Toward Signaling-Driven Biomarkers Immune to Normal Tissue Contamination

Libertas Academica
FREEDOM TO RESEARCH

John C. Stansfield[1], Matthew Rusay[1], Roger Shan[1], Conor Kelton[1], Daria A. Gaykalova[2], Elana J. Fertig[3], Joseph A. Califano[2,4] and Michael F. Ochs[1]

[1]Department of Mathematics and Statistics, The College of New Jersey, Ewing, NJ, USA. [2]Department of Otolaryngology-Head and Neck Surgery, Johns Hopkins Medical Institutions, Baltimore, MD, USA. [3]Division of Oncology Biostatistics and Bioinformatics, Johns Hopkins School of Medicine, Baltimore, MD, USA. [4]Milton J. Dance Jr. Head and Neck Center, Greater Baltimore Medical Center, Baltimore, MD, USA.

ABSTRACT: The goal of this study was to discover a minimally invasive pathway-specific biomarker that is immune to normal cell mRNA contamination for diagnosing head and neck squamous cell carcinoma (HNSCC). Using Elsevier's MedScan natural language processing component of the Pathway Studio software and the TRANSFAC database, we produced a curated set of genes regulated by the signaling networks driving the development of HNSCC. The network and its gene targets provided prior probabilities for gene expression, which guided our CoGAPS matrix factorization algorithm to isolate patterns related to HNSCC signaling activity from a microarray-based study. Using patterns that distinguished normal from tumor samples, we identified a reduced set of genes to analyze with Top Scoring Pair in order to produce a potential biomarker for HNSCC. Our proposed biomarker comprises targets of the transcription factor (TF) HIF1A and the FOXO family of TFs coupled with genes that show remarkable stability across all normal tissues. Based on validation with novel data from The Cancer Genome Atlas (TCGA), measured by RNAseq, and bootstrap sampling, the biomarker for normal vs. tumor has an accuracy of 0.77, a Matthews correlation coefficient of 0.54, and an area under the curve (AUC) of 0.82.

KEYWORDS: gene expression profiling, biomarkers, cancer, biostatistics

CITATION: Stansfield et al. Toward Signaling-Driven Biomarkers Immune to Normal Tissue Contamination. *Cancer Informatics* 2016:15 15–21 doi: 10.4137/CIN.S32468.

TYPE: Original Research

RECEIVED: September 01, 2015. **RESUBMITTED:** December 08, 2015. **ACCEPTED FOR PUBLICATION:** December 10, 2015.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Nine peer reviewers contributed to the peer review report. Reviewers' reports totaled 4282 words, excluding any confidential comments to the academic editor.

FUNDING: This work was funded by the NIH, NLM R01 LM011000 to MFO.The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: ochsm@tcnj.edu

Introduction

Genome-wide gene expression data are now typically available in many cancer studies. The six hallmarks of cancer, sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis, all result from genetic and epigenetic changes and drive changes in gene expression.[1] These hallmarks are the defining features of cancer and are required for tumorigenesis. While the natural way to identify cancer is through invasive capture of tumor cells coupled with genetic and cytologic analysis, this obviously requires previous identification of cancer. In this study, we focus on leveraging gene expression changes driven by cancer-type-specific pathways to identify biomarkers that may lead to minimally invasive detection of cancer.

The vast amounts of data generated through microarrays and sequencing technologies create many challenges for analysis. We had earlier shown the value of matrix factorization techniques to isolate the signatures of pathway activity in the presence of overlapping gene regulation.[2] Nonnegative matrix factorization (NMF) has also been shown to be advantageous over other clustering methods for identifying cancer subclasses.[3] Here, we apply the Bayesian NMF algorithm CoGAPS[4] to isolate the underlying processes of head and neck squamous cell carcinoma (HNSCC).

HNSCC is typically caused by tobacco and alcohol use or by human papillomavirus (HPV). HNSCC is the sixth leading cancer by incidence worldwide, and it is estimated that only 40–50% of patients with HNSCC will survive for five years with the disease, likely due to failure to detect the disease at early stages.[5] Therefore, early diagnosis using a robust biomarker could substantially improve the treatment of patients with HNSCC.

Mapping the signaling networks of interest for the cancer under study is an integral part of our approach. Figure 1 displays the protein signaling network involved in HNSCC, which was constructed based on two reviews by experts in the field.[6,7] The root nodes (IGF-1R, VEGFR, EGFR, and cMet) are receptor tyrosine kinases, which, when activated, drive signaling cascades that lead to the activation or repression of transcription factors (TFs). In individual patients, several different mutations or epigenetic changes have been identified that can change signal propagation in this network. Therefore, copy number and epigenetic measurements on individual patients can provide prior probabilities of TF activity.

