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

Objective: 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 the 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.
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 checkpoints 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 data, SNPs, protein expression, and methylation data. Of these 523 patients 313 OSCC patients were the largest grouping of head and neck cancer patients and thus provided the most power to detect in high risk molecular signature.

**Table 1**

| 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, 40.0%)          | (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, 88.6%)        | (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 100% 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.

**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
pathways predicting treatment response. DGE analysis yielded a list of genes that are expressed differently between two strata. The strata used in this study were vital status within five years of follow-up. The TCGA RNA sequencing data were preprocessed with RSEM software, yielding normalized counts per million (CPM) gene expression counts [16]. The data were filtered to CPM ≥ 2, absolute fold change ≥ 1.5, Fisher’s exact p-value ≤ 0.05 and a false discovery rate ≤ 0.05. The list of genes produced by these filters was used to create a predictive signature comprised of 20 genes selected by the highest absolute fold change value.

100 Runs of differential gene expression analysis using Monte Carlo validation

A defining feature of MCCV is the random selection of observations into test and training sets across multiple iterations [35]. This study did require that some randomness be sacrificed, as a constant proportion of living and deceased were included at each iteration (opposed to a random proportion) to ensure that Cox regression survival analyses could be conducted. DGE analysis was repeated 100 times with a randomly selected (without replacement) set of 100 patients from the 264 total patients. Of the 100 patients selected in each iteration, 66 survived past 5 years and 34 were deceased prior to 5 years. At each iteration the top 20 genes with highest absolute fold change value were chosen to be placed in an additive Cox regression model predicting overall survival in OSCC patients. An AUC was produced for each of the signatures (comprised of 20 genes) created at each of the 100 iterations using the remaining 164 patients as a test set. The selected genes were aggregated to yield a table counting the number of times each gene met filter criteria over all the 100 iterations (Supplemental Table 1). 100 iterations is double the number of iterations used in previous studies applying MCCV for similar purposes [36,37]. The number of genes within the final model was set at 40 to produce more robust estimates of survival than those models with 20 genes. The number of genes included in the signature did not exceed 40 as the model would not converge properly due to sample size restrictions. This application of MCCV has been used in the past to identify genetic predictors of disease status in breast cancer and Parkinson’s disease [36,37]. This study applies a similar method to identify those gene expression patterns that exert the greatest influence in predicting treatment response in OSCC.

The final aggregated model was comprised of counts per million for each gene in the final aggregated signature multiplied by a model weight. Once all 40 of the weighted CPM were summed across all 40 genes a risk score would be generated indicating whether a patient would be set into high (> 1.5) of low (≤ 1.5). The cutoff for high risk and low risk was set as the minimum difference between sensitivity and specificity on the ROC curve. This minimum value was identified using the PROC package in R [38].

In order to provide additional assurance that these results were not reached by chance alone, the study repeated the 100 signature validations using genes that were randomly selected from the 20,530 genes in the dataset. The distribution of AUC across 100 runs of signatures based upon DGE analysis results were compared to the distribution of AUC derived from signatures comprised of genes that were randomly selected. To visualize these results, histograms were created by binning AUC by frequency (Supplemental Fig. 1).

Sensitivity of the aggregated signature

The sensitivity of the aggregated signature was validated by applying it to clinical subsets of all 264 test patients. Kaplan-Meier survival curves were used for this series of validation. Cox regression was used to determine the sensitivity of the aggregated signature when other variables were in the model. The Cox model included race, gender, chemotherapy treatment, and tumor grade. Alcohol consumption and radiation variables were run in the model with dummy variables to address the effect of large amount of missingness within these variables (145 missing variables in alcohol consumption, 45 missing variables for radiation dose). These variables were not found to have a significant impact on the estimates produced for the high risk scores and were excluded from the final cox regression model. Tumor Necrosis was excluded from the model due to the high amount of correlation with the gender variable which led to unstable estimates (There were no female patients with tumor necrosis < 15% present in the sample). Clinical stage was not included within this analysis due to the improved fit offered by the tumor grade variable, and both clinical stage and tumor grade were found to be nonsignificant when included within the model. Similar results for both tumor grade and clinical stage are not unexpected as tumor grade is a component of the clinical staging criteria. All analyses used age as a strata to prevent bias created by any skewness in the distribution of age within each variable. Univariate cox regression was performed to provide context for multivariable analyses (Table 2).

