J. Maxillofac. Oral Surg. (Apr–June 2019) 18(2):320–327
https://doi.org/10.1007/s12663-019-01187-z

SHORT COMMUNICATION

A Six-Gene-Based Prognostic Model Predicts Survival in Head and Neck Squamous Cell Carcinoma Patients

Shrikant Pawar[1,2] · **Aditya Stanam**[3]

Received: 24 November 2018 / Accepted: 14 January 2019 / Published online: 24 January 2019

Abstract

Background and Objective Head and neck cancer is a malignant tumor that begins in the head and neck region, and has the sixth highest incidence worldwide. Previous studies have indicated several prognostic markers for head and neck squamous cell carcinoma (HNSCC), but due to poor accuracy and sensitivity of these clinical characteristic markers attention has been gradually switched to molecular biomarkers. This study aimed to sort out the mRNAs correlated with patient survival time to establish an mRNA combination prognostic biomarker model for HNSCC patient risk stratification, providing optimal therapeutic regimens and improving patient prognosis.

Methods Clinical data and transcriptome sequencing data of HNSCC were retrieved from TCGA database and were allocated into training and validation datasets. The prognostic model was established using the mRNAs, which were sorted out from training dataset by a significant correlation with survival time. Eventually, the prediction property of the model was evaluated by Kaplan–Meier survival analysis and receiver operating characteristic (ROC) curve.

Results An optimal prognostic model by the combination of six mRNAs was established. Kaplan–Meier survival analysis revealed effective risk stratification by this model for patients in the two datasets. The area under ROC curve (AUC) was > 0.65 for training and validation datasets, indicating good sensitivity and specificity of this model. Moreover, prominent superiority of this model to investigate prognostic biomarkers was demonstrated.

Conclusion Our model provided effective prognostication in terms of death risk stratification and evaluation in HNSCC patients. Combination of this prognostic model with current treatment measures is expected to greatly improve the patients' prognosis.

Keywords Biomarkers · Survival time · Cox regression · Kaplan–Meier survival analysis

Introduction

Head and neck squamous cell carcinoma (HNSCC) is a malignant neoplasm that arises from the head and neck region. HNSCC is the sixth most common neoplasm worldwide and third most common cancer in developing countries [1, 2]. More than 500,000 and 50,000 individuals are being diagnosed with HNSCC annually worldwide and in the USA, respectively. It is a significant cause of cancer morbidity and mortality and account for about 13,000 deaths in the USA [1]. Despite advances in treatment, the median overall survival for advanced cancers has been less than 1 year [3]. Therefore, it is critical to find an effective and reliable method for HNSCC risk stratification, improving prognosis of HNSCC patients. Several

Shrikant Pawar and Aditya Stanam have contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12663-019-01187-z) contains supplementary material, which is available to authorized users.

✉ Shrikant Pawar
spawar2@student.gsu.edu

[1] Department of Computer Science, Georgia State University, 34 Peachtree Street, Atlanta, GA 30303, USA

[2] Department of Biology, Georgia State University, 34 Peachtree Street, Atlanta, GA 30303, USA

[3] Department of Pathology, The University of Iowa, 500 Newton Road, #ML 1132, Iowa City, IA 52242-5000, USA

prognostic markers have been reported. Nonetheless, these studies regarding RNA prognostic biomarkers mostly focused on single prognostic marker, showing the shortcomings of low accuracy, poor stability and lack of universality. Although study by Guo et al. [4] discussed a six-mRNA signature as a prognostic factor for HNSCC patient survival, further studies are needed to confirm their findings.

Hence, in this study, RNA expression profiles of HNSCC cancerous tissues and paracancerous tissues from TCGA database were analyzed, and a model of combinatorial RNA prognostic biomarkers was established for HNSCC patient risk stratification, displaying excellent stability and high accuracy. This model was expected to help in the development of optimal treatment regimen and improve patient's prognosis.

Materials and Methods

Data Source

The whole data of head and neck cancer analyzed in this study were obtained from the TCGA database (TCGA-LIHC), with the data deadline till May 2018. The data of 528 specimens were included. These specimens were used to establish head and neck cancer prognostic biomarker model and validation model. The detailed information of head and neck cancer patients is listed in Table 1. The total genes from each of these samples were 20,531. Associated clinical information like patient gender, clinical stage, neoplasm histologic grade, vital status, patient race, patient HPV status, patient's days to death, neoplasm cancer status, margin status, number of lymph nodes positive and perineural invasion, radiation course number and radiation dosage was also retrieved. The pipeline for the analysis is summarized in Fig. 1.

