Published in final edited form as:

ACM BCB. 2015 September ; 2015: 393–402. doi:10.1145/2808719.2808760.

Developing Robust Predictive Models for Head and Neck Cancer across Microarray and RNA-seq Data

Chanchala D. Kaddi, Wallace H. Coulter, and May D. Wang

Department of Biomedical Engineering Georgia Institute of Technology Atlanta, GA

1-404-385-5059

Abstract

Increased understanding of the transcriptomic patterns underlying head and neck squamous cell carcinoma (HNSCC) can facilitate earlier diagnosis and better treatment outcomes. Integrating knowledge from multiple studies is necessary to identify fundamental, consistent gene expression signatures that distinguish HNSCC patient samples from disease-free samples, and particularly for detecting HNSCC at an early pathological stage. This study utilizes feature integration and heterogeneous ensemble modeling techniques to develop robust models for predicting HNSCC disease status in both microarray and RNAseq datasets. Several alternative models demonstrated good performance, with MCC and AUC values exceeding 0.8. These models were also applied to discriminate between early pathological stage HNSCC and normal RNA-seq samples, showing encouraging results. The predictive modeling workflow was integrated into a software tool with a graphical user interface. This tool enables HNSCC researchers to harness frequently observed transcriptomic features and ensembles of previously developed models when investigating new HNSCC gene expression datasets.

Keywords

Bioinformatics; computational biology; transcriptomics; genes; cancer; microarray; RNA-seq; supervised learning

1. Introduction

Head and neck squamous cell carcinoma (HNSCC) is a group of cancers that affects the upper aerodigestive tract, including the oral cavity, tongue, larynx, oropharynx, and several other disease sites [1]. It is the 6th most prevalent cancer overall, and over 600,000 new cases are expected per year [2]. In the U.S. alone, head and neck cancer accounts for 3% of all cancer cases, and in 2015 more than 60,000 new cases and 12,000 deaths are anticipated [3]. Investigation of gene expression patterns in HNSCC is an active area of research, with numerous studies conducted using gene expression microarrays within the last 10 years [4]. More recently, transcriptomic research has shifted towards RNA sequencing (RNA-seq) because of its high sensitivity and dynamic range [5]. However, due to variations in the sample population, small samples sizes, and differences in experimental design and analysis methods, different transcriptomic studies on the same disease may report notably different lists of significant or key genes [6]. For this reason, integrated analysis of multiple transcriptomic studies is necessary for identifying consistent, fundamental gene expression

patterns that indicate HNSCC status. Previous predictive modeling studies have applied gene expression data to various problems related to HNSCC, including predicting metastatic disease [7, 8], the development of cancer in patients with oral premalignant lesions [9, 10], and the risk of recurrence and relapse [11, 12]. A key aspect of HNSCC research is early detection: if the cancer is detected at an early stage, patient response to treatment is relatively high, and five year survival rates for multiple disease subsites exceeds 80% [13, 14]. However, most cases are detected only at locally advanced stages, which are associated with much worse outcomes. For the same disease subsites, survival for locally advanced cases ranged from 49.8-73%. Current screening recommendations for oral cancer are based on conventional visual and tactile examinations [15]. Effective supervised models for predicting HNSCC status – and in particular, for differentiating early-stage HNSCC patients from healthy individuals – could be useful clinical tools.

In this study, we present an integrated transcriptomic analysis of HNSCC, with the goal of developing robust predictive models for determining disease status. Because the lack of early stage samples is an obstacle, models are initially developed for predicting HNSCC status in general, and are then applied to predict early stage HNSCC in particular. The workflow for this study is shown in Figure 1. First, differential expression (DE) analysis was performed on several microarray datasets to identify common DE genes and the extent of variation among datasets. Second, classification models optimized on one microarray dataset were implemented on the others, in order to evaluate within-platform model robustness. Third, individual and ensemble classification models developed using the microarray datasets were applied to RNA-seq data, to test (i) between-platform robustness, i.e., if informative gene feature sets and model structures are consistent across data types and (ii) performance in detecting early stage HNSCC. Finally, well-performing models were integrated into a software tool with a graphical user interface in order to make predictive models more accessible to HNSCC researchers and clinicians.

