Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Original Article

Mining TCGA database for prognostic genes in head and neck squamous cell carcinoma microenvironment

Qiu-Chi Ran [a,b], Sheng-Rong Long [c], Yan Ye [d], Chen Xie [a], Zhuo-Lin XuXiao [a], Yu-Song Liu [a], Hong-Xia Pang [e], Diwas Sunchuri [a], Nai-Chia Teng [f,g*], Zhu-Ling Guo [a,e**]

[a] *School of Dentistry, Hainan Medical University, Hainan, China*
[b] *Department of Dentistry, Stomatological Hospital Affiliated to China Medical University, Shenyang, China*
[c] *Department of Neurosurgery, The First Affiliated Hospital of China Medical University, Shenyang, China*
[d] *Technology Section, Duoyi Network, Wuhan, China*
[e] *Department of Dentistry, The First Affiliated Hospital of Hainan Medical University, Hainan, China*
[f] *School of Dentistry, College of Oral Medicine, Taipei Medical University, Taipei, Taiwan*
[g] *Department of Dentistry, Taipei Medical University Hospital, Taipei, Taiwan*

Received 27 September 2020; Final revision received 28 September 2020
Available online 16 October 2020

KEYWORDS
HNSCC;
Immune scores;
TCGA;
Tumor microenvironment

Abstract *Background/purpose:* Head and neck squamous cell carcinoma (HNSCC) is one of the most common malignant tumors. The aim of this study was to elucidate the effect of tumor microenvironment-related genes on the prognosis of HNSCC and to obtain tumor microenvironment-related genes that can predict poor prognosis in HNSCC patients.
Materials and methods: The ESTIMATE algorithm was applied to the HNSCC transcriptomic data downloaded from the TCGA (The cancer genome atlas), and then the samples were divided into two groups: high and low immune scoring groups, and high and low basal scoring groups to screen for differentially expressed genes (DEGs) associated with poor patient outcomes. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis was performed to explore the potential functions of DEGs, and then to explore the potential prognostic value of individual DEGs. The results of survival analysis between DEGs and overall survival (OS) to explore tumor microenvironment-related genes relevant to the prognosis of HNSCC patients.
Results: Fifty-nine tumor microenvironment-related genes were screened for association of OS with HNSCC ($P < 0.05$). The GO and KEGG enrichment analysis showed that the selected DEGs

* Corresponding author. School of Dentistry, College of Oral Medicine, Taipei Medical University, Taipei, Taiwan. Fax: +886 227362295.
** Corresponding author. Department of Neurosurgery, The First Affiliated Hospital of China Medical University, Shenyang, China.
E-mail addresses: dianaten@tmu.edu.tw (N.-C. Teng), srlong@cmu.edu.cn (Z.-L. Guo).

https://doi.org/10.1016/j.jds.2020.09.017

may mediate immune response, extracellular matrix, and immunoglobulin binding via neutrophil activation in HNSCC. Six of these DEGs, GIMAP6, SELL, TIFAB, KCNA3, P2RY8 and CCR4 were most significantly associated with OS ($P < 0.001$).

Conclusion: We identified six tumor microenvironment-related genes that were significantly associated with poor prognosis in HNSCC. These genes may inspire researchers to discover new targets and approaches for HNSCC treatment.

Introduction

The incidence of malignant tumors of the head and neck ranks sixth in the world, among which laryngeal cancer, tongue cancer, hypopharyngeal cancer and nasopharyngeal cancer are the most common, and squamous cell carcinoma accounts for 90%–95% of the pathological types, which is called Head and neck squamous cell carcinoma (HNSCC).[1] Despite recent advances in the surgical and comprehensive treatment of HNSCC, malignant pathological features such as early lymph node metastasis and aggressive growth are still important factors in its high rate of postoperative recurrence and metastasis, low survival rate, and poor long-term outcome.[2]

The use of targeted drugs and immunotherapy is bringing new hope to patients. Tumor microenvironment (TME), that is the cellular environment in which the tumor is located, is heterogeneous. TME is composed of immune cells, endothelial cells, fibroblasts and myeloid cells, as well as inflammatory mediators and extracellular matrix molecules,[3] which is a key factor in tumor progression and high patient mortality.[4] In TME, immune cells and stromal cells are two major non-neoplastic components that are valuable for tumor diagnosis and prognosis assessment, and tumor microenvironment prognosis-related genes may also be the next new therapeutic target.