Cox Proportional Hazards Regression Model

Cox proportional hazards regression model is widely applied due to no requirement for studying the distribution patterns of survival time and the capability to predict influence of diverse variants on the hazard rate. This regression model is adopted in this study for analyzing the correlation of survival time with relevant variants. The logarithm to base 2 normalized RNA expression levels was considered for all the genes for all the analysis in this article. Cox proportional hazards regression model was applied with survival status and days for all the 528 patients, and only 1476 patients with a significant P value (< 0.05) were selected for further analysis. Libraries "*survival,*" "*Biobase*" and "*pbapply*" in R were utilized for performing Cox proportional hazards regression.

Table 1 Cinico-pathological characteristics of HCC patients from TCGA database

Characteristics	Groups	Patients number Total = 528
Age at diagnosis		
	≥ 60	98
	< 60	89
Gender		
	Male	133
	Female	54
Tumor stage		
	Stage I	76
	Stage II	45
	Stage III	47
	Stage IV	4
	Others/unknown	15
Clinical stage		
	G1	26
	G2	91
	G3	63
	G4	6
	Others/unknown	1
Survival status	Survive	133
	Death	54

Receiver Operating Characteristic (ROC) Curve

A receiver operating characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true-positive rate (TPR) against the false-positive rate (FPR) at various threshold settings [5]. ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones and is related in a direct and natural way to cost–benefit analysis of diagnostic decision making [5]. Library "*survival*" was utilized in R for performing ROC on 20,531 genes in 1476 patients. Area under the ROC curve (AUC) value of > 0.65 was selected as a significant threshold, and AUC represents the probability that a random positive example is positioned to the right of a random negative example. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. Top six genes were identified with this significant threshold.

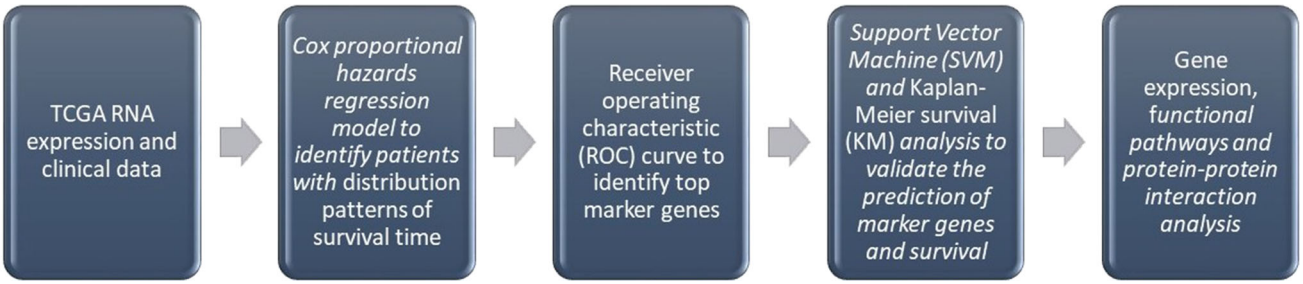

Fig. 1 Pipeline of the analysis

Support Vector Machine (SVM) Analysis

SVM is supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier [6]. An index of 6 selected genes from ROC curve was calculated as sum of RNA expression values. Patients were divided as training and test set for threefold, fivefold and sevenfold validation. Libraries "*caret*" and "*e1071*" were utilized to perform a "C-classification," "linear" kernel model with function "*svm,*" and a "*predict*" function was utilized for calculating prediction accuracies for each of the training and test samples. The accuracies are listed in Table 2.

Kaplan–Meier Survival (KM) Analysis

Finally, Kaplan–Meier survival (KM) curve was applied to detect the differences of survival time for the index genes. Functions "*Surv*" and "*survfit*" were used for making survival objects and a survival fit. Libraries *survival*, *ranger*, *ggplot2*, *dplyr* and *ggfortify* in R statistical package were used for producing survival plots. Libraries "*survminer*" and "*survival*" were used to calculate the survival P values. A significant P value of < 0.001 was found to be associated with the KM curve.

Functional Pathways and Protein–Protein Interaction Analysis

Functional pathway analysis for marker genes was performed using DAVID [7] database. The genes were converted to their respective protein ids and fed into STRING v10.0 [8] database to retrieve confidence values for protein interactions. These data were further fed to Cytoscape [9] to derive an interaction interactome map.

