Research Article

Cancer Prevention Research

Gene Expression Profiling Predicts the Development of Oral Cancer

Pierre Saintigny[1], Li Zhang[2], You-Hong Fan[1], Adel K. El-Naggar[3], Vassiliki A. Papadimitrakopoulou[1], Lei Feng[4], J. Jack Lee[4], Edward S. Kim[1], Waun Ki Hong[1], and Li Mao[1,5]

Abstract

Patients with oral premalignant lesion (OPL) have a high risk of developing oral cancer. Although certain risk factors, such as smoking status and histology, are known, our ability to predict oral cancer risk remains poor. The study objective was to determine the value of gene expression profiling in predicting oral cancer development. Gene expression profile was measured in 86 of 162 OPL patients who were enrolled in a clinical chemoprevention trial that used the incidence of oral cancer development as a prespecified endpoint. The median follow-up time was 6.08 years and 35 of the 86 patients developed oral cancer over the course. Gene expression profiles were associated with oral cancer–free survival and used to develop multivariate predictive models for oral cancer prediction. We developed a 29-transcript predictive model which showed marked improvement in terms of prediction accuracy (with 8% predicting error rate) over the models using previously known clinicopathologic risk factors. On the basis of the gene expression profile data, we also identified 2,182 transcripts significantly associated with oral cancer risk–associated genes (P value < 0.01; univariate Cox proportional hazards model). Functional pathway analysis revealed proteasome machinery, *MYC*, and ribosomal components as the top gene sets associated with oral cancer risk. In multiple independent data sets, the expression profiles of the genes can differentiate head and neck cancer from normal mucosa. Our results show that gene expression profiles may improve the prediction of oral cancer risk in OPL patients and the significant genes identified may serve as potential targets for oral cancer chemoprevention. *Cancer Prev Res; 4(2); 218–29. ©2011 AACR.*

Introduction

Head and neck squamous cell carcinoma (HNSCC) is second only to lung cancer as the most common smoking-related cancer worldwide. Oral squamous cell carcinoma (OSCC) is the most common anatomic site of HNSCC counting for approximately 50% of all HNSCC. Despite the tremendous effort to reduce tobacco use, HNSCC remains one of the leading causes of smoking-attributable mortality in the world—about 438,000 each year (1). Even for HNSCC diagnosed at early stages, surgery (current standard care) is a debilitating, substantially morbid procedure that severely impairs quality of life for many patients. In light of its continuing burden and evasion of substantial control, HNSCC requires new approaches including diagnosis of the disease before the cancerous stage and preventing development of invasive cancers.

Leukoplakia and erythroplakia are the most commonly diagnosed oral premalignant lesions (OPL) with a 17% to 24% rate of malignant transformation over a period of up to 30 years (2, 3). OPLs are associated with hyperkeratosis, dysplasia, or *in situ* carcinoma, but the value of OPL histology as a marker of the risk of OSCC is poor. Recently, we reported that *LOH* profile (4), polysomy (5), p53 (5), overexpression of podoplanin (6), p63 (7), and EGFR, as well as increased *EGFR* gene copy number (8) are associated with increased risk of OSCC.

To systematically study the genes associated with risk of OSCC, we used gene expression profiling on a large cohort of samples of OPL patients. Gene expression profiles or signatures are groups of genes that are differentially expressed among tumors or diseased lesions, reflecting differences in biological features of the tissues. Gene expression profiles have been used to develop prognostic models of cancer outcome and to identify markers for diagnosis and classification of cancers (9–11). However, to assess the value of expression profiles in predicting cancer risk, samples must be collected before cancer diagnosis in a

Authors' Affiliations: Departments of [1]Thoracic/Head and Neck Medical Oncology, [2]Bioinformatics and Computational Biology, [3]Pathology, and [4]Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston, Texas; and [5]University of Maryland Dental School, Baltimore, Maryland

Note: Supplementary data for this article are available at Cancer Prevention Research Online (http://cancerprevres.aacrjournals.org/).

P. Saintigny and L. Zhang contributed equally to this study and report.

Corresponding Author: Li Mao, Department of Oncology and Diagnostic Sciences, University of Maryland Dental School, 650 W Baltimore Street, Baltimore, MD 21201. Phone: 410-706-4339; Fax: 410-706-6115. E-mail: lmao@umaryland.edu

doi: 10.1158/1940-6207.CAPR-10-0155

prospective setting, which takes years with high cost and is therefore difficult to do in practice. We took advantage of a collection of 162 OPL samples that were obtained before cancer development in a chemoprevention clinical trial, which included long-term oral cancer incidence as a prespecified secondary endpoint. During the follow-up time of the trial (median: 7.5 years), 39 of the patients developed cancer. We hypothesized that gene expression profiles in OPLs marked the risk of OSCC development. We measured the gene expression profiles of a subset of the patient samples and searched for their association with oral cancer-free survival (OCFS) time. In this report, we demonstrate that gene expression profile can significantly improve the prediction of OSCC development over clinical and histologic variables in OPL patients and the significant genes may be promising targets for cancer prevention.

Methods

Patients and specimens

From 1992 to 2001, 162 randomized and eligible patients were enrolled in a randomized chemoprevention trial at The University of Texas MD Anderson Cancer Center (MDACC). The patients had been diagnosed with OPL and randomly assigned to intervention with 13-*cis*-retinoic acid (13cRA) versus retinyl palmitate (RP) with or without β-carotene (BC). A total of 153 frozen samples were available at baseline or 3 months after enrollment but before any event (defined as the diagnosis of OSCC). Among 39 samples from patients who developed OSCC, 4 were excluded because of poor RNA quality. Among 114 samples from patients who did not develop OSCC, 8 were excluded because of poor RNA quality. Finally, all the samples from patients who developed oral cancer ($N = 35$) were selected for gene expression profiling, as well as 51 samples (*ad hoc* choice) from patients who did not develop OSCC, randomly selected among 106 patients. The events were oversampled relative to the nonevents to optimize statistical power for finding the significant transcripts. Because the events are rare, we included all of them, as permitted by the quality of the samples. The median follow-up of the 51 patients who did not develop oral cancer was 6.08 years. Clinicopathologic parameters were obtained from the clinical trial database. The follow-up data were obtained from a combination of chart review and a telephone interview. More detailed clinical information has been previously described in Papadimitrakopoulou and colleagues (12). The study was approved by the institutional review board, and written informed consent was obtained from all patients.

