Oral Oncology 78 (2018) 72–79

Contents lists available at ScienceDirect

Oral Oncology

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

Identification of a gene expression signature predicting survival in oral cavity squamous cell carcinoma using Monte Carlo cross validation

John Schomberg, Argyrios Ziogas, Hoda Anton-Culver*, Trina Norden-Krichmar

Department of Epidemiology School of Medicine, University of California, Irvine, Irvine, CA 92617, United States

ARTICLE INFO

Keywords:
Oral cancer
Head and neck cancer
Molecular signature
Cancer survival
Treatment response
Gene signature
Oral cancer pathways
Chemotherapy
Monte Carlo cross validation

ABSTRACT

Objectives: This study aims to identify a robust signature that performs well in predicting overall survival across tumor phenotypes and treatment strata, and validates the application of Monte Carlo cross validation (MCCV) as a means of identifying molecular signatures when utilizing small and highly heterogeneous datasets.

Materials and methods: RNA sequence gene expression data for 264 patient tumors were acquired from The Cancer Genome Atlas (TCGA). 100 iterations of Monte Carlo cross validation were applied to differential expression and Cox model validation. The association between the gene signature risk score and overall survival was measured using Kaplan-Meier survival curves, univariate, and multivariable Cox regression analyses.

Results: Pathway analysis findings indicate that ligand-gated ion channel pathways are the most significantly enriched with the genes in the aggregated signature. The aggregated signature described in this study is predictive of overall survival in oral cancer patients across demographic and treatment strata.

Conclusion: This study reinforces previous findings supporting the role of ion channel gating, interleukin, calcitonin receptor, and keratinization pathways in tumor progression and treatment response in oral cancer. These results strengthen the argument that differential expression of genes within these pathways reduces tumor susceptibility to treatment. Conducting differential gene expression (DGE) with Monte Carlo cross validation, as this study describes, offers a potential solution to decreasing the variability in DGE results across future studies that are reliant upon highly heterogeneous datasets. This improves the ability of studies reliant upon similarly structured datasets to reach results that are reproducible.

Introduction

Head and neck cancers are cancers of the upper airway and/or digestive tract found in the oral cavity, laryngeal, pharyngeal, oropharyngeal, and hypo-pharyngeal tissues. Head and neck cancers make up 3% of cancers diagnosed each year [1,2]. Head and neck cancer incidence has declined from 25 cases per 100,000 at risk in the 1990s to 15 cases per 100,000 at risk in the present day [3]. While the decrease in head and neck cancer incidence may be due to a drop in tobacco use [4,5], the mortality associated with these cancers has not changed significantly in the last twenty years [6]. Human Papilloma Virus (HPV) positive patients have been observed to have an improved survival and response to treatment when compared to HPV negative patients. However, these patients make up the minority of oral squamous cell cancers (OSCC) [7]. Thus, the decline in mortality could be attributed to decrease in smoking, increases in HPV positive cases, or other unknown mechanisms.

Few studies have identified a group of genes predicting treatment response in HPV-negative OSCC patients. To date, the most widely used molecular signature guiding head and neck squamous cell carcinoma (HNSCC) treatment is HPV status. HPV status can be measured directly through polymerase chain reaction analysis, or indirectly through cyclin-dependent kinase inhibitor 2A (CDKN2A) expression. However, HPV preferentially infects oropharyngeal tissues which make up only 15% of HNSCC [8]. There have been multiple studies that have identified the genetic markers that improve prediction of overall survival when HPV status is known [9–12]. Unfortunately, there has been less focus on HPV-negative OSCC patients, an HNSCC subgroup that is known to respond significantly worse to treatment than patients with Oropharyngeal Squamous Cell Carcinoma (OPSCC) [9,12]. OSCC patients have been shown to be less likely to be HPV positive than Oropharyngeal cancer patients and thus are more reflective of the outcomes of HPV negative patients.

Past studies examining molecular signatures in OSCC have found that pathways in cell migration, cell-to-cell signaling and interaction, and cellular growth and proliferation are predictive of overall survival

* Corresponding author.
E-mail address: hantoncu@uci.edu (H. Anton-Culver).

https://doi.org/10.1016/j.oraloncology.2018.01.012
Received 31 October 2017; Received in revised form 20 December 2017; Accepted 18 January 2018
Available online 20 February 2018
1368-8375/

[13,14]. The keratin pathway is also notable in that it has been identified by several studies for its role in predicting the conversion of leukoplakia to malignant tumor, tumor progression, nodal stage, and overall response to treatment [15]. Of the OSCC studies listed the largest sample was 130 patients [14]. A common theme among the reported studies is low reproducibility in the genes identified as predictive of advanced disease or survival.

There has been much success in the production of site specific predictive models that draw upon the rich resource of data in the TCGA [16]. Models predicting survival in glioblastoma, colorectal, ovarian, and even head and neck cancer have drawn upon TCGA data in the past [17–21]. The 2015 study examining head and neck cancer data in the TCGA focused on gene mutations that were observed across all head and neck cancer patients and in those patients that tested HPV positive. While this study did describe treatment response, it did not utilize gene expression data when conducting survival analyses. This study does draw upon gene expression data in the TCGA to produce an aggregated model that predicts survival across strata of tumor behavior, treatment regimen, and gender.