Gene Expression Analysis

RNA sequencing data were utilized to calculate the expression levels of six-gene index to compare the grades, stage, gender, race, HPV infections status, tumor margin, lymph nodes affected, number of radiation settings and dosages in tumor and tumor-free patients. All the analysis code repositories are deposited on authors GitHub account which can be found at: https://github.com/spawar2/Head-Neck-Cancer–ROC–SVM–KM–Expression-Analysis.

Results

Establishment of RNA Prognostic Biomarker Model

Each sample data encompassed the expression of 20,531 RNAs. To find the suitable prognostic biomarker for most patients, a Cox proportional hazards regression model was performed to evaluate the survival time of the patients. Results revealed 1476 patients with a significant P value (< 0.05) were selected for further analysis which showed a significant correlation with survival time. ROC analysis

Table 2 SVM accuracies for validating ROC gene signature

	Sevenfold train	Sevenfold test	Fivefold train	Fivefold test	Threefold train	Threefold test
Accuracy (%)	93.58974	85.38813	94.64286	85.89212	97.22222	86.2069
Train patients	68	180	51	197	34	214
Test patients	5	7	2	10	1	11

was performed on 20,531 genes in 1476 patients, and an AUC value of > 0.65 was selected as a significant threshold for selecting a final biomarker model consisting of top six genes (Fig. 2).

Property Evaluation of the RNA Prognostic Biomarker Model Enclosed in Training and Validation Dataset

By the established RNA prognostic biomarker model, the death risk score of individual patient was calculated based on the expression levels of the 6 RNAs. Patients were divided as training and test set for analyzing how effectively these six genes predict the correct survival status in a patient. A threefold, fivefold and sevenfold validations were generated (Table 2) with SVM accuracies. A significantly reasonable prediction accuracy of 85.38, 85.89 and 86.20% was found with test dataset. This implied that the RNA prognostic biomarker model built in the training dataset was accurate and replicable and can accurately predict the death risk for patients.

Analysis on Differential Expressions of RNA Prognostic Biomarkers

The optimal RNA prognostic biomarker model encompassed six RNAs, *RGMA, KLHL14, DLG2, NOVA1, KRTAP5-8* and *C1orf190*. To explore the potential role of these six RNAs in head and neck cancer initiation and development, analysis of differential expression was performed for obtaining the RNA prognostic biomarkers. Patients with RNA expression data of both cancerous and non-cancerous tissues were chosen, and comparisons of markers expression with stage, grade, sex, race, HPV status, neoplasm margin, lymph nodes, perineural tissue, radiation course and dosages in tumor and tumor-free patients were generated (Figs. 3, 4). Prognostic biomarker was significantly expressed in 156 patients with stage IV compared to 60 patients which were tumor-free, and the same pattern was observed with grade 2 (179 and 66 patients), with males (220 and 83 males), with white race (225 and 94 white race patients) and with negative marginal status (215 and 63 patients). This preliminarily indicated that the 6 RNAs likely played an important part in head and neck carcinogenesis with respect to patient's tumor stage, grade, sex, race and tumor marginal status.

Gene Pathway Enrichment and Protein Interaction Analysis

The DAVID pathway enrichment mapped these genes as involved in repulsive guidance molecule (RGM) family that performs several functions in the developing and adult nervous system, regulation of cephalic neural tube closure, inhibition of neurite outgrowth and cortical neuron branching, formation of mature synapses, involvement in regulation of synaptic stability at cholinergic synapses, mediation of negative regulation of exons splicing, matrix proteins including the high-sulfur and high-glycine-tyrosine keratins and leucine-rich adaptor proteins. With STRING and Cytoscape analysis, RGMA, NOVA1, NEO1 and PTBP2 proteins were found to be interacted as first-degree nodes with confidence score ranging from 0.5 to 0.9 (Fig. 5). PTBP2 is a polypyrimidine tract-binding protein

