Gene expression profiling identifies genes predictive of oral squamous cell carcinoma

Chu Chen[1,2,3], Eduardo Méndez[1,3,4], John Houck[1], Wenhong Fan[5], Pawadee Lohavanichbutr[1], Dave Doody[1], Bevan Yueh[1,3,4,6], Neal D. Futran[3], Melissa Upton[7], D. Gregory Farwell[8], Stephen M. Schwartz[1,2], Lue Ping Zhao[5,9]

[1] Program in Epidemiology, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

[2] Department of Epidemiology, University of Washington, Seattle, WA 98195

[3] Department of Otolaryngology – Head and Neck Surgery, University of Washington, Seattle, WA 98195

[4] Surgery and Perioperative Care Service, VA Puget Sound Health Care System, Seattle, Washington 98108

[5] Program in Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

[6] Department of Otolaryngology – Head and Neck Surgery, University of Minnesota, Minneapolis, MN 55455

[7] Department of Pathology, University of Washington, Seattle, WA 98195

[8] Department of Otolaryngology: Head and Neck Surgery, University of California at Davis, Sacramento, CA 95817

[9] Department of Biostatistics, University of Washington, Seattle, WA 98195

Running Title: Genetic expression profiles of squamous oral cancer

Keywords: oral squamous cell carcinoma, oral cancer, genetic expression profiles, microarrays

Acknowledgments: This study was supported by grants from the U.S. National Institutes of Health (R01CA095419 from the National Cancer Institute, National Research Service Award T32DC00018 from the National Institute on Deafness and Other Communication Disorders, and trans-NIH K12RR023265 Career Development Programs for Clinical Researchers) and by institutional funds from the Fred Hutchinson Cancer Research Center. We thank the study participants and their families. We also thank Carolyn Anderson, Ashley Fahey, Lora Cox, Cynthia Parks and Kathleen Vickers for administrative and technical support.

Corresponding Author: Chu Chen, PhD

Fred Hutchinson Cancer Research Center

1100 Fairview Avenue North, M5-C800

PO Box 19024

Seattle, WA 98109-1024

Phone: (206) 667-6644

Fax: (206) 667-2537

e-mail: cchen@fhcrc.org

Abstract

Oral squamous cell carcinoma (OSCC) is associated with substantial mortality and morbidity. To identify potential biomarkers for early detection of invasive OSCC, we compared gene expression of incident primary OSCC, oral dysplasia, and clinically normal oral tissue from surgical patients without head and neck cancer or pre-neoplastic oral lesions (controls), using Affymetrix U133 2.0 Plus arrays. We identified 131 differentially expressed probe sets using a training set of 119 OSCC patients and 35 controls. Forward and stepwise logistic regression analyses identified 10 successive combinations of genes which expression differentiated OSCC from controls. The best model included *LAMC2,* encoding laminin gamma 2 chain, and *COL4A1,* encoding collagen, type IV, alpha 1 chain. Subsequent modeling without these two markers showed that *COL1A1*, encoding collagen, type I, alpha 1 chain, and *PADI1*, encoding peptidyl arginine deiminase, type 1, also can distinguish OSCC from controls. We validated these two models using an internal independent testing set of 48 invasive OSCC and 10 controls and an external testing set of 42 head and neck squamous cell carcinoma (HNSCC) cases and 14 controls (GEO GSE6791), with sensitivity and specificity above 95%. These two models were also able to distinguish dysplasia (n=17) from control (n=35) tissue. Differential expression of these four genes was confirmed by qRT-PCR. If confirmed in larger studies, the proposed models may hold promise for monitoring local recurrence at surgical margins and the development of second primary oral cancer in OSCC patients.

Introduction

Squamous cell carcinoma of the oral cavity and oropharynx (OSCC) is of considerable public health significance. In the United States, it is estimated that nearly 35,000 new OSCC cases were diagnosed in 2007, and approximately 7,550 OSCC deaths are estimated to occur (http://www.cancer.org). World-wide, OSCC is the 6th most common caner, with an estimated 405,000 new cases and 211,000 deaths annually (http://www-dep.iarc.fr) (1). Despite considerable advances in surgical techniques, and the use of adjuvant treatment modalities, the 5-year survival for OSCC patients is about 60% for U.S. Whites and 36% for U.S. Blacks (http://www.cancer.org). In addition, OSCC is often associated with loss of eating and speech function, disfigurement and psychological distress.

As much as 20% of oral dysplasia undergoes malignant transformation to OSCC (2, 3). Among OSCC patients with histologic positive tumor margins, the likelihood of local recurrence is as high as 70 to 80%. Even among patients with negative margins, the reported probability of recurrence is 30-40% (4), suggesting histologic examination alone is inadequate in predicting recurrence (4-6). There is an urgent need to identify better ways to predict which patients with dysplastic precursor lesions will develop OSCC and which patients treated for OSCC will develop recurrence, so that high-risk patients can be selected for more rigorous treatment and follow-up. We hypothesize that patients who develop local recurrence and/or second primary oral tumors are those whose surgical margins or uninvolved buccal mucosa harbor molecular changes that are found in oral dysplasia or invasive OSCC. In this report, we present results on the differential gene expression profiles between OSCC, oral dysplasia and normal controls and several predictive models t that 1) can potentially be easily used to test biopsies of histologically normal surgical margins and clinically normal oral mucosa of OSCC patients for the prediction

of local recurrence and/or second primary oral cancer; and 2) enhance our understanding of the underlying biological mechanisms of this disease.

Materials and Methods

Study Population. Eligible cases were patients with their first primary OSCC scheduled for surgical resection or biopsy between December 1, 2003 and April 17, 2007 at the University of Washington Medical Center, Harborview Medical Center and the VA Puget Sound Health Care System in Seattle, Washington. We also sought to enroll patients with diagnosed dysplastic lesions at these medical centers during the same period. Eligible controls were patients who had tonsillectomy or oral surgery for treatment of diseases other than cancer, such as obstructive sleep apnea, at the same institutions and during the same time periods in which the OSCC cases were treated. All three groups of patients were 18 years of age or older and capable of communicating in English.

Among 244 eligible OSCC patients, we were able to consent 187 patients. Of these, 171 patients gave permission for medical chart abstraction and provided sufficient tissue to yield GeneChip arrays results that passed our quality control (QC) criteria (see below). Among 21 eligible dysplasia cases, 15 provided consent for the study. Of these, 11 patients had GeneChip results passed QC checks. One dysplasia patient provided dysplasia tissues from two different sites. One OSCC patient provided one piece of cancer tissue and one piece of dysplasia tissue, and assay results from this latter tissue were grouped with the dysplasia patients. Four of the eligible patients originally believed to have OSCC had a final pathology report of dysplasia, and these were included in the dysplasia group, and not in the OSCC group for analyses. In total, 17 dysplasia samples were used for analyses. During the case recruitment period, 47 of 55 eligible

controls consented to participate. Samples from two controls failed QC checks leaving 45 for analyses.

Each participant was interviewed using a structured questionnaire regarding demographic, medical, functional, quality of life, and lifestyle history, including tobacco and alcohol use. Tumor characteristics (site, stage) were obtained from medical records. This study was conducted with written informed consent and Institutional Review Office approvals.

Tissue Collection. Tumor tissue was obtained at time of resection or biopsy from patients with a primary OSCC, or dysplasia. Clinically normal tissue from the oral cavity or oropharynx was obtained from controls. For the small number of controls (~30%) with tonsillitis or tonsil hypertrophy, only mucosa tissue from tonsillar pillar was obtained to avoid potential influence of inflammation on the results. Immediately after surgical removal, the tissue was immersed in RNALater (Applied Biosystems, Inc. Foster City, CA) for a minimum of 12 hours at 4 ° C before being transferred to long term storage at – 80 ° C prior to use.

