

## Research Paper

# Identification of a prognostic 4-mRNA signature in laryngeal squamous cell carcinoma

Cheng Zhang<sup>1</sup>, Bin Shen<sup>2</sup>, Xinwei Chen<sup>2</sup>, Shang Gao<sup>2</sup>, Xinjiang Ying<sup>2</sup>, Pin Dong<sup>2</sup>✉

1. Department of Otorhinolaryngology-head and neck surgery, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China.
2. Department of Otorhinolaryngology-head and neck surgery, Shanghai General Hospital, Shanghai, China.

✉ Corresponding author: Pin Dong, Department of otorhinolaryngology-head and neck surgery, Shanghai General Hospital, No.85 Wujin Road, Shanghai 200080, China. E-mail: dongpin64@aliyun.com.

© The author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>). See <http://ivyspring.com/terms> for full terms and conditions.

Received: 2020.04.29; Accepted: 2021.07.18; Published: 2021.08.03

## Abstract

**Background:** Laryngeal squamous cell carcinoma (LSCC) is one of the most common malignancy in the respiratory tract and could reduce the quality of life seriously like dyspnea, dysphonia and dysphagia. Moreover, 5-year survival rate has decreased over the past 40 years. This study was designed to identify mRNAs that related to prognosis in LSCC to enable early detection and outcome improvement.

**Methods:** Gene expression profiles from Gene Expression Omnibus (GEO) (GSE59102, GSE84957) and The Cancer Genome Atlas (TCGA) were analyzed to identify differentially expressed genes (DEGs) with the help of bioinformatics tools. Functional enrichment analyses including Gene Ontology (GO) and pathway analysis were carried out to investigate the role of those genes and underlying molecular mechanisms in LSCC. Cox's regression analyses (univariate, LASSO and multivariate in order) were utilized to identify DEGs related with patients' overall survival and a 4-mRNA-based prognostic risk score model was established. Univariate and multivariate Cox's regression analyses were then performed on LSCC data (90 patients left) to identify independent predictors of OS, including the signature and clinicopathologic variables. The prognostic value of the gene signature was further validated and the genes were analyzed by GEPIA to get pan-cancer expression profiles.

**Results:** 444 differentially expressed mRNAs (250 up-regulated, 194 down-regulated) were identified based on the threshold of fold change > 2 and adjusted p value < 0.05. Univariate Cox's regression analysis showed that high risk score (HR: 3.056, 95% confidence interval [CI]: 0.135-0.649, p<0.001) and female (HR: 0.296, 95% CI: 2.020-4.624, p=0.002) were associated with relatively poor prognosis. Further multivariate Cox's regression analysis indicated that risk score and gender were independent prognostic factors (p<0.05). The risk score model could stratify patients into high- and low-risk groups, which presents significantly differential overall survival (p= 8.252e-04). The AUCs of 1-, 3- and 5-year OS were 0.724, 0.783 and 0.818, respectively.

**Conclusions:** Our study provides evidence that the four-mRNA signature could serve as a biomarker to predict prognosis in LSCC, especially in long-term.

Key words: Laryngeal squamous cell carcinoma, prognosis, risk model, GEO, TCGA

## Introduction

Laryngeal cancer remains one of the most common tumors of the respiratory tract in which squamous cell cancer is known to be the major pathological type [1]. There are estimated 177,422 new cases diagnosed and approximately 94,771 deaths reported globally in 2018 [2]. Despite advances in treatments, 5-year survival rate has decreased from

66% to 63% over the past 40 years [1]. Although the clinical TNM staging, histopathological grading, history of cigarette smoking and heavy drinking will remain valuable, it is possible to acquire molecular information about host and tumor to optimize the management of laryngeal squamous cell carcinoma (LSCC) [3]. Therefore, it is necessary to stratify

disease-associated risks in patients early and to change treatment strategy, such as modality and intensity in multidisciplinary cancer management, as well as types of medicine administered in different risk groups.

With the increasing application of high-throughput sequencing technology, changes in genomics and transcriptomics have provided essential insights into the molecular-level characteristics of diseases. Many researchers are engaged in exploring potential biomarkers for predicting prognosis early in head and neck squamous cell cancer (HNSCC) patients, one of which is LSCC. Accumulating evidence has shown that mRNA signatures had potential value in predicting prognosis. In the recent studies, a 6-gene prognostic signature (PEX11A, NLRP2, SERPINE1, UPK, CTTN and D2HGDH) was established and area under the curve (AUC) of receiver operating characteristic (ROC) curve in 5-year overall survival (OS) was increased to 0.74 by drawing [4]. Kaplan-Meier (K-M) survival analysis was used to verify that a signature (IGF1R, LAMC2, ITGB1, and IL-6) has an excellent association with poor survival contributed to radioresistance [5]. In laryngeal cancer patients, the prognostic accuracy of the 5-gene signature (EMP1, HOXB9, DPY19L2P1, MMP1 and KLHDC7B) for OS at 5 years was 0.862 [6]. However, more research focused on the extensive category and subsite-specific transcriptional markers for prognosis combining multiple databases are rarely systematically compared and identified. Besides, cost-effectiveness yet to be taken into consideration in clinical practice.

Therefore, in this study we aimed to build a robust prognostic signature with minimal combinations of genes based on Gene Expression Omnibus (GEO) [7] and The Cancer Genome Atlas (TCGA) [8], which might be used to develop a fast detection kit. Subsequently, we identified four potential prognostic mRNAs and confirmed the integrated 4-mRNA signature as a novel prognostic biomarker that might help effectively predict overall survival of LSCC patients.

## Material and Methods

### Data acquisition and preprocessing

Samples selected for study based on the following specific criteria: diagnosed as LSCC; primary, untreated tumor with a source of matched normal tissue; informed consent from patients to donate part of their tumor samples. mRNA expression profile series matrix files of LSCC were retrieved from the GEO. Normalized mRNA expression values, in terms of level-3 fragments per kilobase of transcripts

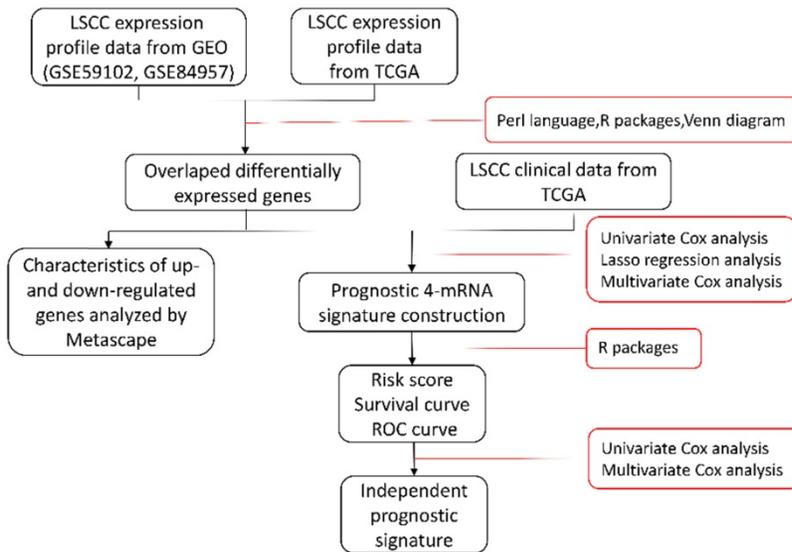
per million mapped reads (FPKM) and clinical meta data files (overall survival times and vital status) of patients with LSCC were downloaded from TCGA datasets via the Genomic Data Commons (GDC) Data Portal of National Cancer Institute and were used for subsequent survival-related analysis (details are listed in Table 1). Data were collected from September 2, 2019 to December 5, 2019. The flow chart of the entire study is displayed in Figure 1.