Yoshihara et al.[5] designed an algorithm called ESTIMATE to explore the potential relationship between the degree of infiltration of immune and stromal cells and tumor prognosis. Subsequent studies soon applied the ESTIMATE algorithm to acute leukemia,[6,7] glioblastoma,[8] renal clear cell carcinoma,[9,10] colon cancer,[11] and breast cancer[12] and showed the effectiveness of this big data-based algorithm. Therefore, this study was based on the ESTATE algorithm to explore the potential relationship between the infiltration of immune cells and stromal cells and the prognosis of HNSCC, and to obtain tumor microenvironment-related genes that can predict poor prognosis in HNSCC patients.

Materials and methods

Data resource

Downloaded from the TCGA data portal (https://portal.gdc.cancer.gov/, accessed August 30, 2019) publicly available HNSCC datasets, including tertiary data on gene expression profiles and survival information. The downloaded RNA expression data were scored by the ESTIMATE algorithm for basal and immune scoring, and cases of lung adenocarcinoma were classified into high and low groups based on the median score. All data involved in this study were downloaded from TCGA, and data collection and application were performed in accordance with TCGA publication guidelines and data access policy without additional approval from the local ethics committee.

Identification of differentially expressed genes (DEGs)

Data analysis was performed using the "limma" package (version 3.8) on R (version 3.6.0). The Wilcoxon rank sum test was used for statistical tests, and the following criteria were used for differentially expressed gene screening: FDR <0.05, $|\log2FC| > 1$.

Heatmaps and clustering analysis

The heat map plot and sample clustering was generated using R "heatmap" package.

Enrichment analysis of DEGs

Functional enrichment analysis of DEGs was performed by R package clusterprofiler[13] to identify gene ontology (GO) categories by their biological processes (BP), molecular functions (MF), or cellular components (CC). The clusterprofiler was also used to perform pathway enrichment analysis with reference from KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways. FDR$<$ 0.05 was used as the cut-off.

Overall survival curve

The survival R package was used to analyze the relationship between DEGs expression levels and the overall survival of HNSCC patients.

Results

We downloaded gene expression profiles and clinical information of all 546 HNSCC patients from the TCGA database. Among them, 63 (11.54%) patients were diagnosed with Grade 1, 310 (56.78%) patients were Grade 2, 125

(22.89%) patients were Grade 3, 7 (1.28%) patients were Grade 4, and 41 (7.51%) patients were of unknown grade. Based on ESTIMATE algorithm, immune scores were distributed between -1057.57 and 2785.18, and stromal scores ranged from -1941.57 to 1947.28, respectively (Fig. 1A, B).

To find out the potential correlation of overall survival with immune scores and/or stromal scores, we divided the 500 HNSCC cases (only 500 of 502 cancer tissues samples have their grade clinical information) into top and bottom halves (high vs. low score groups) based on their scores. Kaplan–Meier survival curves (Fig. 2A) showed that median overall survival of cases with the low score group of immune scores are shorter than the cases in the high score group ($p = 0.107$ in log-rank test). Consistently, cases with lower stromal scores also showed shorter median overall survival compared to patients with higher stromal scores (Fig. 2B, $p = 0.814$ in log-rank test), although it was not statistically significant.

Comparison of gene expression profile with immune scores and stromal scores in HNSCC

To reveal the correlation of global gene expression profiles with immune scores and/or stromal scores, we compared Affymetrix microarray data of these 500 HNCSS cases obtained in TCGA database. Heatmaps in Fig. 3A, B showed distinct gene expression profiles of cases belong to high vs. low immune scores/stromal scores groups. For comparison based on immune scores, 777 genes were upregulated and 161 genes downregulated in the high score than the low score group (fold change > 1.5, $p < 0.05$). Similarly, for the high and low groups based on stromal scores, 986 genes were upregulated and 63 genes were downregulated in the high score group (fold change > 1.5, $p < 0.05$). Moreover, Venn diagrams (Fig. 3C, D) showed that 260 genes were commonly upregulated in the high scores groups, and 15 genes were commonly downregulated. It is worth mentioning that the DEGs extracted from the comparison of high vs. low immune scores groups covered the majority of genes extracted from the comparison based on stromal scores. Thus, we decided to focus on these DEGs for all subsequent analysis in this manuscript. To outline the potential function of the DEGs, we performed functional enrichment analysis of the 260 upregulated genes and 15 downregulated genes in high-immune scores group.