Table 2

| Characteristic          | Univariate |          |          |          |          |          |          |
|-------------------------|------------|----------|----------|----------|----------|----------|----------|
|                         | HR         | 95% CI   | p-value  | HR       | 95% CI   | p-value  |
| No Smoking History      | 0.7        | 0.4-1.2  | 0.24     | 0.6      | 0.1-3.2  | 0.5      |
| Female Gender           | 0.4        | 0.2-0.7  | 0.002    | 0.4      | 0.2-0.07 | 0.004    |
| Tumor                   | 0.7        | 0.5-1.2  | 0.23     | 0.7      | 0.4-1.2  | 0.6      |
| Grade < 2               | 1.0        | 0.4-1.9  | 0.96     | 1.0      | 0.5-2.16 | 0.90     |
| Caucasian Race          | 1.9        | 1.1-3.4  | 0.01     | 1.9      | 1.1-3.5  | 0.01     |
| Chemotherapy Not Received |            |          |          |          |          |          |
| High Risk Signature     | 3.3        | 1.9-5.5  | < 0.0001 | 3.25     | 1.3-6.3  | < 0.0001 |

Pathway enrichment analysis methods

The R packages edgeR, and PA Reactome were used to conduct DGE and pathway enrichment analyses, respectively [39,40]. Pathway analysis tools and annotation databases were used to examine which pathways were enriched with the most frequently identified genes in the signatures produced over one hundred rounds of DGE. It is important to note that false discovery rate (FDR) produced by PA Reactome was not weighted for the frequency we observed genes to be significant over the 100 run DGE analysis, and thus the 0.05 FDR should be considered a conservative threshold. A table of those pathways meeting a Fisher exact p-value threshold of 0.05 was included in the results (Table 3).

Results

Differential gene expression

Each run of the DGE analysis identified differentially expressed genes based upon the gene expression values of randomly selected patients. An AUC reflecting the accuracy of each signature (each comprised of 20 genes) was recorded over 100 runs. These AUCs had a median of 0.84, max of 0.96, minimum of 0.65, mean of 0.83, and a standard deviation of 0.04. Similar analyses were performed on gene signatures of genes randomly selected from the 20,530 genes in the
Pathway analysis produced using Pathway Reactome. The pathways with a p-value less than 0.05 were listed in this table. The false discovery rate (FDR) was also calculated but not shown here. The FDR represents the probability that a gene is significantly enriched in error. The FDR is considered to be a conservative measure of significance, as it is not weighted to adjust for the number of times a gene was identified over 100 runs. Of the pathways listed only the first “Ligand-gated ion channel transport” had an FDR p-value of less than 0.05 (p-value = 3.43E−04).

Table 3
Pathway analysis of aggregated signature.

| Pathway name                                | Number of genes from aggregate signature in pathway | Total number of genes in pathway | Fisher’s exact p-value | Aggregated signature genes found in pathway |
|---------------------------------------------|-----------------------------------------------------|---------------------------------|------------------------|------------------------------------------|
| Ligand-gated ion channel transport          | 2                                                   | 33                              | 2.27E−06               | HTR3C; GLRA4                             |
| Defective pro-SFTPC causes pulmonary surfactant metabolism dysfunction 2 (SMDP2) and respiratory distress syndrome (RDS) | 1                                                   | 2                              | 0.005                  | SFTPC                                   |
| Assembly of active LPL and LIPC lipase complexes | 1                                                   | 30                              | 0.01                   | FGF21                                   |
| Surfactant metabolism                       | 1                                                   | 52                              | 0.01                   | SFTPC                                   |
| Formation of the cornified envelope         | 2                                                   | 130                             | 0.01                   | KRT38; KRT72                            |
| Defective ABCA3 causes pulmonary surfactant metabolism dysfunction type 3 (SMDP3) | 1                                                   | 9                               | 0.02                   | SFTPC                                   |
| Regulation of signaling by NODAL            | 1                                                   | 12                              | 0.03                   | LEFTY2                                  |
| Calcitonin-like ligand receptors            | 1                                                   | 11                              | 0.03                   | CALCRL                                  |
| Plasma lipoprotein remodeling               | 1                                                   | 54                              | 0.03                   | FGF21                                   |
| Class B/2 (Secretin family receptors)       | 2                                                   | 99                              | 0.04                   | CALCRL; GLP2R                           |
| Keratinization                             | 2                                                   | 218                             | 0.04                   | KRT38; KRT72                            |
| POU5F1 (OCT4), SOX2, NANOG repress genes related to differentiation | 1                                                   | 10                              | 0.04                   | CDX2                                    |
| Interleukin-4 and 13 signaling              | 1                                                   | 212                             | 0.04                   | IL17A                                   |