**Table 1.** Clinicopathological characteristics of LSCC patients in TCGA mRNA (n=111)

Categories	n (%)
<b>Age at diagnosis</b>	
< 65	70 (63.06%)
≥ 65	41 (36.94%)
<b>Gender</b>	
Male	91 (81.98%)
Female	20 (18.02%)
<b>Race</b>	
Asian	1 (0.90%)
Black	19 (17.12%)
White	86 (77.48%)
Others	5 (4.50%)
<b>Staging system edition</b>	
5th	4 (3.60%)
6th	20 (18.02%)
7th	87 (78.38%)
<b>Clinical stage</b>	
I	3 (2.70%)
II	11 (9.91%)
III	26 (23.42%)
IV	67 (60.36%)
NA	4 (3.60%)
<b>Pathological stage</b>	
I	2 (1.80%)
II	9 (8.11%)
III	14 (12.61%)
IV	71 (63.96%)
NA	15 (13.51%)
<b>Treatment type</b>	
Pharmaceutical therapy	57 (51.35%)
Radiation therapy	54 (48.65%)
<b>Status</b>	
Alive	61 (54.95%)
Dead	50 (45.05%)

Abbreviations: LSCC: laryngeal squamous cell carcinoma; TCGA: The Cancer Genome Atlas; NA: not available.

Strawberry-perl (version 5.30.1.1) was used to merge data sets of all samples from TCGA into one data set and convert probe IDs into gene symbols. As to TCGA, genome assembly Homo\_sapiens.GRCh38.94.chr.gtf, which was downloaded from Ensembl, was used as a reference to map annotation. For GEO data, Agilent annotation files were utilized to perform the identifier conversion. Expression values were averaged when multiple probes corresponding to the same gene. Downstream analyses were conducted by using R (version 3.5.1).



**Figure 1.** An overview of the gene model construction, red boxes represent methods used in every step. LSCC: laryngeal squamous cell carcinoma; GEO: Gene Expression Omnibus; TCGA: The Cancer Genome Atlas; ROC: receiver operating characteristic.

## Differential expression analysis

The “limma” [9] package was executed for differentially expressed genes (DEGs) analysis in GEO, and Wilcoxon’s rank sum tests were conducted by R function “wilcox.test”. DEGs were defined as fold change (FC) > 2 and adjusted *p* value (adj. *P. val*) < 0.05, where  $\log_2FC > 1$  was regarded as up-regulation and  $\log_2FC < -1$  as down-regulation. Visualization of the DEGs, in form of heatmap, were achieved by package “pheatmap” [10] of R.

## Grouping of DEGs list

Web tool Venny (version 2.1) [11] was used to overlap up- and down-regulated gene lists with Venn’s diagrams respectively.

## Construction of enriched ontology clusters

Aberrantly changed DEGs were submitted to Metascape [12], which incorporates a core set of default ontologies like Gene Ontology (GO) biological processes, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, Reactome gene sets, Canonical pathways and so on, for biological processes and pathways analysis. Default settings were used, including the number of genes included in enrichment terms  $\geq 3$ , *p* value  $\leq 0.01$  and minimum enrichment factor is 1.5.

## Prognostic mRNA signature construction

Combined expression quantity with clinical data, all DEGs genes associated with the OS were selected by using univariate Cox’s proportional hazards regression. “Risky” genes (Hazard ratio [HR] > 1) and “protective” genes ( $0 < HR < 1$ ) were selected in up-

and down-regulated DEGs respectively. Subsequently, least absolute shrinkage and selection operator (LASSO) Cox’s proportional hazards regression method (package “glmnet”) [13] and “survival” [14], with a 10-fold cross-validation, were employed to acquire genes with proper weights. In order to facilitate the clinical application, the results were analyzed by using stepwise multivariable Cox’s regression further and the best fitting COX model was selected based on the lowest Akaike’s information criterion (AIC) [15].

We thus constructed a linear risk classifier model (formula as follows):

$$\text{Risk score} = \sum_{i=1}^n \beta_i \times \lambda_i,$$

where  $\lambda_i$  is the FPKM expression value and  $\beta_i$  is corresponding regression coefficient of each individual gene.

## Identification of independent prognostic factors for LSCC

The independent predictors of OS (risk\_score model, age, gender, cTNM\_stage and pTNM\_stage, treatment\_type) were identified by univariate and multivariate Cox’s proportional hazards regression.

## Performance of the mRNA signature

R package “pheatmap” [10] was utilized to draw risk related dot plots. Patients were divided into high- and low-risk groups according to the median risk score. The survival status of the two groups can also be displayed by dots with different colors separately. By running “survival” and “timeROC” [16] package in R, we obtained K-M survival curves. In addition, log-rank tests were applied to assess prognostic significance. The prediction performance of the model was presented by time-dependent ROC and evaluated based on the AUC in 1-, 3- and 5-year OS. All statistical tests were two-sided and *p*-values < 0.05 were considered statistically significant.

## Changes in genes expression of 33 cancer types

All genes in the signature were input into GEPIA [17] for differential expression analysis. Four-way analysis of variance (ANOVA) method was used to calculate differential expression between tumor and paired adjacent normal samples from TCGA and Genotype-Tissue Expression (GTEx) samples, where genes with  $|\log_2FC|$  values higher than 1 and *q* values lower than 0.01 are considered differentially expressed genes.

## Results

### Identification of DEGs

Twenty nine cancer samples and thirteen margin samples of LSCC from GSE59102 [18], 9 pairs of samples from GSE84957 [19] and 11 paired RNA-sequencing (RNA-seq) data from TCGA were used to identify differential genes respectively (Table 2). According to the critical value mentioned in differential expression analysis, up- and down-regulated differential genes obtained from each datasets are taken into intersection analyses and drawn by Venn diagrams. The hierarchical-clustering heatmaps suggested differences in expression patterns between the two groups of samples in GSE59102, GSE84957 and TCGA (Figure 2A-2C). A total of 444 DEGs were identified, composed of 250 up-regulated mRNAs and 194 down-regulated mRNAs (Figure 2D-2E).

### Functional enrichment analysis

The modulation-specific biological processes and pathways features of these up- and down-regulated DEGs were then analyzed. Results were visualized by bar graphs and clustering networks as shown in Figure 3-4. Among the top 20 output terms of

up-regulated genes, enrichment is mainly found in extracellular matrix organization, cell cycle, cell adhesion and differentiation, and pathways in cancer, etc. Down-regulated genes are enriched in substance processing, leukocyte activation, metabolic processes, transport of substance, and regulation of signaling.