2. Methods

2.1 Data

Gene expression microarray datasets were obtained from the Gene Expression Omnibus (GEO) and ArrayExpress public repositories. To increase consistency in the downstream analysis, datasets selected for study met the following criteria: (i) data was from patient samples, not cell lines; (ii) data from both diseased and normal samples were available; (iii) the raw, unprocessed data was available; (iv) an associated publication was available; and (v) Affymetrix array platforms were used. These filtering steps led to five candidate Affymetrix datasets, of which three shared the Human Genome U133 Plus 2.0 (54,675 probes) or U133A (27,777 probes) arrays, enabling direct comparison of probes [16-18]. Only the 22,277 common probes were used for analysis. These datasets are described in Table 1. To obtain gene expression values, the raw CEL files were processed with RMA using the Affymetrix Expression Console software. RNA-seq data (Version 2, meaning it has been processed using MapSplice for alignment and RSEM for quantification) for HNSCC was obtained from The Cancer Genome Atlas (TCGA), along with associated clinical data [19, 20]. Count data for 20,531 genes are available in this dataset. Un-normalized data was used

for DE analysis, and normalized data for classification. At the time of analysis, matched tumor and normal RNA-seq data were available for 40 patients. Of these, 17 patients were categorized as early stage (pathological stages I and II).

2.2 Differential Expression Analysis and Feature Selection

Both the two-sample t-test and Wilcoxon rank-sum test were used to identify DE genes between the HNSCC and normal samples in the microarray datasets. Multiple testing correction was implemented by controlling the False Discovery Rate (FDR = 0.05) or by implementing Bonferroni correction ($\alpha_{\text{Bonferroni}} = 0.05$). DE analysis was performed on the RNA-seq data using edgeR (FDR = 0.05) [21]. To evaluate consistency in gene expression patterns across datasets, the DE gene lists were compared to each other. For classification, features were selected via mRMR (minimum redundancy maximum relevance) from the microarray datasets, implemented using the FEAST toolbox in MATLAB [22-24]. The performance of each classification model was optimized for between 1 and up to the top 50 features.

2.3 Binary Classifiers

Three binary classification methods were tested: k-nearest neighbors (KNN), support vector machine (SVM), and decision tree (DT). Optimal parameters for each model were selected via grid search from the ranges shown in Table 2, using a nested cross-validation scheme as shown in Figure 2. For the microarray datasets, $K = 3$ (E-GEOD-6791 and E-GEOD-9844) and $K = 2$ (E-GEOD-23036) for K-fold cross-validation. For all microarray datasets, the number of outer loop repetitions was $n = 3$, and the number of inner loop repetitions was $m = 5$. Optimization was performed with respect to the Matthews correlation coefficient (MCC), and the area under the ROC curve (AUC) is also reported for optimized models. For the RNA-seq data, which was a heldout test set, model performance is reported across 100 repetitions of 3-fold cross-validation. Both the MCC and the F1 score are reported. All analyses were performed using MATLAB (Mathworks, Natick MA).

2.4 Evaluation of Model Robustness across Microarray Datasets

The robustness of each model was evaluated by testing the model on the other microarray datasets, i.e., those which were not used in its development. In order to avoid the issue of batch effects entirely, models were not applied to other datasets directly. Instead, the model parameters and feature set associated with the top performing model of each classifier type for dataset i were used to train a model on dataset j ($\forall j \neq i$), with $i, j \in [1, 2, \dots, n]$, where n is the number of microarray datasets. In addition, this comparison was also carried out after combining the feature sets for alternative models. For example, such a model to be tested on E-GEOD-6791 would combine the optimized feature sets for both E-GEOD-9844 and E-GEOD-23036. The rationale behind this experiment is to test whether a composite feature library, defined as the union of optimal feature sets from multiple datasets, would help to improve average predictive performance on new incoming datasets.