Sample preparation, amplification, labeling, and microarray hybridization

All steps leading to generation of raw microarray data were processed at the University of Texas MDACC Genomics Core Facility. Human Gene 1.ST platform was used to generate gene expression profiling. Gene expression profiling was obtained from the whole biopsy, including both the epithelial cells and the underlying stroma. A detailed method is provided in Supplementary Material 1.

Statistical methods

Data analysis was performed using the Bioconductor packages in the R language (http://www.bioconductor.org; ref. 13). Raw data of microarrays were processed using quantile normalization and RMA algorithm (14). Univariate Cox proportional hazards model (Coxph) was used to identify transcripts associated with the development of oral cancer. To address the multiple testing problems, false discovery rates (FDR) of genes were calculated according to the BUM model (15).

The multivariate analysis was performed using CoxBoost (16), a model for identifying prognostic markers from microarray data. The algorithm is based on boosting, which constructs a prognostic model by maximizing the partial log likelihood function (logplik) that imposes a penalty for each non-zero coefficients utilized in the model. There are two main parameters that are relevant: penalty score and number of boosting steps. Both the penalty score and boosting steps can be optimized using the functions provided in the CoxBoost package under a cross-validation scheme. We tested the performance of the CoxBoost model using computer-simulated microarray data and survival data. The computer program used in the analysis is available in Supplementary Material 2. We built the predictive models with and without clinical covariates, which includes age, histology at baseline (hyperplasia vs. dysplasia), and podoplanin and deltaNp63 expression. To evaluate the performance of the models, we used the 0.632+ bootstrap method (17) and prediction error curve estimates (16). The latter provides an estimation of the type I (false positive) and the type II (false negative) error rate, or misclassification rate across time. Missing values for deltaNp63 ($N = 5$) and podoplanin ($N = 5$) were imputed using the nearest neighbor hot-deck imputation method (function rrp.impute in rrp-package).

As an alternative to CoxBoost, we also used diagonal linear discriminant analysis (DLDA) model method (18), which is a frequently used method for class discrimination in microarray studies. The patients of our data set were dichotomized into short cancer-free and long cancer-free patients based on a follow-up cutoff point of 5 years. Seventeen patients with 5 or less than 5 years of follow-up time had to be omitted in the analysis. We used a standard 10-fold cross-validation scheme to assess the performance of the prediction models. Specifically, 9 of 10 of the samples ($N = 62$) were used (i) to identify the most significant 50 transcripts that are associated with OCFS time and (ii) to build a DLDA model using the 50 transcripts as the predictors. Then, the model is tested in the remaining 1 of 10 of the samples ($N = 7$) to test the accuracy of the DLDA model. The process was repeated 100 times and the results are aggregated to compute the misclassification rate, the sensitivity and specificity, the positive and negative predictive values. The choice of 50 genes as the number of genes to use for prediction was *ad*

hoc. The prediction accuracy was not sensitive to the number of significant transcripts we chose.

The oral cancer index was computed as the average level of expression of the transcripts associated with a hazard ratio greater than 1 minus the average level of expression of the transcripts that have a hazard ratio less than 1. Our hypothesis was that oral cancer index would be able to discriminate HNSCC from normal mucosa.

Functional analyses were performed using Gene Set Enrichment Analysis (GSEA) software v2.0.4 (19). Functional analyses were performed using GSEA software v2.0.4. GSEA is a robust computational method that determines whether an *a priori* defined set of genes shows statistically significant, concordant differences between 2 biological states (e.g., high risk vs. low risk). GSEA aims to interpret large-scale expression data by identifying pathways and process. The main advantage of this method is its flexibility in creating molecular signature database of gene sets, including ones based on biological pathways, chromosomal location, or expression profiles in previously generated microarray data sets. The input data for GSEA procedure were the following: (i) a complete table of genes ranked according to the $\log_2$ transformed Cox model hazards ratio associated with the development of OSCC, (ii) a mapping file for identifying transcripts in HG-1.ST platform, and (iii) a catalogue of functional gene sets from Molecular Signature Database (MSigDB, version 2 January 2007 release, www.broad.mit.edu/gsea/msigdb/msigdb_index.html). A total of 1,436 curated gene sets (canonical pathway, chemical and genetic perturbations, BioCarta, GenMAPP, and KEGG gene sets) were included in the analysis. Default parameters were used. Inclusion gene set size was set between 15 and 500 and the phenotype was permutated 1,000 times. Gene sets that met the FDR $\leq$ 0.25 criterion were considered (Supplementary Material 3).

External data sets

Nine independent data sets were used to validate our findings and downloaded from Gene Expression Omnibus (GEO), GSE9844 (20), GSE6791 (21), GSE3524 (22), GSE6631 (23), GSE13601 (24), GSE2379 (25), and GSE686 (9), which compared HNSCC and normal mucosa; GSE10774, which studied normal keratinocyte and various HNSCC cell lines (26); and GSE4115, which studied normal bronchial cells in smokers with suspected lung cancer (27).

Data availability

The microarray data analyzed in this study have been deposited in the NIH Gene Expression Omnibus database at www.ncbi.nlm.nih.gov/geo under the accession number GSE26549 (Raw data is also available in Supplementary Material 5). Complete annotation of the samples is provided (Supplementary Material 6).

Results

We performed microarray gene expression on 86 leukoplakia sample of OPL patients. Table 1 shows the clinical and pathologic characteristics of the patients, along with podoplanin and deltaNp63 protein expression measured by immunohistochemistry. These 86 patients were selected from the 162 patients who were involved in the chemoprevention trial. This subset contained a higher cancer incidence because we attempted to include all patients who eventually developed cancer but only a fraction of the cancer-free patients (see more details in the Methods section). The enriched cancer incidence in the subset was designed to increase the statistical power of our analysis for the given sample size.

Developing multivariate predictive models

We used the CoxBoost algorithm to develop multivariate predictive models of OCFS time for OPL patients (16). CoxBoost was designed to develop prognostic models from microarray data using a boosting approach. The algorithm assumes that most transcripts are not associated with OCFS time, hence having zero coefficients in the prediction model. For each transcript with a non-zero coefficient in the prediction model, an explicit penalty score was added to the fitness function, which is a logplik function. In this way, the overparameterization problem was controlled, which is typically encountered in search of biomarkers from microarray data.

The results of using CoxBoost were summarized in Table 2 and Figure 1. Three models were considered. In model 1, only microarray expression data were used as predictors. In model 2, age, histology, and deltaNp63 and podoplanin expression were used as mandatory covariates along with the microarray data. In model 3, only age, histology, and deltaNp63 and podoplanin expression were used as predictors. Table 2 showed the genes found in Model 1. The CoxBoost procedure was repeated 100 times, each time yielding a different set of predictive marker transcripts. The column "Frequency" in Table 2 showed the frequency of occurrences of the transcripts. Among the 29 transcripts, 21 have frequencies greater than 80%, showing the lists of marker transcripts are mostly consistent between different runs of the algorithm. Similarly, Table 2 shows the transcripts found in Model 2; 15 genes of the 23 in Table 2 have frequencies greater than 80%. There are 9 transcripts shared in Table 2.