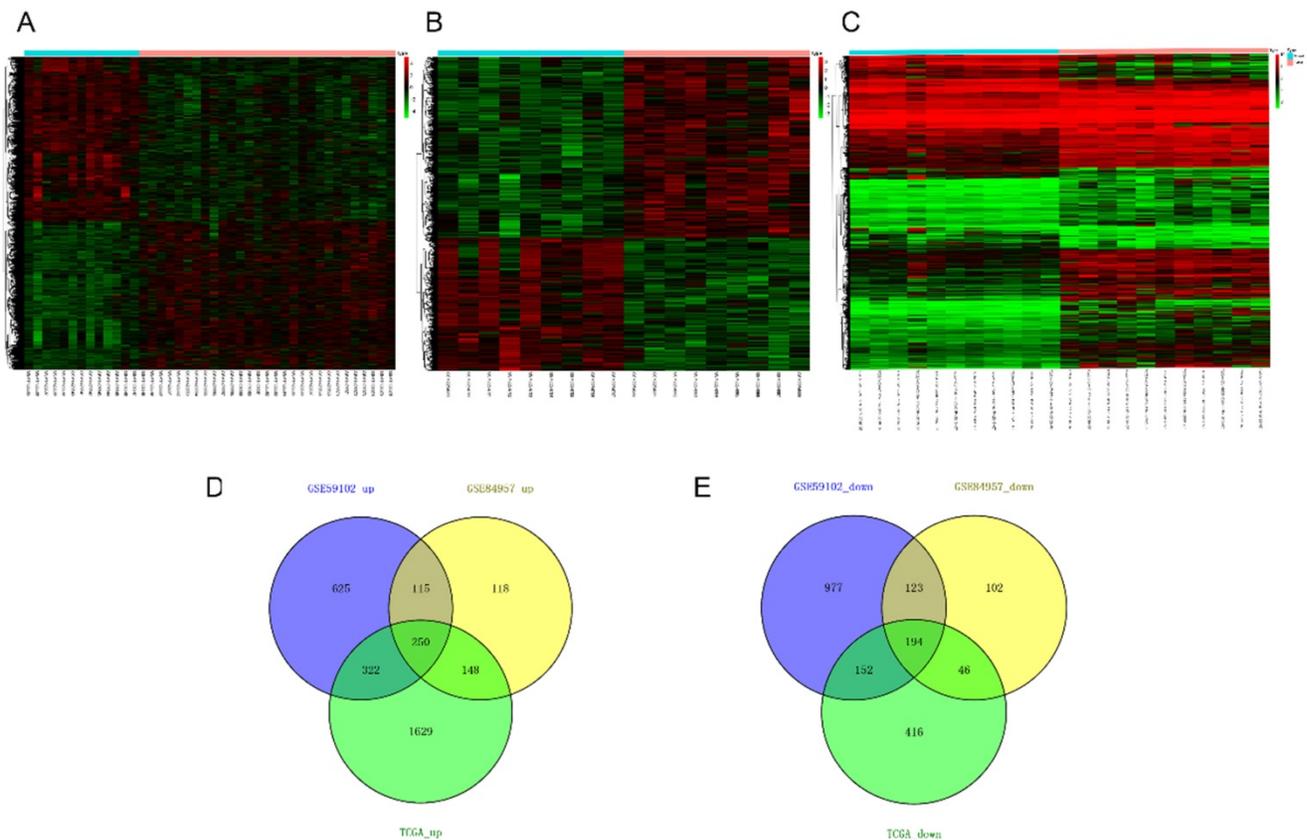
**Table 2.** Overview of each datasets associated with LSCC from GEO and TCGA

Data source	Cases of tumor	Cases of normal	Platform	Scanned items	Clinical files
GSE59102	29	13	GPL6480	mRNA	No
GSE84957	9	9	GPL17843	lncRNA, mRNA	No
TCGA	111	12	NA	mRNA	Yes

Abbreviations: LSCC: laryngeal squamous cell carcinoma; GEO: Gene Expression Omnibus; TCGA: The Cancer Genome Atlas; GSE: GEO Series; GPL: GEO Platform; NA: not available.

### Construction of 4-mRNA prognostic signature

All differential genes were analyzed by univariate Cox’s proportional hazards regression. We obtained 32 OS-related genes in up-regulation gene sets and 7 in down-regulation gene sets (Table S1-2). In view of the classification criteria of “risk” and “protection” genes above, 2 up-regulated genes (CHTF18, SPC24) in tumor samples were excluded



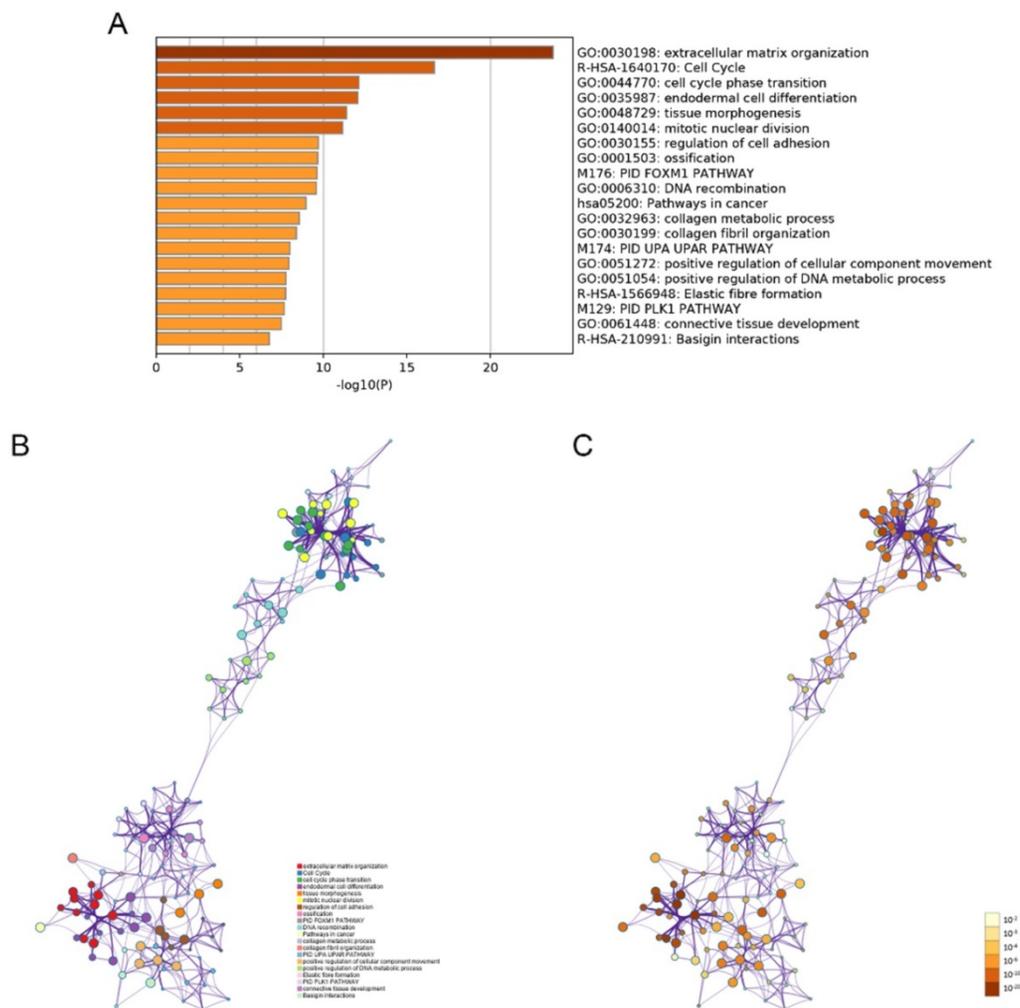
**Figure 2.** Heatmaps and Venn’s diagrams of DEGs. (A-C) The heatmaps of 2758, 1096 and 3157 DEGs in GSE59102, GSE84957 and TCGA respectively (fold change > 2 and adjusted p value < 0.05), the blue bar represents that the tissue type is normal and blue is tumor. (D-E) The Venn diagrams indicate the overlapping of DEGs in three datasets mentioned above, where 250 up- and 194 down-regulated DEGs are showed. DEGs: differential expression genes.

