentially Expressed Genes

Our study follows the flowchart shown in **Figure 1**. The details of the 10 GEO data sets are shown in **Table 1**. From the GSE6631, GSE13601, GSE30784, GSE33205, GSE37991, GSE51985, GSE59102 and GSE138206 data sets, 145, 1140,

1582, 2082, 420, 1848, 620, 2758, and 901 differential genes were selected, respectively. After a comprehensive study of nine GSE datasets based on the RRA means, 193 DEGs were identified, including 69 upregulated and 124 downregulated genes. The top 20 upregulated and downregulated DEGs found in the comprehensive analysis were shown in the volcano diagram (**Figure 2A**). For the TCGA-HNSCC dataset, 3033 DEGs (1359 up-regulated and 1673 down-regulated) were selected, the representative heat map of DEG showed that DEG could effectively distinguish between tumor and normal tissue (**Figure 2C**). Finally, after the intersection of GEO and TCGA results, 257 reliable DEGs were identified (**Figure 2B**).

Functional Analysis of Genes

GO enriched sets ($p < 0.0005$) were analyzed relative to genes involved in serine hydrolase activity, cytokine activity, muscle structure component, serine type peptidase activity, serine-type endopeptidase activity, collagen binding, CXCR chemokine receptor binding, receptor-ligand activity, actin-binding, heparin-binding, extracellular matrix structural constituent (**Figure 3A**). The KEGG pathway analysis of the differential genes mainly involved 3 pathways ($p < 0.05$): tyrosine, retinol metabolism, and drug metabolism cytochromes (**Figure 3B**).

Establishment of a Prognostic Model

A total of 491 TCGA patients with a survival greater than one month were incorporated into this model analysis. The basic clinical parameters of the included patients are shown in **Table 2**. Univariate analysis identified 166 genes that were significantly associated with OS. Through Lasso Cox regression analysis, then, seven genes were selected to build a predictive HNSCCSig.
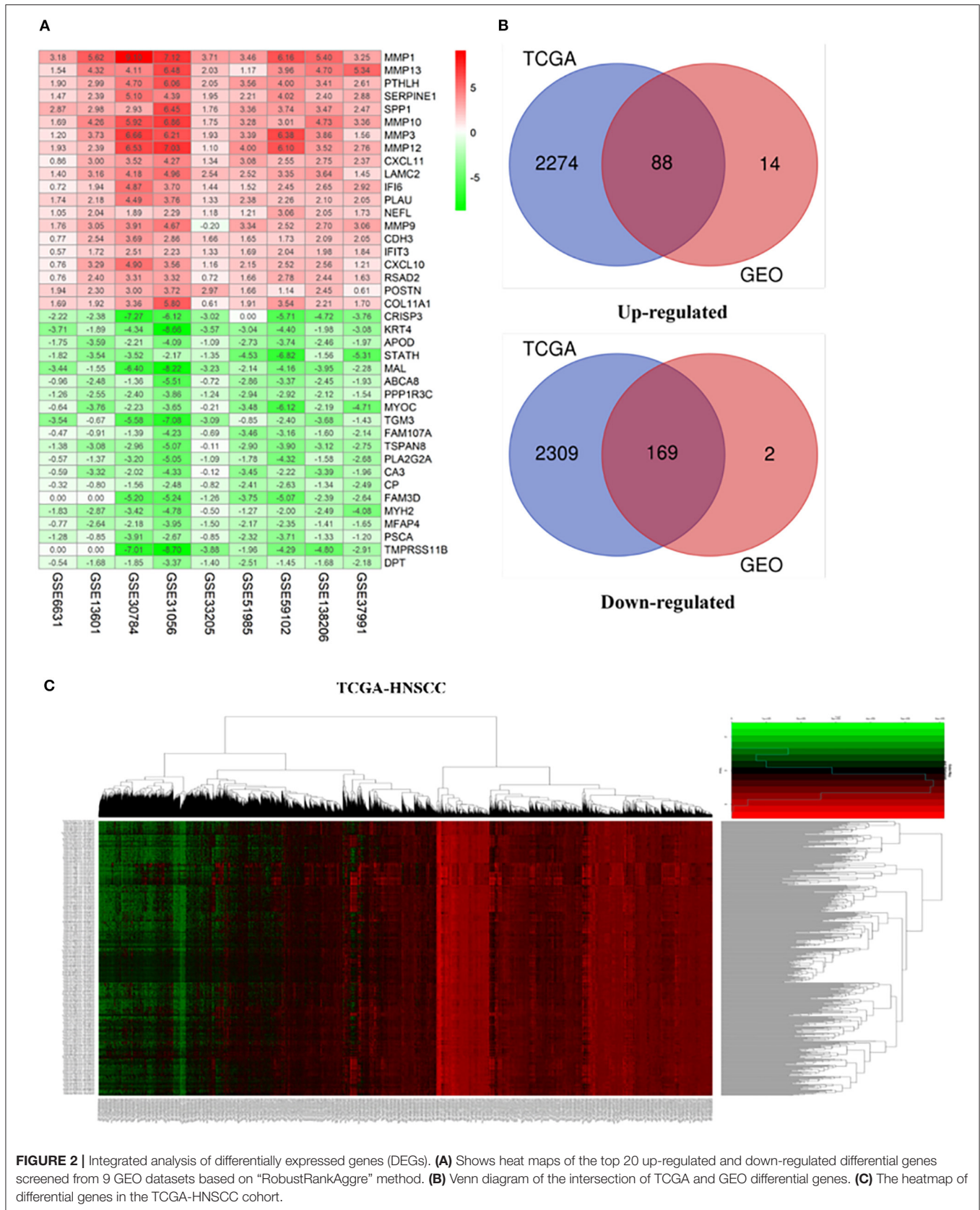
FIGURE 2 | Integrated analysis of differentially expressed genes (DEGs). **(A)** Shows heat maps of the top 20 up-regulated and down-regulated differential genes screened from 9 GEO datasets based on "RobustRankAggre" method. **(B)** Venn diagram of the intersection of TCGA and GEO differential genes. **(C)** The heatmap of differential genes in the TCGA-HNSCC cohort.