The HR and Wald test *P* values obtained from using univariate Coxph model for each of the 29 and 23 transcripts in Models 1 and 2 are also shown in Table 2. The vast majority of the transcripts selected by the CoxBoost approach were highly significant. Furthermore, the CoxBoost coefficient was always consistent with the HR provided by the Coxph model (positive and negative CoxBoost coefficients being associated with a HR > 1 and <1, respectively).

Figure 1 showed the prediction error curves of the prediction models, which were used to evaluate the performance of the models. The prediction error was computed as squared difference between predicted survival probability at time *t* and the true state (0 for being still under risk, and 1 if an event of cancer occurred). Lower prediction

Table 1. Characteristics of the 86 patients included in the gene expression study, and the whole population of the trial that included 162 patients

Variable	Whole population of the trial (%)	Patients included in the present study (%)	Patients not included in the present study (%)	*P* value
All patients	162 (100)	86 (100)	76 (100)	<0.0001
No oral cancer	123 (76)	51 (59)	72 (95)	
Oral cancer	39 (24)	35 (41)	4 (5)	
Follow-up time of the censored observations				0.48
Median	7.47	7.11	7.71	
Range, y	0.19–15.31	0.92–15.31	0.19–12.73	
Sex				0.97
Female	77 (48)	41 (48)	36 (47)	
Male	85 (52)	45 (52)	40 (53)	
Race				0.6
White	145 (89)	78 (91)	67 (88)	
Other	17 (11)	8 (9)	9 (12)	
Alcohol history				0.49
Current	93 (57)	49 (57)	44 (58)	
Former	19 (12)	8 (9)	11 (14)	
Never	50 (31)	29 (34)	21 (28)	
Smoking history				0.009
Current	56 (35)	22 (25)	34 (45)	
Former	65 (40)	35 (41)	30 (39)	
Never	41 (25)	29 (34)	12 (16)	
Age				0.57
Median	56	57.5	55	
Range	23–90	23–90	27–81	
Treatment arm				0.43
BC-RP	45 (28)	21 (24)	24 (31)	
13cRA	81 (50)	47 (55)	34 (45)	
RP only	36 (22)	18 (21)	18 (24)	
Histology at baseline dysplasia				0.19
Dysplasia	53 (33)	32 (37)	21 (28)	
Hyperplasia	109 (67)	54 (63)	55 (72)	
DeltaNp63				0.43
Low	109 (73)	57 (70)	54 (76)	
High	40 (27)	24 (30)	17 (24)	
Podoplanin				0.0009
Low	94 (63)	41 (51)	53 (77)	
High	56 (37)	40 (49)	16 (23)	

Abbreviation: NA, not available.

errors suggest better performance. Following Binder and colleagues (16), we computed the prediction error curve using boostrap samples and aggregated into 0.632+ estimates. The 0.632+ method was invented by Efron and colleagues as an improvement over conventional cross-validation schemes for assessing model performance (17). The advantage of 0.632+ method is that it allows the use of all observations to train the prediction model, but nonetheless results in an accurate assessment of prediction error.

The prediction error curves in Figure 1 demonstrated that the expression profiling data can markedly improve the prediction accuracy over model 3 that used only the previously known factors (i.e., age, histology, and deltaNp63 and podoplanin expression). Models 1 and 2 have similar performance (model 2 is slightly better) with prediction error around 8% beyond 2 years of follow-up time. The prediction error of model 3 started to show higher values after year 1 in follow-up time and the difference increased over time. For comparison, the null model, which only

Table 2. Models generated by the CoxBoost approach

Probeset ID	*Gene symbol*[a]	CoxBoost coefficient	Frequency, %	Cox *P* value	Cox hazard ratio
Model 1					
8095441	*CSN1S2A*	0.22	100	1.80E-06	160.02
8023314	*CCDC11*	0.19	100	1.80E-06	3.20
7986442	ENST00000391004	0.13	100	3.40E-07	14.70
8062842	ENST00000387867	0.10	100	2.50E-06	101.75
8084002	*KCNMB2*	0.08	99	6.60E-05	5.76
7915846	*MKNK1*	0.06	100	2.50E-08	39.03
8165709	NC_001807	0.05	99	6.80E-05	9.71
8122200	ENST00000385892	0.05	79	2.10E-04	3.82
8046408	*PDK1*	0.04	93	3.00E-03	5.41
8153223	*PTK2*	0.04	90	1.30E-03	6.40
8172119	*MED14*	0.03	98	2.30E-06	26.81
8061092	NA	0.02	96	4.00E-04	8.35
7927106	ENST00000387096	0.02	85	8.50E-07	3.90
7948894	*RNU2-1*	0.02	59	8.50E-04	3.93
8083939	*AK128090*	−0.02	20	1.10E-01	0.19
7939865	*OR4B1*	−0.02	37	4.00E-03	0.08
7916777	hsa-mir-101-1	−0.02	51	1.00E-05	0.01
7964360	*STAT6*	−0.02	72	3.70E-05	0.01
8101762	*SNCA*	−0.02	75	2.20E-03	0.39
7962489	*PLEKHA9*	−0.03	99	9.80E-05	0.07
7901361	ENST00000387793	−0.04	86	5.50E-04	0.12
8044682	*SNRPA1*	−0.04	96	1.40E-02	0.18
8028950	*CYP2G1P*	−0.04	69	1.70E-02	0.06
7977480	ENST00000386651	−0.04	95	7.00E-05	0.04
8067983	ENST00000387011	−0.08	100	7.50E-05	0.01
8097743	ENST00000410285	−0.10	100	9.90E-06	0.07
8093957	*CNO*	−0.12	100	6.50E-07	0.02
8086536	ENST00000365398	−0.15	100	2.00E-05	0.02
8121943	ENST00000384255	−0.18	100	8.90E-05	0.03
Model 2					
8061746	*DNMT3B*	0.26	100	4.30E-06	7.73
8092638	ENST00000384774	0.21	100	2.40E-10	18.56
8165709	NC_001807	0.19	100	6.80E-05	9.71
7949019	ENST00000365219	0.14	100	4.40E-08	9.30
7978905	*SDCCAG1*	0.13	100	4.20E-09	79.79
7959891	ENST00000384123	0.09	91	3.20E-05	10.63
8023314	*CCDC11*	0.09	95	1.80E-06	3.20
7907769	*FAM163A*	0.07	98	1.50E-03	18.37
8095441	*CSN1S2A*	0.06	98	1.80E-06	160.02
8175119	ENST00000410882	0.05	40	1.00E-02	9.54
7918757	*DENND2C*	0.03	22	4.20E-05	6.62
8063839	*SS18L1*	0.03	23	8.50E-04	30.51
8078600	*TCEA1*	0.02	20	2.40E-04	11.92
8046408	*PDK1*	0.02	12	3.00E-03	5.41
7925939	*AKR1C4*	0.02	10	3.60E-04	4.35
8134599	*CPSF4*	−0.02	4	1.50E-03	0.06
7964183	*GLS2*	−0.02	54	3.50E-04	0.04
7901361	ENST00000387793	−0.08	98	5.50E-04	0.12
8121943	ENST00000384255	−0.12	99	8.90E-05	0.03
7916777	hsa-mir-101-1	−0.15	100	1.00E-05	0.01