There are a host of methods that can be applied in the identification of a predictive molecular signature. When composing a signature that is predictive and prognostic, there are several quality checkmarks that must be addressed. Model building of any kind must go through an internal validation process where data is divided between test and training data. While model simplicity or complexity improve model usability, they are superseded in importance by measures of model performance [22]. Internal validation is an acceptable form of validation only when the test data set is completely untouched and no aspect of test data plays a part in model development. A drawback to splitting data in this way is the decrease in model efficiency due to the use of only a subset of the total data. One method addressing this inefficiency is to split a dataset into training and test data many times in a Monte Carlo validation (MCCV) or leave-one-out cross validation. These methods lead to nearly unbiased estimates of model performance (in the case of leave-one-out cross validation), and do not require sacrifice of sample size [23,24]. These methods have been applied by other studies in the successful identification of predictive models in many different types of cancer using leave-one-out cross validation [25–29] and MCCV [30,31]. The application of MCCV involves random sampling without replacement which means that subsets of the population with gene expression values with strong effect have a greater opportunity to have that effect detected. MCCV differs from k-folds cross validation in that in MCCV an observation may be chosen to be included in a test set multiple times over the total number of iterations over all analyses opposed to one time in K-fold validation. MCCV is also viewed as a more conservative approach to cross validation as it overestimates the model prediction error in comparison to a k-fold cross validation which tends to underestimate prediction error [32]. External validation is an important and often costly task required for measuring a model's exportability. It is for this reason that robust internal validation measures should be adopted by those studies that lack the funding to carry out external validation in early stages of analysis.

Methods

Datasets

The Cancer Genome Atlas (TCGA) is a large, multi-dimensional, multi-center project compiling genomics data for over 29 cancer types into one central database [33]. TCGA contains clinical and demographic variables, gene expression profiling data, SNPs, protein expression, and methylation data. Clinical data on radiation dose, demographic variables, exposures (tobacco, alcohol, and HPV), chemotherapy type, and measures of overall and disease progression-free survival are included in the TCGA database (Table 1). Data accessed for this study were publicly available through the TCGA genomic data commons data portal. 523 head and neck cancer cases were downloaded from the data portal with all corresponding gene expression counts and corresponding clinical data. Of these 523 patients 313 OSCC patients were selected. 264 of the remaining 313 OSCC patients were included as only these patients possessed full survival data.

Table 1
Patient demographics stratified by low and high risk molecular signature.

Characteristics	ALL (264, 100%)	Low Risk (n = 151, 57%)	High Risk (n = 113, 42%)
Vital status			
Alive	189	(130, 86.6%)	(59, 52.2%)
Deceased	75	(21, 13.9%)	(54, 47.7%)
Age			
Age greater than 60	152	(85, 56.2%)	(67, 59.2%)
Age less than 61	112	(66, 43.7%)	(46, 40.7%)
Gender			
Female	88	(55, 36.4%)	(33, 29.2%)
Male	176	(96, 63.5%)	(80, 70.7%)
Tumor grade			
G1	34	(19, 12.6%)	(15, 13.3%)
G2	153	(92, 61.3%)	(61, 54.4%)
G3	59	(29, 19.3%)	(30, 26.7%)
G4	5	(3, 2.0%)	(2, 1.7%)
GX	11	(7, 4.6%)	(4, 3.5%)
Race			
White	224	(127, 86.3%)	(97, 88.1%)
Not White	33	(20, 13.6%)	(13, 11.8%)
Clinical stage			
Stage I	8	(4, 2.7%)	(4, 3.6%)
Stage II	57	(28, 19.1%)	(29, 26.1%)
Stage III	58	(38, 26.0%)	(20, 18.0%)
Stage IVA	126	(71, 48.6%)	(55, 49.5%)
Stage IVB	6	(4, 2.7%)	(2, 1.8%)
Stage IVC	2	(1, 0.6%)	(1, 0.9%)
Alcoholic Drinks > 2 consumed per day			
TRUE	57	(37, 50.6%)	(20, 41.6%)
FALSE	64	(36, 49.3%)	(28, 58.3%)
History of smoking			
TRUE	195	(113, 74.8%)	(82, 72.5%)
FALSE	69	(38, 25.1%)	(31, 27.4%)
Tumor Necrosis Greater than or equal to 15%			
TRUE	115	(60, 41.0%)	(55, 50.4%)
FALSE	140	(86, 58.9%)	(54, 49.5%)
Radiation > 66 Gy			
TRUE	23	(14, 11.4%)	(9, 9.2%)
FALSE	196	(108, 88.5%)	(88, 90.7%)
Receiving chemotherapy			
TRUE	95	(60, 39.7%)	(35, 30.9%)
FALSE	169	(91, 60.2%)	(78, 69.1%)

"Chemotherapy" is not specific to a given chemotherapeutic agent. This merely reflects whether a patient was assigned to chemotherapy treatment or not. History of Smoking stratifies patients into "never" or "ever" smokers. High Grade includes G1 and G2 patients, while low grade includes G3, G4, GX tumor grades. Not all characteristics total to 264 as some variables were incomplete (Tumor Grade NA = 2, Clinical Stage NA = 7, alcohol consumption per day NA = 143, Tumor Necrosis NA = 9, Radiation NA = 45)

Differential expression analyses

Differential Gene Expression (DGE) analysis is a method of identifying genes that are expressed differently across time, tissue, and conditions, such as disease states [34]. This method of analysis uses fold change and significance criterion to select the genes in a molecular signature for predicting tumor phenotype, clinical subtype, or treatment response. All patients with cancer in tongue, lip, alveolar ridge, hard palate, floor of mouth, maxilla, and buccal mucosa were included. OSCC patients were the largest grouping of head and neck cancer patients and thus provided the most power to detect influential genetic