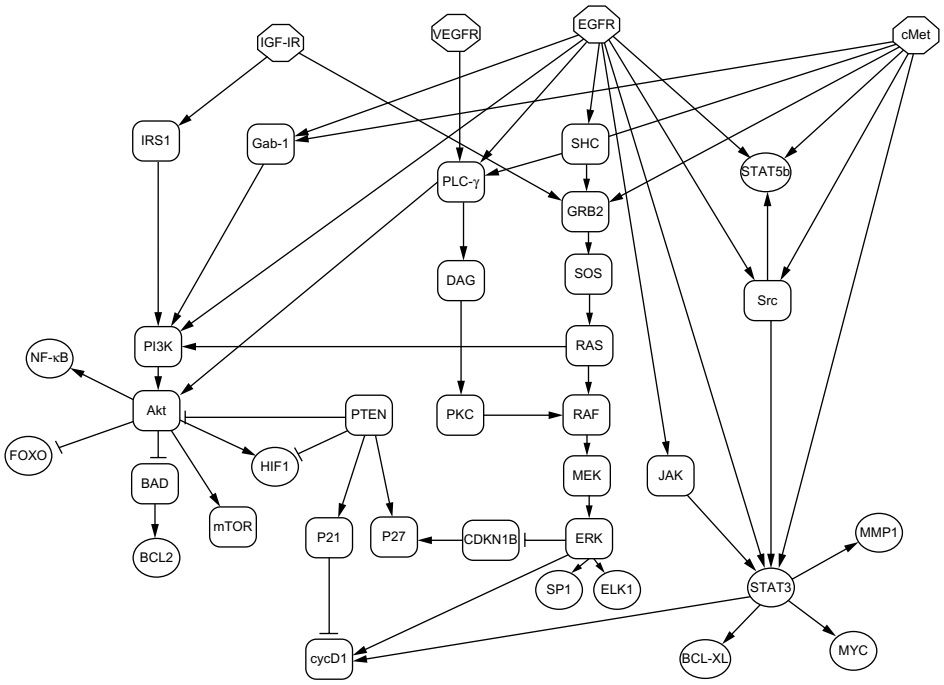

Figure 1. Diagram of the signaling network involved in HNSCC. The root nodes (octagons) of this diagram represent the receptors that are activated and then drive the rest of the network. The leaf nodes (circles) represent the TFs that activate a large number of genes involved in HNSCC. A pointed arrow represents activation of the target and a T represents repression of the target. Rounded rectangles represent signaling proteins.

CoGAPS permits encoding of this information as prior probabilities of the expression of target genes of the TFs.

Biomarkers provide an easily measured indicator of hidden biological processes of interest, and the identification of biomarkers has proven to be essential for disease diagnosis and for determining the treatment strategies for cancer.[8] Our goal is to identify mRNA biomarkers related to specific deregulated signaling known to drive cancer development. Here, we utilize CoGAPS to isolate patterns associated with HNSCC and Top Scoring Pair (TSP) to generate biomarkers robust to normalization artifacts.[9] The advantage of TSP lies in the inclusion of internal controls by looking only at relative expression between two genes. Unlike the sets of genes that tend to rely on relative levels, TSP relies only on ranks. A marker solely based on rank with the same number of genes may be equally effective, but the threshold would be harder to implement because $n/2$ pairwise comparisons for n genes would increase to $n(n-1)/2$ pairwise comparisons. Importantly, our application of TSP aims to identify gene pairs that consist of one gene, which is a target of a TF involved in HNSCC, and one gene from a set of reference genes, which we have found to have extremely stable expression values in all normal cell types. This provides a path to a biomarker that is immune to normal tissue contamination.

Methods

Summary. The overall analysis plan is summarized in Figure 2. Multiple molecular data types are downloaded and, if not preprocessed by the provider, processed to create properly normalized data sets. Expression data are filtered based on the known targets of TFs in the network, while other data (ie, mutation, copy number, methylation) are filtered to include only network members. The nonexpression data provide prior relative probabilities of the activity of different proteins in the signaling network, and these prior probabilities are propagated through a graphical model to a probability of the expression of the TF targets. The expression data are then analyzed with these prior relative probabilities using CoGAPS. The results of analysis include patterns that are reviewed for association with tumor status. Patterns with such an association are then analyzed for significance of TF activity, and targets of these TFs are captured. The TSP algorithm is run on these genes and the reference gene list to identify biomarkers with one gene from the targets of significantly active TFs and one gene from the stable reference gene list.