FIGURE 3 | Results of GO function **(A)** and KEGG pathway enrichment analyses **(B)**.
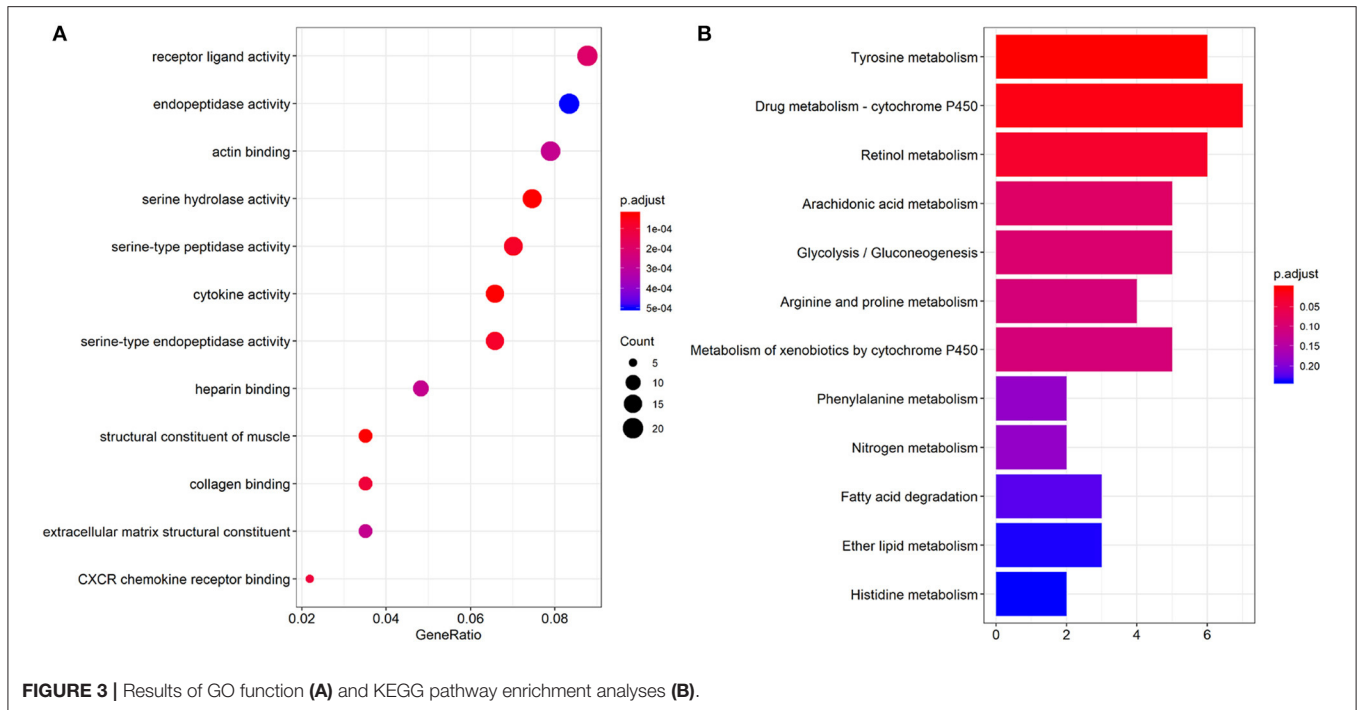
The formula for the risk score is as follows: risk model = $(-0.00173 * \text{SLURP1}) + (-0.21582 * \text{SCARA5}) + (-0.0458 * \text{CLDN10}) + (0.030114 * \text{MYH11}) + (-0.00725 * \text{CXCL13}) + (-0.0839 * \text{HLF}) + (0.005721 * \text{ITGA3})$. Next, Kaplan–Meier analysis disclosed a significantly better prognosis in the low-risk group ($P < 0.0001$) (**Figure 4A**). In addition, the ROC and C-indexes also evaluated the good predictive power of the HNSCCSig. The time-dependent AUCs of risk scores were 0.666 at 3-year and 0.739 at 5-year OS (**Figure 4C**). The C-index of the risk score was 0.623 (95% CI, 0.538–0.708).

External Validation Set and Performance

The GSE65858 dataset with 267 HNSCC patients was selected to validate the HNSCCSig. We calculated risk scores with the same risk formula and divided the patients in the GSE65858 dataset into high-risk and low-risk groups. The model established in this study was also meaningful in the validation set, the results show that significant differences in OS between high- and low-risk groups ($p < 0.0001$) (**Figure 4B**). The predicted AUC for the 3- and 5-year OS were 0.706 and 0.618, respectively, indicating good diagnostic performance (**Figure 4D**). The C-index of the gene signature was 0.641(95% CI, 0.528 – 0.754).

Subgroup Analysis on the Seven-Gene Signature

We performed subgroup analysis on the TCGA data set to further verify the validity of HNSCCSig and the results revealed that the model also had certain diagnostic performance in different subgroups (**Figures 5A–D**). They were stage T1-2 patients ($p <$

0.001), T3-4 patients ($p < 0.001$), N0-1 patients ($p < 0.001$), N2-3 patients ($p = 0.004$), G1-2 patients ($p < 0.001$), G3-4 patients ($p < 0.001$), and stage III-IV ($p < 0.001$).

Independent Predictability of Prognostic Models

Univariate and multivariate Cox regression were performed to assess the independent predictive capacity of the seven genes HNSCCSig based on 378 HNSCC patients with available clinicopathological parameters. The p-value for Stage, T-stage, N-stage, M-stage and the riskscore was <0.05 in univariate Cox regression analysis (**Figure 6A**). Next, multivariate Cox regression analysis demonstrated that N-stage and the riskscore can independently predict OS for head and neck cancer (**Figure 6B**). The same approach was used for the validation set, and results also indicated that the risk score was an independent prognostic factor (**Figures 6C,D**).

Establishment and Verification of the Nomogram

We constructed a nomogram to predict 1-, 3-, and 5-year OS of HNSCC patients based on the TCGA-HNSCC cohort using prognostic factors including T-stage, N-stage, M-stage and risk score (**Figure 7A**). The C-index of the nomogram was 0.728 (95% CI, 0.568–0.734). The calibration curve intuitively demonstrated that the predicted value of the nomogram was close to the true value and had reliable prediction performance (**Figure 7B**). The group with a lower risk score had a significantly better prognosis (**Figure 7C**). The DCA results further suggested that the nomogram was more clinically useful than the gene-based risk model or staging system alone (**Figure 7D**).

TABLE 2 | Patients' information of this study from the TCGA and GEO databases.