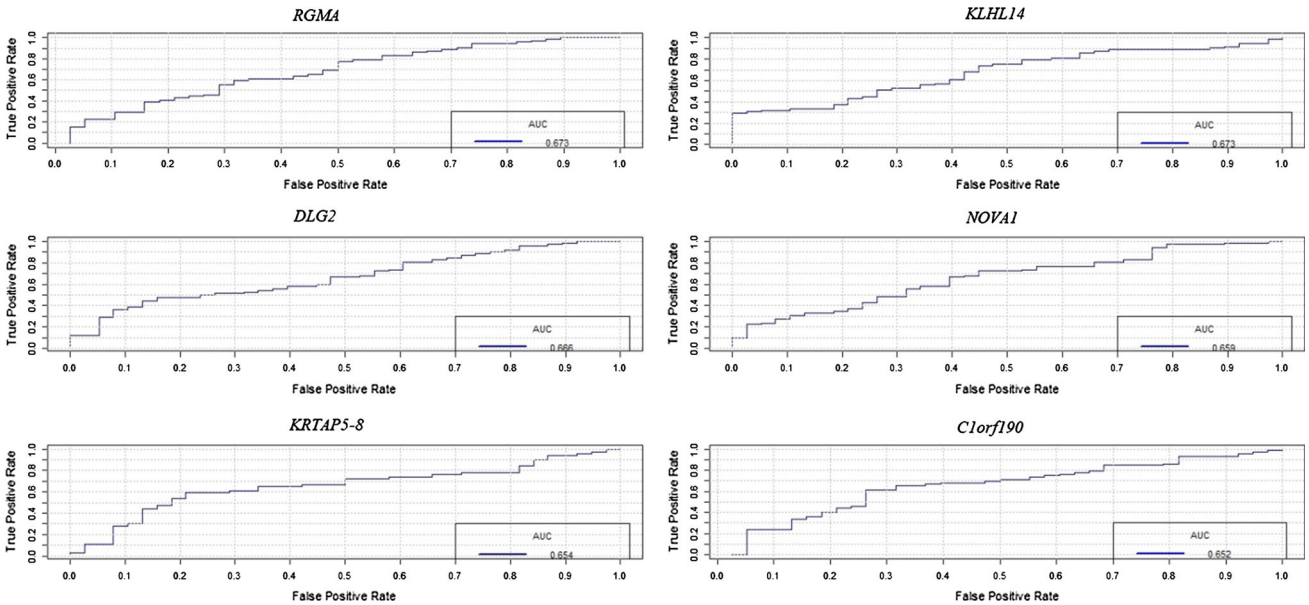

Fig. 2 ROC curves, AUC values for the selected biomarker six genes

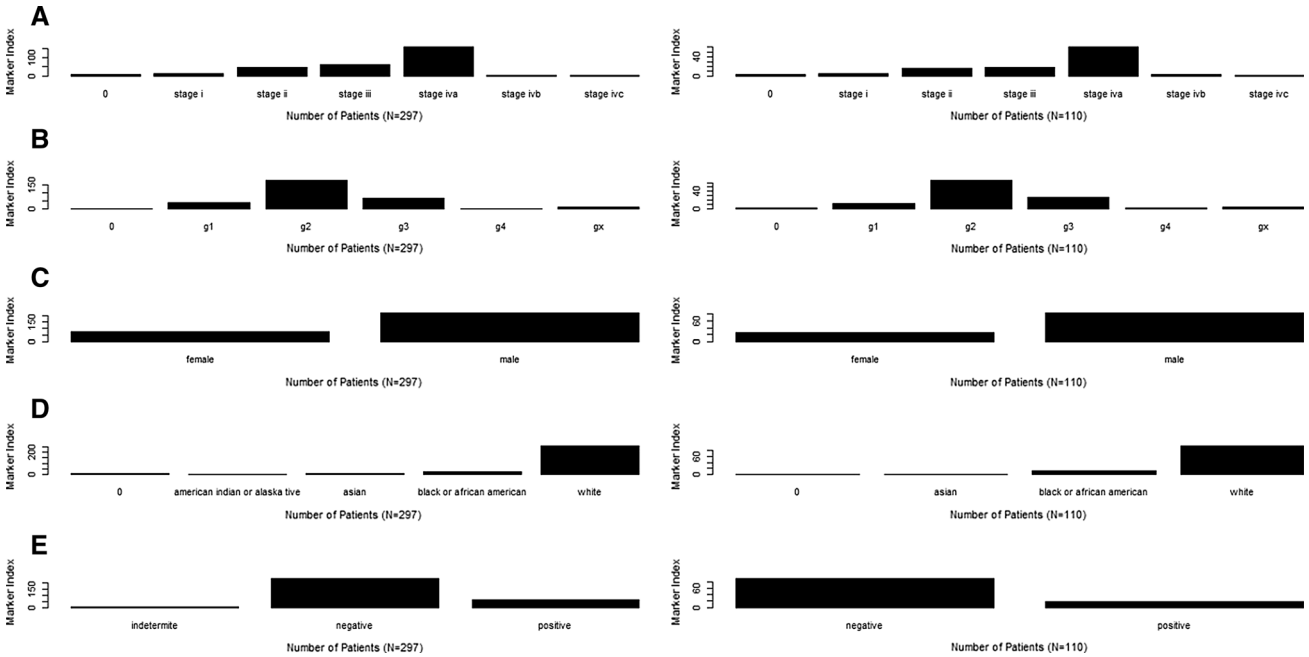

Fig. 3 Comparison of six-gene expression index with stage (**a**), grade (**b**), sex (**c**), race (**d**) and HPV status (**e**) in tumor (left panel) and tumor-free patients (right panel)

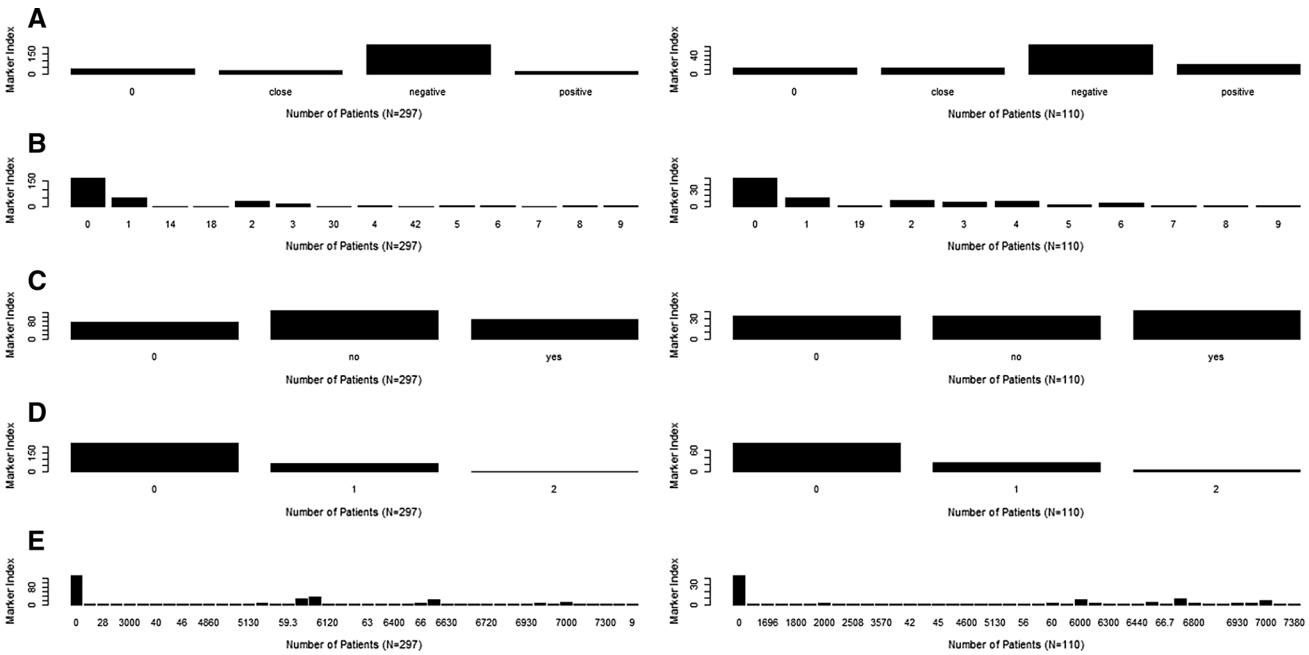

Fig. 4 Comparison of six-gene expression index with neoplasm margin (**a**), lymph nodes (**b**), perineural tissue (**c**), radiation course (**d**) and dosages (**e**) in tumor (left panel) and tumor-free patients (right panel)

2, and it binds to intronic polypyrimidine tracts and mediates negative regulation of exons splicing. It may antagonize in a tissue-specific manner the ability of NOVA1 to activate exon selection. In addition to its function in pre-mRNA splicing, it also plays a role in the regulation of translation.