Enrichment analysis of DEGs

We found 260 genes were commonly upregulated in the high scores groups, and 15 genes were commonly downregulated (Supplementary Table 1). In order to reveal the potential functions of differentially expressed genes, a functional enrichment analysis was performed on genes that were generally up- and down-regulated in immune and matrix scores. Functional enrichment clustering of these genes showed strong association with immune response as well. Top GO terms identified neutrophil activation involved in and mediated immune response, extracellular matrix, and immunoglobulin binding (Fig. 4A). In addition, all the pathways that were yielded from the KEGG identified Cytokine–cytokine receptor interaction, Complement and coagulation cascades and B cell receptor signal pathway (Fig. 4B).

Correlation of expression of individual DEGs in overall survival

To explore the potential roles of individual DEGs in overall survival, we generated Kaplan–Meier survival curves from TCGA database. Among the 275 DEGs in the high-immune scores group, a total of 59 DEGs (Supplementary Table 2) were shown to significantly predict poor overall survival in log-rank test ($p < 0.05$, selected 6 genes which most significant relate to OS ($p < 0.001$) are shown in Fig. 5).

Discussion

Data mining of TCGA databases has been widely used to predict cancer prognosis, and recent studies have shown that the tumor microenvironment plays a critical role in tumor growth and progression. The immune cells and stromal cells infiltrated in TME are composed of many different cell types. On the one hand, fibroblasts, as the most abundant stromal cells, form a physical barrier to avoid immune recognition and elimination of tumor cells; on the other hand, they promote tumor proliferation and

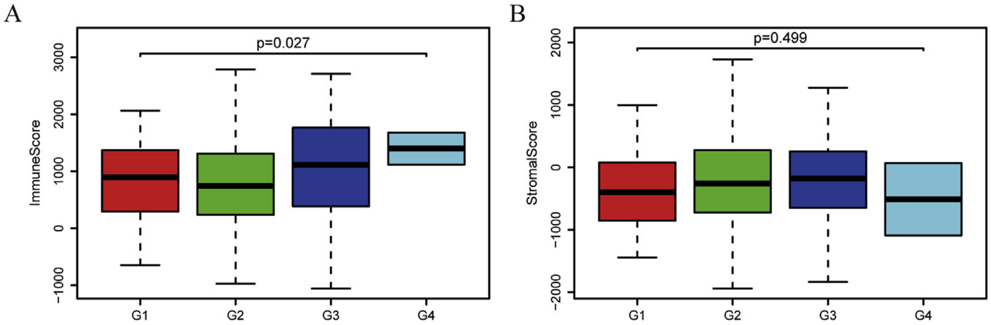

Figure 1 Immune scores and stromal scores are associated with HNSCC grades and their overall survival. A) Distribution of immune scores of HNSCC grades. Boxplot shows that there is significant association between HNSCC grades and the level of immune scores ($p < 0.05$). B) Distribution of stromal scores of HNSCC grades. Box-plot shows that there is not significant association between HNSCC grades and the level of stromal scores ($p < 0.05$).