Characteristic		TCGA training datasets (n = 491)	GSE65858 (n = 270)
Age (years)	<=50	87	47
	>50	404	223
Survival Status	Living	302	176
	Dead	189	94
Gender	Female	130	47
	male	361	223
Grade	G1	60	\
	G2	293	\
	G3	117	\
	G4	2	\
	GX	19	\
PathologicT	T1	44	35
	T2	129	80
	T3	96	58
	T4	166	97
	Unknow	56	\
PathologicN	N0	167	94
	N1	65	32
	N2	159	132
	N3	7	12
	NX/unknow	93	\
PathologicM	M0	366	263
	M1	12	7
	MX	113	\
Tumor stage	Stage I	25	18
	Stage II	69	37
	Stage III	78	37
	Stage IV	251	178
	Unknow	68	\

Gene Set Enrichment Analysis

To investigate the possible underlying pathobiological mechanisms of the prognostic genes, Gene Set Enrichment Analysis (GSEA) was used to analyze important enrichment pathways in different risk groups of TCGA. A total of 491 TCGA-HNSCC patients were enrolled in the GSEA comparison high-risk and low-risk cohort. The results indicated that 1 KEGG pathway, 2 GO terms, and 3 oncological signatures were enriched in the high-risk group. The identified enriched KEGG pathway and GO terminology mainly involved the proteasome pathway, proteasome accessory and endopeptidase signal transduction pathway, and three oncological signatures included early serum response (CSR), glioma-associated oncogene (GLI1) and B cell-specific Moloney murine leukemia virus integration site (BMI1) (**Figures 8A–C**).
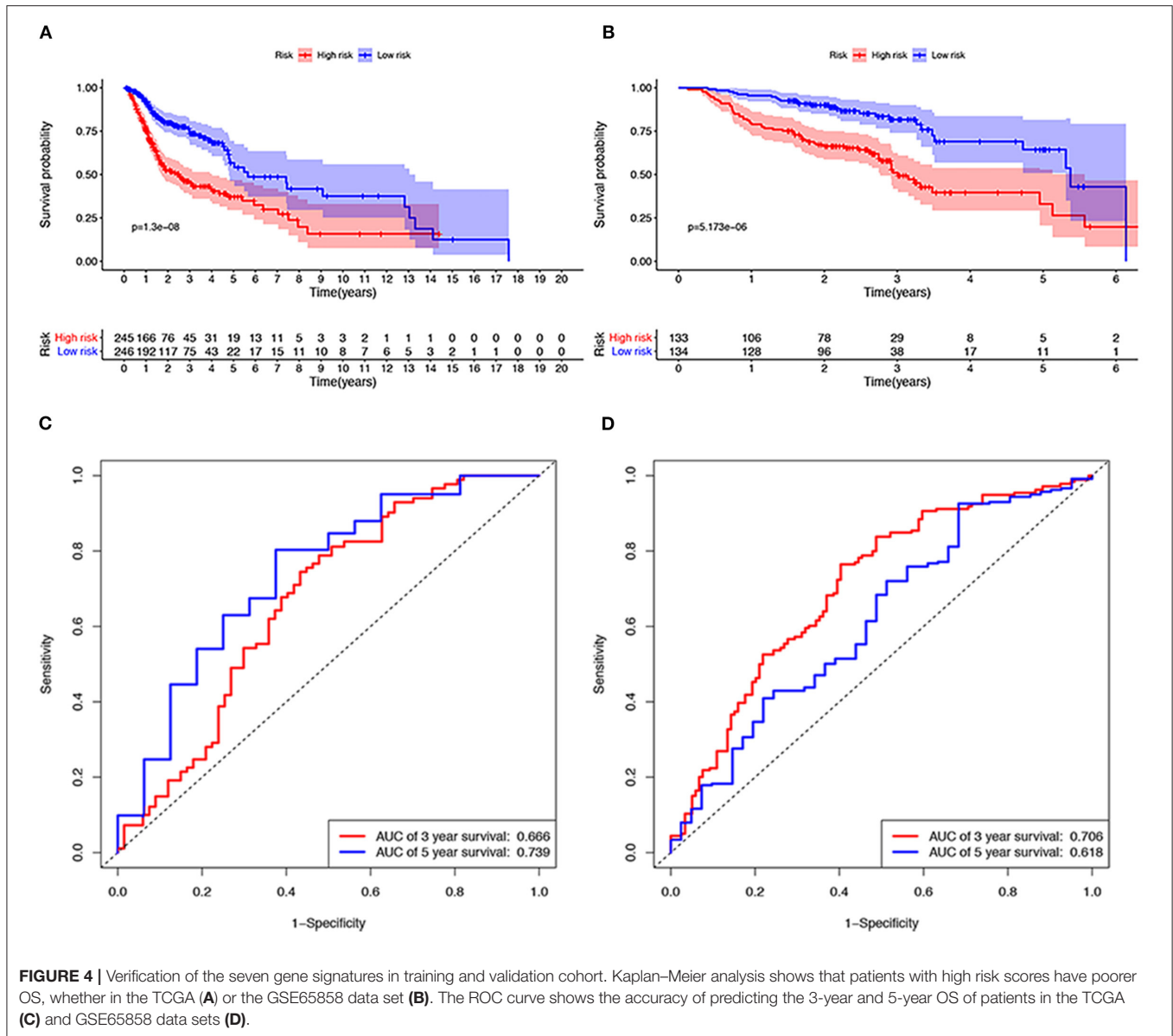
Tumor Immune Mechanism of Prognostic Genes

Expression data obtained from the TCGA HNSCC dataset were used to calculate the matrix, immunity, and estimated scores using an estimation algorithm. The results revealed that the immunity scores and estimated scores of the high-risk group were significantly lower, indicating that there was less infiltration of immune cells in the cancerous tissue ($p < 0.05$; **Figure 8D**). The difference in the matrix score between the different risk groups was not statistically significant ($p > 0.05$). To further study the relationship between the seven prognostic genes and immune infiltration in head and neck carcinoma, we used the ssGSEA method to assess the correlation between these genes and the level of immune cell infiltration, including CD4+ T, CD8+ T, B cell, dendritic cells, neutrophils, macrophages and other immune cells. The results suggested that there was a relatively higher level of CD8+ T cells, B cells, neutrophils, and NK cell infiltration in the low-risk group (**Figure 8E**).

DISCUSSION

HNSCC is a highly heterogeneous tumor with a very poor prognosis and its incidence is increasing every year. The five-year survival rate is only 40–50% (2). With the expanding research in the field of gene sequencing technology, gene expression profiling has attracted increasing attention, and has been used to identify

FIGURE 4 | Verification of the seven gene signatures in training and validation cohort. Kaplan–Meier analysis shows that patients with high risk scores have poorer OS, whether in the TCGA (**A**) or the GSE65858 data set (**B**). The ROC curve shows the accuracy of predicting the 3-year and 5-year OS of patients in the TCGA (**C**) and GSE65858 data sets (**D**).

prognostic markers associated with the heterogeneity of tumors. Therefore, screening for prognostic molecular markers that fully reflect the tumor's biological characteristics may provide clinicians with novel tools to treat HNSCC patients.