DNA Microarray. Total RNA was extracted using a TRIzol method (Invitrogen, Carlsbad, CA), purified with an RNeasy mini kit (Qiagen, Valencia, CA), processed using a GeneChip Expression 3'-Amplification Reagents Kit (Affymetrix), and interrogated with an Affymetrix U133 2.0 Plus GeneChip arrays (see Supplemental Material for experimental details).

QC Checks of GeneChip Results. We conducted two rounds of QC checks to evaluate whether to include results from each of the GeneChips. In the first round, recommendations made by Affymetrix (http://www.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf) were followed. In the second round, we used the "affyQCReport" and "affyPLM" software in the Bioconductor package (http://www.bioconductor.org) to search for poor quality chips. In

total, 172 chips from 165 patients (119 OSCC patients, 35 controls and 11 dysplasia patients passed two rounds of QC evaluation.

Preprocessing and Probe Set Filtering. For those GeneChip arrays that passed QC checks, we used gcRMA algorithm from Bioconductor to extract gene expression values and perform normalization. Next, to limit the multiple testing penalty in the statistical testing step, we eliminated the probe sets that either showed no variation across the samples being compared (inter quartile range (IQR) of expression levels less than 0.1 on log2 scale) or were expressed at very low magnitude (any probe set in which the maximum expression value for that probe set in any of the samples was less than 3 on log2 scale). After these criteria were applied, ~21,000 probe sets remained for differential expression analyses.

Differential Gene Expression Analyses. To examine differential gene expression and to build prediction models, we divided our samples into a training set of 119 OSCC cases and 35 controls and a testing set of 48 OSCC cases and 10 controls. The division of study subjects into training and testing sets was based on the calendar date that patients were enrolled into the study. Gene expression values from gcRMA were analyzed using a regression-based, estimating equations, approach implemented in GenePlus software (http://www.enodar.com/) (7, 8). Age and sex were included as covariates in the analyses of the training set. To control type I errors, we declared a particular group of genes either "upregulated/overexpressed" or "downregulated/underexpressed" based on a fixed number of false discoveries (NFD), i.e., the number of false discoveries in a list of discovered genes is controlled at the pre-specified NFD (9). The choice of NFD, with an appropriate account for the number of genes under investigation (*J*), dictates the threshold for individual gene-specific p-values as NFD/*J*. Using NFD<1 as a statistical testing criterion, we identified 7,604 probe sets as being differentially

expressed between controls and cases. To build predictive models and substantially reduce the number of comparisons, we further narrowed this list of candidate probe sets using the following criteria that retained only those probe sets that showed large difference in signal intensity between cases and controls: 1) absolute Z-score >6 in the differential gene expression analysis, implying exceptionally high statistical significance; 2) a 1.5-fold or greater difference in gene expression between controls and cases. A large difference is needed to provide good predictive ability. And, 3) the mean expression value summarized by Affymetrix Microarray Suite 5.0 across samples >300 (with the scaled mean expression value of 1000). Probe sets with such expression values are more likely to be suitable for validation by alternative methodologies such as qRT-PCR. A total of 131 probe sets were selected by these three criteria.

Biological Pathway Analyses and Hierarchical Clustering of Differentially Expressed Genes. We analyzed the 7,604 differentially expressed probe sets between OSCC and controls using Ingenuity Pathway Analysis 4.0 (Ingenuity®Systems, www.Ingenuity.com) and performed hierarchical clustering of all the samples based on their expression of the 131 probe sets using Affymetrix GeneSpring software GX7.3.1.

Prediction Models. The selected 131 probe sets were analyzed using both forward and hybrid of forward-backward logistic regression procedures (SAS PROC LOGISTIC). For the one OSCC case with results from 5 replicate tissues and one control with results from duplicate tissues, the respective average of the replicate results was used. In the forward stepwise selection, probe sets were processed in the logistic regression model: one probe set at a time until no probe set could be added based on the significance level of 0.01. When the hybrid stepwise selection was adopted, the probe set with the smallest p-values and $p< 0.01$ entered first, and significance levels for other selected probe sets were evaluated for possible removal if their p-values were

greater than 0.05 in the current model. We compared the performance of the two models (results from the forward and hybrid stepwise procedures) using receiver operating characteristic (ROC) curves. An ROC curve is a plot of true positive rate (sensitivity) on the Y-axis against false positive rate (1-specificity) on the X-axis for each possible value (in our case, the logistic score for each individual for a given model) representing a positive test,. A model with perfect discrimination between cases from controls will have a ROC curve that passes through the upper left corner, with 100% sensitivity, 100% specificity, and area under the curve (AUC) of 1. An AUC=0.5 represents a test that is no better than chance at discriminating between cases and controls (10-12).

Validating Prediction Models. We validated the selected prediction models with our own independent validation dataset and an external validation dataset from GEO (Gene Expression Omnibus, www.ncbi.nlm.nih.gov/geo, GSE6791 containing 42 HNSCC cases and 14 controls) (13). CEL files from these datasets were extracted using gcRMA algorithm. ROC curves were drawn by applying the expression results to the prediction models.

Comparison of Gene Expression of the Prediction Models in Different Tissues to Test the Specificity of the Models for OSCC. We downloaded gene expression data from GEO GSE6791 for normal and tumor cervical tissue samples and GSE6044 for normal and tumor lung samples. We chose these datasets because: 1) they were generated using the same Affymetrix U133 GeneChip platform as ours, facilitating testing the tissue specificity of our predictive models; and 2) OSCC share some of the same risk factors as cervical cancer and lung cancer; Human Papillomavirus in the case of cervical cancer and cigarette smoking in the case of both cervical and lung cancer. We extracted gene expression values using gcRMA and, for each tissue

type, calculated the scores for each of the prediction models derived from analysis of our training dataset.

Comparison of Gene Expression Profiles in Controls, Dysplastic Lesions and Invasive Cancer. While the expression of some genes may be continuously increasing or decreasing from the moment normal oral tissue begins its oncogenic process, it is also possible that some genes get turned on or off during the conversion from dysplasia to invasive cancer. To explore this hypothesis and to identify genes that may be specific for the conversion of dysplasia to OSCC, we compared gene expressions of invasive cancer (n=167) with those of normal oral tissue (from 45 controls) and dysplastic lesions (n=17) combined using ~21,000 filtered probe sets. From those probe sets that were differentially expressed between OSCC samples and the combination of controls and dysplastic lesions, we further excluded those that were differentially expressed between controls and dysplasia using NFD=1 (see Supplemental Material for schematic representation of the method for selecting the differentially expressed genes specific to OSCC). The resulting gene list contained the genes that were up- or downregulated in OSCC but not in dysplasia. Conversely, we combined dysplastic lesions and OSCC samples and compared them with the controls. For those probe sets showing differential expression, we excluded the genes that were also differentially expressed between dysplasia and cancer. The resulting gene list contained genes that showed up- or downregulation (relative to normal tissue) as early as dysplasia.

Validation of Gene Expression of *LAMC2, COL4A1, COL1A1,* and *PADI1* by qRT-PCR. qRT-PCR was performed in triplicate on a subset of 30 OSCC cases and 30 controls using a QuantiTect SYBR Green RT-PCR kit (Qiagen, Valencia, CA) and bioinformatically validated

QuantiTect primers (Qiagen, Valencia, CA) on a 7900HT Sequence Detection System (ABI, Foster City, CA) (See experimental details in Supplemental Material).