Superiority of the RNA Prognostic Biomarker Model

Different types of RNAs have been reported as molecular prognostic biomarkers for different cancers. Our results with AUC of > 0.65 demonstrate superiority of the RNA prognostic biomarker model over other prognostic

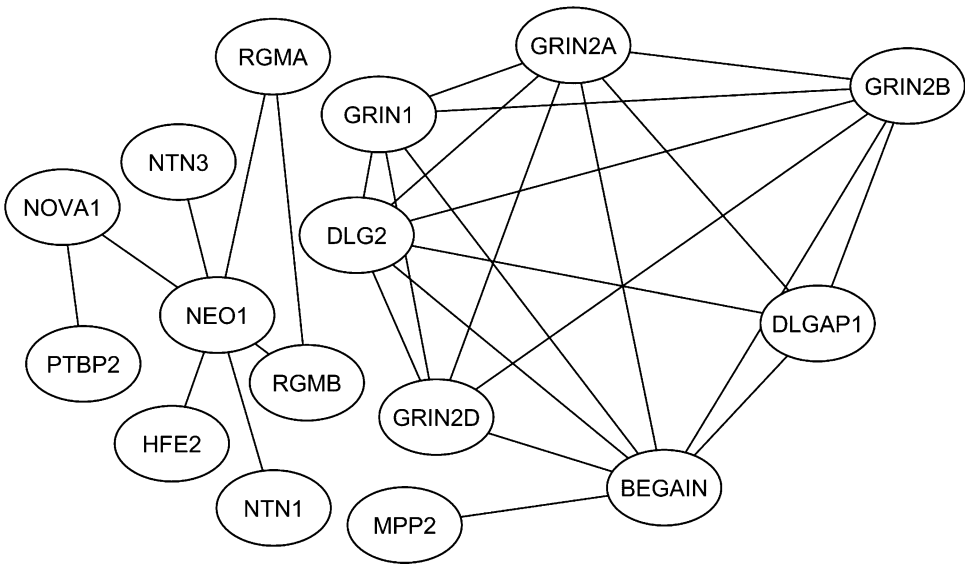

Fig. 5 Interactome analysis of biomarker proteins and its first-degree interacting partners

biomarkers or their combinations in prediction accuracy of patient survival status. We further did a Kaplan–Meier survival (KM) curve comparing survival probability of patients with high six-gene expression index in tumor and tumor-free patients. Results displayed a significant survival time difference between the two groups in both datasets ($P < 0.001$) (Fig. 6).

Discussion

Among diverse treatment measures for head and neck cancer, medication remains the relatively important part. However, it is usually accompanied with adverse events. In our study, head and neck cancer prognostic biomarker model constituted by few RNAs was established to stratify the risk hazards for these patients. Also the medical staff using this model could administer proper therapy regimens to the patients with different risk grades depending on the risk score and would sufficiently minimize the blind administration of anticancer drugs and reduce adverse events, improving the life quality of patients.

Six molecular biomarkers were encompassed in the current prognostic biomarker model: *RGMA, KLHL14, DLG2, NOVA1, KRTAP5-8* and *C1orf190. RGMA* gene is a member of the repulsive guidance molecule (RGM) family that performs several functions in the developing and adult nervous system. It regulates cephalic neural tube closure and inhibits neurite outgrowth and cortical neuron