In this study, we screened DEGs in head and neck cancer tissues and adjacent non-tumor tissues, and performed Univariate, Lasso Cox analysis to establish HNSCCSig. We identified seven DEGs: SLURP1, SCARA5, CLDN10, MYH11, CXCL13, HLF, and ITGA3, and among them MYH11 and ITGA3 have defined risk factors and the remainder as protective factors. Five of the seven gene signatures have previously been linked to head and neck cancer.

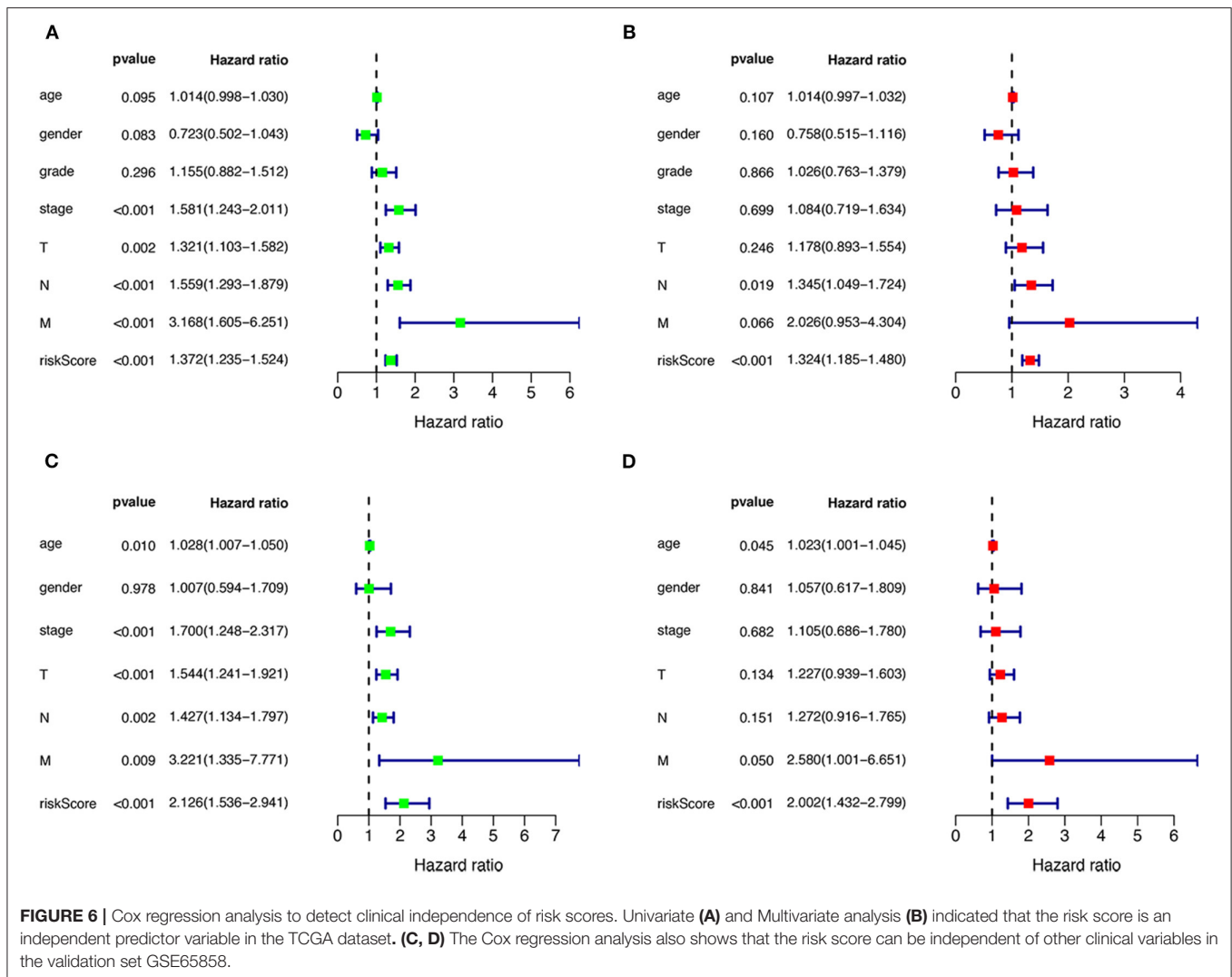SCARA5, a member of the scavenger receptor Type A family, can bind lipopolysaccharide, bacteria, and nucleotides to charge residues in the cysteine domain. In addition, the SCARA5 receptor can recognize and engulf pathogens, and then transmit intracellular signals to generate an immune response. When the number of receptors is reduced, the cell is unable to mount an immune-induced defense. The reduction of SCARA5 expression can promote invasiveness and proliferation of oral tumor cells, and the expression of SCARA5 was found to be down-regulated in oral tumors (14). SCARA5 expression is also usually downregulated in hepatocellular carcinoma due to high promoter methylation and allele imbalance (15).