Pathway analysis produced using Pathway Reactome. The “Fisher’s exact p-value” represents the probability that the genes would be selected if they were selected by chance alone. Only pathways with a p-value less than 0.05 were listed in this table. The false discovery rate (FDR) was also calculated but not shown here. The FDR represents the probability that a gene is significantly enriched in error. The FDR is considered to be a conservative measure of significance, as it is not weighted to adjust for the number of times a gene was identified over 100 runs. Of the pathways listed only the first “Ligand-gated ion channel transport” had an FDR p-value of less than 0.05 (p-value = 3.43E−04).

Validation of aggregate signature across tumor phenotypes and clinical strata

This model was applied to subsets of the 264 patient test dataset. When overall survival difference was measured using all patients in the test set, it was found that there was a significant difference in patient survival outcomes when stratifying by the molecular signature risk score (p-value = 2.66E−08) (Fig. 1c). When stratifying by tumor grade, the signature was predictive of survival in those patients with high grade (Greater than G2) tumors and low grade (Less than G3) tumors (p-value 0.0008, 8.8e−06), respectively (Fig. 2a and b).

The log rank survival by molecular signature risk score in only those patients receiving chemotherapy was (p-value = 0.002). The significance of difference by risk score in those patients not receiving chemotherapy was (p-value = 0.002) (Fig. 3a and b). This signature continues to be predictive when all women were removed from the sample and the prediction of survival in men alone was tested (p-value = 9.7e−07). However, this signature was not predictive in women and was found to be only marginally significant (p-value = 0.04) (Fig. 3c and d).

Univariate and multivariable cox regression

After adjusting for confounding variables, the signature risk score continued to be predictive of treatment response in both multivariable and univariate analyses (Table 2). High risk score was associated with an HR of 3.2 (95% CI 1.3 to 6.3, p-value < 0.0001) times greater odds of death when compared to patients with low risk score in a univariate model. No significant effect was discovered when this model was applied to women alone. It was observed that both the signature and tumor necrosis lost effect size when performing multivariable adjustment. As this seemed indicative of possible correlation between the two variables, a Spearman correlation test was applied and yielded a 21% correlation significant with a (p-value = 9e−04). Our results showed that in addition to the signature risk score, gender and chemotherapy treatment were also predictive of overall survival.

Pathway analysis results

Significant pathways enriched with genes in the original signature were Interleukin, Calcitonin, ligand-gated ion channel transport, keratinization, and cornified envelope pathways (Table 3). There were no pathways that were enriched with greater than 2 genes from our signature. The most significantly enriched pathway was the ion channel transport pathway. In total 11 genes from the 40 genes within the aggregated signature were identified as being enriched in the aforementioned pathways. The ligand gated ion channel transport pathway passed both Fisher exact test and false discovery rate thresholds for significant enrichment (Fisher’s exact p-value = 2.3 2−06, False discovery p-value = 3.4 e−04). Genes within the ligand gated ion channel transport pathway were GLRA4 and HTR3C which were identified as significantly differentially expressed in 17% and 13% of the MCCV respective replications. All pathways listed in Table 3 meet a Fisher’s exact p-value of 0.05 or less.

Discussion

Interpretation of signature validation

The aggregated signature was shown to be predictive of treatment response in OSCC patients regardless of chemotherapy treatment status, or tumor grade. In addition to the identification of a signature that predicts overall survival in OSCC patients, this study also validated the use of Monte Carlo cross validation in producing gene signatures that are more likely to be reproduced across multiple studies. This method can be adopted by other researchers that wish to apply free and publicly available data to the testing of hypotheses in a manner that has the greatest likelihood of reproducibility across datasets.
Interpretation of pathway enrichment