owing to their HR < 1. LASSO Cox's regression was performed on the recognized 37 DEGs. The optimal value (minimum error) of tuning parameter  $\lambda$ , a constant controlling the degree of penalty, is 0.144 (Figure 5). Thus, 6 genes (STC2, ITGA5, AQP9, EPHX2, TCEA3, and MMP1) were selected out and further analyzed by multivariate Cox's proportional hazards regression to determine which one or more candidate genes could exhibit better predictive role. Consequently, we constructed a prognostic signature model based on the expression value of those mRNAs and their regression coefficients. The result is as follows: Risk score = 0.0494 × Expr (STC2) + 0.0866 × Expr (AQP9) + 0.0006 × Expr (MMP1) - 0.0721 × Expr (TCEA3).

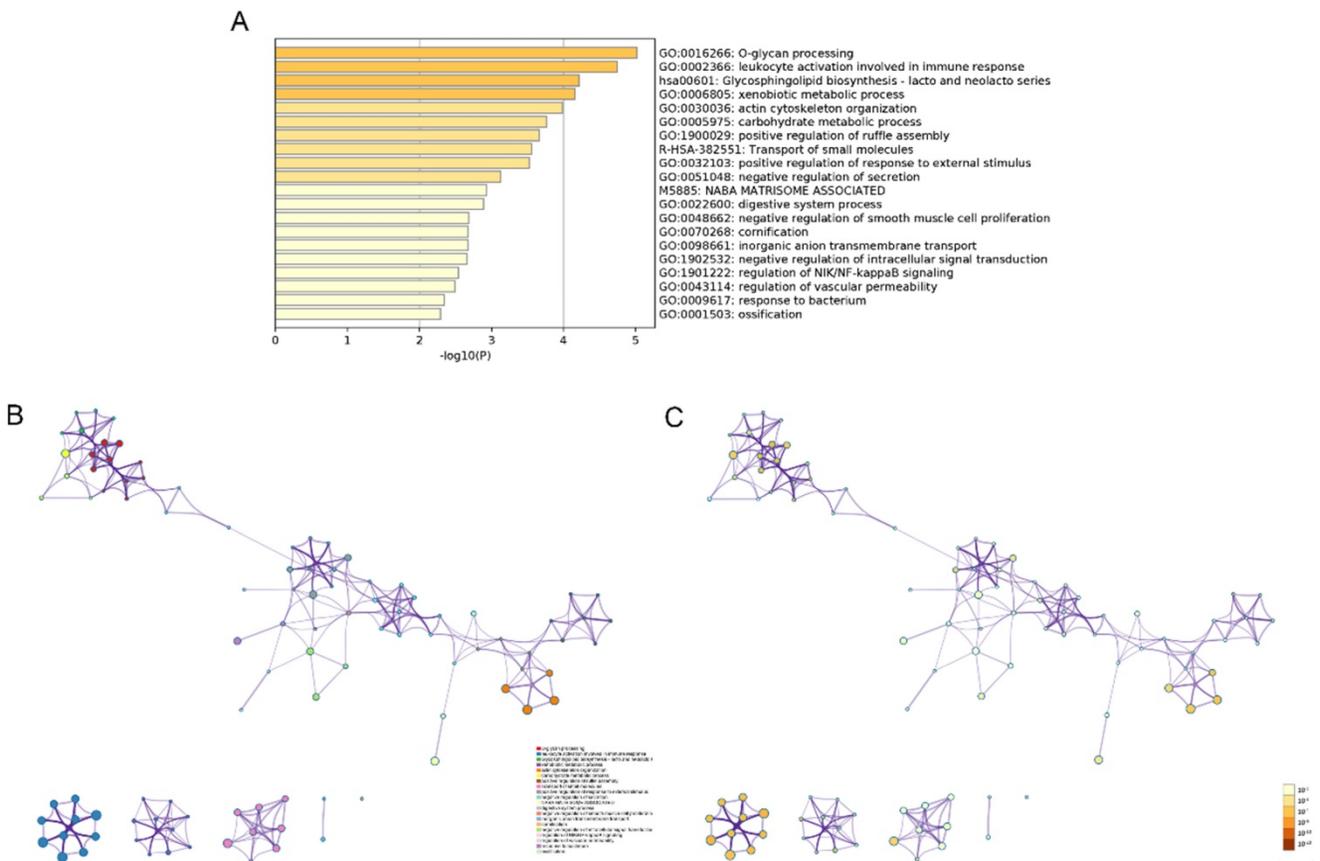
### Univariate and multivariate Cox's regression analyses of the prognostic signature and clinical characteristics predictive of OS

Given that the AJCC cancer staging standards of laryngeal cancer in the 6<sup>th</sup> edition [20] are the same as