Data. The HNSCC data used as a training set for this study were from a public domain data set generated at Johns Hopkins University, containing microarray expression, promoter methylation, and copy number data (Gene Expression Omnibus (GEO) accession: GSE33232), from 44 subjects with HNSCC tumors (HPV+ 13, HPV– 31) and from 25 subjects from uvulopalatopharyngoplasty surgery. The normal samples were taken from different individuals to avoid any contamination due to field cancerization, which can lead to nonlocalized premalignant transformation of tissues in the head and neck area. The expression data were normalized using RMA,[10] copy number data were summarized using CRLMM,[11] and methylation data were normalized based on their natural beta distribution.

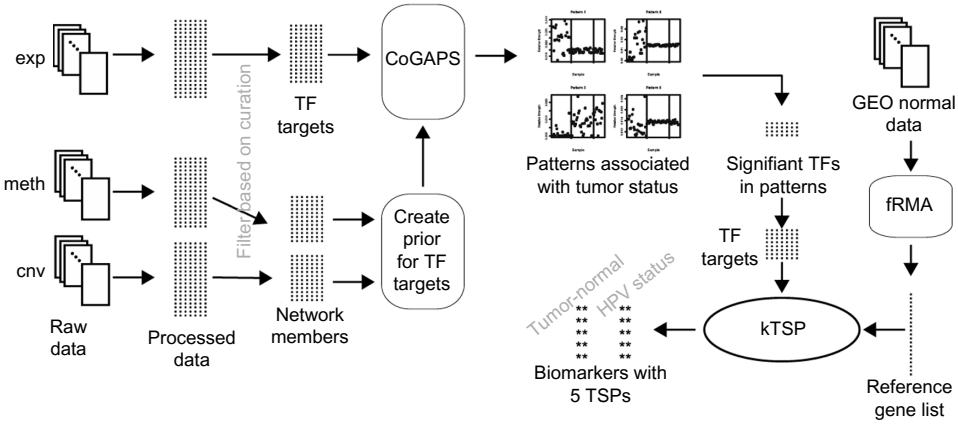

Figure 2. The overall analysis path for the creation of robust biomarkers. The diagram shows the plan from initial data gathering to biomarker identification and is described in detail in the text.

For validation, level 3 data from TCGA, comprising 515 tumor samples with 44 normal samples, were downloaded on November 17, 2015.[12] The measurements for the genes in the biomarker were extracted from the complete gene-level summaries.

Pathway curation. In order to encode prior information from methylation and copy number measurements on signaling proteins, the model of the signaling network shown in Figure 1 is used. The network drives transcriptional changes through the TFs, so the final link to expression is to identify the targets of the TFs (shown as circles in Fig. 1). The identification of TF targets was done using the TRANSFAC database[13] and Elsevier's MedScan software, which is part of the Pathway Studio tool.

For the TFs ELK1, the FOXO family, and MYC, targets were curated by identifying the abstracts of papers with MedScan, as the TRANSFAC data were limited. All identified abstracts were manually reviewed to classify the TF–target interaction, confirm a direct regulatory relationship, and thus complete the link from signaling pathway to transcripts. For other TFs in the network, TRANSFAC was used exclusively.

Determining priors for expression analysis. In order to set priors on the potential expression of genes that are targets of HNSCC network shown in Figure 1, information on protein activity is needed. For this, an outlier analysis was performed on the methylation and copy number data. Outliers were counted for the hypomethylation of promoters or amplification of genes that coded signaling proteins. A rank outlier method was used,[14] where an outlier for a gene was defined such that the methylation of a tumor was below the normal by at least 0.1 or the copy number of the tumor was above the normal by at least 0.5. For each gene, this resulted in a count, C, for each tumor capturing how many normals it exceeded in methylation and copy number. We converted this to an empirical P-value with $P = (N - C + 1)/N$, so the more times a tumor exceeded the normals, the lower the P-value. We did this separately for methylation and copy number and then counted the number of significant P-values for each gene across the 44 tumors and two molecular types at the significance level of $\alpha = 0.05$. This method of counting outliers was shown to be robust to changes in the minimum difference for copy number and methylation level previously.[14] The number of outliers was then linearly scaled to provide a value for each protein between 0.9 (many outliers) and 0.5 (no outliers).

The network of Figure 1 was then propagated with these values to the TFs as follows. For receptors and other root nodes with no parents, the relative probability of activity was set equal to the value. For any node x with only activating parents pa(x),

$$p(x) = \max\left(p_{pa(x)}, p_p\right)$$

where $p_{pa(x)}$ is the maximum relative probability of all parent nodes and p_p is the value calculated from outliers. For cases including the repressors of x, which compete with the activators, the relative probability was given by

$$p(x) = \max\left(p_{pa(x)}, p_p\right) \times \left(1 - \max\left(p_{pr(x)}, p_p\right)\right)$$

where $p_{pr(x)}$ is the maximum relative probability of the repressors being active. This provided for repressors dominating activators overall and for a single activation or repression step to tend to have a dominant effect.