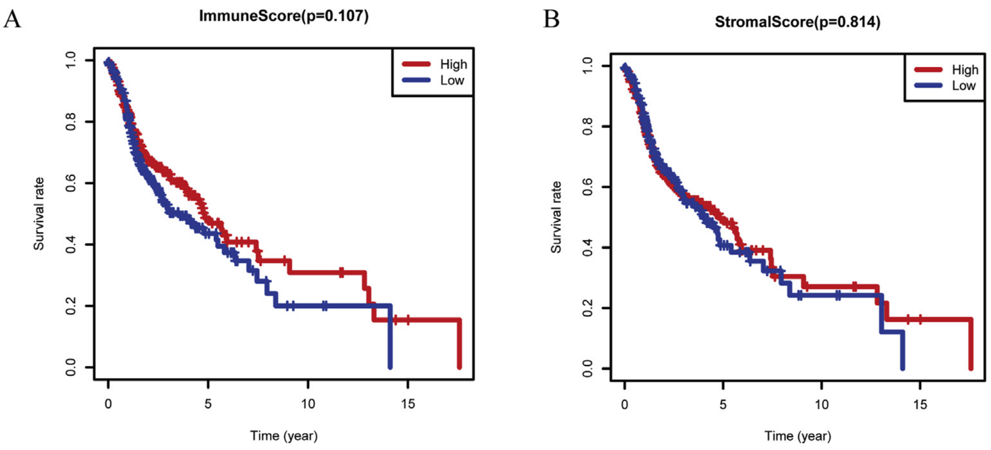

Figure 2 Immune scores and stromal scores are associated with HNSCC overall survival. A) HNSCC cases were divided into two groups based on their immune scores: the top half of cases with higher immune scores and the bottom half of cases with lower immune scores. As shown in the Kaplan—Meier survival curve, median survival of the low score group is shorter than high score group, as indicated by the log-rank test, p value is 0.107. B) HNSCC cases were divided into two groups based on their stromal scores: the top half of 209 cases and the bottom half of 208 cases. The median survival of the low score group is shorter than the high score group, however, it is not statistically different as indicated by the log-rank test p value is 0.814.

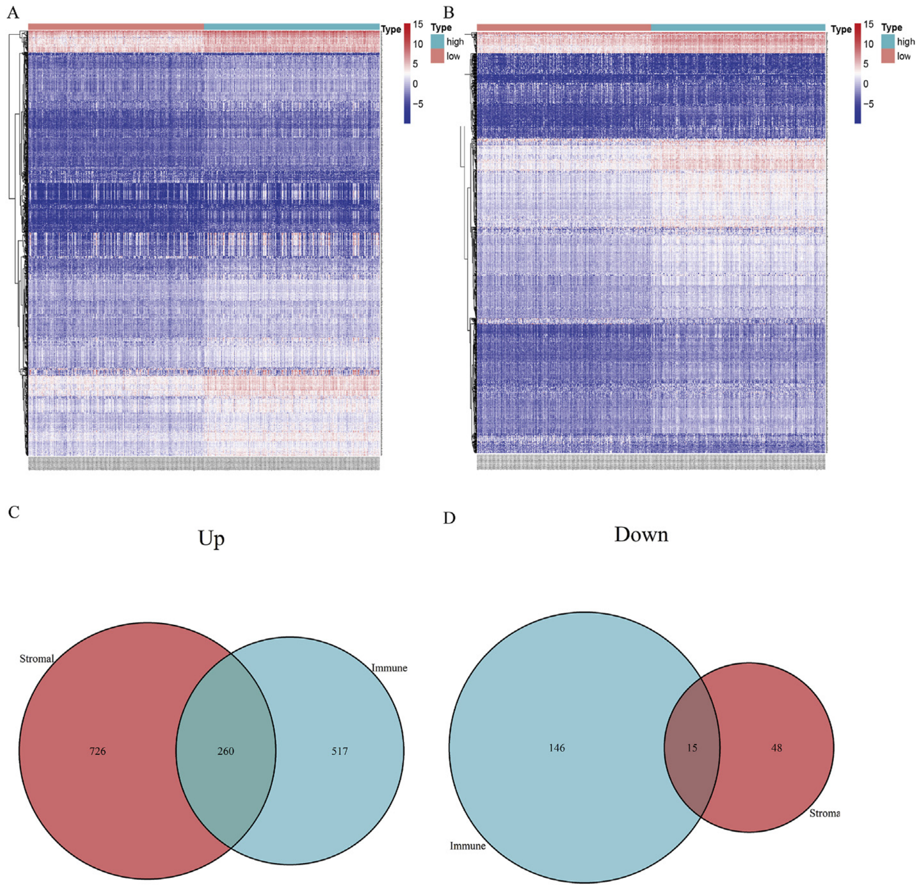

Figure 3 Comparison of gene expression profile with immune scores and stromal scores in HNSCC. Heatmaps were drawn based on the average linkage method and Pearson distance measurement method. Genes with higher expression are shown in red, lower expression are shown in blue, genes with same expression level are in white. A) Heatmap of the DEGs of immune scores of top half (high score) vs. bottom half (low score). $p < 0.05$, fold change >1.5). B) Heatmap of the DEGs of stromal scores of top half (high score) vs. bottom half (low score). $p < 0.05$, fold change >1.5). Venn diagrams showing the number of commonly upregulated C) or downregulated D) DEGs in stromal and immune score groups.