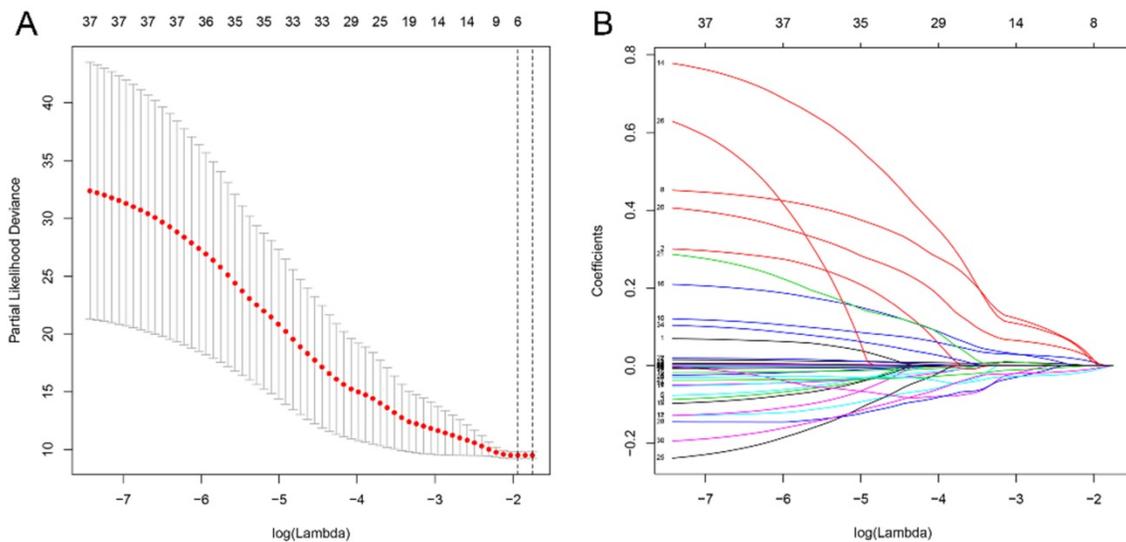
those in the 7<sup>th</sup> edition [21], we excluded 4 cases clinical data of the 5<sup>th</sup> edition [22] and integrated the former two editions. Furthermore, we also excluded 4 cases and 13 cases respectively for which cTNM and pTNM staging information was not available. Univariate and multivariate Cox's regression analyses were then performed on LSCC data (90 patients left) to identify independent predictors of OS, including the signature and clinicopathologic variables. Univariate Cox's regression analysis showed that risk score and gender were significantly associated with the prognosis of LSCC patients (Figure 6A). High risk score was associated with poor prognosis (HR: 3.056, 95% confidence interval [CI]: 0.135-0.649,  $p < 0.001$ ). Male were associated with relatively good prognosis (HR: 0.296, 95% CI: 2.020-4.624,  $p = 0.002$ ). Further multivariate Cox's regression analysis indicated that risk score and gender were independent prognostic factors ( $p < 0.05$ ) (Figure 6B).



**Figure 3.** Functional enrichment analysis of up-regulated DEGs performed by Metascape. (A) Bar graph of top 20 non-redundant enrichment clusters and the color represents statistical significance. (B) Networks of enriched terms linked by edges representing Kappa similarity > 0.3. A Node represents each term and its size is proportional to the number of input genes fall into that term. Terms with the same cluster identity are marked corresponding color. (C) The same enrichment network as Figure B, darkness of the color indicated the p-value.



**Figure 4.** Functional enrichment analysis of down-regulated DEGs carried out by Metascape. (A) Bar graph of top 20 non-redundant enrichment clusters and the color represents statistical significance. (B) Networks of enriched terms linked by edges representing Kappa similarity > 0.3. A Node represents each term and its size is proportional to the number of input genes fall into that term. Terms with the same cluster identity are marked corresponding color. (C) The same enrichment network as Figure B, darkness of the color indicated the p-value.



**Figure 5.** Analysis of 37 prognosis-associated DEGs analyzed by LASSO Cox's regression. (A) The left vertical bar indicates the minimum error, the right shows the largest value of  $\lambda$  such that the error is within one standard deviation of the minimum. (B) Coefficient paths of the 37 DEGs that corresponded by value of  $\lambda$ . LASSO: least absolute shrinkage and selection operator.

### Estimations of the prognostic signature in the TCGA datasets

We employed 111 of cases LSCC patients' clinical files from TCGA RNA-Seq to verify the

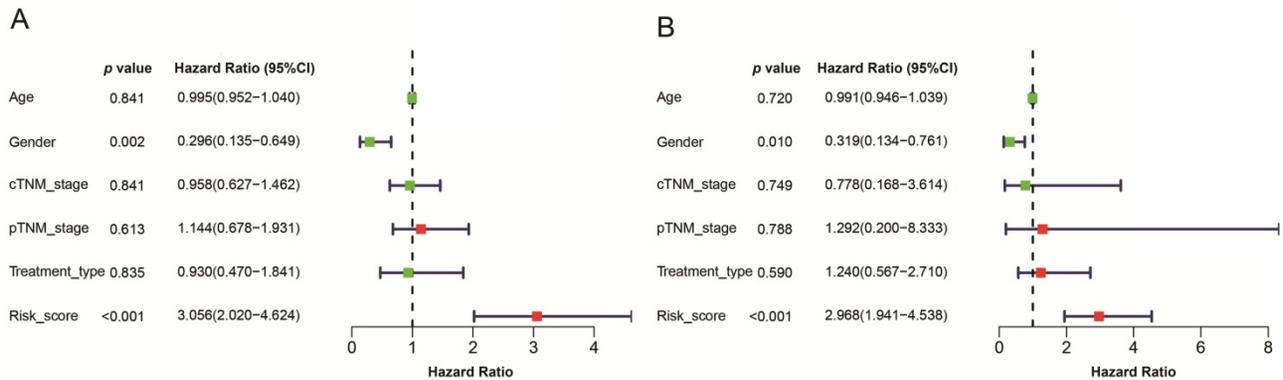
prognostic value of the signature. The median risk score was used as a cutoff value to divide the cases into two types of risk groups, of which 63.6% (35/55) patients were deceased in the high-risk group and 26.8% (15/56) in the low-risk group (Figure 7A). K-M

survival curves indicate that the signature could clearly distinguish different risk levels, where high-risk group had worse OS ( $p= 8.252e-04$ ) (Figure 7B). Time-dependent ROC analysis showed that the AUC values of 1, 3 and 5 year survival were 0.724, 0.783 and 0.818 respectively (Figure 7C).

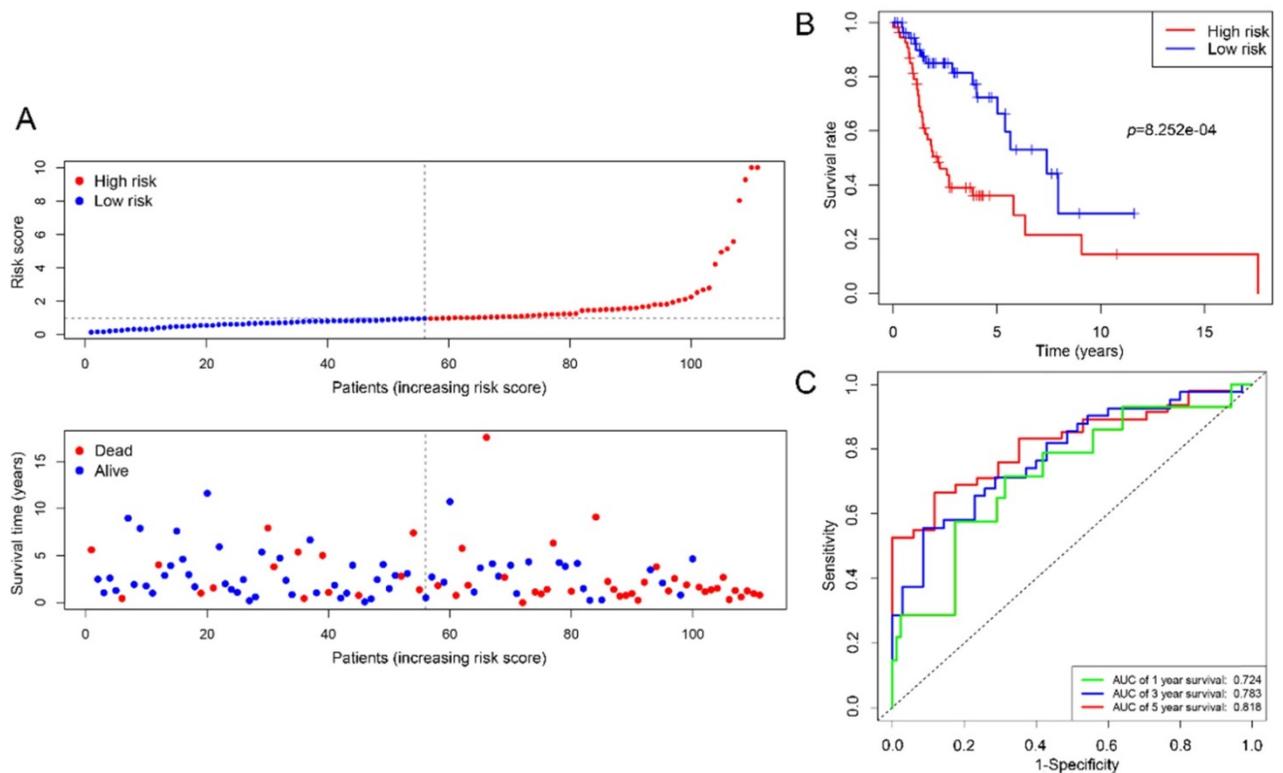
**Pan-cancer expression profile of genes in signature**