Finally, the relative probability of a TF being active was then used as the prior relative probability of a target being expressed. The implementation of the prior scaled all values to have equal overall prior probability assigned to each pattern, so these values effectively just set the relative probability within one pattern (one column of the **A** matrix – see next section).

Analysis of gene expression data with CoGAPS. CoGAPS is an NMF algorithm that utilizes Bayesian statistics

and Markov Chain Monte Carlo (MCMC) sampling. NMF works to factor a data matrix, **D**, into a pair of matrices (**A**, **P**) that best approximate **D** as follows:

$$D_{ij} \approx \sum_{k=1}^{F} A_{ik} P_{kj} \quad (1)$$

where F indicates the number of dimensions or factors, i indexes the gene, and j the sample. The matrix **A** provides an assignment of genes to patterns, while the matrix **P** provides an indication of which patterns are associated with samples, and nonnegativity serves to reduce the nonidentifiability problem. Eqn. 1 allows for handling multiple regulation of genes by different TFs. Nonnegativity is generally not sufficient to eliminate nonidentifiability, so the sparseness inherent in gene regulation (eg, all genes are not to be expressed in all processes) is often leveraged as well. A full explanation of the methods used in CoGAPS has been published.[15]

Estimation of the dimensionality of the data (or the number of factors needed to recover the data within the noise) is an outstanding problem in all analyses of expression data, including clustering methods, principal component analysis (PCA), and NMF. To determine the best dimensionality, we reviewed the patterns generated for the separation of HPV+, HPV–, and normal samples. As the final goal of this study is a validated biomarker unrelated to the CoGAPS factorization, the exact dimensionality determined may not be critical so long as the signaling processes are successfully identified, thus providing a biomarker that withstands validation.

Estimating TF activity. The patterns generated by CoGAPS were analyzed to infer TF activity using a Z-score statistic with an empirical null.[16] In brief, the Z-score for each TF is estimated as the mean Z-score of all its R target genes. CoGAPS provides a mean and standard deviation for every element in the **A** matrix from MCMC sampling, which are easily calculated. The Z-score of the TF is then compared to the empirical null distribution generated by 500 random draws of R genes from the pattern, and an empirical P-value is generated.

TSP and biomarker discovery. In order to identify biomarkers robust to normalization, we applied the TSP algorithm.[9] TSP finds pairs of genes chosen by how well the statistic can distinguish the two classes based on the inversion of the relative values between the classes. One limitation of TSP is that it searches all possible gene pairs, which can produce pairs driven by noise, because there are many more gene pairs than samples. We avoided this limitation by limiting the genes being input into TSP.

To limit the TSPs to genes expected to change expression due to HNSCC signaling by HNSCC, we only included curated targets of the TFs in the pathways of interest for HNSCC (Fig. 1). While TFs will not themselves generally show expression changes, their targets should change expression based on the TF activity changes driven by the signaling pathways. Because the patterns from CoGAPS are correlated with disease status, strong TF activity in a pattern determined by the TF Z-score is also correlated with tumor status.

To make the TSPs robust to tissue contamination, we also required each TSP to include one gene related to HNSCC signaling and one gene from a reference gene list. The reference gene list was generated by gathering all normal tissues measured on the U133plus2 Affymetrix array and deposited in GEO. All genes with medium expression levels in all samples (log2 expression as determined by frozen RMA[17] of 5–7) were ranked for low variance. The genes with the least variance were retained for inclusion in TSPs. The R package switchBox was used for the TSP analysis,[18] which yielded a biomarker composed of five paired genes with one gene from the target list and one gene from the reference list.

Validation. Fivefold cross validation was performed on our original data set to determine the error rate of our model at predicting the tumor status of a patient. The biomarker was then tested on the TCGA data set.

Results

The pathway curation allowed us to produce a list of targets for the TFs of interest for HNSCC. Targets of the terminal TFs in the network were first identified in TRANSFAC. For TFs with limited information, further curation was done with MedScan. Targets for ELK1, FOXO1, FOXO2, FOXO3, FOXO4, and MYC were extended with MedScan, and all targets were integrated into the network shown in Figure 1 as leaf nodes. The combined list of FOXO family targets was taken as targets of FOXO in the network.