2.5 Application of Microarray-Developed Models to RNA-seq data

Next, model robustness across data formats was investigated by applying microarray-developed models to RNA-seq data from TCGA. As described for the microarray-only cross-performance tests, the model parameters and feature set associated with a given model was used to train a model on the RNA-seq data. In order to transfer the optimized microarray feature sets, microarray probes were mapped to RNA-seq features on the level of gene symbols. The performance of the nine individual and nine feature combination-based (generated by considering three microarray dataset optimized parameter sets for the three classifiers, with one composite feature set per classifier) models on RNA-seq data. In addition, two ensemble modeling frameworks – majority voting and stacking [25] – were tested. Majority voting is the simplest ensemble framework; for binary classification, given a set of predicted labels from alternative models, the ensemble-predicted label is the most-commonly predicted label. All possible combinations of at least three of the nine microarray-derived models (three classifier types for three datasets) were considered for voting-based ensembles, resulting in 511 alternative models. Stacking involves a two-step classification process. In the first step, label predictions are obtained from alternative models on the training data. This set of predicted labels serves as the features for a second classification model, which is used to generate the final predictions. Three stacking models were developed, using all nine microarray-derived models as the first-level predictors and SVM, KNN, and DT as the three possible second-level classifiers.

2.6 Tool Design

The developed predictive models were integrated into a software tool with a graphical user interface (GUI) to make them more accessible to HNSCC researchers and clinicians, and for easy application to new datasets. Users can apply previously developed individual or ensemble models to process incoming datasets, and then visualize and export the results.

3. Results

3.1 mRMR Gene Selections and Comparison of DE Gene Lists

Notable differences were observed among the DE genes selected in each microarray dataset, as shown in Table 3. The more conservative Bonferroni method resulted in only 5 common DE genes among the three datasets: MMP1, ABCA8, MYO1B, ARHGEF10L, and SASH1; all of these have been associated with HNSCC in recent literature [16, 18, 26-28]. When applying FDR, more than 600 common DE genes were identified across the three datasets for both statistical tests. Functional analysis was performed on the FDR gene lists using DAVID [29]. Although not statistically significant, the top 10 Gene Ontology (GO) terms selected for both sets included GO:0006915~apoptosis GO:0008219~cell death, GO:0012501~programmed cell death, GO:0016265~death, GO:0043588~skin development. Overall, these results indicate that while there is substantial variation across the microarray datasets, the commonly-selected DE genes are relevant to HNSCC. This variation was also observed for the RNA-seq data: 10,239 DE genes were identified in the RNA-seq data through edgeR. 526 and 610 of these genes overlapped with the microarray common DE gene lists (FDR) for the t-test and rank-sum test, respectively. Some of the commonly selected microarray DE features were also represented in the mRMR feature lists. The

common DE genes selected using either with FDR were compared with the top 50 mRMR-selected features for the three microarray datasets. The number of features in the intersection of these lists ranged from 9 to 23.

3.2 Model Performance across Microarray Datasets

Table 4 shows the performance of KNN models across datasets. The top three rows show the performance of the KNN model optimized through nested CV for each dataset on its own testing data split and on the other two datasets. For example, the KNN model optimized for the dataset E-GEOD-6791 resulted in perfect performance on the testing data in the best outer-loop data split, with MCC and AUC values of 1 ± 0 . The same model (i.e., KNN with a given parameter set and feature set) also performed well on another dataset, E-GEOD-9844, giving MCC and AUC values of 0.89 ± 0.10 and 0.94 ± 0.06 . However, when applied to the third dataset, E-GEOD-23036, very poor performance was observed, with the mean MCC near zero and the mean AUC near 0.5. This example demonstrates the lack of model robustness across datasets. Similar patterns were observed for the other datasets for KNN, as well as for the DT and SVM models in Tables 5 and 6, respectively. The lower three rows of Table 4 show the model performances resulting from combining the optimal feature sets of two models and predicting on the third. For example, the bottommost row shows that using the combined optimal feature sets of the EGEOD-9844 and E-GEOD-23036 KNN models to develop a KNN model for E-GEOD-6791 resulted in MCC and AUC values of 0.84 ± 0.17 and 0.95 ± 0.05 , respectively. While this is lower than the performance of the optimal model developed for E-GEOD-6791 itself, it is higher than the average performance ($0.52 = \text{mean}(0.71, 0.33)$) observed from applying either of the other models to this dataset. For this particular group of datasets, EGEOD-6791 and E-GEOD-23036 show poor performance during cross-prediction tests with the other, while E-GEOD-9844 shows more stable performance. Notably, for the DT analysis of EGEOD-9844, application of the model to the testing data yielded very poor results; a follow-up analysis using the entire dataset (indicated by an asterisk in Table 5) showed better performance. Other than this outlier case, similar trends for cross-prediction were observed across the cases shown in Tables 4, 5 and 6. Overall, utilizing a composite feature set aggregated from multiple models appears to yield more robust predictions for previously unseen data.