The ion gate channel pathway was one of the pathways enriched with genes in the aggregated signature identified in this study. Ion gate channel pathway genes within the aggregated signature that were found to be significantly enriched were 5-Hydroxytryptamine Receptor (serotonin receptor) (HTR3C) and Glycine Receptor Alpha (4GLRA4). HTR3C has also been reported to be associated with other upper GI cancers such as esophageal adenocarcinoma [41]. Other ion channel regulators like voltage-gated potassium channel Kv3.4 mRNA expression have been found to affect the progression of OSCCs, and inhibition of Kv3.4 inhibits growth of OSCC [42–44]. POU5F1, OCT4, SOX2, NANOG gene repression pathways were also found to be significantly enriched. These genes play a role in chemosensitivity to platinum based chemotherapies [45,46]. The keratin pathway is also notable in that it has been identified for its role in predicting the conversion of leukoplakia to malignant tumor [47,48]. The MCCV approach did not detect all pathways typically associated with the development of OSCC. Pathways associated with HPV negative OSCC development include AKT, JNK, IL-6/STAT3, ILK, RAS, MAPK/ERK, p38/PAK, TGFβ, PI3K/mTOR and WNT signaling. The research questions focused upon by this study were which pathways were associated with treatment response. Thus, pathways associated with disease progression were not identified. Evidence of supporting literature is provided (Supplemental Table 2) in a matrix of gene names and search terms related to OSCC, head and neck cancer, and cancer treatment response produced by Pubmatrix [49] (Supplemental Table 2). The Pubmatrix results show that 65% of genes identified in this study are supported by existing literature reporting these genes’ roles in treatment response, survival, and progression.

Strengths and limitations

This study had several limitations, TCGA data are known to be biased towards patients with later stage cancers with tumor sizes that are greater than 200 g [21,50]. Additionally, samples in TCGA are contributed by multiple academic medical centers where collection methods may vary. When studying rare cancers it is common to have analysis curtailed by sample size, which is the limitation that this study hopes to specifically address through the application of MCCV. OSCC
occurs more often in men than women and thus women only make up approximately 1/3 of our sample. The Monte Carlo validation approach is well suited to address these sample size limitations and is meant to serve as a model for other studies utilizing similar datasets. A drawback of the MCCV approach is that it necessitates discarding signatures identified as predictive in single runs. Such sacrificed signatures may indeed point to true biological mechanisms which the other iterations of analysis did not detect due to their unique mix of patients. MCCV is designed to exclude all but the strongest effects. In many cases a combination of weak effects of genes may produce a predictive signature that can classify patients with accuracy but makes interpretation of biological mechanisms difficult. This study provides support for greater adoption of MCCV when conducting genomic or transcriptomic research in less common cancers.

**Conclusion**

The role of ion gate channel pathway in OSCC and its role in a molecular signature predicting treatment response is supported by this study. The ion channel gate pathway was the only pathway to pass both fisher exact test and false discovery rate significance thresholds. These results provide evidence that applying a MCCV approach to DGE model creation is a suitable method to control variability in results when using heterogeneous datasets, and offers a method of validation prior to devoting time and funding required for additional sequencing. The robustness of this signature was supported by the finding that the distribution of AUC for random signatures and signatures selected through MCCV were completely separate. Those researchers adopting heterogeneous datasets combined over multiple studies must address issues of result variability if they truly wish to contribute to the advancement of this field. This study describes and validates one approach that may be applied towards this goal.
Conflict of interest

The authors of this manuscript have no financial and or personal relationships with other people or organizations that could inappropriately influence (bias) their work. Therefore the authors have no conflicts of interest to report.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.oraloncology.2018.01.012.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin 2016;66:7–30. http://dx.doi.org/10.3322/caac.21332.
[2] Siegel R, Miller K, Jemal A. Cancer statistics, 2015. CA Cancer J Clin 2015;65:29. http://dx.doi.org/10.3322/caac.21254.
[3] Cooper JS, Porter K, Mallin K, Hoffman HT, Weber RS, Ang KK, et al. National cancer database report on cancer of the head and neck: 10-Year update. Head Neck 2009;31:748–58. http://dx.doi.org/10.1002/hed.21022.
[4] Sinha P, Logan HL, Mendenhall WM. Human papillomavirus, smoking, and head and neck cancer. Am J Otolaryngol 2012;33:130–6. http://dx.doi.org/10.1016/j.amjoto.2011.02.001.
[5] Sturgis EM, Cinciripini PM. Trends in head and neck cancer incidence in relation to smoking prevalence: an emerging epidemic of human papillomavirus-associated cancers? Cancer 2007;110:1429–35. http://dx.doi.org/10.1002/cncr.22963.
[6] Yamamoto N, Shibahara T. Epidemiology of the oral cancer. Oral Cancer Diagnosis Ther 2015;1:1–21. http://dx.doi.org/10.1007/978-4-331-54938-3_1.
[7] Ragin CCR, Taioli E. Survival of squamous cell carcinoma of the head and neck in