(Continued on the following page)

Table 2. Models generated by the CoxBoost approach (Cont'd)

Probeset ID	*Gene symbol*[a]	CoxBoost coefficient	Frequency, %	Cox *P* value	Cox hazard ratio
8045339	ENST00000363848	−0.15	100	7.00E-07	0.10
8067983	ENST00000387011	−0.19	100	7.50E-05	0.01
8093957	*CNO*	−0.26	100	6.50E-07	0.02
	Age	0.012	100	7.60E-02	1.03
	Histology at baseline	0.078	100	2.00E-01	0.64
	Podoplanin	0.43	100	3.30E-03	3.10
	DeltaNp63	1.47	100	5.69E-05	4.31

NOTE: Model 1 includes microarray data only (29 transcripts), whereas model 2 includes microarray data (23 transcripts) as well as age, histology at baseline, and deltaNp63 and podoplanin expression at baseline (clinical and pathologic covariates were mandatory); *P* values and HR are from the univariate Cox model. The CoxBoost procedure was repeated 100 times, each time yielding a different set of predictive marker genes. The column "Frequency" showed the frequency of occurrences of the genes.

[a]When a gene symbol was not available, RefSeq, ENSEMBL, GenBank or miRBase Micro RNA Database identifier was provided; NA: information not provided by the manufacturer

used random numbers as predictors, was also shown in the Figure 1. As expected, null model has the worst performance. Models 1 and 2 showed comparable performance, suggesting that the previously known factors may be substituted by expression profiles as alternative predictors.

In Figure 2, we showed that predicted oral cancer risk according to model 2 is strongly associated with OCFS time. The median of the oral cancer risk was used to dichotomize the patients between low and high index groups. The Kaplan–Meier survival curves of the 2 groups and the accompanying curves based on 50 bootstrap samples display marked differences in survival time (log-rank *zP* value = 1.03E-14). The strong association can also be observed if model 1 is used instead of model 2 (details not shown). It should be noted, however, that because the survival data were used to identify and to optimize the model, Figure 2 merely represents a strong association based on the training data, and such strong association may be reduced in independent samples. It is more reasonable to use the prediction error curves in Figure 1 for model evaluation.

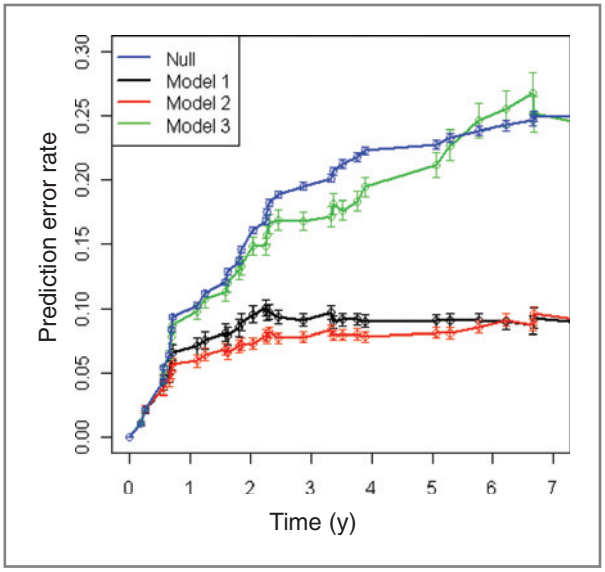

Figure 1. Model comparison using prediction error curves. Null model used random number data as predictors. Model 1 used only microarray data as predictors. Model 2 used microarray data as well as age, histology, and podoplanin and deltaNp63 expression as predictors. Model 3 used only age, histology, and podoplanin and deltaNp63 expression as predictors. The vertical lines shows the error bars obtained from 100 runs of the procedures.

Checking and testing the significant genes

Because overfitting and fragile inference is a well-recognized concern when microarray data are used to develop prediction models (28–30), we used various approaches to examine if our results are robust and reproducible. A common symptom of fragile inference may be that the results are highly dependent on the particular scheme or parameter choices taken in the modeling process. To check if this could have happened to our analysis, we used DLDA method as an alternative approach to CoxBoost for building the multivariate prediction models. We found the misclassification rate to be 16% (Supplementary Material 7). The sensitivity was 91% (95% CI: 88%–93%), specificity was 76% (95% CI: 72%–78%), positive predictive value was 80% (95% CI: 78%–82%), and negative predictive value was 89% (95% CI: 85%–92%). Because of the nature of censored data, these are estimates, with some assumptions, of the accuracy of the predictors. This result appears to be poorer than that from CoxBoost. We thought this was understandable because we

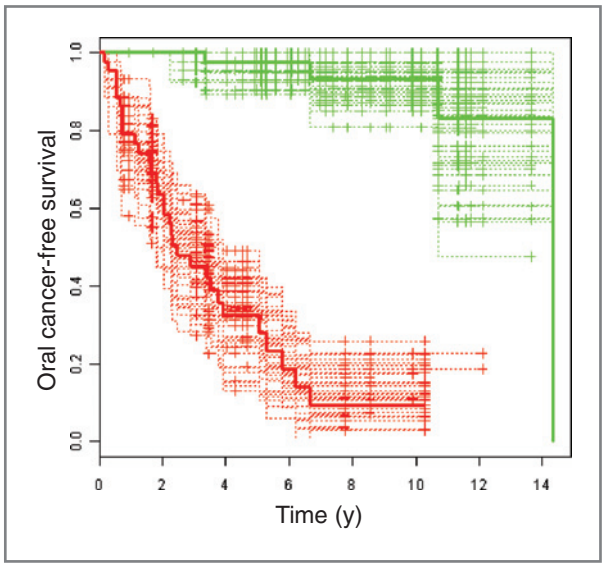

Figure 2. OCFS dichotomized by oral cancer index. The red solid curve showed the patients with above median cancer risk index (median = −0.42) whereas the green solid curve showed the patients with below median oral cancer index. The oral cancer risk indices were computed as the hazard values according to model 2 (age, histology at baseline, and podoplanin and deltaNp63 expression, and 23 probesets) using CoxBoost-optimized parameters. Accompanying curves are based on 50 bootstrap samples.