3.3 Performance of Microarray-Developed Models on RNA-seq Data

As noted in Figure 1, the RNA-seq data were held-out, and were not used for model development. Overall, the nine individual (non-ensemble) models developed using microarray data performed reasonably on RNA-seq data, as shown in Figure 3. This experiment tracks the distribution of mean MCC values of each model category across 100 repetitions of 3-fold CV on the RNA-seq dataset. The median performance of the nine individual KNN, DT, and SVM models when applied to the RNA-seq data was 0.73, with the best results for any CV repetition approaching 0.86 (result set (a)). Compared to the individual models, the feature combination approach (b) showed a slight increase in median performance, but also resulted in many low-performing outliers. Voting using single-classifier models (c-e) showed slight increases in median performance – the KNN-only and DT-only ensembles achieved median performances around 0.78, and the SVM-only reached 0.81 – and also increases in the minimum performance. The combination voting approaches

(f-l) had a trend of slightly increasing median performance and of lower variation as the number of models in the combination increased. The seven-, eight-, and nine-member ensembles had best overall median performances, near 0.85-0.87. Among the stacking ensembles (m-o), the best median performance of 0.85 was observed when KNN was the secondary classifier. In total in this experiment, 42 models were developed which had better performance than the best-performing individual model (mean MCC = 0.8599) in at least 50 of the 100 CV repetitions.

These models were part of the five-, six-, seven- and nine-member voting ensembles, which also have smaller amounts of variation than many of the other model categories. Most of the models had instances of statistically significant improvement ($p < 0.05$) over the best-performing individual model, as assessed by Steiger's Ztest [30]. Similar performance trends were also observed in terms of the F1 score, also shown in Figure 3. Figure 4 shows the performance of the same models when applied to the early stage vs. normal RNA-seq data, across 100 repetitions of 3-fold CV. Overall, performances are slightly lower and also more variable, reflecting the more challenging nature of the classification problem. Otherwise, similar trends were observed across the model categories. The median performance for the individual models (a) was 0.68. The median performances of the combination voting ensembles were in the range of 0.77-0.78. Unlike in the previous experiment, no models had better performance than the best-performing individual model across any CV repetition (MCC = 0.89) for more than 50 of the 100 CV repetitions. This is due in part to the overall lower model performances in most categories, as well as the slightly higher value of the maximum individual performance for this experiment. However, all of the combination voting and stacking models (f-o) had instances – and many had first quartile values – exceeding the median performance of the individual models. Again, similar patterns of performance were observed in terms of the F1 score.

3.4 System Design

The suite of microarray models was integrated into a MATLAB GUI that allows users to (i) load a new dataset of interest, (ii) select a model developed using previously examined datasets, or compare all models, and (iii) specify options including the number of cross-validation partitions and implement the selected model, and (iv) visualize and export the results. Figure 5 shows a screenshot of the interface. The goal of developing this system is to enable HNSCC researchers, particularly those from more clinical-oriented, non-computational backgrounds, to take advantage of predictive modeling resources developed by the computational research community. In particular, by gathering multiple models from different datasets together in a single tool, it becomes easier to implement ensemble approaches that improve overall performance. This tool will be made freely available to the research community through the website of the Bio-medical Informatics and Bio-Imaging Laboratory (<http://www.bio-miblab.org/>).