MYH11, a smooth muscle myosin, may be involved in cell migration and adhesion, intracellular transport, and signal transduction, and as a contractile protein, it converts the chemical energy hydrolyzed by adenosine triphosphate (ATP)
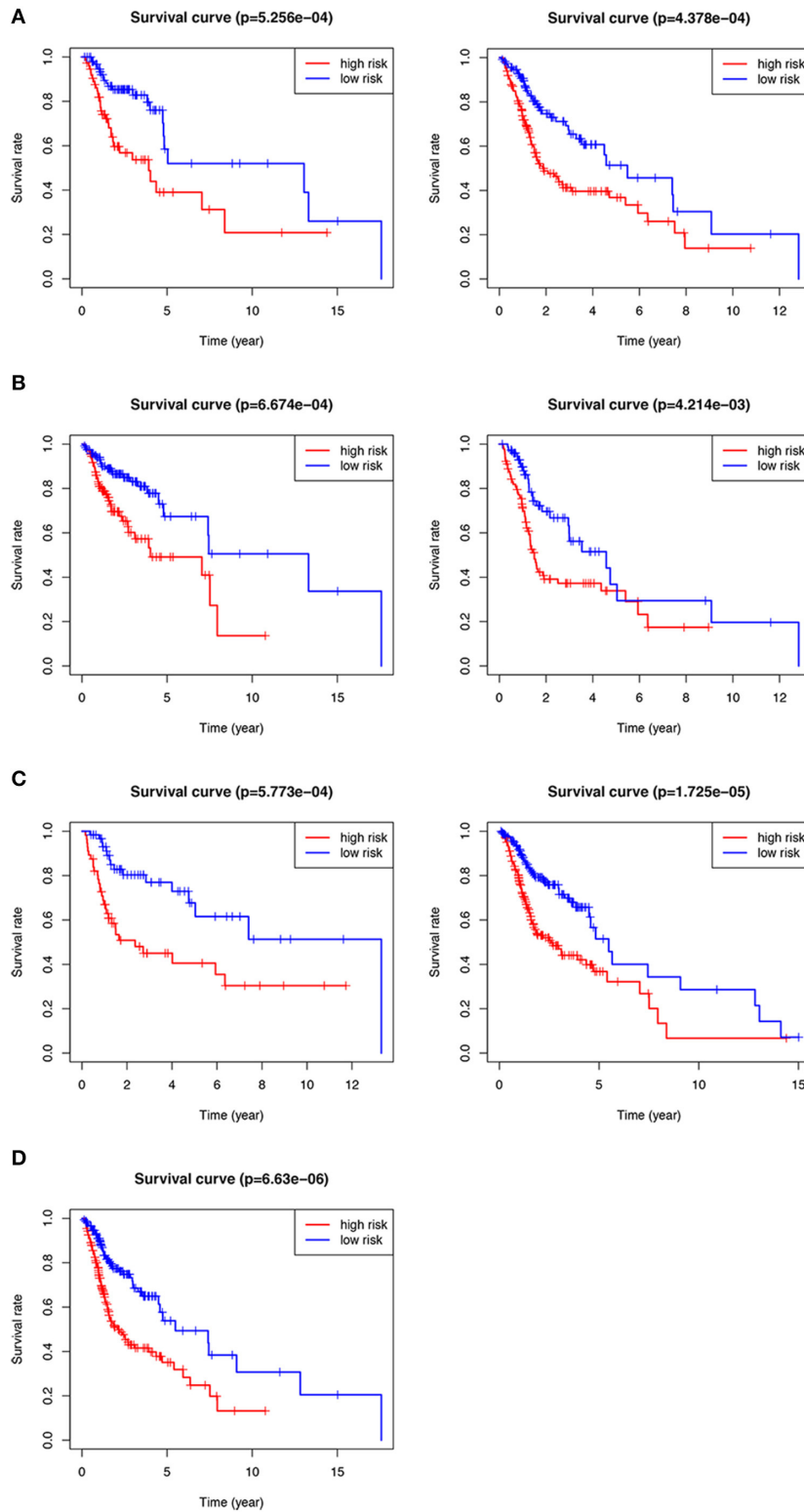
FIGURE 5 | Kaplan–Meier analysis of different subgroups in the TCGA test set. **(A)** T1-2 & T3-4 patients. **(B)** N0-1 & N2-3 patients. **(C)** G1-2 & G3-4 patients. **(D)** Stage III-IV patients.
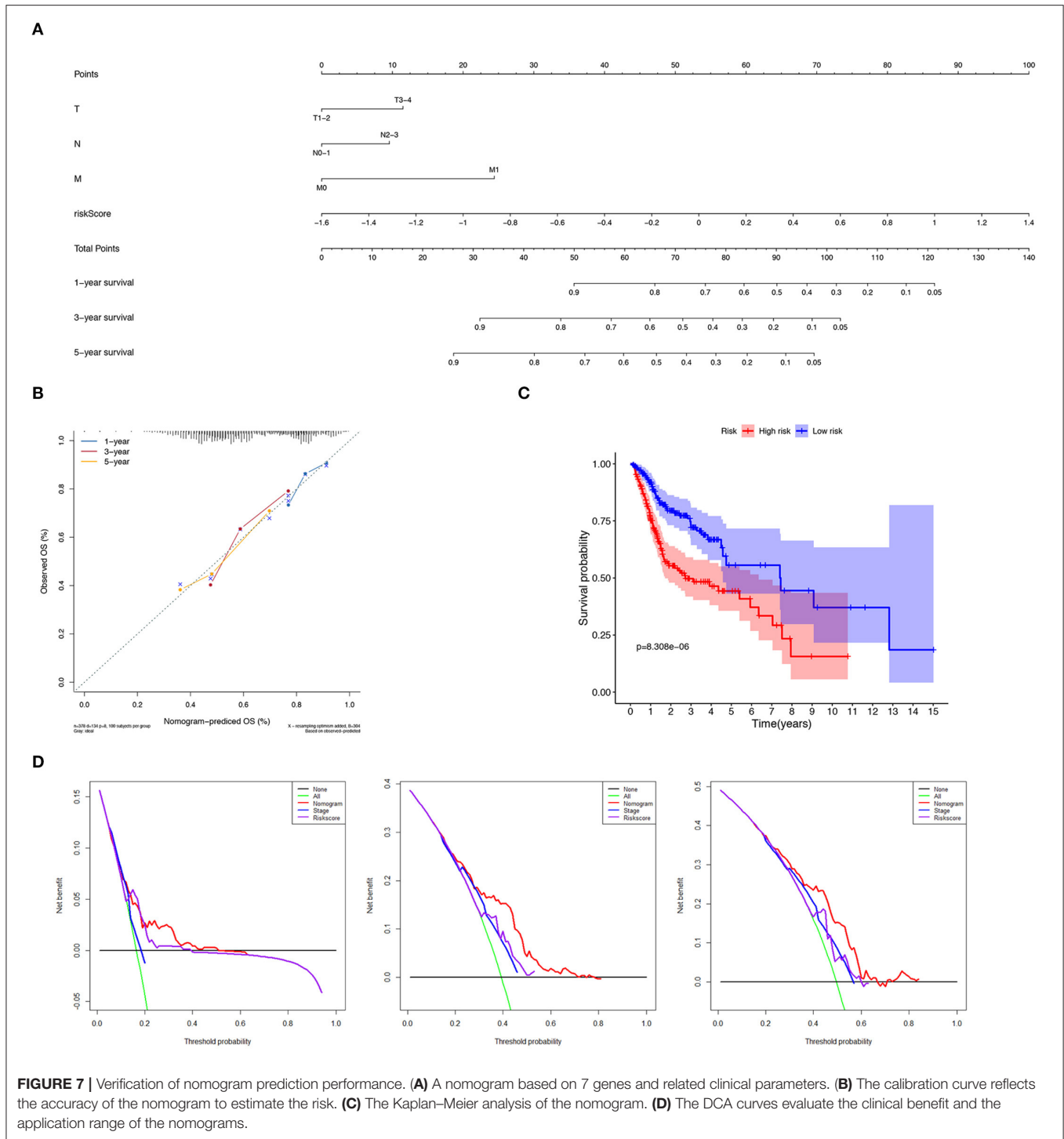
FIGURE 6 | Cox regression analysis to detect clinical independence of risk scores. Univariate **(A)** and Multivariate analysis **(B)** indicated that the risk score is an independent predictor variable in the TCGA dataset. **(C, D)** The Cox regression analysis also shows that the risk score can be independent of other clinical variables in the validation set GSE65858.

into mechanical energy (16). MYH11 is closely related to the survival of HNSCC, acute myeloid leukemia, colorectal cancer, bladder cancer, and other tumors (17–19).