dichotomized the samples into short and long cancer-free patients with DLDA, which led to partial loss of information from the survival data. Nevertheless, the results from DLDA model is highly significant when compared with null model, which assigns short survivor and long cancer-free labels randomly (i.e., randomly permuting the labels). We estimated the P value to be $\leq 10^{-16}$ according to Fisher's exact test. Such results strongly suggest that the underlying gene expression profiles are predictive of oral cancer risk.

Functional pathway analysis of the genes associated with OCFS time

To explore pathways that are associated with OCFS time, it is desirable to obtain a comprehensive list of transcripts associated with OCFS time. We applied the univariate Coxph s model to identify the significant transcripts. A P value (Wald test) was computed for each of the transcripts.

We found 2,182 significant transcripts that have a value of $P < 0.01$ and the FDR was estimated to be 11%. Supplementary Material 8 shows a histogram of the P values, and Supplementary Material 9 provides the complete list of the transcripts. The sharp spike on the left indicates that there is a large group of transcripts having significant association with OCFS time. Had no significant association existed (i.e., the null hypothesis), the histogram were supposed to shape like a uniform distribution from 0 to 1. We used the BUM model (15) to estimate the FDR of the significant transcripts, which assumes that distribution of the P values of the nonsignificant transcripts follows a uniform distribution whereas the distribution of the P values of significant genes follows a β distribution. Among the 2,182 significant transcripts, 1,262 were associated with a high risk to develop oral cancer (HR > 1), and 920 were associated with low risk (HR < 1). All but 3 transcripts included in the CoxBoost model 1 and all but 1 transcript included in the model 2 were included in list of 2,182 transcripts, with very significant P values (Table 2). As well, HRs were always consistent with the coefficient sign provided by the CoxBoost models (Table 2).

To identify pathways associated with oral cancer development, we performed functional analyses using the GSEA agorithm (19) that seeked significant associations between the HRs, which we calculated with the predefined gene sets in GSEA database. The detailed results of these analyses are presented in Supplementary Material 3. Gene sets associated with proteasome machinery, and *MYC* upregulation as well as ribosomal components, the latest being mainly regulated by *MYC* (31), were associated with a high risk to develop oral cancer. Similarly, genes commonly upregulated in cancer relative to normal tissue, and genes upregulated in undifferentiated stem cells or cancer cells were associated with a high risk to develop oral cancer. The enrichment in the proteasome pathway is shown as an example in Supplementary Material 4.

Assessing the relationship between the significant genes and cancer

We found that the significant transcripts identified from our current study tend to be differentially expressed between normal and cancer cells in multiple data sets. We took the 2,182 significant transcripts, found them correspond to 1,270 gene symbols according to annotation provided by the manufacturer. We then extracted gene expression data that have matched gene symbols from multiple data sets composed of HNSCC, and normal mucosa samples, including one data set comparing HNSCC and normal keratinocytes cell lines. We computed the oral cancer indices as described in the method section. Figure 3 shows that oral cancer indices were consistently lower in HNSCC compared with normal mucosa across 7 independent data sets, and lower in HNSCC cell compared with normal keratinocytes in one cell line. All these differences were highly significant, with few overlap between cancer and normal samples in most of the studies. GSE6791 (21) data set included the information on a human papillomavirus infection (HPV) status; the difference between cancer and normal samples was very significant in both the subgroup of HNSCC HPV^+ (26 HNSCC vs. 14 normal mucosa; P value = 4E-06) and in the subgroup of HNSCC HPV^- (16 HNSCC vs. 14 normal mucosa; P value = 1E-13).

Similarly, we found that the gene expression profiles of the 1,270 genes that we identified were differentially expressed in normal bronchial epithelial samples from smokers with versus without lung cancer. We used the data set published by Spira and colleagues (27), who demonstrated that gene expression profile in histologically normal large airway epithelial cells obtained at bronchoscopy from