MMP1 was over-expressed in all tumors where differential expression is exhibited. For instance, it is highly expressed in breast invasive carcinoma (BRCA), lung neoplasms and HNSC, etc. and lowly expressed in all normal tissues (Figure S1). STC2 was

over-expressed in tumors such as colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), glioblastoma multiforme (GBM) and HNSC, etc. and under-expressed in acute myeloid leukemia (LAML) and skin cutaneous melanoma (SKCM) (Figure S2). AQP9 was observed more down-regulated in all tumors found differential expressed like lung neoplasms, thyroid carcinoma (THYM) and up-regulated in ovarian serous cystadenocarcinoma (OV) and pancreatic adenocarcinoma (PAAD) (Figure S3). As to TCEA3, it was more under-expressed in tumors like HNSC, OV and PAAD, etc. (Figure S4).



**Figure 6.** A model was identified based on univariate (A) and multivariate (B) analysis of independent predictors of OS. Bar lengths are hazard ratios (HR) for variables with 95% confidence intervals (95% CI), red boxes indicate HR>1, and the green boxes are the opposite. p-values < 0.05 were considered statistically significant.



**Figure 7.** Estimations of the 4-mRNA prognostic signature in the TCGA datasets. (A) Vertical dotted line divides patients into low- and high-risk group based on the median risk score. Upper: curve of risk scores of all patients ranked in order of its increasing value. Lower: corresponding survival time with status for each patient. (B) Kaplan-Meier survival analysis of the 4-mRNA signature. (C) AUC values of 1-, 3-, and 5-years overall survival by drawing ROC curves. AUC: area under ROC curve.

## Discussion

Great efforts have been made to develop an optimal tool for predicting prognosis in LSCC patients, but no consensus has been reached. In this integrated analysis, candidate DEGs were selected by univariate and LASSO Cox's regression and were analyzed by multivariable Cox regression further to identify the best prognostic gene signature. Based on the expression value of these DEGs and their regression coefficients, a prognostic risk formula was constructed. The prediction accuracy of the model was analyzed by time-dependent ROC and evaluated by the area under the curve. All AUC values are all greater than 0.7 and show an increasing trend as the survival time increases. High expression of "risky" mRNAs (STC2, AQP9 and MMP1) and low expression of "protective" mRNA (TCEA3) was found to be significantly associated with poor prognosis. Furthermore, K-M analysis confirmed their prognostic role in LSCC.

Taken together, the following are the issues we encountered and then addressed throughout the analysis. First of all, there are 11 tumor-adjacent samples that have their paired LSCC samples in the TCGA database, we thus used the 11 tumor samples instead of all to perform differential analyses. In this way, the inefficiency caused by a significant imbalance in the number of samples between groups is reduced [17]. Secondly, the common significant DEGs were obtained from the intersection analyses of the results from multiple datasets sources, which could reduce systematic and random errors due to sequencing platforms and sampling, therefore, improves the robustness of the results and leads to the precise interpretation of molecular landscape of LSCC. Thirdly, we only selected data for LSCC to make the signature more specific. Fourthly, prognostic signatures based on multiple deregulated mRNAs have gained much attention recently and have shown their potential in prognosis prediction in different kinds of cancer [23, 24]. Whereas, most single-gene biomarker researches focused more on basic experiments; ROC or other methods are rarely used to test their prediction accuracy. Last but not least, the basic concept of LASSO is that a penalty is used to shrink variable weights towards zero, with the result that small weights may get shrunken to zero and thus to prevent overfitting and improve model interpretability. Lambda is a hyperparameter that controls the strength of the penalty [25].

According to gene expression profiling studies, some aberrantly expressed mRNAs, such as HMGA2 [26], FSCN1 [27] and LAMA3 [28], had been reported to be related to clinicopathological features in LSCC

tissues. Considering that the Metascape could reduce the redundancy among ontology terms and monthly updates ensured by adopting a novel two-phase approach, we use the web-based portal to analyze the underlying biological processes and pathways of these DEGs involved in the genesis and development of LSCC.

In the results of the functional enrichment analysis of the 444 DEGs, MMP1 showed strong associations with biological processes related to extracellular matrix (ECM) organization, collagen metabolic process and pathways related to cancer, uPA/uPAR pathway and basigin interactions. Except for MMP1, the other three signature genes were not identified in the following top 20 enrichment terms. Matrix metalloproteinases (MMPs) represent a family of zinc-dependent proteinases, which can degrade ECM components, such as collagens and proteoglycans, and mediate tumor invasion and metastasis. Moreover, many MMPs identified in human are expressed and contribute to HNSCC progression. [29]. Over-expression of MMP-1 has been found in all various cancers with differential expression [30, 31], which might indicate its underlying roles in those tumor phenotypes (Supplementary). Wang et.al found that Astrocyte elevated gene-1 (AEG-1) modulates the phosphorylation at serine 536 of the p65 subunit of NF- $\kappa$ B and enhances p65 binding to the MMP1 promoter, and subsequently increase the expression of the downstream gene in HNSCC [32]. Some reports demonstrated that MMPs can be triggered by plasmin, which converted from plasminogen during urokinase plasminogen activator (uPA) /uPA receptors (uPAR) signaling pathway [33, 34]. In addition, a study showed that co-localization of uPA with MMP-1, -2, -9 was observed in advanced epithelial ovarian primary tumors and metastatic lesions. [35]. Kanekura and colleagues reported that coculture of basigin-expressing human malignant melanoma cells with dermal fibroblasts could induce the production of MMPs including MMP-1, MMP-2, MMP-3, etc. and induce invasion through a reconstituted basement membrane [36]. Perturbation of basigin may have potential therapeutic uses in the prevention of MMP-2 and MMP-1-dependent cancer metastasis [37].