4. Discussion

Consistency among different studies lends support to research findings; in the same way, consistent performance among datasets increases confidence in a predictive model, and in

the functional importance of the features that it utilizes. In this paper, we developed predictive models for HNSCC using gene expression data that exhibit robust performance both within and between transcriptomic data types. Other recent transcriptomic HNSCC studies have also compared results across several datasets. De Cecco and colleagues used three microarray datasets to develop their model, and tested it on six other datasets, including TCGA RNA-seq data [11]; however, the endpoint of interest in their study was risk of relapse, not diagnosis. Saintigny and colleagues tested their model for risk of oral cancer development in leukoplakia patients on nine other microarray datasets [10]. However, neither of these studies considered multiple classification approaches and ensemble methods, as in this study. Ye and colleagues performed a meta-analysis across 63 HNSCC transcriptomic studies, considering premalignant lesions vs. normal samples, primary tumors vs. normal samples, and primary tumors vs. metastatic disease [18]; however, the study focused on identifying key genes and pathways, and did not build predictive models. This study contributes to the existing literature on transcriptomic analysis for HNSCC by considering several alternative modeling frameworks for the endpoint of disease state, with an application of early diagnosis. Overall, multiple models with good performance (MCC 0.8, AUC 0.8) were identified. In addition, we compared and identified ensemble strategies that increased model performance for differentiating both general and early HNSCC from normal samples. One direction for further research is in the integration of protein and gene expression data for early HNSCC detection. Several recent studies have investigated the use of salivary RNA and/or proteins for detecting oral cancer [31-33]. Some of the salivary RNA markers validated in [31] – IL-1B, IL-8, and H3F3A – were also selected in the DE gene lists in this study, and IL-8 was one of the mRMR-selected features for E-GEOD-9844. This observation is promising in both directions: applying other feature selection methods to the current group of datasets may reveal more previously-validated markers, and future validation studies may support the clinical relevance of features used in the current models. In addition, we have recently demonstrated that combining transcriptomic and proteomic models increases performance when predicting HNSCC pathological stage [34]. Thus, models combining multiple –omic data types may also improve performance for early disease detection. While current results are encouraging, more systematic testing, comparison, and refinement of models will be possible with additional and larger datasets. The three microarray datasets investigated here collectively include only 158 samples, and the number of matched tumor-normal RNA-seq samples currently available in TCGA is also limited. However, data availability – particularly for early stage disease – is always a limitation in cancer research. One method for further testing the model performance patterns observed here is to investigate whether similar trends hold for different cancer types – particularly those which have been found to have molecular-level similarities to head and neck cancer [35]. Future research can also consider the behavior of different classifiers, feature selection techniques, and even different performance metrics. One direction is testing other established ensemble techniques, as well as developing new ensemble construction approaches suited to this application. In addition, in this study, only mRMR was used for feature selection; the performance of wrapper methods and other filter methods can also be compared. A related direction is the development of composite feature selection methods, such as those recently proposed involving mRMR [36, 37]. Given the

class-imbalanced datasets, examining other evaluation metrics, such as precision and recall, may also provide additional insight into model robustness patterns.

In addition, one of the design goals of the modeling tool is to continually update its collection of individual and ensemble models as additional labeled transcriptomic data becomes available. In this sense, it can serve to help accelerate translational research. In the process, the tool will also be expanded to accommodate data and models for other prediction endpoints, such as length of survival, recurrence, and response to alternative therapies. It can also consider specific subsets of HNSCC, such as HPV+ vs. HPV- disease [38]. Thus, it can become a central component of a future clinical decision support system for assisting in HNSCC diagnosis and treatment planning.

Acknowledgments

This research has been supported by grants from The Parker H. Petit Institute for Bioengineering and Bioscience (IBB), Johnson & Johnson, Bio Imaging Mass Spectrometry Initiative at Georgia Tech, National Institutes of Health (Bioengineering Research Partnership R01CA108468, Center for Cancer Nanotechnology Excellence U54CA119338), Georgia Cancer Coalition (Distinguished Cancer Scholar Award to MDW), Microsoft Research, the National Science Foundation (GRFP to CDK), and P.E.O. International (Scholar Award to CDK). The authors thank Dr. John H. Phan, Dr. Chih-Wen Cheng, and Po-Yen Wu for their helpful comments during the preparation of this manuscript.