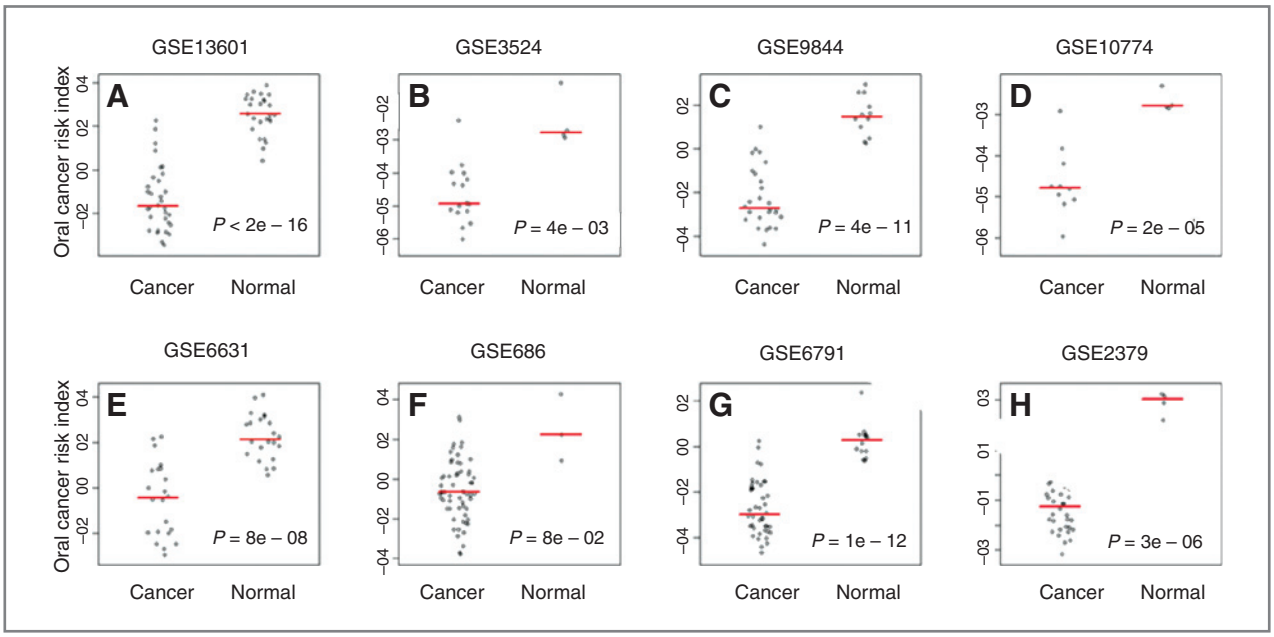

Figure 3. Oral cancer risk index in head and neck cancers. Samples under comparison are HNSCC versus normal mucosa in both human tumors (A–C, E–H) and head and neck cell lines (D). The oral cancer index was computed as the average level of expression of the transcripts associated with a HR > 1 minus the average level of expression of the transcripts that have a HR < 1. A–C compared HNSCC and normal mucosa from the oral cavity. E–G compared HNSCC and normal mucosa from various anatomic locations. Panel H compared hypopharynx squamous cell carcinoma from normal hypopharynx mucosa. The microarray data sets were downloaded from GEO. The GEO accession numbers associated with the data sets were shown at the top of each panel.

smokers with suspicion of lung cancer could be used as a lung cancer biomarker. We computed the oral cancer risk indices for the 163 samples in the data set and found them to be significantly different between lung cancer and the rest (Fig. 4A). Spira and colleagues developed a biomarker score computed from 80 genes that can distinguish lung cancer from normal lung in the data set. Interestingly, there was a strong correlation between the oral cancer index and the reported biomarker scores for these samples (Fig. 4B).

Discussion

In this report, we demonstrate that gene expression profile can significantly improve the prediction of OSCC development over clinical and histologic variables in OPL patients. Multiple prediction models were developed and compared using CoxBoost algorithm. We observed a marked improvement in prediction accuracy when a gene expression profile was used. With the gene expression profile only, we developed a 29-trancript prediction model that had prediction error rate around 8%. Using the profile in combination with the previously known risk factors, the model showed a similar prediction error rate as the expression profile alone. Because using the previously known risk factors alone had a clear inferior performance (Fig. 1) compared with models 1 and 2, it is clear that the expression profiles have a predictive value beyond the known risk factors. As an alternative way to assess the misclassification rate of genomic predictors in general, we employed a simpler approach, which used DLDA algorithm to develop prediction models and the standard 10-fold cross-validation scheme to evaluate the models. We obtained 16% misclassification rate, which is highly statistically significant ($P < 1.0E-16$, compared with null hypothesis) compared with other risk factors alone. These results suggest that the gene expression profile may robustly predict oral cancer development in patients with OPL.

Because no prospective cohort is currently available to validate our finding, we acknowledge that our study only represents the first step in the development of a biomarker that could be used in clinical practice. However, we consider it as a proof-of-principle that a gene expression signature developed in patients with preneoplasia may improve our prediction of cancer development over clinical and pathologic factors. It also provides a list of transcripts that could facilitate future efforts to better understand the disease and intervene in its progress. To move this work into a clinical tool, the next step will be (i) to refine the signature and adapt it on a CLIA (Clinical Laboratory Improvement Amendments)-certifiable platform and (ii)- to identify an independent cohort of patients for the validation of our models. This work is ongoing and clearly beyond the scope of this study.

It is important to note that we used tissue samples collected prior to cancer development in this study, which is different from most gene expression–based studies in which cancers were used. A number of studies have shown a value of gene expression profiles in cancer prognosis. For

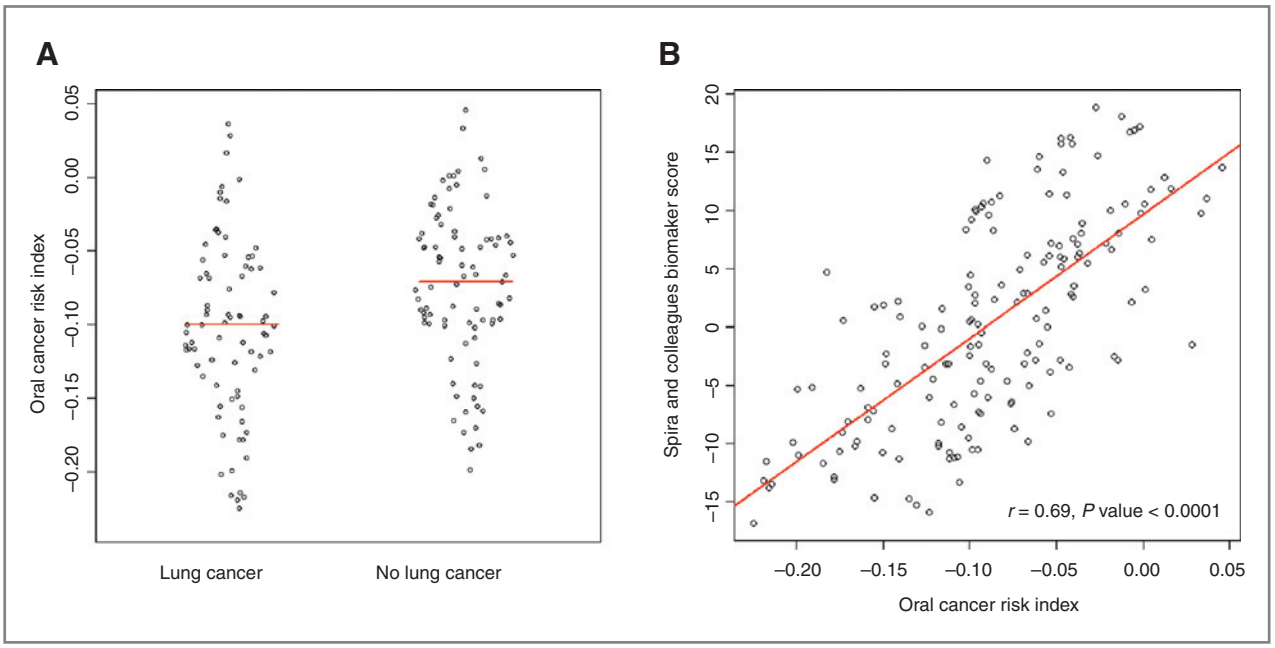

Figure 4. Correlation between oral cancer risk index (OCRI; average level of expression of the transcripts associated with a HR > 1 minus the average level of expression of the transcripts that have a HR < 1) and Spira's biomarker score (from GSE4115) in 163 patients with a suspicious lung lesion. A, OCRI calculated in patients included in the study of Spira and colleagues, with or without lung cancer (P value = 0.003). B, correlation between OCRI and Spira biomarker score (r = 0.69, P value < 0.0001).

example, Shedden and colleagues performed a multisite, blinded validation study to assess several prognostic models based on gene expression profiles of 442 lung adenocarcinomas (10). Several of the models being evaluated produced risk scores that significantly correlated with outcome, and the models worked better with clinical data. However, cancer prognosis remains to be a difficult problem because the tumors are heterogeneous and they evolve over time. Samples collected from a particular site at a particular time may not be able to provide adequate information to predict behavior of the cancer. In comparison, the samples used in this study may be less complex because they were in the early tumorigenic process.