STC2 is a glycoprotein hormone involved in many biological processes, especially calcium and phosphate homeostasis, and it can also regulate the progression of malignant tumors [38]. More recent researches revealed that downregulation of LINC00460 and HOTAIR could decrease STC2 via up-regulating microRNA-206 (miR-206) and promotes autophagy, proliferation, invasion and

migration in HNSCC [39, 40]. Moreover, a study showed that STC2 could upregulate the phosphorylation of AKT and enhance HNSCC metastasis via Snail-mediated increase of vimentin and the decrease of E-cadherin [41], which was proved that STC2 could activate PI3K/AKT signaling pathway by down-regulating miR-206 [40]. Note that STC2 was found over-expressed (Supplementary Figure) and might be directly regulated by HMGA2 at the transcriptional level in high-grade serous ovarian cancer [42]. Furthermore, HMGA2 overexpression appeared to be a strong feature of larynx carcinoma [26].

Evidence showed that AQP9 played a critical role in the transmembrane transport of As<sub>2</sub>O<sub>3</sub> and modulating arsenite sensitivity in leukemia [43, 44]. In addition, patients treated with chemotherapy in AQP9 high expression subgroup showed significantly better disease-free survival in colorectal cancer, demonstrating AQP9 functions as a drug transporter and further sensitized tumor cells to chemotherapy drugs associated with RAS signaling activation [45]. Recently AQP9 has also been found to play opposite roles in progression of different cancers. A report by Liao et al. identified that AQP9 could inhibit growth and metastasis of hepatocellular carcinoma cells via Wnt/ $\beta$ -catenin pathway [46]. On the contrary, AQP9 would promote astrocytoma cell invasion and motility [47] and its increase in mRNA-level was significantly correlated with aggressive progression and poor survival in clear cell renal cell carcinoma patients [48].

As a member of the transcription elongation factor TFIIS family in vertebrates, TCEA3 was significantly downregulated in cancer tissues compared with paired normal tissues; its upregulation could induce apoptosis in gastric cancer, ovarian cancer and rhabdomyosarcoma cell lines [49-51].

Of course, all genes in our signature have been reported to be involved and played crucial roles in the development of many other tumors, these results could provide potential new insights for LSCC research. There are still some limitations in our study. Firstly, due to public databases and our center currently lack enough clinical samples, more predictive gene candidates could exist but were missed from our study of limited scope. Furthermore, in terms of the inconsistency of staging system annotations among samples and the absence of smoking history has further reduced the number of prognostic factors and samples to be comprehensively analyzed. Data used in the study only consisted of OS and still belong to the training set category. More external validation that covering risk score and other

clinical variables in independent cohorts is required. Secondly, in addition to changes at the RNA level, changes of the protein and in-depth molecular-level mechanism could be investigated further.

## Conclusion

This study revealed deregulated mRNAs in LSCC, and discussed their possible roles in tumor progression. To our knowledge, the four-mRNA signature model has not been reported previously and could be promising to be a supplement to the individualized classification of LSCC patients, especially for the long-term survival. Nevertheless, further investigations of those DEGs and the signature are warranted.

## Supplementary Material

Supplementary figures and tables.

<http://www.jcancer.org/v12p5807s1.pdf>

## Acknowledgements

The authors would like to thank funding and public data (TCGA, GEO and Ensembl) and software (R and Strawberry-perl) providers.

## Funding

This study was supported by funding from the National Natural Scientific Foundation of China (No. 81570902) and the Clinical Research Plan of SHDC (16CR2029B).

## Competing Interests

The authors have declared that no competing interest exists.

## References

1. Steuer CE, El-Deiry M, Parks JR, Higgins KA, Saba NF. An update on larynx cancer. *CA Cancer J Clin.* 2017; 67: 31-50.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018; 68: 394-424.
3. Almadori G, Bussu F, Cadoni G, Galli J, Paludetti G, Maurizi M. Molecular markers in laryngeal squamous cell carcinoma: towards an integrated clinicobiological approach. *Eur J Cancer.* 2005; 41: 683-93.
4. Wang J, Chen X, Tian Y, Zhu G, Qin Y, Chen X, et al. Six-gene signature for predicting survival in patients with head and neck squamous cell carcinoma. *Aging (Albany NY).* 2020; 12: 767-83.
5. You GR, Cheng AJ, Lee LY, Huang YC, Liu H, Chen YJ, et al. Prognostic signature associated with radioresistance in head and neck cancer via transcriptomic and bioinformatic analyses. *BMC Cancer.* 2019; 19: 64.
6. Zhang G, Fan E, Yue G, Zhong Q, Shuai Y, Wu M, et al. Five genes as a novel signature for predicting the prognosis of patients with laryngeal cancer. *J Cell Biochem.* 2019; [Epub ahead of print].
7. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002; 30: 207-10.
8. Wang Z, Jensen MA, Zenklusen JC. A Practical Guide to The Cancer Genome Atlas (TCGA). *Methods Mol Biol.* 2016; 1418: 111-41.
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.
10. Kolde R. pheatmap: Pretty Heatmaps. version 1.0.10. 2018.
11. Oliveros JC. Venny. An interactive tool for comparing lists with Venn's diagrams. 2007-2015.

12. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun.* 2019; 10: 1523.
13. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw.* 2011; 39: 1-13.
14. Therneau T. A Package for Survival Analysis in S. version 2.38. 2015.
15. Liu G, Locascio JJ, Corvol JC, Boot B, Liao Z, Page K, et al. Prediction of cognition in Parkinson's disease with a clinical-genetic score: a longitudinal analysis of nine cohorts. *Lancet Neurol.* 2017; 16: 620-9.
16. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med.* 2013; 32: 5381-97.
17. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 2017; 45: W98-W102.
18. Placa JR, Bueno Rde B, Pinheiro DG, Panepucci RA, de Araujo LF, Mamede RC, et al. Gene expression analysis of laryngeal squamous cell carcinoma. *Genom Data.* 2015; 5: 9-12.
19. Feng L, Wang R, Lian M, Ma H, He N, Liu H, et al. Integrated Analysis of Long Noncoding RNA and mRNA Expression Profile in Advanced Laryngeal Squamous Cell Carcinoma. *PLoS One.* 2016; 11: e0169232.
20. Greene FL, Page DL, Fleming ID, Fritz AG. *AJCC Cancer Staging Manual.* 6th ed. New York: Springer-Verlag; 2002.
21. Edge SB, Byrd DR, Compton CC, Fritz AG. *AJCC Cancer Staging Manual.* 7th ed. New York: Springer-Verlag; 2010.
22. Fleming ID, Cooper JS, Henson DE, Hutter RVP. *AJCC Cancer Staging Manual.* 5th ed. Philadelphia, New York: Lippincott-Raven; 1997.
23. Xie X, Wang J, Shi D, Zou Y, Xiong Z, Li X, et al. Identification of a 4-mRNA metastasis-related prognostic signature for patients with breast cancer. *J Cell Mol Med.* 2019; 23: 1439-47.
24. Chen L, Luo Y, Wang G, Qian K, Qian G, Wu CL, et al. Prognostic value of a gene signature in clear cell renal cell carcinoma. *J Cell Physiol.* 2019; 234: 10324-35.
25. Wiley FJ. *R Deep Learning Essentials: Build automatic classification and prediction models using unsupervised learning.* Birmingham, UK: Packt Publishing Ltd.; 2016.
26. Palumbo A, Jr., De Martino M, Esposito F, Frassetto F, Neto PN, Valverde Fernandes P, et al. HMGA2, but not HMGA1, is overexpressed in human larynx carcinomas. *Histopathology.* 2018; 72: 1102-14.
27. Gao W, Zhang C, Li W, Li H, Sang J, Zhao Q, et al. Promoter Methylation-Regulated miR-145-5p Inhibits Laryngeal Squamous Cell Carcinoma Progression by Targeting FSCN1. *Mol Ther.* 2019; 27: 365-79.
28. Ni RS, Shen X, Qian X, Yu C, Wu H, Gao X. Detection of differentially expressed genes and association with clinicopathological features in laryngeal squamous cell carcinoma. *Oncol Lett.* 2012; 4: 1354-60.
29. Iizuka S, Ishimaru N, Kudo Y. Matrix metalloproteinases: the gene expression signatures of head and neck cancer progression. *Cancers (Basel).* 2014; 6: 396-415.
30. Poola I, DeWitty RL, Marshallck JJ, Bhatnagar R, Abraham J, Leffall LD. Identification of MMP-1 as a putative breast cancer predictive marker by global gene expression analysis. *Nat Med.* 2005; 11: 481-3.
31. Sauter W, Rosenberger A, Beckmann L, Kropp S, Mittelstrass K, Timofeeva M, et al. Matrix metalloproteinase 1 (MMP1) is associated with early-onset lung cancer. *Cancer Epidemiol Biomarkers Prev.* 2008; 17: 1127-35.
32. Wang YP, Liu IJ, Chiang CP, Wu HC. Astrocyte elevated gene-1 is associated with metastasis in head and neck squamous cell carcinoma through p65 phosphorylation and upregulation of MMP1. *Mol Cancer.* 2013; 12: 109.
33. Sugioka K, Kodama-Takahashi A, Yoshida K, Aomatsu K, Okada K, Nishida T, et al. Extracellular Collagen Promotes Interleukin-1beta-Induced Urokinase-Type Plasminogen Activator Production by Human Corneal Fibroblasts. *Invest Ophthalmol Vis Sci.* 2017; 58: 1487-98.
34. Pins GD, Collins-Pavao ME, Van De Water L, Yarmush ML, Morgan JR. Plasmin triggers rapid contraction and degradation of fibroblast-populated collagen lattices. *J Invest Dermatol.* 2000; 114: 647-53.
35. Wang L, Madigan MC, Chen H, Liu F, Patterson KI, Beretov J, et al. Expression of urokinase plasminogen activator and its receptor in advanced epithelial ovarian cancer patients. *Gynecol Oncol.* 2009; 114: 265-72.
36. Kanekura T, Chen X, Kanzaki T. Basigin (CD147) is expressed on melanoma cells and induces tumor cell invasion by stimulating production of matrix metalloproteinases by fibroblasts. *Int J Cancer.* 2002; 99: 520-8.
37. Sun J, Hemler ME. Regulation of MMP-1 and MMP-2 production through CD147/extracellular matrix metalloproteinase inducer interactions. *Cancer Res.* 2001; 61: 2276-81.
38. Zhang C, Chen S, Ma X, Yang Q, Su F, Shu X, et al. Upregulation of STC2 in colorectal cancer and its clinicopathological significance. *Oncol Targets Ther.* 2019; 12: 1249-58.
39. Xue K, Li J, Nan S, Zhao X, Xu C. Downregulation of LINC00460 decreases STC2 and promotes autophagy of head and neck squamous cell carcinoma by up-regulating microRNA-206. *Life Sci.* 2019; 231: 116459.
40. Li T, Qin Y, Zhen Z, Shen H, Cong T, Schiferle E, et al. Long non-coding RNA HOTAIR/microRNA-206 sponge regulates STC2 and further influences cell biological functions in head and neck squamous cell carcinoma. *Cell Prolif.* 2019; 52: e12651.
41. Yang S, Ji Q, Chang B, Wang Y, Zhu Y, Li D, et al. STC2 promotes head and neck squamous cell carcinoma metastasis through modulating the PI3K/AKT/Snail signaling. *Oncotarget.* 2017; 8: 5976-91.
42. Wu J, Lai M, Shao C, Wang J, Wei JJ. STC2 overexpression mediated by HMGA2 is a biomarker for aggressiveness of high-grade serous ovarian cancer. *Oncol Rep.* 2015; 34: 1494-502.
43. Leung J, Pang A, Yuen WH, Kwong YL, Tse EW. Relationship of expression of aquaglyceroporin 9 with arsenic uptake and sensitivity in leukemia cells. *Blood.* 2007; 109: 740-6.
44. Bhattacharjee H, Carbrey J, Rosen BP, Mukhopadhyay R. Drug uptake and pharmacological modulation of drug sensitivity in leukemia by AQP9. *Biochem Biophys Res Commun.* 2004; 322: 836-41.
45. Huang D, Feng X, Liu Y, Deng Y, Chen H, Chen D, et al. AQP9-induced cell cycle arrest is associated with RAS activation and improves chemotherapy treatment efficacy in colorectal cancer. *Cell Death Dis.* 2017; 8: e2894.
46. Liao S, Chen H, Liu M, Gan L, Li C, Zhang W, et al. Aquaporin 9 inhibits growth and metastasis of hepatocellular carcinoma cells via Wnt/beta-catenin pathway. *Aging (Albany NY).* 2020; 12: 1527-44.
47. Lv Y, Huang Q, Dai W, Jie Y, Yu G, Fan X, et al. AQP9 promotes astrocytoma cell invasion and motility via the AKT pathway. *Oncol Lett.* 2018; 16: 6059-64.
48. Xu WH, Shi SN, Xu Y, Wang J, Wang HK, Cao DL, et al. Prognostic implications of Aquaporin 9 expression in clear cell renal cell carcinoma. *J Transl Med.* 2019; 17: 363.
49. Li J, Jin Y, Pan S, Chen Y, Wang K, Lin C, et al. TCEA3 Attenuates Gastric Cancer Growth by Apoptosis Induction. *Med Sci Monit.* 2015; 21: 3241-6.
50. Kazim N, Adhikari A, Oh TJ, Davie J. The transcription elongation factor TCEA3 induces apoptosis in rhabdomyosarcoma. *Cell Death Dis.* 2020; 11: 67.
51. Cha Y, Kim DK, Hyun J, Kim SJ, Park KS. TCEA3 binds to TGF-beta receptor I and induces Smad-independent, JNK-dependent apoptosis in ovarian cancer cells. *Cell Signal.* 2013; 25: 1245-51.