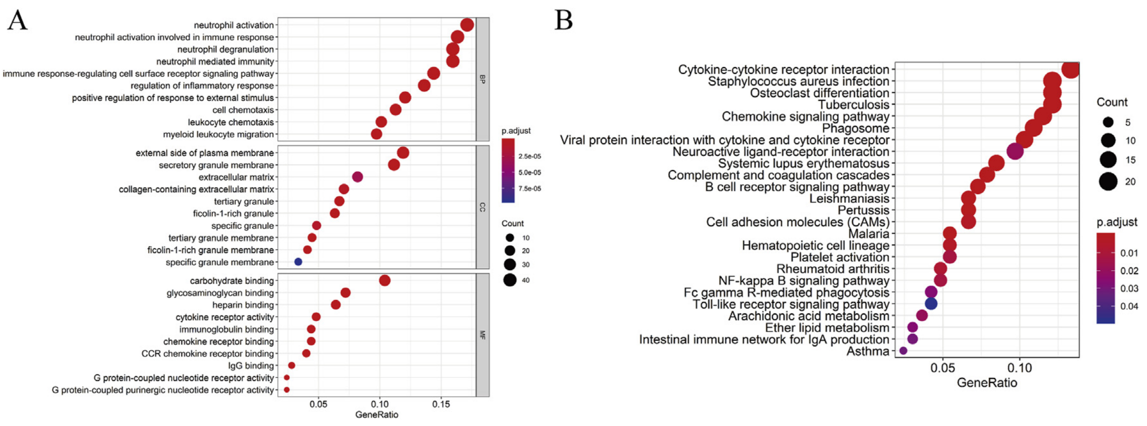

Figure 4 A: Top 10 BP, CC, MF terms and B: top 25 KEGG pathways enrichment analysis was performed by R package clusterprofiler (*p*. adjusted <0.05).