CXCL13 is an independent and cloned B lymphocyte chemokine named Angie, which is an antimicrobial peptide. CXC chemokines are highly expressed in spleen follicles, lymph nodes, and Peyer's patches. It promotes B-cell migration by stimulating chemotaxis into cells expressing Burkitt lymphoma receptor 1(Blr-1) and calcium influx (20). Previous studies have shown that CXCL13 is associated with the prognosis of various cancers. For example, oral squamous cell carcinoma, breast cancer, and prostate carcinoma (21–24).

Hlf-encoded proteins are a member of the proline and acid-rich (PAR) protein family, which activate transcription by forming homotypic or heterotypic dimers with other PAR family members and binding specific promoter elements. It has been reported that HLF is closely related to the prognosis in Oral Cancer, liver cancer, gastric cancer, and lung cancer among others (25–28).

The ITGA3-encoded protein belongs to the integrin family. Integrin is an isomeric membrane protein composed of chains and chains, which acts as a cell surface adhesion molecule. the down-regulation of ITGA3 reduces the phosphorylation of AKT, extracellular signal-regulated kinases 1 and 2 (ERK1/2), and focal adhesion kinase (FAK) in SAS cells and significantly inhibited migration of cancer cells and invasion of HNSCC cells. High expression of ITGA3 predicted poor survival in patients with HNSCC (29). Moreover, several studies have confirmed that ITGA3 is a marker of glioblastoma, pancreatic cancer, and thyroid cancer (30–32).

The role of SLURP1 and CLDN10 in head and neck cancer has not been described. The protein encoded by SLURP1, a member of the Ly6/uPAR family, has anti-tumor activity (33). SLURP1 is related to the occurrence and development of pancreatic cancer. First, it can not only reduce invasion of cancer cells by controlling AKT, ERK, and NF-kB signaling, but also attenuates nicotine-mediated migration and invasion possibly through competing binding sites (34). Furthermore, mutations

FIGURE 7 | Verification of nomogram prediction performance. **(A)** A nomogram based on 7 genes and related clinical parameters. **(B)** The calibration curve reflects the accuracy of the nomogram to estimate the risk. **(C)** The Kaplan–Meier analysis of the nomogram. **(D)** The DCA curves evaluate the clinical benefit and the application range of the nomograms.

in SLURP1 can increase the incidence of melanoma and mucosal skin cancer (35).

Studies have shown that CLDN10 is highly expressed in thyroid papillary carcinoma, and can affect cell proliferation, migration, and invasion *in vitro*; further, it plays the role of a tumor promoter, and the up-regulation of CLDN10 is related to lymph node metastasis (36). In contrast, low expression of CLDN10 indicated a poor prognosis in lung cancer patients. The expression of CLDN10 was negatively correlated with the expression of c-fos. c-fos is considered a recognized oncogene, CLDN10 may control the invasion and metastasis of lung cancer cells by inhibiting the c-fos pathway (37). In our study, the down-regulation of CLDN10 was associated with poor survival rates for HNSCC, which was consistent with the latter view. The
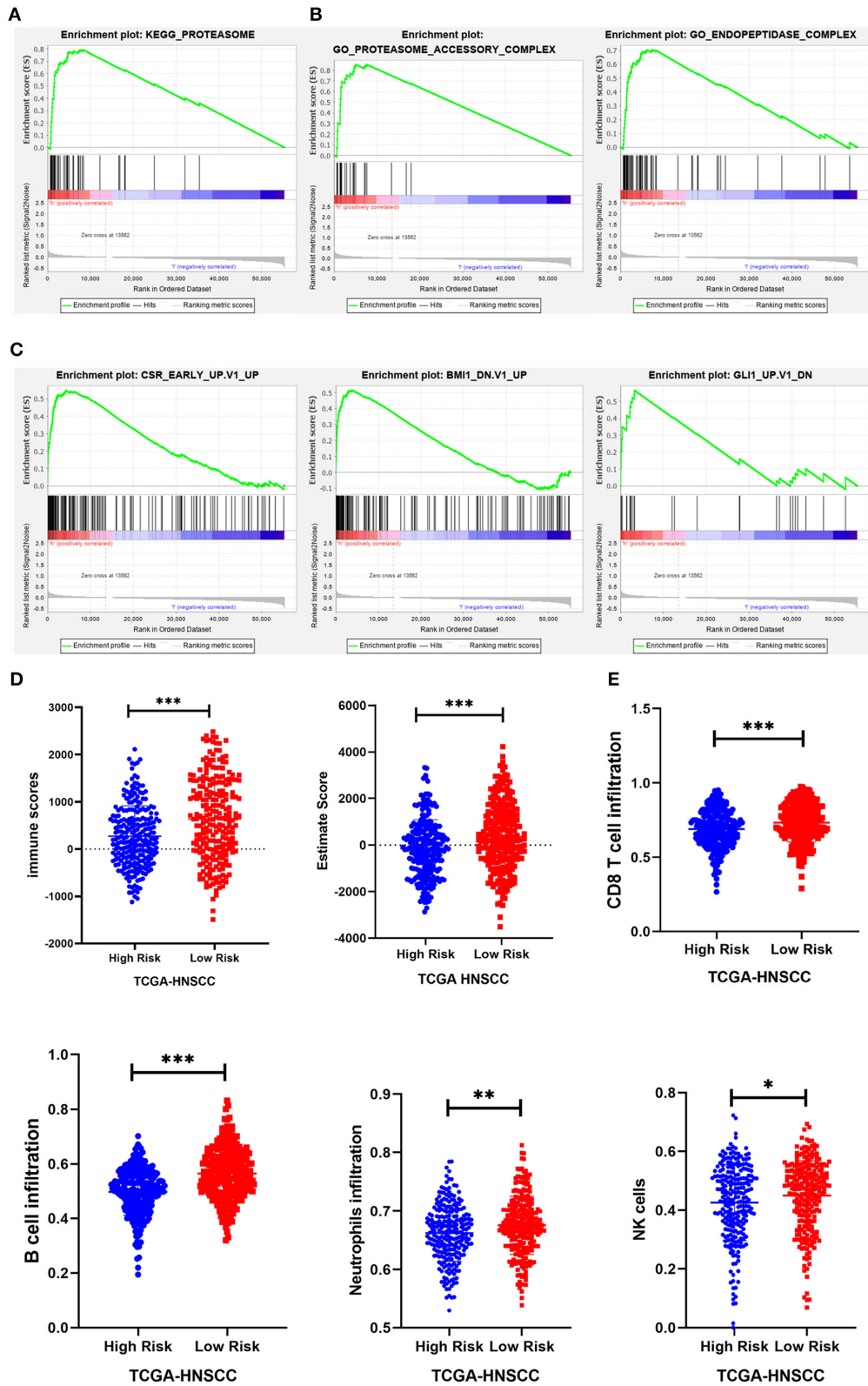
FIGURE 8 | GSEA and immunity correlation analysis of seven prognostic genes. (A–C) The results showed that 1 KEGG pathway, 2 GO terms, and 3 oncological signatures were enriched in the high-risk group. (D) The immune score and estimated score of the high-risk group were significantly lower in the TCGA cohort. (E) The infiltration level of immune cells (including CD8+ T cells, B cells, neutrophils, and NK cells) in the low-risk group was significantly higher than that in the high-risk group in the TCGA cohort. *$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$ respectively.