Results

The cases in both the training and testing sets tended to be older than the controls. Compared to controls, cases were more likely to be male, white, and current smokers. Approximately two thirds of the cases had AJCC stage III or IV disease with about 50% of the cases presenting with metastasis to the neck. Oral cavity tumors accounted for 74% and 60% and oropharyngeal tumors account for 26% and 40% of the OSCC cases in the training and testing sets, respectively. Most of the dysplasia subjects were White males whose lesions were located in the oral cavity (see Supplemental Table 1).

Results obtained with the Ingenuity Pathway Analyses tool showed that the JAK/STAT signaling pathway and the IFG-γ signaling pathway were the top two biological pathways associated with the differentially expressed genes. Figure 1 shows genes that were up- or downregulated in these two pathways in our training dataset.

Table 1 lists the 131 probe sets differentially expressed between OSCC and controls based on the criteria described in the Methods. Among the 131 probe sets were transforming growth factor *TGFB1*, cell signaling molecule *STAT1*, immune markers *IL1β*, chemokines *CXCL2, 3, 9*, and genes encoding for extracellular matrix proteins and collagens that have previously been shown to be involved in the motility and invasion of tumor cells. Hierarchical clustering of gene expression using the 131 probe sets showed that invasive OSCC and normal control formed two main clusters. About half the dysplasia tissues clustered with OSCC samples and half clustered with the controls. Compared to invasive OSCC, oral dysplasia tissue appeared to have a set of

genes that were not yet upregulated and another set of genes that were not yet downregulated (see heat map in Supplemental Material).

Table 2 lists the top 10 models from the logistic regression analyses of the 131 probe sets in our training data set. The model with *LAMC2* (probe set 207517_at, encoding laminin $\gamma 2$) and *COL4A1* (211980_at, encoding collagen type IV, $\alpha 1$) had the most discriminating power to separate OSCC from controls (AUC=0.99952). The power to distinguish OSCC from controls was very slightly reduced if expression of only one of these two probe sets was used (AUC=0.99424 with *COL4A1* alone). After removing *LAMC2* and *COL4A1* from subsequent modeling, *COL1A1* (202310_s_, encoding for collagen type I, $\alpha 1$) and *PADI1* (220962_s_, encoding for peptidyl arginine deimminase type 1) emerged as the next set of markers that best separated OSCC from controls (AUC=0.99976).

When we applied the expression values from the testing datasets to the predictive models derived from our training dataset, the model with *LAMC2* (probe set 207517_at) and *COL4A1* (211980_at) had the most discriminating power to separate OSCC from controls: AUC=0.997 in our independent testing set and AUC=0.976 in the external testing set (GEO GSE6791), respectively (Table 2). The model with *COL1A1* and *PADI1* also was strongly predictive (AUC=0.99167 in our testing set, and AUC=0.97789 in the external GEO GSE6791 data set (Table 2). Results on the testing of the other eight models against the internal and external datasets indicate that they also performed well in distinguishing OSCC from controls (Table 2). Results of qRT-PCR on *LAMC2, COL4A1, COL1A1* and *PADI1* confirmed the differential expression of these genes between OSCC and controls at the transcript level (Table 3).

We next examined whether the top two models that were particularly effective in discriminating OSCC from controls were specific to OSCC (or HNSCC) and not other epithelial

cancer types with overlapping risk factors. For each of these two predictive models, we compared the scores for cases and controls calculated from our testing dataset to the scores from the GEO HNSCC dataset (GSE6791) and from the GEO cervical cancer and lung cancer data sets (GSE6044) and their controls. The model containing *LAMC2* and *COL4A1* distinguished HNSCC from controls, but not cervical cancer nor lung cancer from their respective controls (Figure 2, top panel); *COL1A1 and PAD*I1 also performed well for HNSCC and, to a lesser extent, for lung cancer, but not cervical cancer (Figure 2, bottom panel). Furthermore, our results showed that these two models could not only distinguish invasive cancer from controls, but also distinguish oral dysplasia from controls. The respective AUC was 0.98 for *LAMC2* and *COL4A1* and 0.99477 for *COL1A1* and *PADI1*. However, the effect we observed here for the model *LAMC2* and *COL4A1* was driven by *COL4A1,* suggesting *COL4A1* up-regulation occurs earlier than *LAMC2* up-regulation in oral carcinogenesis (data not shown).

Comparison of gene expressions of invasive cancer with those of normal oral tissue (from controls) and dysplasia combined using ~21,000 filtered probe sets, followed by elimination of those probe sets that were differentially expressed between dysplasia and controls, showed the differential expression of 6544 probe sets, including 3988 upregulated and 2666 downregulated probe sets in invasive OSCC. Table 4 lists among the 131 probe sets the 49 probe sets that may be specific for the conversion of oral dysplasia to OSCC. Sixty-seven probe sets that may be specific for the development of dysplasia from normal are provided in the Supplemental Material.

Discussion

We have identified 131 probe sets, corresponding to 108 known genes, which are highly effective in distinguishing invasive OSCC and normal oral tissue, as well as a list of genes that might be involved in the transformation of normal oral tissue to dysplasia, and of oral dysplasia to invasive OSCC. Although prior studies, including our own, have described global changes in gene transcription that distinguish normal oral epithelium from carcinoma, there is considerable heterogeneity among the lists of genes that have been reported and, to our knowledge, few studies have produced a limited combinations of genes as in the current study with high sensitivity and specificity in distinguishing OSCC from normal oral tissue through rigorous statistical testing and validation with independent datasets, and none had provided prediction models (14). The current study provides prediction models that were generated using rigorous statistical analyses, and the differences in gene expression detected using microarray technology was validated by qRT-PCR, and by testing against independent internal and external genome-wide gene expression datasets. The ultimate goal of our work has been to generate candidate markers that can be easily applied to the testing of biopsies or surgical margins to aid diagnosis and prognosis of OSCC. It is our hope that the signals we identify will be strong enough to use in a clinical test without resorting to the isolation of the tumor cells and stromal cells, knowing that both cell populations play important role in OSCC development and progression. Thus, we have deliberately choose not to use laser capture microdissection to isolate tumor cells for this investigation. We believe that our current prediction models and the 131 genes that we identified warrant testing in subsequent studies for their utility in predicting local recurrence at surgical margins or the development of second primary cancer of OSCC patients, or for selective screening of individuals who are at high risk of OSCC. It is possible that histologically- negative margins harbor microscopic original tumor as residual disease. If so, the gene expression profile would more likely resemble that of the resected invasive OSCC, and measurement of one or more of the 131 genes we identified and

application of one of our top models could potentially be of use for its detection. For individuals who are at high risk of OSCC, their oral epithelium could contain cells that are molecularly abnormal and primed for the development of cancer. As such, the molecular profile might be more similar to that of a pre-neoplastic oral lesion than that of an invasive OSCC. The list of genes that we generated that distinguishes invasive OSCC from dysplasia and controls could potentially be used to gauge malignant potential of these molecular changes. Recently, p53 and eIF4E have been evaluated to augment histologic assessment of surgical margins (4, 15). eIF4E expression, but not P53 mutation and overexpression, in histologically negative surgical margins was a significant predictor of recurrence and shorter disease-free survival of HNSCC patients (16-18).