Using methylation and copy number measurements for the members of the signaling network shown in Figure 1, an outlier analysis generated a ranking of each pathway member by the total number of hypomethylated promoters and gene amplifications. The range of values was linearly scaled to a range from 0.9 for the most outliers to 0.5 for the fewest. These values were propagated through the network shown in Figure 1 as detailed in Methods section, and the relative probabilities for the TFs were taken as prior relative probabilities of the expression of their targets. These provided a modified probability of a gene being associated with the first pattern in the matrix factorization. There was no effect on the other patterns, which retain flat priors across all genes.

CoGAPS was run seeking three to nine patterns. Six patterns provided the best factorization of the HNSCC data based on the visual separation of normal, HPV+, and HPV– groups.

This factorization produced two flat patterns and four patterns showing differing levels in the **P** matrix between subjects. In order to determine if the patterns provided a separation of tumors from normals, we clustered the pattern data using hierarchical clustering with average linkage and Euclidean distance (Fig. 3). The two clusters of patients defined by the first split were then tested for the separation of tumors

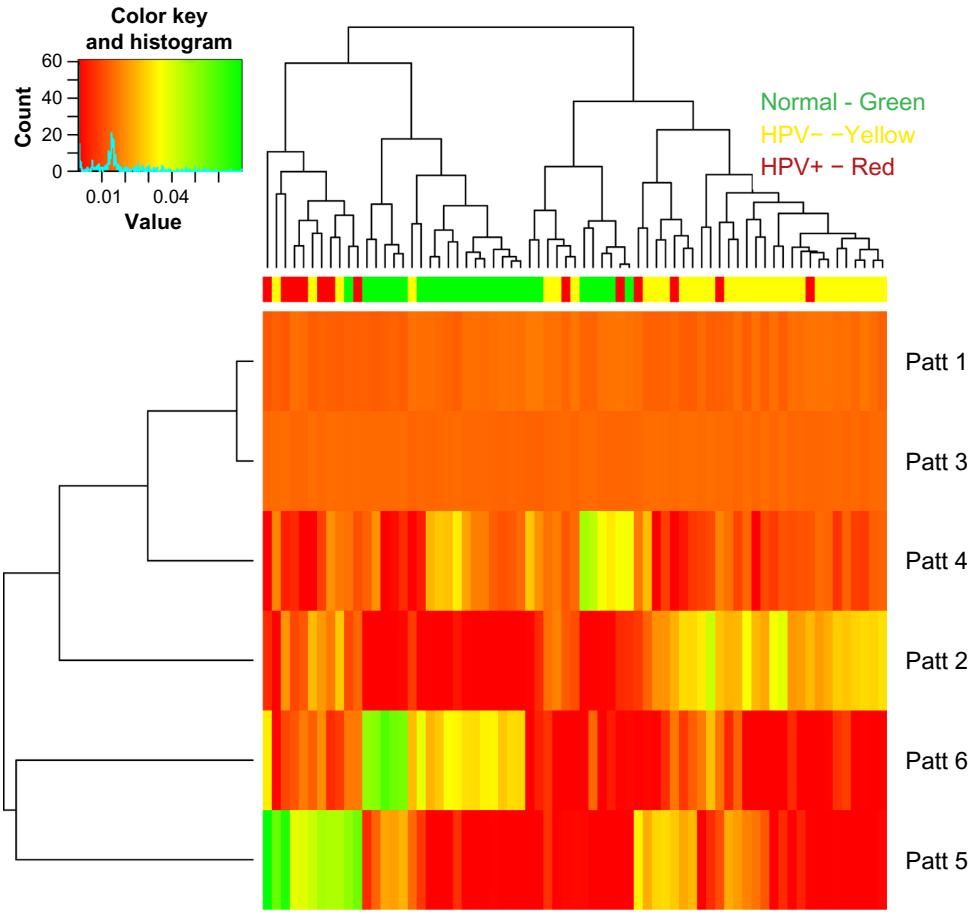

Figure 3. Heat map showing hierarchical clustering of subject types across the patterns. The values in the heat map provide the level of association of a sample with a pattern. Class labels are presented in the top bar: HPV+ tumors (red), HPV– tumors (yellow), or normal tissue (green).

and normals by Fisher's exact test, which provided a *P*-value of 0.06. This suggests that there is separation of tumors and normals beyond chance, although not to the typically applied α level.

The four patterns with interpatient variation (Fig. 4) showed differing statistics for the TF activities. ELK1 showed low activity in tumor samples, while HIF1A, SP1, and FOXO all showed strong activity in HPV– tumor samples. MYC showed low activity in the HPV– and normal samples and some slight activation in the HPV+ samples. Overall, HIF1A and FOXO provided the strongest *Z*-scores in the four patterns with minimal overlap, so we focused on the targets of these TFs for generating a TSP-based biomarker.