mechanism of SLURP1 and CLDN10 in head and neck tumors deserves further study.

Many studies have shown that the immune system acts to control tumor growth and progression, and the prognosis of the tumor is related to lymphocyte infiltration. As an emerging anti-tumor force, immunotherapy has shown great therapeutic potential in many cancers, HNSCC is no exception. In addition, HNSCC patients with highly infiltrated CD8 T cells have a better prognosis, especially in patients who are HPV positive [38]. Our results also suggest that the infiltration levels of CD8+ T cells, B cells, neutrophils, and NK cells are lower in the high-risk group.

Our research has many advantages. Firstly, the number of samples was more than in previously published studies, we integrated 9 GEO data sets and TCGA data sets based on the RRA method, which obtain more reliable differential genes. Secondly, we use the Lasso analysis method to establish HNSCCSig to prevent the model from overfitting. HNSCCSig has good predictive accuracy and is verified by an external cohort. Furthermore, C-index, Calibration curves, Kaplan–Meier analysis, DCA curve were applied to assess the predictability of the prognostic nomogram. Finally, we also analyzed the relevance between prognostic gene expression and the immune microenvironment. Nonetheless, the present study also includes certain limitations. First, this study only included clinical parameters such as age, gender, grade, AJCC-staging, T staging, N staging and M staging, and because the TCGA database lacks other host-related factors and complete histopathological variables. Second, the results obtained by bioinformatics analysis alone are not sufficient and need to be confirmed by experimental verification. Therefore, further identification of prognostic biomarkers requires experimental studies with larger samples and experimental validation.

CONCLUSION

In conclusion, we established and validated the HNSCCSig consisting of SLURP1, SCARA5, CLDN10, MYH11, CXCL13, HLF, and ITGA3. Our results suggested that the seven genetic markers were closely related to the prognosis and progression of HNSCC, both in the training set and the validation set. A nomogram combining HNSCCSig and some clinical parameters was construct to identify high-risk HNSCC patients with poor prognosis. Notably, our research shows that the high-risk group has a low degree of immune cell infiltration, which may provide new insights into the research of hnscc's immune microenvironment.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

HZ: study design and the drafting of the manuscript. HZ and HLiu: data acquisition, analysis, and manuscript modification. HZ, YL, and HLi: the research oversight and manuscript reviews. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin.* (2020) 70:7–30. doi: 10.3322/caac.21590

2. Leemans CR, Braakhuis BJ, Brakenhoff RH. The molecular biology of head and neck cancer. *Nat Rev Cancer.* (2011) 11:9–22. doi: 10.1038/nrc2982

3. Gildener-Leapman N, Ferris RL, Bauman JE. Promising systemic immunotherapies in head and neck squamous cell carcinoma. *Oral Oncol.* (2013) 49:1089–96. doi: 10.1016/j.oraloncology.2013.09.009

4. Ohaegbulam KC, Assal A, Lazar-Molnar E, Yao Y, Zang X. Human cancer immunotherapy with antibodies to the PD-1 and PD-L1 pathway. *Trends Mol Med.* (2015) 21:24–33. doi: 10.1016/j.molmed.2014.10.009

5. Lin W, Chen M, Hong L, Zhao H, Chen Q. Crosstalk between PD-1/PD-L1 blockade and its combinatorial therapies in tumor immune microenvironment: a focus on HNSCC. *Front Oncol.* (2018) 8:532. doi: 10.3389/fonc.2018.00532

6. Shen S, Bai J, Wei Y, Wang G, Li Q, Zhang R, et al. A seven-gene prognostic signature for rapid determination of head and neck squamous cell carcinoma survival. *Oncol Rep.* (2017) 38:3403–11. doi: 10.3892/or.2017.6057

7. Liu G, Zheng J, Zhuang L, Lv Y, Zhu G, Pi L, et al. A prognostic 5-lncRNA expression signature for head and neck squamous cell carcinoma. *Sci Rep.* (2018) 8:15250. doi: 10.1038/s41598-018-33642-1

8. She Y, Kong X, Ge Y, Yin P, Liu Z, Chen J, et al. Immune-related gene signature for predicting the prognosis of head and neck squamous cell carcinoma. *Cancer Cell Int.* (2020) 20:22. doi: 10.1186/s12935-020-1104-7

9. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* (2015) 43:e47. doi: 10.1093/nar/gkv007

10. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med.* (1997) 16:385–95. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

11. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics.* (2000) 56:337–44. doi: 10.1111/j.0006-341X.2000.00337.x

12. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* (2006) 26:565–74. doi: 10.1177/0272989X06295361

13. Li Y, Lu Z, Che Y, Wang J, Sun S, Huang J, et al. Immune signature profiling identified predictive and prognostic factors for esophageal squamous cell carcinoma. *Oncoimmunology.* (2017) 6:e1356147. doi: 10.1080/2162402X.2017.1356147

14. Liu HL, Hu JW, Wei R, Zhou LF, Pan H, Zhu HC, et al. SPAG5 promotes hepatocellular carcinoma progression by downregulating SCARA5 through modifying beta-catenin degradation. *J Exper Clin Cancer Res.* (2018) 37:1–14. doi: 10.1186/s13046-018-0891-3

15. Huang J, Zheng DL, Qin FS, Cheng N, Chen H, Wan BB, et al. Genetic and epigenetic silencing of SCARA5 may contribute to human hepatocellular carcinoma by activating FAK signaling. *J Clin Invest.* (2010) 120:223–41. doi: 10.1172/JCI38012