In the current study, we found that the expressions of two pairs of genes (*LAMC2* and *COL4A1; COL1A1* and *PADI1*) were particularly effective in distinguishing OSCC from normal oral tissue in independent testing sets. The sensitivity and specificity were close to 100%. Because of the stringent criteria we applied to select candidate markers, it is expected that there are other probe sets among the 131 probe sets with a similar predictive property. We previously observed the differential expression of many of the 131 probe sets, including *LAMC2, COL1A1 and COL4A1* (19). Overexpression of laminin gamma 2 in HNSCC, particularly in the invasive front of tumors, has been reported by others (20, 21). A study by Pyeon et al (13) that used normal controls (n=14) and the same Affymetrix GeneChip arrays also found highly expressed *LAMC2*, *COL4A1* and *COL1A1* in OSCC (n=42), compared to controls. A study by Ziober et al (22), using Affymetrix U133 GeneChip arrays to compare gene expression of oral cavity tumors and paired adjacent clinically normal oral tissue from 13 patients, produced a list of 25 genes that showed 86-89% accuracy in distinguishing OSCC from controls in three small testing datasets

that contained 13, 18 and 5 tumor samples and even fewer controls. Only seven of the 25 probe sets, encoding for COL1A1, 4A1, 5A1, 5A2, microtubule, periostin and podoplanin, were among our list of 131 probe sets. Given the differences between their study and ours, i.e., sample size, tumor site, source of control samples, analytical methods and the sample size of the testing sets, the common observation of differential expression of collagen genes and genes involved in cell shape and movement underscores the potential importance of these genes in oral carcinogenesis. Another study of gene expression signature (23), involving comparison of oropharyngeal tumor samples from three patients with adjacent normal nonmalignant mucosa using a 9,350 EST cDNA array, reported differential expression of nine genes (23). Only periostin in their list was among our 131 top candidate markers.

Our results were adjusted for age and sex. Although life style characteristics, such as tobacco use and infection with Human Papillomavirus (HPV) play an important role in OSCC development, we did not observe any appreciable difference in gene expression on the genome-wide level according to smoking status (former/current vs. never) or HPV status (positive vs. negative). Only when we examined oropharyngeal cancers alone, did we find differential gene expression between HPV-positive and HPV-negative tumors. The latter results have been submitted for review in a separate manuscript (Lohavanichbutr et al).

Laminin binds to Type IV collagen and to many cell types via cell surface laminin receptors (24). Following attachment to laminin in the basement membrane, tumor cells secrete collagenase IV that specifically breaks down type IV collagen thus facilitate cell spreading and migration (25). In addition, laminin fragments generated by post-translational proteolytic cleavage bind to cell surface integrins and other proteins to trigger and modulate cellular motility (26). Increased levels of laminin have been associated with a number of carcinoma (27-35). In

some of these studies, laminin was associated with tumor aggressiveness, metastasis and poor prognosis. Results from mouse models showed that tumor cells with high levels of laminin and low level of unoccupied laminin receptor are resistant to killing by natural cytotoxic T cells and are highly malignant (36) and that treatment with low concentrations of laminin receptor binding fragments of laminin blocked lung metastasis of hematologenously introduced tumor cells (37). A large number of unoccupied laminin receptors have been observed for breast and colon cancer cells (25); no similar reports have appeared on OSCC or HNSCC cells. Further studies of laminin and its receptors should be pursued for its role in OSCC etiology and progression.

The gene products of *COL4A1* and *COL4A2* are assembled into type IV collagen that form the scaffold of basement membrane integrating other extracellular molecules, including laminin, to produce a highly organized structural barrier. Collagen IV also plays an important role in the interaction of basement membrane with cells (38, 39). Immune cells, migrating endothelial cells and metastatic tumor cells have been reported to produce and tightly regulate type IV collagen-specific collagenase (40-42). Degradation of Type IV collagen could compromise basement membrane integrity and facilitate tumor cell spreading and migration. It is possible that the observed overexpression of *COL4A1* by our study and by Pyeon et al is the net result of overproduction and degradation. Whether COL4A1 contributes to OSCC development is unknown and awaits investigation.

Peptidyl arginine deiminases (EC 3.5.3.15) catalyze post-translational modification of proteins through conversion of arginine residues to citrullines. Although their physiological functions are not well understood, they have been implicated in the genesis of multiple sclerosis, rheumatoid arthritis, and psoriasis (43). The isoform peptidyl arginine deiminases type 1 (PADI1) is present in the keratinocytes of all layers of human epidermis. It has been reported that deimination of

filaggrin by PADI1 is necessary for epidermal barrier function and deimination of keratin K1 may lead to ultrastructural changes of the extracellular matrix (43). We found the expression of *PADI1* to be downregulated in both dysplasia and OSCC when compared to controls. If deimination of arginine residues of proteins in the keratinocytes of oral mucosa by PADI1 forms an epidermis barrier, downregulation of PADI1 may allow the growth, expansion and movement of tumor cells. Given the strength of our observation, it would be important to examine the function of PADI1 in cell lines and animal model systems.

Among the biological pathways we identified to be prominently involved in OSCC were the JAK/STAT and interferon gamma (IFN-γ) signaling pathways. A wide array of cytokines and growth factors, including EGFR, transmit signals through the JAK/STAT pathway (44, 45). EGFR overexpression has been reported in up to 90% of HNSCC tumors (46). Single modality therapeutics that target against EGFR, such as small molecule tyrosine kinase inhibitors, monoclonal antibodies, antisense therapy or immunotoxin conjugates, however, were only effective in 5-15% of patients with advanced HNSCC (47). These observations suggest that there are other proteins and pathways driving the growth of some of these tumors. To our knowledge, this is the first study to show a strong association between IFN- γ signaling pathway and OSCC. Interestingly, IFN- γ signaling also involves the JAK/STAT pathway (44, 48). It is unclear whether the upregulation of the IFN- γ pathway is intrinsic to the tumor cells or is due mainly to the immune cells present in the stroma. Further studies utilizing laser capture microdissection to address this question are warranted.

We identified a set of genes that are possibly involved in, and specific for, the malignant transformation of oral dysplasia into invasive OSCC. These genes include those that encode for proteins that are known for cell-matrix and cell-cell interaction, cellular migration or invasion,

such as *LAMC2* and, *SERPINE1 (PAI-1)*; for directed-cellular movement, such as *CXCL2, 3*, and *9*, as well as for immune function, such as *IL1β* and *IFIT3*. Due to the small number of dysplasia cases we studied, however, we were not able to separate the samples into a training set and a testing set. Another limitation is that the comparisons were made between dysplasia samples collected from the oral cavity and OSCC from both the oral cavity and oropharynx, and the controls from mucosa of oropharynx or tonsillar pillar. Thus, our results await confirmation or refutation by others. Kondoh et al (49) reported the differential expression of 27 genes between 27 OSCC and 19 leukoplakia tissues based on their IntelliGene Human Expression cDNA array and qRT-PCR. Among those 27 genes, only *LAMC2, IFIT3* and *USP18* were on our list. The observed discrepancy is not surprising, given the large number of differences between the two studies: 1) Kondoh et al compared OSCC with leukoplakias, while we compared OSCC with dysplastic lesions; and 2) that study used microdissected samples to remove stroma while we did not, and they assayed the samples with a 16,600 probe set cDNA array, as opposed to our ~50,000 probe set oligonucleotide array. Nonetheless, their study and ours show that *LAMC2, IFIT3* and *USP18* are worthy of further investigation as predictors of the development of OSCC among patients with oral dysplasia. It is interesting to point out that, among our 131 probe sets, a large number of collagen genes were among the probe sets that may be associated with the conversion of oral tissue to dysplasia (Supplemental Table) and were absent among the probe sets that may be involved in the conversion of dysplasia to invasive OSCC (Table 4). These observations suggest that collagen genes may play an important role early in the oral carcinogenesis process.

