Oral Oncology 85 (2018) 44–51

Contents lists available at ScienceDirect

Oral Oncology

journal homepage: www.elsevier.com/locate/oraloncology

Three distinct genomic subtypes of head and neck squamous cell carcinoma associated with clinical outcomes

Dong Jin Lee[a,b,*], Young-Gyu Eun[a,c], Young Soo Rho[b], Eui Hyun Kim[a,d], Sun Young Yim[a,e], Sang Hee Kang[a,f], Bo Hwa Sohn[a], Gee Hwan Kwon[b], Ju-Seog Lee[a]

[a] Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States
[b] Department of Otolaryngology-Head and Neck Surgery, Hallym University Medical Center, Seoul, Republic of Korea
[c] Department of Otolaryngology-Head and Neck Surgery, School of Medicine, Kyung Hee University, Seoul, Republic of Korea
[d] Department of Neurosurgery, Severance Hospital, Brain Tumor Center, Yonsei University College of Medicine, Seoul, Republic of Korea
[e] Division of Gastroenterology and Hepatology, Department of Internal Medicine, Korea University College of Medicine, Seoul, Republic of Korea
[f] Department of Surgery, Korea University College of Medicine, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Molecular subtypes
Pan-SCCs
HNSCCs
TCGA
RNAseq

ABSTRACT

Objectives: Heterogeneity of head and neck squamous cell carcinomas (HNSCCs) results in unpredictable outcomes for patients with similar stages of cancer. Beyond the role of human papilloma virus (HPV), no validated molecular marker of HNSCCs has been established. Thus, clinically relevant molecular subtypes are needed to optimize HNSCC therapy. The purpose of this study was to identify subtypes of HNSCC that have distinct biological characteristics associated with clinical outcomes and to characterize genomic alterations that best reflect the biological and clinical characteristics of each subtype.

Materials and methods: We analyzed gene expression profiling data from pan-SCC tissues including cervical SCC, esophageal SCC, lung SCC, and HNSCC (n = 1346) to assess the similarities and differences among SCCs and to identify molecular subtypes of HNSCC associated with prognosis. Subtype-specific gene expression signatures were identified and used to construct predictive models. The association of the subtypes with prognosis was validated in two independent cohorts of patients.

Results: Pan-SCC analysis identified three novel subtypes of HNSCC. Subtype 1 had the best prognosis and was similar to cervical SCC, whereas subtype 3 had the worst prognosis and was similar to lung SCC. Subtype 2 had a moderate prognosis. The 600-gene signature associated with the three subtypes significantly predicted prognosis in two independent validation cohorts. These three subtypes also were associated with potential benefit of immunotherapy.

Conclusion: We identified three clinically relevant HNSCC molecular subtypes. Independent prospective studies to assess the clinical utility of the subtypes and associated gene signature are warranted.

Introduction

Head and neck squamous cell carcinoma (HNSCC) is usually classified according to pathological features and location, which are used to predict prognosis [1]. However, even among tumors with the same clinical stage and location, some are indolent and progress slowly, whereas others are aggressive and progress quickly. Traditionally, clinicopathological features, such as extracapsular nodal spread, positive margins, multiple positive nodes, or perineural/vascular invasion, have been used as prognostic factors in HNSCC [2]. But the clinical use of these factors is limited, especially in patients who undergo chemotherapy, radiation therapy, or targeted therapy without surgical treatment.

Genomic analysis has led to the proposal of four molecular subtypes of HNSCCs—atypical, basal, classical, and mesenchymal—that share some molecular features with lung squamous cell carcinoma (LSCC) [3]. Molecular features of these four subtypes include deregulation of the KEAP1/NFE2L2 oxidative stress pathway, differential utilization of the lineage markers of SOX2 and TP63, and preference for the oncogenes *PIK3CA* and *EGFR*. However, in contrast to subtypes in LSCC that effectively reflect patient prognosis [4], these four subtypes are not associated with prognosis in HNSCC [3], indicating the need to identify

* Corresponding author at: Department of Systems Biology, Unit 1058, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030, United States.
E-mail address: jlee@mdanderson.org (D.J. Lee).

https://doi.org/10.1016/j.oraloncology.2018.08.009
Received 28 January 2018; Received in revised form 7 June 2018; Accepted 14 August 2018

clinically relevant molecular subtypes of HNSCC.

A large-scale genomic study by The Cancer Genome Atlas (TCGA; Bethesda, MD) identified molecularly distinct subtypes of several types of cancer that reflect clinical differences [5–13]. These data have been used to analyze genomic data from different cancer types to identify shared genetic features among different cancers [14–17]. In this study, we analyzed genomic data from four different types of squamous cell carcinoma (SCC) of different origins (HNSCC, esophageal SCC [ESCC], LSCC, and cervical SCC [CSCC]) to explore their molecular similarities and differences. Unexpectedly, we discovered three novel HNSCC subtypes; two of these subtypes have genomic and molecular characteristics that are similar to CSCC and LSCC, respectively. Our findings may provide clinically relevant insights into the molecularly different subtypes of HNSCC.

Materials and methods

Genomic and clinical datasets

TCGA genomic data of four SCCs (HNSCC, ESCC, LSCC, and CSCC) were obtained from the data portal (https://tcga-data.nci.nih.gov) and cancer browser (https://genome-cancer.ucsc.edu). Gene-level expression data from mRNA-seq (n = 1346), copy number variation data (n = 1308), somatic mutation data (n = 795), and clinical data (n = 1346) were included in our analyses. Clinical data included survival data, sex, age, TNM stage, primary sites, HPV status, alcohol and smoking habits, and margin status (Appendix data 1 and 2). Samples were classified as HPV-positive using an empiric definition of detection of > 1000 mapped RNA-Seq reads, primarily aligning to viral genes E6 and E7 [5]. All genomic and clinical data used in this study were data released in February 2, 2015.