16. Islam T, Rahman R, Gov E, Turanli B, Gulfidan G, Haque A, et al. Drug targeting and biomarkers in head and neck cancers: insights from systems biology analyses. *OMICS.* (2018) 22:422–36. doi: 10.1089/omi.2018.0048

17. Faber ZJ, Chen X, Gedman AL, Boggs K, Cheng J, Ma J, et al. The genomic landscape of core-binding factor acute myeloid leukemias. *Nat Genet.* (2016) 48:1551–6. doi: 10.1038/ng.3709

18. Hu J, Zhou L, Song Z, Xiong M, Zhang Y, Yang Y, et al. The identification of new biomarkers for bladder cancer: a study based on TCGA and GEO datasets. *J Cell Physiol.* (2019) 234:15607–18. doi: 10.1002/jcp.28208

19. Zhao B, Baloch Z, Ma Y, Wan Z, Huo Y, Li F, et al. Identification of potential key genes and pathways in early-onset colorectal cancer through bioinformatics analysis. *Cancer Control.* (2019) 26:1073274819831260. doi: 10.1177/1073274819831260

20. Liu XJ, Asokan SB, Bear JE, Haugh JM. Quantitative analysis of B-lymphocyte migration directed by CXCL13. *Integrative Biology.* (2016) 8:894–903. doi: 10.1039/c6ib00128a

21. El-Haibi CP, Singh R, Sharma PK, Singh S, Lillard JW. CXCL13 mediates prostate cancer cell proliferation through JNK signalling and invasion through ERK activation. *Cell Proliferation.* (2011) 44:311–9. doi: 10.1111/j.1365-2184.2011.00757.x

22. Sambandam Y, Sundaram K, Liu A, Kirkwood KL, Ries WL, Reddy SV. CXCL13 activation of c-Myc induces RANK ligand expression in stromal/preosteoblast cells in the oral squamous cell carcinoma tumor-bone microenvironment. *Oncogene.* (2013) 32:97–105. doi: 10.1038/onc.2012.24

23. Chen L, Huang Z, Yao G, Lyu X, Li J, Hu X, et al. The expression of CXCL13 and its relation to unfavorable clinical characteristics in young breast cancer. *J Transl Med.* (2015) 13:168. doi: 10.1186/s12967-015-0521-1

24. Barstad B, Tveitnes D, Dalen I, Noraas S, Ask IS, Bosse FJ, et al. The B-lymphocyte chemokine CXCL13 in the cerebrospinal fluid of children with Lyme neuroborreliosis: associations with clinical and laboratory variables. *Infect Dis.* (2019) 51:856–63. doi: 10.1080/23744235.2019.1671988

25. Hou JY, Wang YG, Ma SJ, Yang BY, Li QP. Identification of a prognostic 5-Gene expression signature for gastric cancer. *J Cancer Res Clin Oncol.* (2017) 143:619–29. doi: 10.1007/s00432-016-2324-z

26. Ge Y, Li W, Ni Q, He Y, Chu JJ, Wei PM. Weighted gene co-expression network analysis identifies hub genes associated with occurrence and prognosis of oral squamous cell carcinoma. *Med Sci Monitor.* (2019) 25:7272–88. doi: 10.12659/MSM.916025

27. He R, Zuo S. A robust 8-gene prognostic signature for early-stage non-small cell lung cancer. *Front Oncol.* (2019) 9:693. doi: 10.3389/fonc.2019.00693

28. Xiang DM, Sun W, Zhou T, Zhang C, Cheng Z, Li SC, et al. Oncofetal HLF transactivates c-Jun to promote hepatocellular carcinoma development and sorafenib resistance. *Gut.* (2019) 68:1858–71. doi: 10.1136/gutjnl-2018-317440

29. Koshizuka K, Hanazawa T, Kikkawa N, Arai T, Okato A, Kurozumi A, et al. Regulation of ITGA3 by the anti-tumor miR-199 family inhibits cancer cell migration and invasion in head and neck cancer. *Cancer Sci.* (2017) 68:1681–92. doi: 10.1111/cas.13298

30. Jiao Y, Li YQ, Liu SY, Chen QM, Liu YH. ITGA3 serves as a diagnostic and prognostic biomarker for pancreatic cancer. *Onco Targets Ther.* (2019) 12:4141–52. doi: 10.2147/OTT.S201675

31. Wang Z, Gao L, Guo X, Feng C, Lian W, Deng K, et al. Development and validation of a nomogram with an autophagy-related gene signature for predicting survival in patients with glioblastoma. *Aging.* (2019) 11:12246–69. doi: 10.18632/aging.102566

32. Zhai TY, Muhanhali D, Jia X, Wu ZY, Cai ZQ, Ling Y. Identification of gene co-expression modules and hub genes associated with lymph node metastasis of papillary thyroid cancer. *Endocrine.* (2019) 66:573–84. doi: 10.1007/s12020-019-02021-9

33. Campbell G, Swamynathan S, Tiwari A, Swamynathan SK. The secreted Ly-6/uPAR related protein-1 (SLURP1) stabilizes epithelial cell junctions and suppresses TNF-alpha-induced cytokine production. *Biochem Biophys Res Commun.* (2019) 574:729–34. doi: 10.1016/j.bbrc.2019.07.123

34. Throm VM, Mannle D, Giese T, Bauer AS, Gaida MM, Kopitz J, et al. Endogenous CHRNA7-ligand SLURP1 as a potential tumor suppressor and anti-nicotinic factor in pancreatic cancer. *Oncotarget.* (2018) 9:11734–51. doi: 10.18632/oncotarget.24312

35. Radiono S, Pramono ZAD, Oh GGK, Surana U, Widiyani S, Danarti R. Identification of novel homozygous SLURP1 mutation in a Javanese family with Mal de Meleda. *Int J Dermatol.* (2017) 56:1161–8. doi: 10.1111/ijd.13776

36. Zhou Y, Xiang J, Bhandari A, Guan Y, Xia E, Zhou X, et al. CLDN10 is associated with papillary thyroid cancer progression. *J Cancer.* (2018) 9:4712–7. doi: 10.7150/jca.28636

37. Zhang Z, Wang A, Sun B, Zhan Z, Chen K, Wang C. Expression of CLDN1 and CLDN10 in lung adenocarcinoma in situ and invasive lepidic predominant adenocarcinoma. *J Cardiothorac Surg.* (2013) 8:95. doi: 10.1186/1749-8090-8-95

38. de Ruiter EJ, Ooft ML, Devriese LA, Willems SM. The prognostic role of tumor infiltrating T-lymphocytes in squamous cell carcinoma of the head and neck: A systematic review and meta-analysis. *Oncoimmunology.* (2017) 6:e1356148. doi: 10.1080/2162402X.2017.1356148