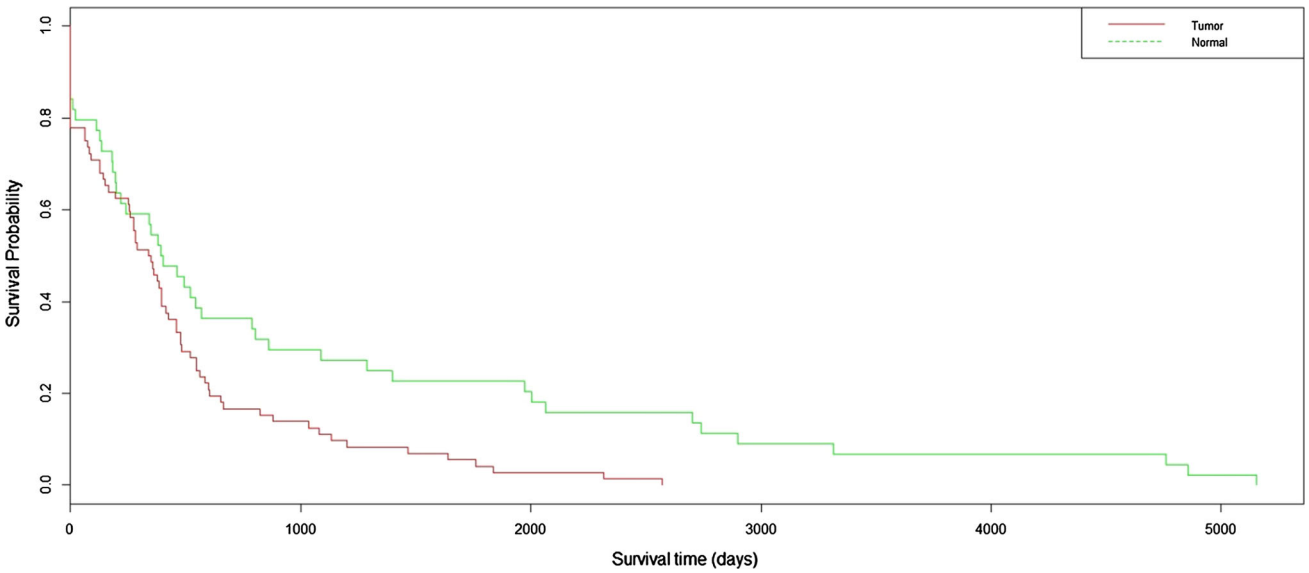

Fig. 6 Kaplan–Meier survival (KM) curve comparing survival probability of patients with high six-gene expression index in tumor and tumor-free patients (P value < 0.001)

branching, and the formation of mature synapses. Binding to its receptor neogenin induces activation of RHOA-ROCK1/Rho-kinase signaling pathway through UNC5B-ARHGEF12/LARG-PTK2/FAK1 cascade, leading to collapse of the neuronal growth cone and neurite outgrowth inhibition [10]. *DLG2* disks, large homolog 2 is required for the perception of chronic pain through NMDA receptor signaling, regulates surface expression of NMDA receptors in dorsal horn neurons of the spinal cord and interacts with the cytoplasmic tail of NMDA receptor subunits as well as inward rectifying potassium channels. It is also involved in the regulation of synaptic stability at cholinergic synapses [11]. *NOVA1* or neuro-oncological ventral antigen 1 regulates RNA splicing or metabolism in a specific subset of developing neurons, *KRTAP5-8* gene forms keratin-associated protein in the hair cortex, and hair keratin intermediate filaments are embedded in an interfilamentous matrix, consisting of hair keratin-associated protein (KRTAP), which are essential for the formation of a rigid and resistant hair shaft through their extensive disulfide bond cross-linking with abundant cysteine residues of hair keratins [12, 13]. *LURAP1* or leucine-rich adaptor protein 1 is activator of the canonical NF-kappa-B pathway and drives the production of proinflammatory cytokines. It promotes the antigen (Ag)-presenting and priming function of dendritic cells via the canonical NF-kappa-B pathway. In concert with *MYO18A* and *CDC42BPA*, it is involved in modulating lamellar actomyosin retrograde flow that is crucial to cell protrusion and migration [14].

The human immune system is highly antigen-specific that can precisely differentiate the normal and malignant ones and identify the "non-self" molecules or cells with high sensitivity and specificity. Therefore, tumor immunotherapy has gained increasing attention for intensive research on tumor therapy. Enrichment pathway analysis demonstrated that the primary pathway relevant to this study was B cell activation. Thus, the antitumor therapy strategy could achieve optimal clinical efficacy by activating, repairing, modifying or even rebuilding patient antitumor B cellular immune response. Nonetheless, current immunotherapy faces some challenges; for instance, determination of the duration of immunotherapy, selection of the therapy target, screening for individual variations and combinatory therapy regimen are yet to be elucidated. The current results would provide insights for establishing a successful immunotherapy for liver cancer with a precise therapeutic strategy.

Conclusion

In this study, we contributed an RNA combination prognostic biomarker model for head and neck cancer patient risk stratification using the RNAs picked out from training dataset by a significant correlation with survival time. The evaluation by Kaplan–Meier survival analysis and receiver operating characteristic (ROC) curve [15–23] showed that our prognostic model presented very good property in most of the situations. Armed with this model, we can recognize the high-risk patients immediately after the operative treatment accurately and combination of this model with current treatment measures is expected to greatly improve the patients' prognosis.

Author Contributions AS and SP contributed to the conception and design, Cox proportional hazard regression analysis, Kaplan–Meier analysis and ROC analysis of the data, as well as the drafting of the manuscript. All authors read and approved the final paper.

Compliance with Ethical Standards

Conflict of interest The author reports no conflicts of interest in this work.