The TSP analysis of HIF1A and FOXO targets and reference genes (Table 1) produced five pairs of genes that could serve as a biomarker. These pairs are listed in Table 2. The genes HMOX1, TF, and HIF3A are the targets of HIF1A, and the genes BLNK and SELL are the targets of FOXO. The set of genes paired with these TF targets is from our reference gene list. Because the reference genes have stable expression throughout all subjects, using these TSPs as biomarkers will allow us to detect HNSCC even if a sample is contaminated with normal tissues.

A receiver operator characteristic (ROC) analysis of the TSP-based tumor vs. normal biomarker was performed, and the sensitivity and specificity for a threshold of three votes from the five TSPs were 0.91 and 0.92, respectively. Figure 5A shows the full ROC curve for this model generated by changing the number of votes needed to generate a tumor call.

The fivefold cross validation of the biomarker for tumor vs. normal generated an error rate of 28.5%. We applied the biomarker to predict the cancer status in the TCGA data. We obtained a sensitivity of 0.855, a specificity of 0.674, an accuracy of 0.773, and an MCC of 0.54 using the biomarker on the TCGA data. To address the issue that there were 515 tumor samples but only 44 normal samples in the TCGA data, we used a balanced bootstrap to estimate this result. We generated 100 bootstrap samples, comprising 44 normal samples and 44 tumor samples, and generated the measures from these samples. Then, we also generated an ROC curve for the measurements and estimated the AUC at 0.84. The ROC curve is shown in Figure 5B.

Discussion

HNSCC is a heterogeneous disease, which has contributed to a lack of accurate prognostication, treatment planning, and

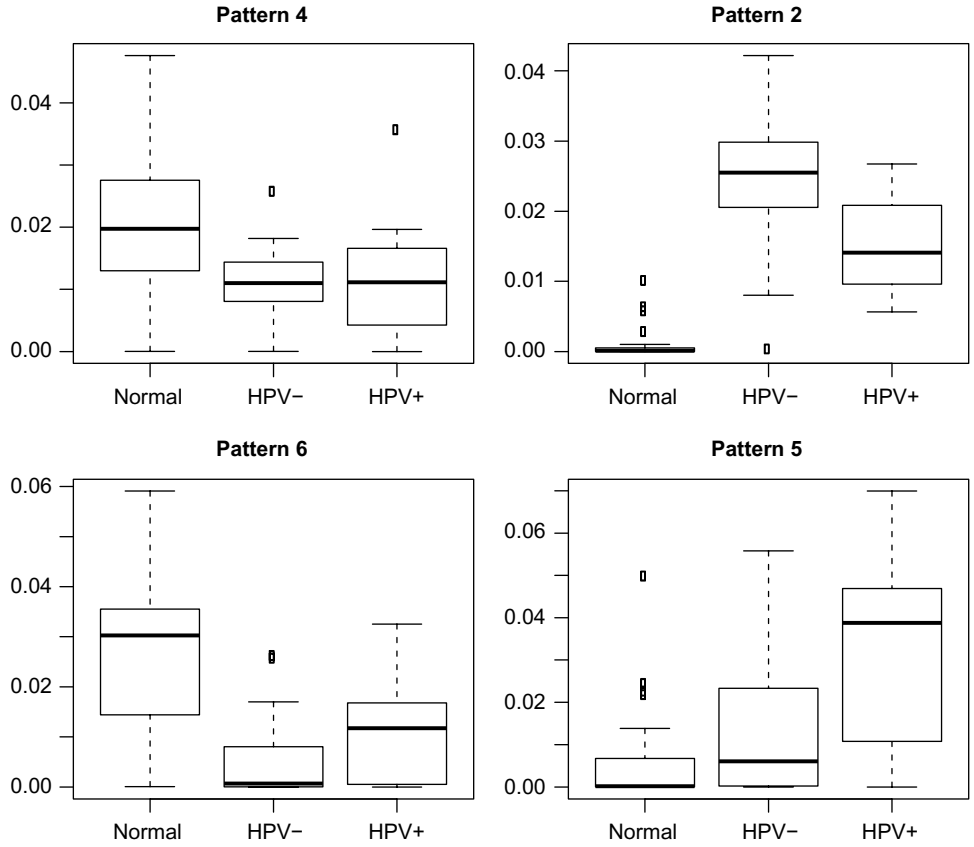

Figure 4. Boxplots of the strength of each sample in the patterns related to disease status produced by running CoGAPS with the HNSCC network prior.

identification of pivotal genes as the cause of tumor growth.[5] It is possible to distinguish several subclasses of HNSCC through histological studies, and RNA and DNA profiling studies have helped to identify further subtypes of the disease. A thorough review of expression studies in HNSCC is provided in Ochs and Califano.[19] The current study aimed to provide an approach to generate robust, minimally invasive biomarkers that could be used to identify the presence of disease.