Analysis of the gene expression data and unsupervised clustering

BRB-ArrayTools (http://linus.nci.nih.gov/BRB-ArrayTools.html) software was used to analyze gene expression data [18]. ConsensusClusterPlus (Bioconductor) [19] was used to perform unsupervised clustering of gene expression data (6856 genes, 2-fold difference in at least 134 cases relative to the median value across tissues) from 1346 pan-SCC tumors and to find the optimal number of clusters (Supplementary Fig S1). A heatmap was generated using the Cluster and TreeView programs [20]. Other statistical analyses were performed using the R language environment (http://www.r-project.org). To select genes that were differentially expressed between subtypes, we performed multiple two-sample *t*-tests for all possible combinations of the three subtypes with a stringent significance cutoff of $P < 0.001$ and 1.5-fold difference. Genes were then ranked according to fold-ratios, and the top 200 genes were selected for each subtype (Supplementary Fig S2 and Appendix data 3). Pathway analysis was carried using Ingenuity Pathways Analysis (IPA, Ingenuity, Redwood City, CA, USA). Genes associated with canonical pathway in the Ingenuity Pathways Knowledge Base were considered for analysis. The significance of association between 200 genes of each group and the canonical pathway was measured using Fischer's exact test ($P < 0.001$)

Prediction models with genomic signatures

The gene expression signature from the TCGA cohort was used to stratify patients with HNSCC from two validation cohorts from the Gene Expression Omnibus database: GSE39366 [3] and GSE65858 [21]. Expression data from 200 subtype-specific genes in the TCGA set were combined to form a classifier according to a Bayesian compound covariate predictor (BCCP), as described previously [22–26]. The BCCP classifier estimated the likelihood of an individual patient being in one of three subtypes. Briefly, gene expression data for each subtype gene signature from the TCGA cohort (i.e., the 200 significant genes for each subtype) were used to generate the Bayesian probability of each tissue sample belonging to a particular subtype, generating three probability scores for each tumor. Samples in the validation cohorts were assigned to one of the three subtypes according to the highest probability scores. We used equal prior probability option with 0.5 of uncertainty threshold to decide whether the case was subtype 1 or non-subtype1 (subtype 2 or subtype 3). We also performed this analysis for subtype 2 and subtype 3. As for HPV state assessment in validation cohorts, in situ hybridization was used in GSE39366 [3] and both HPV16 DNA status and HPV16 RNA status were detected using analysis of E6 transcripts by RT-PCR in GSE65858 [21].

The potential response of each patient to immunotherapy was estimated by a previously established immune signature score predictor with 105 gene [27]. The immune signature score ranged from 0 to 1, and 0.5 was used as the cutoff for potential responders (> 0.5) and non-responders (< 0.5).

Analysis of copy number alteration and somatic mutation

Multiple two-sample *t*-tests were performed for all possible combinations of the three subtypes to select subtype-specific genes with copy-number alterations. Of the 127 most frequently mutated cancer genes in 12 cancer types identified in a previous study [28], the most frequently mutated genes in HNSCC (21 genes) were selected for analysis. Somatic mutation data were analyzed and visualized using OncoPrint (https://cbioportal.org).

Statistical analysis

The association of each subtype with overall survival (OS) and recurrence-free survival (RFS) in the TCGA cohort and validation cohorts was assessed by Kaplan-Meier plots and the log-rank test. OS was defined as the time from surgery to death, and RFS was defined as the time from surgery to the first confirmed recurrence. Data were censored for patients who were alive without recurrence at the last follow-up. *P* values < 0.05 were considered statistically significant.

Results

Analysis of pan-SCC genomic data and novel molecular subtypes of HNSCC

To explore the molecular profiles of SCC, we carried out unsupervised clustering analysis with mRNA expression data from four different types of SCC, including HNSCC, ESCC, LSCC, and CSCC (Appendix data 1). As expected, most tumors were clustered together according to the origin of the tumors (Fig. 1a). Most CSCCs were in cluster 1, whereas most HNSCCs were in cluster 2. LSCCs and ESCCs were in cluster 3 (Fig. 1a and Supplementary Table S1), suggesting potential similarities in the underlying biology between LSCC and ESCC. Subsets of each tumor type were grouped with other tumor types. For example, a subset of CSCCs was clustered with LSCC, whereas a small subset of LSCC was clustered with HNSCC, indicating molecular heterogeneity within each tumor type. Interestingly, a substantial subset of HNSCCs was clustered with either CSCC or LSCC, indicating that these HNSCCs share more molecular similarity with other tumor types than most HNSCCs (Supplementary Table S1).

As expected, the clinical outcomes of patients in the three clusters were significantly different and reflected the different tumor types in each cluster. LSCC had a typically worse prognosis than other SCCs [29] (Fig. 1b). Surprisingly, the prognostic significance of the three clusters remained the same even when only HNSCCs were analyzed (Fig. 1c). The OS rate of patients with HNSCC in cluster 1 was similar to that of patients with CSCC, whereas the OS rate of patients with HNSCC in cluster 3 was similar to that of patients with LSCC. It is important to point out that the distribution of clinical stages in the three clusters was not significantly different (Table 1, Appendix data 2), suggesting that