References

1. Matta A, Ralhan R. Overview of current and future biologically based targeted therapies in head and neck squamous cell carcinoma. *Head Neck Oncol.* 2009; 1:6. [PubMed: 19284526]
2. Leemans CR, Braakhuis BJ, Brakenhoff RH. The molecular biology of head and neck cancer. *Nat Rev Cancer.* Jan.2011 11:9–22. [PubMed: 21160525]
3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. *CA Cancer J Clin.* Jan.2015 65:5–29. [PubMed: 25559415]
4. Yu YH, Kuo HK, Chang KW. The Evolving Transcriptome of Head and Neck Squamous Cell Carcinoma: A Systematic Review. *PLoS ONE.* 2008; 3:e3215. [PubMed: 18791647]
5. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews Genetics.* 2009; 10:57–63.
6. Ramasamy A, Mondry A, Holmes CC, Altman DG. Key Issues in Conducting a Meta-Analysis of Gene Expression Microarray Datasets. *PLoS Medicine.* Sep 02.2008 5:e184. [PubMed: 18767902]
7. Clatot F, Gourant S, Mareschal S, Cornic M, Berghian A, Choussy O, et al. The gene expression profile of inflammatory, hypoxic and metabolic genes predicts the metastatic spread of human head and neck squamous cell carcinoma. *Oral Oncology.* 2014; 50(3):200–207. [PubMed: 24387976]
8. Roepman P, Wessels LF, Kettelarij N, Kemmeren P, Miles AJ, Lijnzaad P, et al. An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. *Nat Genet.* Feb.2005 37:182–6. [PubMed: 15640797]
9. Kondoh N, Ohkura S, Arai M, Hada A, Ishikawa T, Yamazaki Y, et al. Gene expression signatures that can discriminate oral leukoplakia subtypes and squamous cell carcinoma. *Oral Oncol.* May. 2007 43:455–62. [PubMed: 16979924]
10. Saintigny P, Zhang L, Fan YH, El-Naggar AK, Papadimitrakopoulou VA, Feng L, et al. Gene Expression Profiling Predicts the Development of Oral Cancer. *Cancer Prevention Research.* Feb 1.2011 4:218–229. 2011. [PubMed: 21292635]
11. De Cecco L, Bossi P, Locati L, Canevari S, Licitra L. Comprehensive gene expression meta-analysis of head and neck squamous cell carcinoma microarray data defines a robust survival predictor. *Annals of Oncology.* May 14.2014 2014.
12. Fountzilias E, Markou K, Vlachtsis K, Nikolaou A, Arapantoni-Dadioti P, Ntola E, et al. Identification and validation of gene expression models that predict clinical outcome in patients