The overall poor prognosis of HNSCC, especially HPV– disease, has been linked to the lack of early detection. Therefore, the development of minimally invasive biomarkers could substantially improve prognosis. We tested our biomarker comprising five TSPs developed from a microarray-based study to the TCGA HNSCC data set, where RNAseq was used. Despite the change in measurement platform, the biomarker performed well with an accuracy of 77.3%, which reflects the design of the TSP method to use internal normalization through seeking a change in a relative expression of just two genes at a time.

Table 1. The target genes of HIF1A and FOXO and the reference gene list from which the biomarker of Table 2 was developed.

HIF1A AND FOXO TARGETS	REFERENCE GENE LIST
ANGPTL2 IGFBP1 FBXO32 RBL2 GALT NR2C2 TNFRSF10A TNFRSF10B ESR1 ID1 BLNK CCL20 CTGF G6PC GADD45A NOS3 PRL RAG1 RAG2 SEPP1 SIRT1 ATG12 CCR7 EDN1 GABARAPL1 INS KLF2 RUNX2 SCN5A SELL AKT1 BCL2L11 BECN1 MAP1LC3B PIK3CA TNFSF10 TRIM63 IFNB1 MMP9 EGR1 FSHB MYOCD TNF VEGFA TSC22D3 PGK1 LDHA TERT HIF3A PPARA ENO1 HMOX1 BACE1 EPO EDN1 SERPINE1 TF TFRC	TOP3A ACTR8 PTCD1 ZFYVE27 IRGQ MAPK11 NDOR1 MUL1 TBC1D25 SSH3 HOXB4 COPS7B UBIAD1 POLR3H MYBBP1A ZNF74 ST7L RHBDD1 RNF26 MLL2 CIAO1 RUNDC3A TMEM161A GRWD1 NCAPH2 FAM192A C7orf49 SAP130 UBOX5 EDC3 ADC BAP1 ATAD3A ZNF408 SLC25A42 TAF5L C6orf47 HDGFRP2 TCEB2 PMS2P1 PPIL2 AKAP8 TUBA3C PPIL2 TGFBRAP1 GIGYF2 SLC41A3 FOXK2

Table 2. Table of TSPs produced from the analysis of the targets of HIF1A and FOXO to find a biomarker for differentiating HNSCC from normal tissue. Column one is the gene from the reference gene list, while column 2 provides the target of the TF identified by CoGAPS. The third column contains the score of the TSP.

GENE 1 (REFERENCE GENES)	GENE 2 (TF TARGETS)	TSP SCORE
MYBBP1A	HMOX1	0.470
ZNF74	TF	0.448
UBOX5	HIF3A	0.225
COPS7B	BLNK	0.806
RHBDD1	SELL	0.669

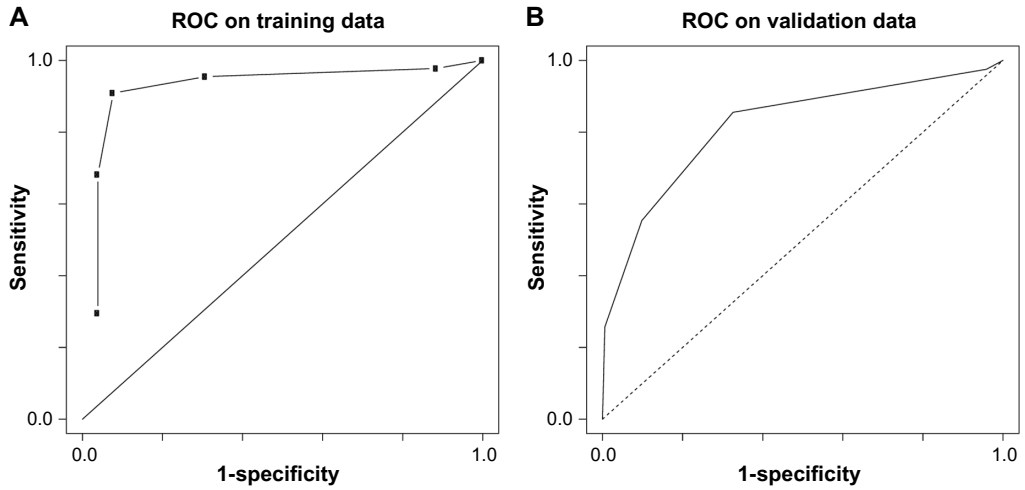

Figure 5. ROC curves for the results of the TSPs as predictors for cancer in the original data set (**A**) and in the TCGA data generated by bootstrapping (**B**). Six thresholds (0–5) for the number of votes required to determine the case vs. control were used for producing these plots.