Although our sample size is substantially larger than other microarray articles published on HNSCC, it is nonetheless very small when compared to the number of genome-wide

comparisons we were making. Furthermore, the sample sizes of the internal and external testing sets that we used to test the predictive power of our proposed models were also small. Although we validated the differential expressions of the four markers in the top two models, whether these four markers will continue to exhibit the greatest predictive power remains to be seen when they are further tested in independent studies with a much larger sample size.

References

1. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. CA Cancer J Clin 2005;55:74-108.

2. Silverman SJ, Gorsky M, Lozada F. Oral leukoplakia and malignant transformation. A follow-up study of 257 patients. Cancer 1984;53:563-568,.

3. Reibel J. Prognosis of oral pre-malignant lesions: significance of clinical, histopathological, and molecular biological characteristics. Crit Rev Oral Biol Med 2003;14:47-62.

4. Upile T, Fisher C, Jerjes W et al. The uncertainty of the surgical margin in the treatment of head and neck cancer. Oral Oncol 2007;43:321-6.

5. Batsakis JG. Surgical excision margins: a pathologist's perspective. Adv Anatomic Path 1999;6:140-8.

6. Brandwein-Gensler M, Teixeira MS, Lewis CM et al. Oral squamous cell carcinoma: histologic risk assessment, but not margin status, is strongly predictive of local disease-free and overall survival. Am J Surg Pathol 2005;29:167-178.

7. Thomas JG, Olson JM, Tapscott SJ, Zhao LP. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Res 2001;11:1227-36.

8. Zhao LP, Prentice R, Breeden L. Statistical modeling of large microarray data sets to identify stimulus-response profiles. Proc Natl Acad Sci USA 2001;98:5631-6.

9. Xu XL, Olson JM, Zhao LP. A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington's disease transgenic model. Hum Mol Genet 2002;11:1977-85.

10. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283-98.

11. Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. Principles and applications. Ann Intern Med 1981;94:557-92.

12. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561-77.

13. Pyeon D, Newton M.A, Lambert PF et al. Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. Cancer Res 2007;67:4605-4619.

14. Choi P, Chen C. Genetic Expression Profiles and Biologic Pathway Alterations in Head and Neck Squamous Cell Carcinoma. Cancer 2005;104:1113-28.

15. Black C, Marotti J, Zarovnaya E, Paydarfar J. Critical evaluation of frozen section margins in head and neck cancer resections. Cancer 2006;107:2792-2800.

16. Nathan CO, Franklin S, Abreo FW, Nassar R, De Benedetti A, Glass J. Analysis of surgical margins with the molecular marker eIF4E: a prognostic factor in patients with head and neck cancer. J Clin Oncol 1999;17:2909-2914.

17. Nathan CO, Sanders K, Abreo FW, Nassar R, Glass J. Correlation of p53 and the proto-oncogene eIF4E in larynx cancers: prognostic implications. Cancer Res 2000;60:3599-604.

18. Nathan CA, Amirghahri N, Rice C, Abreo FW, Shi R, Stucker FJ. Molecular analysis of surgical margins in head and neck squamous cell carcinoma patients, Laryngoscope 2002;112:2129-40.

19. Mendez E, Cheng C, Farwell DG et al. Transcriptional expression profiles of oral squamous cell carcinomas. Cancer 2002;95:1482-94.

20. Patel V, Aldridge K, Ensley JF et al. Laminin-gamma2 overexpression in head-and-neck squamous cell carcinoma. Int J Cancer 2002;99:583-8.

21. Lindberg P, Larsson A, Nielsen BS. Expression of plasminogen activator inhibitor-1, urokinase receptor and laminin gamma-2 chain is an early coordinated event in incipient oral squamous cell carcinoma. Int J Cancer 2006;118:2948-56.

22. Ziober AF, Kirtesh RP, Faizan A et al. Identification of a Gene Signature for Rapid Screening of Oral Squamous Cell Carcinoma. Clin Cancer Res 2006;12:5960-71.

23. Gonzalez HE, Gujrati M, Frederick M et al. Identification of 9 genes differentially expressed in head and neck squamous cell carcinoma. Arch Otolaryngol Head Neck Surg 2003;129:754-9.

24. McCarthy JB, Basara ML, Palm SL, Sas DF, Furcht LT. The role of cell adhesion proteins--laminin and fibronectin--in the movement of malignant and metastatic cells. Cancer Metastasis Rev 1985;4:125-52.

25. Liotta LA, Wewer U, Rao NC et al. Biochemical mechanisms of tumor invasion and metastases. Prog Clin Biol Res 1988;256:3-16.

26. Hintermann E, Quaranta V. Epithelial cell motility on laminin-5: regulation by matrix assembly, proteolysis, integrins and erbB receptors. Matrix Biol 2004;*23:*75-85.

27. Yamamoto H, Itoh F, Iku S, Hosokawa M, Imai K. Expression of the gamma(2) chain of laminin-5 at the invasive front is associated with recurrence and poor prognosis in human esophageal squamous cell carcinoma. Clin Cancer Res 2001;7:896-900.

28. Kagesato Y, Mizushima H, Koshikawa N et al. Sole expression of laminin gamma 2 chain in invading tumor cells and its association with stromal fibrosis in lung adenocarcinomas, Jpn J Cancer Res 2001;92:184-192.

29. Barsky SH, Rao CN, Hyams D, Liotta LA. Characterization of a laminin receptor from human breast carcinoma tissue. Breast Cancer Res Treat 1984;4:181-8.

30. Haslam SZ, Woodward TL. Host microenvironment in breast cancer development: epithelial-cell-stromal-cell interactions and steroid hormone action in normal and cancerous mammary gland. Breast Cancer Res 2003;5:208-15.

31. Kaklamani VG, Gradishar WJ. Gene expression in breast cancer. Current Treat Options Oncol 2006;7:123-8.

32. Aishima S, Matsuura S, Terashi T et al. Aberrant expression of laminin gamma 2 chain and its prognostic significance in intrahepatic cholangiocarcinoma according to growth morphology. Mod Pathol 2004;17:938-45.

33. Soini Y, Maatta M, Salo S, Tryggvason K, Autio-Harmainen H. Expression of the laminin gamma 2 chain in pancreatic adenocarcinoma. J Pathol 1996;180:290-4.

34. Olsen J, Kirkeby LT, Brorsson MM et al. Converging signals synergistically activate the LAMC2 promoter and lead to accumulation of the laminin gamma 2 chain in human colon carcinoma cells. Biochem J 2003;371:1-21.

35. Gontero P, Banisadr S, Frea B, Brausi M. Metastasis markers in bladder cancer: a review of the literature and clinical considerations. Eur Urol, 2004;46:296-311.

36. Malinoff HL, McCoy JP Jr, Varani J, Wicha MS. Metastatic potential of murine fibrosarcoma cells is influenced by cell surface laminin. Int J Cancer 1984;33:651-655.

37. Barsky SH, Rao CN, Williams JE, Liotta LA. Laminin molecular domains which alter metastasis in a murine model. J Clin Invest 1984;74:843-8.

38. Kuhn K. Basement membrane (type IV) collagen. Matrix Biol 1995;14:439-445.

39. Van Waes C, Carey TE. Overexpression of the A9 antigen/alpha 6 beta 4 integrin in head and neck cancer. Otolaryngol Clin North Am 1992;25:1117-39.

40. Bosman FT, Havenith M, Cleutjens JP. Basement membranes in cancer. Ultrastruct Pathol 1985;8:291-304.

41. Schmidt C, Pollner R, Poschl E, Kuhn K. Expression of human collagen type IV genes is regulated by transcriptional and post-transcriptional mechanisms. FEBS Lett 1992;312: 174-8.