- with early-stage laryngeal cancer. *Annals of Oncology*. Aug 1.2012 23:2146–2153. 2012. [PubMed: 22219018]
13. Pulte D, Brenner H. Changes in survival in head and neck cancers in the late 20th and early 21st century: a period analysis. *Oncologist*. 2010; 15:994–1001. [PubMed: 20798198]
 14. Worsham MJ. Identifying the risk factors for latestage head and neck cancer. *Expert Rev Anticancer Ther*. Sep.2011 11:1321–5. [PubMed: 21929305]
 15. Rethman MP, Carpenter W, Cohen EE, Epstein J, Evans CA, Flaitz CM, et al. Evidence-based clinical recommendations regarding screening for oral squamous cell carcinomas. *J Am Dent Assoc*. May.2010 141:509–20. [PubMed: 20436098]
 16. Pavón M, Parreño M, Téllez-Gabriel M, Sancho F, López M, Céspedes M, et al. Gene expression signatures and molecular markers associated with clinical outcome in locally advanced head and neck carcinoma. *Carcinogenesis*. Sep 1.2012 33:1707–1716. 2012. [PubMed: 22696598]
 17. Pyeon D, Newton MA, Lambert PF, den Boon JA, Sengupta S, Marsit CJ, et al. Fundamental Differences in Cell Cycle Deregulation in Human Papillomavirus–Positive and Human Papillomavirus–Negative Head/Neck and Cervical Cancers. *Cancer research*. 2007; 67:4605–4619. [PubMed: 17510386]
 18. Ye H, Yu T, Temam S, Ziober B, Wang J, Schwartz J, et al. Transcriptomic dissection of tongue squamous cell carcinoma. *BMC Genomics*. 2008 Feb 06.9:1–11. 2008. [PubMed: 18171476]
 19. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*. Feb 15.2010 26:493–500. [PubMed: 20022975]
 20. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*. Oct.2010 38:e178. [PubMed: 20802226]
 21. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. Jan 1.2010 26:139–40. [PubMed: 19910308]
 22. Brown G, Pocock A, Ming-Jie Z, Luján M. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *Journal of Machine Learning Research*. 2012; 13:27–66.
 23. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*. Apr.2005 3:185–205. [PubMed: 15852500]
 24. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*. Aug.2005 27:1226–38. [PubMed: 16119262]
 25. Zhou, ZH. Chapman and Hall/CRC. 1st. 2012. Ensemble Methods: Foundations and Algorithms.
 26. Ohmura G, Tsujikawa T, Yaguchi T, Kawamura N, Mikami S, Sugiyama J, et al. Aberrant Myosin 1b Expression Promotes Cell Migration and Lymph Node Metastasis of HNSCC. *Mol Cancer Res*. Apr.2015 13:721–31. [PubMed: 25421751]
 27. Stanam A, Love-Homan L, Joseph TS, Espinosa Cotton M, Simons AL. Upregulated interleukin-6 expression contributes to erlotinib resistance in head and neck squamous cell carcinoma. *Molecular Oncology*.
 28. Zhang C, Song X, Zhu M, Shi S, Li M, Jin L, et al. Association between MMP1 -1607 1G>2G polymorphism and head and neck cancer risk: a metaanalysis. *PLoS One*. 2013; 8:e56294. [PubMed: 23441173]
 29. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009; 4:44–57. [PubMed: 19131956]
 30. Steiger JH. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*. 1980; 87:245–251.
 31. Elashoff D, Zhou H, Reiss J, Wang J, Xiao H, Henson B, et al. Prevalidation of Salivary Biomarkers for Oral Cancer Detection. *Cancer Epidemiology Biomarkers & Prevention*. Apr 1.2012 21:664–672. 2012.
 32. Franzmann EJ, Reategui EP, Pereira LHM, Pedrosa F, Joseph D, Allen GO, et al. Salivary protein and solCD44 levels as a potential screening tool for early detection of head and neck squamous cell carcinoma. *Head & Neck*. 2012; 34:687–695. [PubMed: 22294418]

33. Ying S, Wei D, Chunguang Z, You Z, Zhongbo C, Yuan T, et al. A Computational Method for Prediction of Saliva-Secretory Proteins and Its Application to Identification of Head and Neck Cancer Biomarkers for Salivary Diagnosis. *NanoBioscience, IEEE Transactions on*. 2015; 14:167–174.
34. Kaddi CD, Wang MD. Models for Predicting Stage in Head and Neck Squamous Cell Carcinoma using Proteomic and Transcriptomic Data. *Manuscript Under Review*. 2015
35. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*. 158:929–944.
36. El Akadi A, Amine A, El Ouardighi A, Aboutajdine D. A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. *Knowledge and Information Systems*. 2011 Mar 01;26:487–500. 2011.
37. Zhang Y, Ding C, Li T. Gene selection algorithm by combining reliefF and mRMR. *BMC Genomics*. 2008; 9(2):S27. [PubMed: 18831793]
38. Marur S, D'Souza G, Westra WH, Forastiere AA. HPV-associated head and neck cancer: a virus-related cancer epidemic. *The Lancet Oncology*. 2010; 11(8):781–789. [PubMed: 20451455]

Figure 1. Modeling workflow describing roles for microarray, RNA-seq (all stages) and early-stage RNA-seq datasets

Figure 2. Schematic of nested cross-validation

Figure 3. Comparison of alternative individual and ensemble models developed from microarray data when applied to predict HNSCC vs. normal samples from RNA-seq data, in terms of MCC (top) and F1 (bottom) values

Figure 4. Comparison of alternative individual and ensemble models developed from microarray data when applied to predict early stage HNSCC vs. normal samples from RNA-seq data, in terms of MCC (top) and F1 (bottom) values