References

1. Siegel R, Miller K, Jemal A (2015) Cancer statistics. Cancer J Clin 65(1):29
2. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ (2009) Cancer statistics. Cancer J Clin 59(4):1–25
3. Price KAR, Cohen EE (2012) Current treatment options for metastatic head and neck cancer. Curr Treat Options Oncol 13(1):35–46
4. Guo W, Chen X, Zhu L, Wang Q (2017) A six-mRNA signature model for the prognosis of head and neck squamous cell carcinoma. Oncotarget 8(55):94528–94538
5. Detector performance analysis using ROC curves—MATLAB & Simulink example. www.mathworks.com. Retrieved 11 Aug 2016
6. Cortes C, Vapnik VN (1995) Support-vector networks. Mach Learn 20(3):273–297. https://doi.org/10.1007/BF00994018
7. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44–57
8. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J et al (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43(Database issue):D447–D452. https://doi.org/10.1093/nar/gku1003
9. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504. https://doi.org/10.1101/gr.1239303
10. Xu X, Gao Y, Shan F, Feng J (2016) A novel role for RGMa in modulation of bone marrow-derived dendritic cells maturation induced by lipopolysaccharide. Int Immunopharmacol 33:99–107. https://doi.org/10.1016/j.intimp.2016.02.008
11. Shao YW, Wood GA (2018) Cross-species genomics identifies DLG2 as a tumor suppressor in osteosarcoma. Oncogene. https://doi.org/10.1038/s41388-018-0444-4
12. Ludlow AT, Wong MS (2018) NOVA1 regulates hTERT splicing and cell growth in non-small cell lung cancer. Nat Commun. https://doi.org/10.1038/s41467-018-05582-x

13. Berens EB, Sharif GM (2017) Keratin-associated protein 5–5 controls cytoskeletal function and cancer cell vascular invasion. Oncogene 36(5):593–605. https://doi.org/10.1038/onc.2016.234
14. Shiraishi A, Uruno T (2017) DOCK8 protein regulates macrophage migration through Cdc42 protein activation and LRAP35a protein interaction. J Biol Chem 292(6):2191–2202. https://doi.org/10.1074/jbc.M116.736306
15. Pawar S, Davis CD, Rinehart CA (2011) Statistical analysis of microarray gene expression data from a mouse model of toxoplasmosis. BMC Bioinform 12(Suppl 7):A19
16. Lahiri C, Pawar S, Sabarinathan R, Ashraf MI, Chand Y, Chakravortty D (2014) Interactome analyses of Salmonella pathogenicity islands reveal SicA indispensable for virulence. J Theor Biol 363:188–197
17. Pawar P, Donthamsetty S, Pannu P, Rida P, Ogden A, Bowen N, Osan R, Cantuaria G, Aneja R (2014) KIFCI, a novel putative prognostic biomarker for ovarian adenocarcinomas: delineating protein interaction networks and signaling circuitries. J Ovarian Res 7(1):53
18. Ashraf MI, Ong SK, Mujawar S, Pawar P, More P, Paul S, Lahiri C (2018) A side-effect free method for identifying cancer drug targets. Sci Rep 8(1):6669. https://doi.org/10.1038/s41598-018-25042-2
19. Pawar S, Ashraf MI, Mujawar S, Mishra R, Lahiri C (2018) In silico identification of the indispensable quorum sensing proteins of multidrug resistant proteus mirabilis. Front Cell Infect Microbiol 8:269. https://doi.org/10.3389/fcimb.2018.00269
20. Pawar S, Ashraf M, Mehata K, Lahiri C (2017) Computational identification of indispensable virulence proteins of Salmonella Typhi CT18. Curr Top Salmonella Salmonellosis. https://doi.org/10.5772/66489
21. Mittal K, Choi DH, Klimov S, Pawar S, Kaur R, Mitra AK, Gupta MV, Sams R, Cantuaria G, Rida PC, Aneja R (2016a) A centrosome clustering protein, KIFC1, predicts aggressive disease course in serous ovarian adenocarcinomas. J Ovarian Res 9:17
22. Lahiri C, Shrikant P, Sabarinathan P, Ashraf MI, Chakravortty D (2012) Identifying indispensable proteins of the type III secretion systems of Salmonella enterica serovar Typhimurium strain LT2. BMC Bioinform 13 (S12)
23. Mittal K, Choi DH, Klimov S, Pawar S, Kaur R, Mitra A, Gupta MV, Sams R, Cantuaria G, Rida PCG, Aneja R (2016b) Evaluation of centrosome clustering protein KIFC1 as a potential prognostic biomarker in serous ovarian adenocarcinomas. J Clin Oncol 34(15_suppl):e17083–e17083