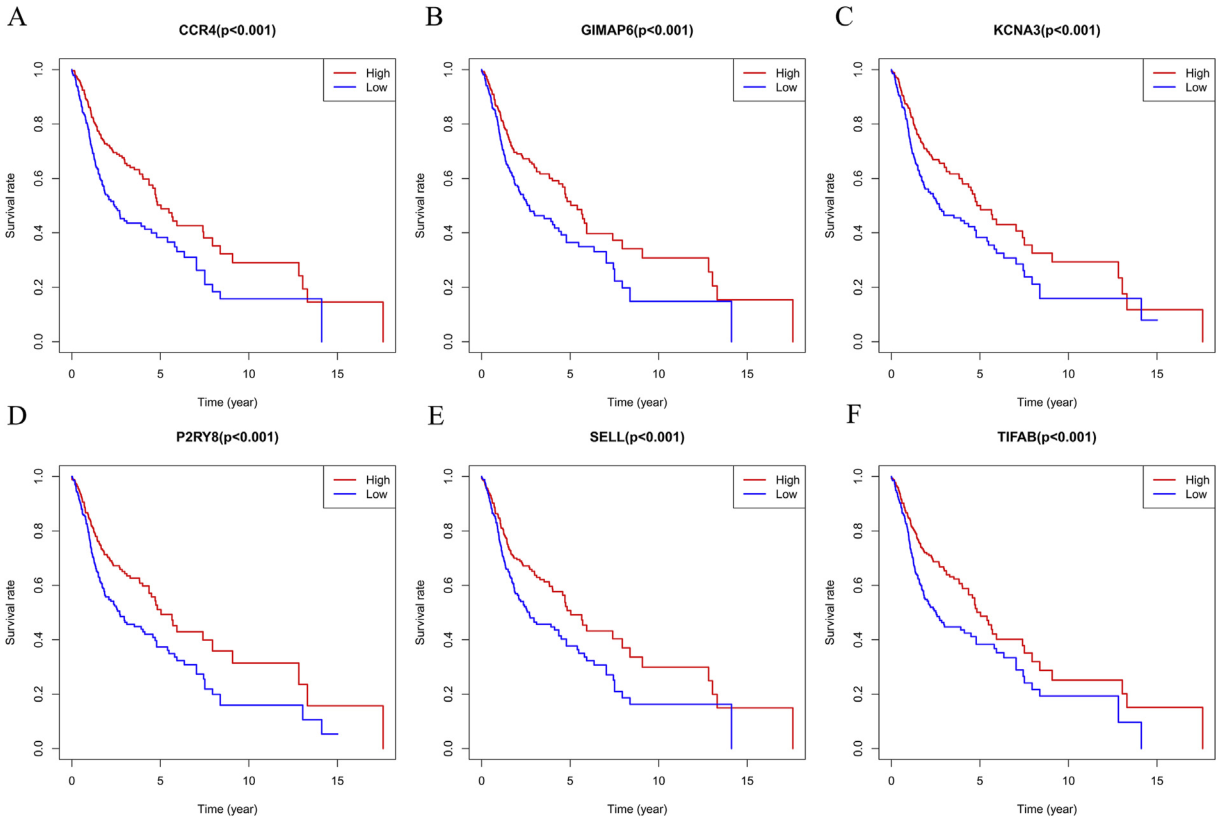

Figure 5 DEGs extracted from TCGA database with overall survival. Kaplan–Meier survival curves were generated for selected DEGs extracted from the comparison of groups of high (red line) and low (blue line) gene expression. $p < 0.001$ in Log-rank test.

metastasis by modulating the extracellular matrix and secreting relevant cytokines or growth factors.[14,15]

In this work, we attempted to identify tumor microenvironment related genes contributing to HNSCC overall survival in the TCGA database. In particular, by comparing global gene expression in a large number of cases with high vs. low immune scores, we extracted 59 genes involved in extracellular matrix and immune response. By analyzing 275 differentially expressed genes yielded from comparison of high vs. low immune scores (or stromal scores) groups, we found that many of them were involved in tumor microenvironment, as shown by GO term analysis (Fig. 2). This is in accordance with previous reports that the functions of immune cells and ECM molecules are closely interrelated in building tumor microenvironment in HNSCC.[16–19] Finally, we analyzed overall survival analysis of these 275 genes and found that 59 genes were associated with poor outcomes in HNSCC patients. There are 6 genes, GIMAP6, SELL, TIFAB, KCNA3, P2RY8 and CCR4 which most significant relate to OS of HNSCC. This demonstrates the prognostic value of our big data analysis of lung adenocarcinoma microenvironment-associated genes in the TCGA database of HNSCC based on the ESTIMATE algorithm.

We are particularly interested in CCR4, GIMAP6 and SELL. CCR4 has been proven that it may have an important role in HNSCC progression, regional lymph node metastasis and recurrence. Significant progress has been made on the correlation of overall survival with gene expression in HNSCCs.[20−22] Many of these experiments were done in tumor cell lines, animal models, or patients' tumor samples. GIMAP6 has also been reported to have a potential role in tumor evolution mechanisms related to inflammation and microenvironment.[23] SELL also has been identified overexpressed in HNSCC that may modulate metastasis and affect survival.[24] However, the complexity of HNSCC and HNSCC microenvironment demands more comprehensive analysis consisting of larger cohorts. The interaction between HNSCC and its tumor microenvironment critically affects tumor evolution, which subsequently impacts tumor subtype classification, recurrence, drug resistance, and the overall prognosis of patients.

In summary, these 6 TME-associated DEGs were significantly associated with poor prognosis of HNSCC using HNSCC transcriptome data in this paper. Further study of these TME-associated genes may provide new insights into the potential relationship between TME and HNSCC prognosis, and these genes may also inspire researchers to discover new therapeutic targets and approaches for HNSCC.

Declaration of competing interest

No conflicts of interest relevant to this article were disclosed.

Acknowledgments

All the authors thank Dr. Guangyu Li for supporting this paper. This research was funded by Scientific Research Fund of Hainan Provincial Education Department (Hnky2020ZD-19), Young Talents' Science and Technology Innovation Project of Hainan Association for Science and Technology (QCXM201817), Educational scientific Research of Hainan Medical University (HYYB202014), Marxism Theory Discipline Research Project of Hainan Medical University (QMYKYLX201807), Research Project of Hainan Provincial Humanities Medical Research Base (QRYZH201811(YB)).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jds.2020.09.017.