Figure 5. Screenshot of tool interface, displaying import, analysis, visualization, and export capabilities

Table 1
Description of gene expression microarray datasets examined in study

Dataset	Affymetrix Array Platform	Samples	Reference
E-GEOD-9844	Human Genome U133 Plus 2.0	25 cancer, 12 normal	[18]
E-GEOD-6791	Human Genome U133 Plus 2.0	42 cancer, 11 normal	[17]
E-GEOD-23036	Human Genome U133A 2.0	63 cancer, 5 normal	[16]

Table 2
Classification model parameters examined via nested cross-validation

Classification Method	Parameters	Set of values
KNN	Number of neighbors (K)	$K \in [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$
SVM	Kernel Soft margin cost (C) γ for GBRF	Kernels; linear, Gaussian radial basis function (GRBF) $C \in 2^m$, $m \in [-2, -1, 0, 1, 2]$ $\gamma \in 2^m$, $m \in [-2, -1, 0, 1, 2]$
Decision Tree (DT)	Splitting criterion	Criteria: Gini diversity index (GDI), Twoing rule, Maximum deviance reduction (MDR)

Table 3

Comparison of DE Genes across Microarray Datasets

	E-GEOD-6791		E-GEOD-9844		E-GEOD-23036		Common DE Genes	
	Bonferroni	FDR	Bonferroni	FDR	Bonferroni	FDR	Bonferroni	FDR
T-test	213	5763	84	2451	273	3154	5	682
Rank-sum	14	6836	5	2759	0	3401	0	785

Table 4
Multi-Dataset Performance of KNN Models in terms of MCC (AUC)

Data for Model Development (E-GEOD-#)			Prediction Dataset (E-GEOD-#)		
6791	9844	23036	6791	9844	23036
•			1±0 (1±0)	0.89±0.10, (0.94±0.06)	-0.08±0.03, (0.54±0.01)
	•		0.71±0.14, (0.98±0.02)	1±0, (1±0)	1±0, (1±0)
		•	0.33±0.27, (0.68±0.19)	0.75±0.11, (0.85±0.04)	1±0, (1±0)
			Average Cross-Dataset Performance		
			0.52	0.82	0.46
•	•		-	-	0.90±0.18, (1±0)
•		•	-	1±0, (1±0)	-
	•	•	0.84±0.17, (0.95±0.05)	-	-

Table 5
Multi-Dataset Performance of DT Models in terms of MCC, (AUC)

Data for Model Development (E-GEOD-#)				Prediction Dataset (E-GEOD-#)	
6791	9844	23036		6791	9844
•				$0.88 \pm 0.21, (0.97 \pm 0.06)$	$0.88 \pm 0.10, (0.94 \pm 0.06)$
	•			$0.93 \pm 0.12, (0.94 \pm 0.10)$	$0 \pm 0, (0.5 \pm 0) \ 0.76 \pm 0.11, (0.86 \pm 0.10) (*)$
		•		$-0.07 \pm 0.09, (0.62 \pm 0.11)$	$0.82 \pm 0.19, (0.90 \pm 0.10)$
				Average Cross-Dataset Performance	
				0.43	0.85
•	•			-	-
•		•		-	$0.88 \pm 0.10, (0.94 \pm 0.06)$
	•	•		$0.93 \pm 0.12, (0.94 \pm 0.10)$	-

Table 6
Multi-Dataset Performance of SVM Models in terms of MCC, (AUC)

Data for Model Development (E-GEOD-#)			Prediction Dataset (E-GEOD-#)		
6791	9844	23036	6791	9844	23036
•			1±0, (1±0)	0.89±0.10 (0.54±0.07)	0.09±0.31 (0.5±0)
	•		0.85±0.14 (0.98±0.03)	1±0, (1±0)	0.90±0.18 (1±0)
		•	0.51±0.28 (0.82±0.12)	0.88±0.10 (0.99±0.02)	1±0, (1±0)
			Average Cross-Dataset Performance		
			0.68	0.89	0.50
•	•		-	-	0.90±0.18 (1±0)
•		•	-	0.89±0.10 (0.5±0)	-
	•	•	0.76±0.22 (0.98±0.03)	-	-