Gene expression profiling was obtained from the whole biopsy. The absence of microdissection to isolate the epithelial cells from the underlying stroma, did not allow us to differentiate the respective contribution of these 2 compartments. Therefore, the genes we identified may include genes expressed by both the epithelial cells and stromal cells. Our objective in this work was to improve our prediction accuracy over clinical and histologic markers. We believe that capturing the information from both compartments may be important to achieve this goal (32).

The samples used in this study were collected at baseline or at 3 months after the inclusion. The conclusion of the trial was that the drugs used in the trial, even if they induced some clinical responses, were ineffective in preventing oral cancer development (7, 12). Therefore, the influence of the drugs used in the trial on gene expression is likely, but is peripheral in the context of our study, which objective was to identify genes associated with oral cancer development. Similarly, we did not consider other factors, such as gender and ethnicity, which may influence gene expression but were not associated with oral cancer development.

Our set of patients was enriched in patients who developed oral cancer and in never smokers compared with the remaining patients not included in the trial. Tobacco has been established as a significant risk factor in the development of oral leukoplakia and oral cancer. However, the population with leukoplakia is heterogeneous, and although never smokers as well as women often represent a small proportion of the patients with oral leukoplakia, the risk of oral cancer development has been reported to be higher than in smokers. With a mean follow-up of 7.2 years, Silverman and colleagues reported a transformation rate of oral leukoplakia of 24% in never smokers versus 16% and 12% in current and former smokers, respectively (2). Einhorn and colleagues reported an 8-fold risk for patients who never smoked with oral leukoplakia (33). Because the incidence of HPV infection in oral cancer is low, as opposed to oropharyngeal cancer (34), further studies are needed to better understand the development of oral cancer in never smokers.

It is a well-recognized challenge to develop prognostic models from microarray gene expression profile data. Subramaniam and Simon identified a number of statistical issues in the design and evaluation of the prognostic models in recent studies, which casts some doubts on the readiness of the models for practical clinical use (29). To ensure that our results are reproducible, we documented the script used in our analysis in detail (Supplementary Material 2). The CoxBoost algorithm fits a Coxph

model by component-wise, likelihood-based boosting. It is especially suited for models with a large number of predictors and allows for mandatory covariates. Binder and colleagues demonstrated the utility of the method using both simulated data and real microarray data from patients with bladder cancer (16, 35). It was shown that microarray features selected by the CoxBoost approach can improve prediction performance over a purely clinical model. The algorithm has also been recently used as along with 3 other popular methods to compare gene-based with pathway-based procedures for the identification of prediction models (36). Thus, we thought that CoxBoost is an appropriate tool to identify biomarkers beyond clinical variables from microarray gene expression profiling data. The consistency between the new CoxBoost approach and the more common Coxph model as shown in Table 2, was also reassuring.

Microarray gene expression profiling has become a mature and widely used high-throughput technology. Even though it is typical that real-time PCR is used for validating the finding in microarray studies, we did not think cherry-picking some of the transcripts included in models 1 and 2 is effective or adequate. Instead, we used 8 different data sets generated from different microarray platforms to test whether the oral cancer index, which summarizes the information from a comprehensive list of transcripts associated with oral cancer development, can differentiate cancer from normal cells. Because we are able to find significant association between the oral cancer index and cancer status, it greatly enhances our confidence in our results. Furthermore, this list of transcripts may provide key biological factors associated with oral cancer development.

In a recent study, Bhutani and colleagues demonstrated that oral epithelium could serve as a surrogate tissue for assessing smoking-induced molecular alterations in the lungs (37). They studied promoter methylation of *p16* and *FHIT* genes in oral and bronchial brush specimens from smokers enrolled in a randomized placebo-controlled chemoprevention trial. They showed that bronchial methylation was correlated with oral tissue methylation. These results suggest the possibility of oral tissues as a molecular mirror of lung carcinogenesis (38). On the other hand, Spira and colleagues studied gene expression profiles of normal bronchial samples of smokers (27). The authors developed a multigene index that can distinguish smokers with or without lung cancer from noncancer samples with high sensitivity and specificity. They proposed that this index may also predict lung cancer risk in smokers. Because our study also predicts cancer risk, as we expected, we found that the risk index calculated according to our list of significant transcripts also correlated with the lung cancer risk index given by Spira and colleagues (Fig. 4).

Because many of the significant transcripts have been shown altered in cancers, it suggests that gene expression profiles may evolve progressively toward cancer before the cells become cancers. Consistently, we observed a significant upregulation of several gene sets associated with the proteasome machinery using functional pathway analysis of the significant genes. Protein synthesis and degradation is a tightly regulated process that is essential for normal cellular homeostasis (39). Many proteasome target proteins are involved in important processes of carcinogenesis and cancer survival, such as *TP53* and *CDKN1B p27* (39). Downregulation of these genes were also significantly associated with the development of oral cancer in our study (Supplementary Material 8).

Consistent with our previous results using deltaNp63 protein expression, tumor protein p63 (TP63) mRNA expression was also associated with a high risk to develop OSCC (HR: 4.4, Wald test $P = 3.6E\text{-}4$; ref. 7). Among other very significant genes were 4 of the 5 small integrin-binding ligand N-linked glycoproteins (SIBLING), that are cell adhesion modulators, were among the transcripts most significantly associated with oral cancer development [dentin sialophosphoprotein (*DSPP*), dentin matrix protein 1 (*DMP1*), secreted phosphoprotein 1 (*SPP1*), and integrin-binding sialoprotein (*IBSP*)]. The genes encoding the SIBLINGs are located within a cluster on chromosome 4. They deserve further studies to define their functional role in oral cancer development (ref. 40; Supplementary Material 8).