References

1. Galbiatti A, Padovani-Junior J, Maníglia J, Rodrigues C, Pavarino É, Goloni-Bertollo E. Head and neck cancer: causes, prevention and treatment. *Braz J Otorhinolar* 2013;79:239−47.
2. Zhang S, Lu Z, Luo X, et al. Retrospective analysis of prognostic factors in 205 patients with laryngeal squamous cell carcinoma who underwent surgical treatment. *PloS One* 2013;8:e60157.
3. Hanahan D, Weinberg R. Hallmarks of cancer: the next generation. *Cell* 2011;144:646−74.
4. Kothari A, Mi Z, Zapf M, Kuo P. Novel clinical therapeutics targeting the epithelial to mesenchymal transition. *Clin Transl Med* 2014;3:35.
5. Yoshihara K, Shahmoradgoli M, Martínez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 2013;4:2612.
6. Yan H, Qu J, Cao W, et al. Identification of prognostic genes in the acute myeloid leukemia immune microenvironment based on tcga data analysis. *Cancer Immunol Immun : CII* 2019;68:1971−8.
7. Huang S, Zhang B, Fan W, et al. Identification of prognostic genes in the acute myeloid leukemia microenvironment. *Aging* 2019;11:10557−80.
8. Jia D, Li S, Li D, Xue H, Yang D, Liu Y. Mining tcga database for genes of prognostic value in glioblastoma microenvironment. *Aging* 2018;10:592−605.
9. Chen B, Chen W, Jin J, Wang X, Cao Y, He Y. Data mining of prognostic microenvironment-related genes in clear cell renal cell carcinoma: a study with tcga database. *Dis Markers* 2019;2019:8901649.
10. Xu W, Xu Y, Wang J, et al. Prognostic value and immune infiltration of novel signatures in clear cell renal cell carcinoma microenvironment. *Aging* 2019;11:6999−7020.
11. Alonso M, Aussó S, Lopez-Doriga A, et al. Comprehensive analysis of copy number aberrations in microsatellite stable colon cancer in view of stromal component. *Br J Cancer* 2017;117:421−31.
12. Priedigkeit N, Watters R, Lucas P, et al. Exome-capture rna sequencing of decade-old breast cancers and matched decalcified bone metastases. *JCI Insight* 2017;2.
13. Yu G, Wang L, Han Y, He Q. Clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics* 2012;16:284−7.
14. Chen X, Song E. Turning foes to friends: targeting cancer-associated fibroblasts. *Nat Rev Drug Discov* 2019;18:99−115.
15. Seino T, Kawasaki S, Shimokawa M, et al. Human pancreatic tumor organoids reveal loss of stem cell niche factor dependence during disease progression. *Cell Stem Cell* 2018;22:454−67. e6.
16. Lu P, Takai K, Weaver V, Werb Z. Extracellular matrix degradation and remodeling in development and disease. *Cold Spring Harb Perspect Biol* 2011;3:a005058.
17. Bonnans C, Chou J, Werb Z. Remodelling the extracellular matrix in development and disease. *Nat Rev Mol Cell Biol* 2014;15:786−801.
18. Pickup M, Mouw J, Weaver V. The extracellular matrix modulates the hallmarks of cancer. *EMBO Rep* 2014;15:1243−53.
19. Peranzoni E, Rivas-Caicedo A, Bougherara H, Salmon H, Donnadieu E. Positive and negative influence of the matrix architecture on antitumor immune surveillance. *Cell Mol Life Sci* 2013;70:4431−48.
20. González-Arriagada W, Lozano-Burgos C, Zúñiga-Moreta R, González-Díaz P, Coletta R. Clinicopathological significance of chemokine receptor (ccr1, ccr3, ccr4, ccr5, ccr7 and cxcr4) expression in head and neck squamous cell carcinomas. *J Oral Pathol Med* 2018;47:755−63.
21. Sun W, Li W, Wei F, et al. Blockade of mcp-1/ccr4 signaling-induced recruitment of activated regulatory cells evokes an antitumor immune response in head and neck squamous cell carcinoma. *Oncotarget* 2016;7:37714−27.
22. Tsujikawa T, Yaguchi T, Ohmura G, et al. Autocrine and paracrine loops between cancer cells and macrophages promote lymph node metastasis via ccr4/ccl22 in head and neck squamous cell carcinoma. *Int J Cancer* 2013;132:2755−66.

23. Sanati N, Iancu O, Wu G, Jacobs J, McWeeney S. Network-based predictors of progression in head and neck squamous cell carcinoma. *Front Genet* 2018;9:183.

24. Liu C, Liu T, Kuo L, et al. Differential gene expression signature between primary and metastatic head and neck squamous cell carcinoma. *J Pathol* 2008;214:489—97.