This work provides the initial methodology of utilizing multiple biomolecular measurements for prior information on the signaling network, deduction of the key TFs related to the signaling activity, curation of the targets of the TFs as potential expression markers, use of a reference set of genes that are stably expressed in most normal tissues, and use of TSP to build a robust biomarker. Future studies will focus on adding the consideration of overall expression levels in tumors, so that we can refine the biomarker to one likely to find an adequate signal even in the case where the tumor sample is highly diluted relative to normal tissue, and on further curation of genes associated with specific TFs in the network. An ideal biomarker for other cancers would also be circulating in blood, allowing a noninvasive test. As such, seeking signaling-driven secreted proteins or stable miRs that show the same relative changes between tumors and normals would be desirable although of greater difficulty.

Author Contributions

Conceived and designed the experiments: MFO. Analyzed the data: JCS, MFO. Wrote the first draft of the manuscript: JCS. Contributed to the writing of the manuscript: MFO. Agreed with the manuscript results and conclusions: JCS, MR, RS, CK, DAG, EJF, JAC, MFO. Jointly developed the structure and arguments for the paper: EJF, MFO. Made critical revisions and approved the final version: MFO. All the authors reviewed and approved the final manuscript.

Supplementary File

The reduced data sets and their analysis are presented with the R script and Rda files in the supplementary material: Analysis_and_Data.zip.

REFERENCES

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646–74.
2. Kossenkov AV, Ochs MF. Matrix factorization for recovery of biological processes from microarray data. *Methods Enzymol*. 2009;467:59–77.
3. Gao Y, Church G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*. 2005;21(21):3970–5.
4. Fertig EJ, Ding J, Favorov AV, et al. CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. *Bioinformatics*. 2010;26(21):2792–3.
5. Leemans CR, Braakhuis BJ, Brakenhoff RH. The molecular biology of head and neck cancer. *Nat Rev Cancer*. 2011;11(1):9–22.
6. Morgan S, Grandis JR. ErbB receptors in the biology and pathology of the aerodigestive tract. *Exp Cell Res*. 2009;315(4):572–82.
7. Ratushny V, Astsaturov I, Burtness BA, et al. Targeting EGFR resistance networks in head and neck cancer. *Cell Signal*. 2009;21(8):1255–68.
8. Kafetzopoulou LE, Boocock DJ, Dhondalay GK, et al. Biomarker identification in breast cancer: beta-adrenergic receptor signaling and pathways to therapeutic response. *Comput Struct Biotechnol J*. 2013;6:e201303003.
9. Edelman LB, Toia G, Geman D, et al. Two-transcript gene expression classifiers in the diagnosis and prognosis of human diseases. *BMC Genomics*. 2009;10:583.
10. Irizarry RA, Bolstad BM, Collin F, et al. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*. 2003;31(4):e15.
11. Scharpf RB, Irizarry RA, Ritchie ME, et al. Using the R Package crlmm for genotyping and copy number estimation. *J Stat Softw*. 2011;40(12):1–32.
12. Parfenov M, Pedamallu CS, Gehlenborg N, et al. Characterization of HPV and host genome interactions in primary head and neck cancers. *Proc Natl Acad Sci U S A*. 2014;111(43):15544–9.
13. Matys V, Kel-Margoulis OV, Fricke E, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*. 2006;34(Database issue):D108–10.
14. Ochs MF, Farrar JE, Considine M, et al. Outlier analysis and top scoring pair for integrated data analysis and biomarker discovery. *IEEE/ACM Trans Comput Biol Bioinform*. 2014;11(3):520–32.
15. Ochs MF. Bayesian decomposition. In: Parmigiani G, Garrett E, Irizarry R, Zeger S, eds. *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer Verlag; 2003:p.388–408.
16. Ochs MF, Rink L, Tarn C, et al. Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. *Cancer Res*. 2009;69(23):9125–32.
17. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). *Biostatistics*. 2010;11(2):242–53.
18. Afsari B, Fertig EJ, Geman D, et al. switchBox: an R package for k-Top Scoring Pairs classifier development. *Bioinformatics*. 2015;31(2):273–4.
19. Ochs MF, Califano JA. Molecular determinants of head and neck cancer. In: Golemis EA, Burtness BA, eds. *Molecular Determinants of Head and Neck Cancer*. New York: Springer; 2014:325–42.