42. Schmidt C, Fischer G, Kadner H, Genersch E, Kuhn K, Poschl,E. Differential effects of DNA-binding proteins on bidirectional transcription from the common promoter region of human collagen type IV genes COL4A1 and COL4A2. Biochim Biophys Acta 1993;1174: 1-10.

43. Nachat R, Mechin MC, Takahara H et al. Peptidylarginine deiminase isoforms 1-3 are expressed in the epidermis and involved in the deimination of K1 and filaggrin. J Invest Dermatol 2005;124:384-93.

44. O'Shea JJ, Gadina M, Schreiber RD. Cytokine signaling in 2002: new surprises in the Jak/Stat pathway. Cell 2002;109:S121-31.

45. Rawlings JS, Rosler KM, Harrison DA. The JAK/STAT signaling pathway. J Cell Sci 2004;117:8-3.

46. Kalyankrishna S, Grandis JR. Epidermal growth factor receptor biology in head and neck cancer. J Clin Oncol 2006;24:2666-72.

47. Choong NW, Cohen EE. Epidermal growth factor receptor directed therapy in head and neck cancer. Crit Rev Oncol-Hematol 2006;57:25-43.

48. Hebenstreit D, Horejs-Hoeck J, Duschl A. JAK/STAT-dependent gene regulation by cytokines. Drug News Persp 2005;18:243-249.

49. Kondoh N, Ohkura S, Arai M et al. Gene expression signatures that can discriminate oral leukoplakia subtypes and squamous cell carcinoma. Oral Oncol 2007;43:455-462.

Table 1. One hundred and thirty one differentially expressed genes between OSCC and controls in training set (See attachment “Table 1”)

Table 1. One hundred and thirty one differentially expressed genes between OSCC and controls in training set

Up-regulation in OSCC						Down-regulation in OSCC		
Probe Set	Gene	Z Score	Probe Set	Gene	Z Score	Probe Set	Gene	Z Score
202311_s_at	COL1A1	22.3	212365_at	MYO1B	12.0	1553212_at	KRT78	-18.6
202404_s_at	COL1A2	20.6	212012_at	PXDN	11.9	220149_at	FLJ22671	-18.1
202310_s_at	COL1A1	20.3	229860_x_at	LOC401115	11.8	241233_x_at	C21orf81	-16.4
211980_at	COL4A1	19.5	219211_at	USP18	11.8	1569608_x_at	LOC643187	-15.2
202267_at	LAMC2	18.8	219863_at	HERC5	11.8	220962_s_at	PADI1	-14.5
204415_at	IFI6	18.6	204619_s_at	CSPG2	11.7	210868_s_at	ELOVL6	-14.0
225681_at	CTHRC1	18.0	203968_s_at	CDC6	11.5	205319_at	PSCA	-13.7
212488_at	COL5A1	17.0	208156_x_at	EPPK1	11.5	204754_at	HLF	-13.7
211924_s_at	PLAUR	16.8	210797_s_at	OASL	11.4	218779_x_at	EPS8L1	-13.5
203256_at	CDH3	16.3	1568765_at	SERPINE1	11.4	221665_s_at	EPS8L1	-13.2
221729_at	COL5A2	16.2	204972_at	OAS2	11.2	220016_at	AHNAK	-12.8
213869_x_at	THY1	15.8	223541_at	HAS3	11.2	218885_s_at	GALNT12	-12.7
217312_s_at	COL7A1	15.7	218888_s_at	NETO2	11.1	231118_at	ANKRD35	-12.7
1555778_a_at	POSTN	15.6	209949_at	NCF2	11.1	225548_at	SHROOM3	-12.1
212489_at	COL5A1	15.3	204779_s_at	HOXB7	11.0	206094_x_at	UGT1A6	-11.9
221730_at	COL5A2	15.2	41037_at	TEAD4	11.0	206093_x_at	TNXB	-11.9
212354_at	SULF1	15.1	209800_at	KRT16	10.9	218935_at	EHD3	-11.9
207517_at	LAMC2	15.0	217519_at	MACF1	10.8	207126_x_at	UGT1A4	-11.8
212344_at	SULF1	15.0	202238_s_at	NNMT	10.7	230740_at	---	-11.7
204715_at	PANX1	14.7	221898_at	PDPN	10.7	204532_x_at	UGT1A4	-11.6
208851_s_at	THY1	14.2	201108_s_at	THBS1	10.7	242417_at	LOC283278	-11.5
222693_at	FNDC3B	13.9	209969_s_at	STAT1	10.4	213421_x_at	PRSS3	-10.6
204647_at	HOMER3	13.9	203921_at	CHST2	10.2	205200_at	CLEC3B	-10.5
213668_s_at	SOX4	13.7	204103_at	CCL4	10.2	1552283_s_at	ZDHHC11	-10.4
205574_x_at	BMP1	13.3	241872_at	SGIP1	10.1	220037_s_at	XLKD1	-10.1
1555420_a_at	KLF7	13.3	207850_at	CXCL3	10.0	1553861_at	TCP11L2	-10.0
217430_x_at	COL1A1	13.3	204747_at	IFIT3	9.7	204378_at	BCAS1	-9.7
210809_s_at	POSTN	13.2	219725_at	TREM2	9.6	242009_at	---	-9.7
205157_s_at	KRT17	12.9	203915_at	CXCL9	9.6	207206_s_at	ALOX12	-9.6
203695_s_at	DFNA5	12.9	204879_at	PDPN	9.6	205730_s_at	ABLIM3	-9.4
203325_s_at	COL5A1	12.9	1554008_at	OSMR	9.3	238715_at	ARHGAP27	-9.1
209900_s_at	SLC16A1	12.8	204051_s_at	SFRP4	9.2	205428_s_at	CALB2	-8.6
203085_s_at	TGFB1	12.7	227697_at	SOCS3	9.1	1565661_x_at	FUT6	-8.4
229225_at	NRP2	12.7	210001_s_at	SOCS1	9.0	208609_s_at	TNXA /TNXB	-8.3
225288_at	---	12.5	235276_at	---	8.5	227782_at	ZBTB7C	-8.3
202235_at	SLC16A1	12.5	222344_at	C5orf13	8.4	226303_at	PGM5	-8.1
204114_at	NID2	12.5	225520_at	MTHFD1L	8.4	240000_at	---	-8.0
229554_at	---	12.4	218404_at	SNX10	8.3	201497_x_at	MYH11	-7.8
214453_s_at	IFI44	12.4	229055_at	GPR68	8.1	227419_x_at	PLAC9	-7.3
212472_at	MICAL2	12.3	209774_x_at	CXCL2	8.0	230104_s_at	TPPP	-7.2
205483_s_at	ISG15	12.2	39402_at	IL1B	7.2	212224_at	ALDH1A1	-7.1
226997_at	---	12.2	219300_s_at	CNTNAP2	6.9	243718_at	---	-6.7
212473_s_at	MICAL2	12.0	229947_at	PI15	6.7	209975_at	CYP2E1	-6.5
225292_at	COL27A1	12.0	238581_at	GBP5	6.3			

Table 2. Validation of predictive models using internal and external (GSE6791) testing datasets

Model with Gene Name & Affymetrix Probe Set ID	Model from Logistic Regression	Area Under the Curve	
		Own Testing	GSE6791 Testing
Model 1 LAMC2, 207517_at COL4A1, 211980_at	7.8739*LAMC2+7.6269*COL4A1	0.99792	0.97619
Model 2 COL1A1, 202310_s_at PADI1, 220962_s_at	2.4377*COL1A1-2.8841*PADI1	0.99167	0.97789
Model 3 C21orf81, 241233_x_	-2.1042* C21orf81	0.98540	0.97450
Model 4 KRT17, 205157_s_at PRSS3, 213421_x_at	2.5638* KRT17-2.4506* PRSS3	0.97710	0.97450
Model 5 COL1A2, 202404 EST, 230740_at	1.9345* COL1A2-1.5931*230740_at	0.98960	0.95920
Model 6 COL1A1, 202311_s_at XLKD1, 220037_s_at	2.2372*COL1A1-1.3377* XLKD1	0.99170	0.95070
Model 7 THY1, 208851_s_at FLJ522671, 220149_at HAS3, 223541_at	2.4643* THY1-1.6340* FLJ522671 +1.5310* HAS3	0.99790	0.96260
Model 8 POSTN, 1555778_a_at TIA2, 221898_at	1.4909* POSTN+1.8340* TIA2	0.98960	0.90820
Model 9 MGC40368, 1553861_at GIP3, 204415_at COL27A1, 225288_at	-2.2659* MGC40368+1.0718* GIP3+1.7854* COL27A1	0.97290	0.95410
Model 10 CDH3, 203256_at ELOVL6, 210868_s_at	1.9861*CDH3-2.1743*ELOVL6	0.99380	0.89800

Table 3. qRT-PCR results comparing RNA transcripts for four genes between OSCC cases and controls

	Mean (S.D.) Ct*	95% CI**	p
LAMC2			
Case	2.83 (1.02)	2.44-3.21	≤0.0001
Control	7.38 (0.54)	7.18-7.59	
COL4A1			
Case	5.13 (0.86)	4.81-5.45	≤0.0001
Control	8.58 (0.78)	8.29-8.87	
COL1A1			
Case	2.28 (1.14)	1.85-2.71	≤0.0001
Control	6.94 (0.64)	6.70-7.18	
PADI1			
Case	10.86 (2.34)	9.99-11.73	≤0.0001
Control	5.06 (0.94)	4.71-5.41	

*Ct (threshold cycle) values are inversely associated with the amount of RNA transcripts in the sample. Based on analyses of 30 OSCC cases and 30 controls.

**CI: confidence interval

Table 4. Differentially expressed genes between OSCC and dysplasia plus normal controls that overlap with the 131 genes

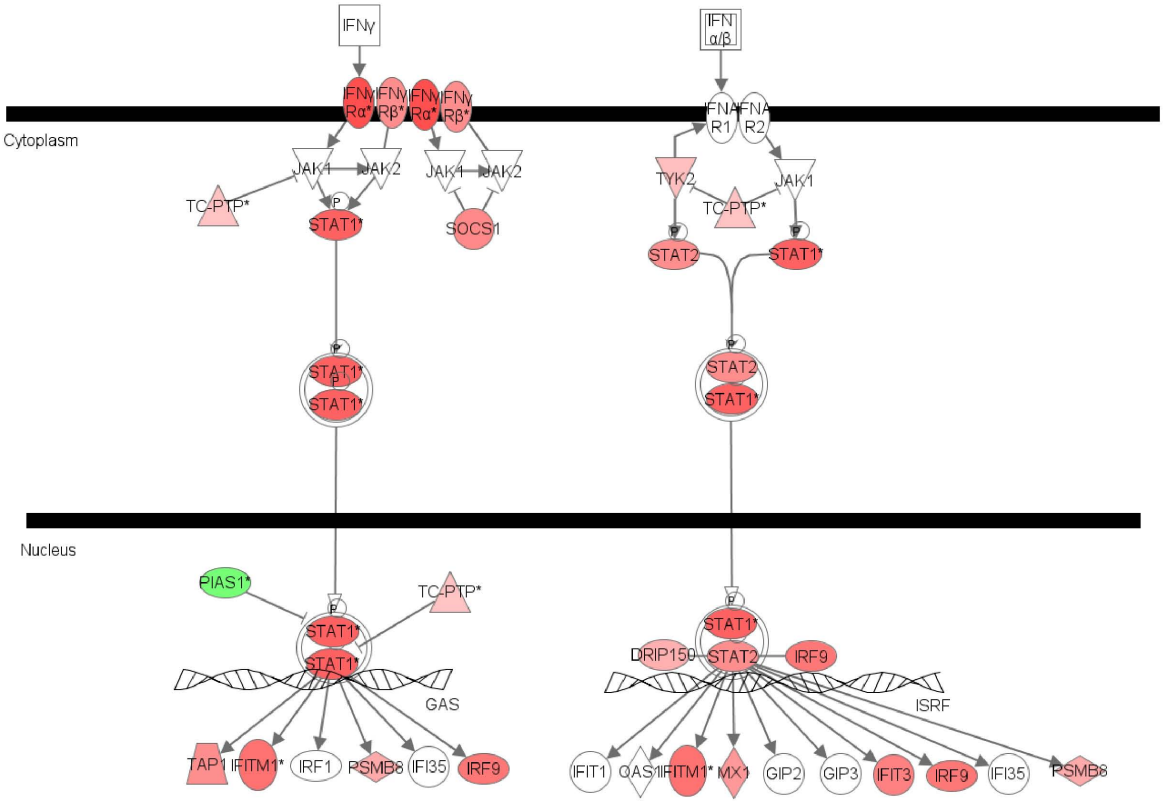
Up-regulation in OSCC			Down-regulation in OSCC		
Probe Set	Gene	Z Score	Probe Set	Gene	Z Score
202267_at	LAMC2	17.6	205200_at	TNA	-14.7
207517_at	LAMC2	17.5	231118_at	FLJ25124	-14.0
1568765_at	SERPINE1	14.9	206093_x_at	TNXB	-13.1
1555420_a_at	KLF7	13.8	227782_at	ZBTB7C	-11.3
222693_at	FAD104	13.6	1552283_s_at	ZDHHC11	-11.2
207850_at	CXCL3	12.8	213421_x_at	PRSS3	-10.8
210001_s_at	SOCS1	11.9	238715_at	ARHGAP27	-10.6
229225_at	NRP2	11.8	208609_s_at	TNXB	-10.4
227697_at	SOCS3	11.6	242009_at		-10.3
209949_at	NCF2	11.5	226303_at	PGM5	-9.9
204103_at	CCL4	11.5	207206_s_at	ALOX12	-9.7
218404_at	SNX10	11.4	230104_s_at	TPPP	-9.6
203695_s_at	DFNA5	11.4	220037_s_at	XLKD1	-8.6
212354_at	SULF1	11.2	243718_at		-8.6
229860_x_at	LOC4115	10.9	227419_x_at	PLAC9	-8.5
209774_x_at	CXCL2	10.8	201497_x_at	MYH11	-8.0
241872_at	SGIP1	10.7	204532_x_at	UGT1A10	-7.8
203968_s_at	CDC6	10.6	206094_x_at	UGT1A1,1A4,1A6	-7.6
225520_at	FTHFSDC1	10.1	207126_x_at	UGT1A10	-7.5
229947_at	PI15	9.3	212224_at	ALDH1A1	-6.4
204747_at	IFIT3	8.9			
39402_at	IL1B	8.9			
235276_at	EPSTI1	8.8			
203915_at	CXCL9	8.7			
204779_s_at	HOXB7	8.6			
219211_at	USP18	8.5			
238581_at	GBP5	7.6			
219300_s_at	CNTNAP2	5.0			
204051_s_at	SFRP4	4.2			

Figure 1 (See attachment “Figure 1”)

Legend for Figure 1:

Figure 1. Most prominently involved biological pathways in OSCC. Top: JAK/STAT pathway. Bottom: IFN-γ signaling pathway, antigen-presenting pathway. Red denotes up-regulation and green denotes down-regulation of the gene. Ingenuity®Systems, version 4.0

Interferon Signaling



JAK/STAT Signaling

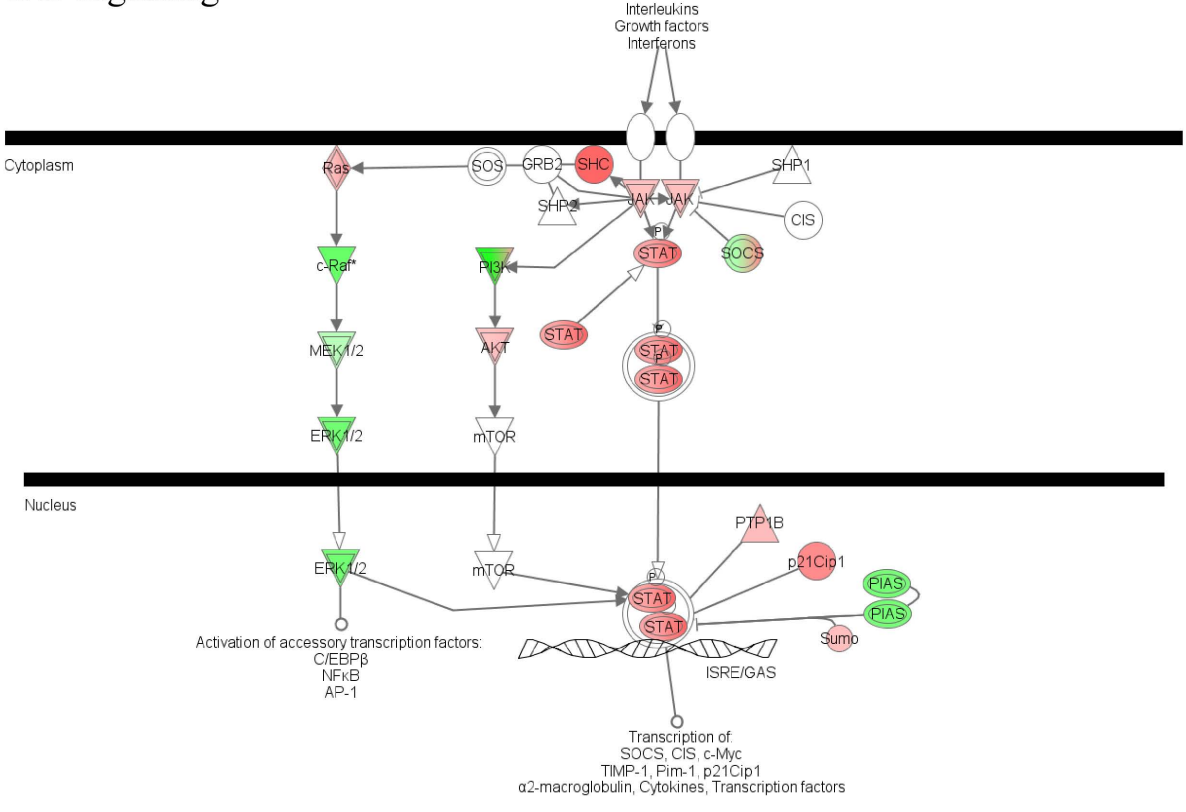


Figure 2

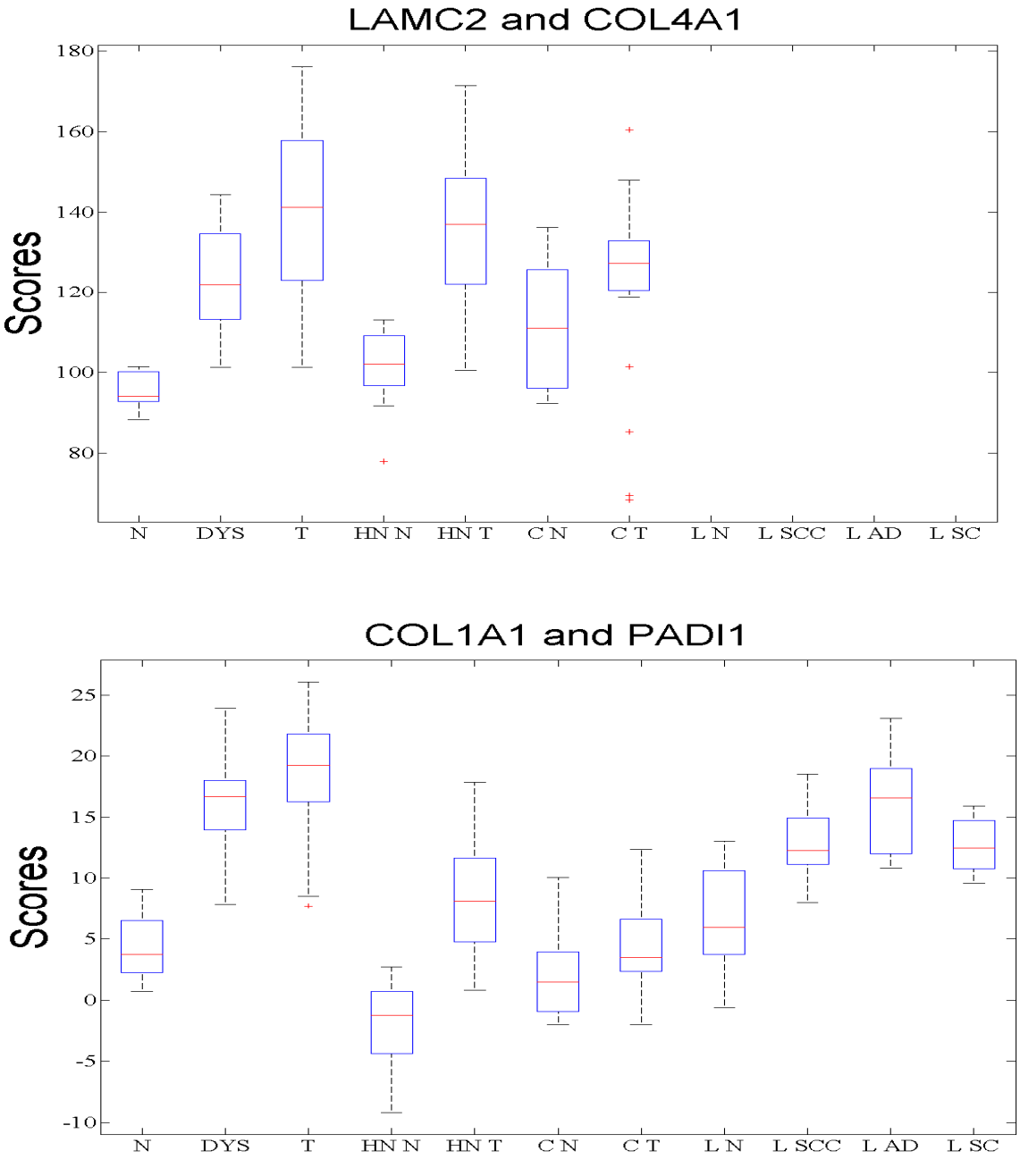


Figure 2. Tissue specificity of model *LAMC2* and *COL4A1* (top) and model *COL1A1* and *PADI1* (bottom). Box Whisker plots of logistic regression scores (y axis) for normal controls and cases in our own testing set (N: normal, DYS: dysplasia, T: OSCC), GEO GSE6791 head and neck normal controls (HN N) and cases (HN T), GEOGSE 6791 cervical normal controls (C N)and cases (C T), and GEO GSE6044 lung normal controls (L N), lung squamous cell carcinoma (L SCC), lung adenocarcinoma (L AD) and lung small cell cancer (L SC).