Our study may provide valuable information for designing cancer prevention strategies. One may consider to use proteasome inhibitors (e.g., bortezomib) for oral cancer prevention. As a single agent or in combination with standard therapy, its limited inhibition activity in HNSCC and other solid tumors (41) may be related to an upregulation of both proapoptotic proteins and antiapoptotic proteins. Recent studies have shown that combining bortezomib with cetuximab (an EGFR-directed antibody) or STAT3 inhibitors, might enhance its efficacy (42, 43). However, bortezomib toxicity and its intravenous mode of administration preclude its evaluation in the chemoprevention setting (41). Less toxic and orally active proteasome inhibitors are under development (44). Several natural compounds with proteasome-inhibitory effects have also been investigated in chemoprevention (41). Green tea consumption has produced promising effects against development of prostate cancer, without inducing major toxicities (45). Based on the results of our study, those compounds deserve further evaluation in preclinical models of oral carcinogenesis. Tsao and colleagues reported recently the results of a phase II randomized, placebo-controlled trial of green tea extract (GTE) in patients with high-risk OPLs. The OPL clinical response rate was higher in all GTE arms ($n = 28$; 50%) versus placebo ($n = 11$; 18.2%; $P = 0.09$) but did not reach statistical significance. However, the 2 higher dose GTE arms [58.8% (750 and 1,000 mg/m^2), 36.4% (500 mg/m^2), and 18.2% (placebo); $P = 0.03$] had higher responses, suggesting a dose–response effect (46).

DNMT3B transcript, which is one of the most significant risk factors in our list (HR: 7.7, Wald test P = 4.3E-6) and part of model 2, may deserve particular attention for its role in epigenetic tumorigenesis. It is possible that epigenetic tumorigenesis mediated by *DNMT3B* could be an early

event in oral tumorigenesis. The role of *DNMT3B* in tumorigenesis has been recently highlighted in various cancer (47, 48). Variant forms of *DNMT3B* transcripts have been described to play a major role in non–small cell lung cancer and may deserve further studies in HNSCC (49). Some *DNMT3B* polymorphisms have been associated with HNSCC risk in non-Hispanic whites (50). A recent study of the combination of a DNA demethylating drug and all-trans-retinoic acid has shown a reduction of oral cavity cancer induced by the carcinogen 4-nitroquinoline 1-oxide in a mouse model (51). We compared *DNMT3B* expression levels in 3 publicly available data sets and found *DNMT3B* was overexpressed in HNSCC versus normal mucosa, consistent with the role of *DNMT3B* overexpression in head and neck tumorigenesis (details not shown). One possible mechanism of regulation of *DNMT3B* expression involves noncoding RNAs. Micro RNA-29 family has been demonstrated to revert aberrant methylation in lung cancer by targeting *DNMT3A* and *DNMT3B* (52). Our microarray platform also measured the precursor forms of miRNA. Consistent with this hypothesis, hsa-miR-29b-1 was found to be the most protective marker in our univariate Cox model analysis (HR: 0.0008, Wald test $P = 0.0002$). A significant negative correlation was observed between hsa-miR-29b-1 and *DNMT3B* expression ($R = -0.38$, $P = 0.0002$).

Our results showed that hsa-miR-101-1 (Table 2) was another micro RNA associated with a low risk to develop oral cancer. hsa-miR-101 expression was also reported to be reduced in early-stage neoplastic transformation in the lungs of F344 rats chronically treated with the tobacco carcinogen NNK [4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone; ref. 53]. It has been associated in these studies with the upregulation of *COX2*, and enhancer of zeste homologue 2 (*EZH2*), a mammalian histone methyltransferase that contributes to the epigenetic silencing of target genes and regulates the survival and metastasis of cancer cells (54, 55). However, in our study, *COX2* and *EZH2* gene expression were not significantly associated with OSCC development. Other genes might be regulated by this micro RNA.

The micro RNA–based strategies might therefore be considered in future chemoprevention studies, especially for OPLs, which is easily accessible and frequently involves only one or a few lesions.

In summary, we have demonstrated the value of gene expression profiles in predicting oral cancer development in OPL patients, beyond previously reported clinical and pathologic biomarkers. If validated in future studies, the profiles may serve as biomarkers to classify OPLs for oral cancer risk in routine clinical practice. Interestingly, certain transcripts in the profiles may be important in oral tumorigenesis and should be considered as potential targets for oral cancer prevention.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Grant Support

This work was supported in part by National Cancer Institute grants P01 CA106451, P50 CA97007, and P30 CA16672.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received July 8, 2010; revised December 16, 2010; accepted December 22, 2010; published online February 3, 2011.

References

1. Kim J, Raz D, Jablons D. Unmet need in lung cancer: can vaccines bridge the gap?Clin Lung Cancer 2008;9 Suppl 1:S6–12.
2. Silverman S Jr, Gorsky M, Lozada F. Oral leukoplakia and malignant transformation. A follow-up study of 257 patients. Cancer 1984; 53:563–8.
3. Warnakulasuriya S, Johnson NW, Van Der Waal I. Nomenclature and classification of potentially malignant disorders of the oral mucosa. J Oral Pathol Med 2007;36:575–80.
4. Mao L, Lee JS, Fan YH, Ro JY, Batsakis JG, Lippman S, et al. Frequent microsatellite alterations at chromosomes 9p21 and 3p14 in oral premalignant lesions and their value in cancer risk assessment. Nat Med 1996;2:682–5.
5. Lee JJ, Hong WK, Hittelman WN, Mao L, Lotan R, Shin DM, et al. Predicting cancer development in oral leukoplakia: ten years of translational research. Clin Cancer Res 2000;6:1702–10.
6. Kawaguchi H, El-Naggar AK, Papadimitrakopoulou V, Ren H, Fan YH, Feng L, et al. Podoplanin: a novel marker for oral cancer risk in patients with oral premalignancy. J Clin Oncol 2008;26:354–60.
7. Saintigny P, El-Naggar AK, Papadimitrakopoulou V, Ren H, Fan YH, Feng L, et al. DeltaNp63 overexpression, alone and in combination with other biomarkers, predicts the development of oral cancer in patients with leukoplakia. Clin Cancer Res 2009;15:6284–91.
8. Taoudi Benchekroun M, Saintigny P, El-Naggar AK, Papadimitrakopoulou V, Ren H, Lang W, et al. Epidermal growth factor receptor expression and gene copy number in the risk of oral cancer. Cancer Prev Res 2010;3:800–9.
9. Chung CH, Parker JS, Karaca G, Wu J, Funkhouser WK, Moore D, et al. Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. Cancer Cell 2004;5:489–500.
10. Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med 2008;14:822–7.
11. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999–2009.
12. Papadimitrakopoulou VA, Lee JJ, William WN Jr, Martin JW, Thomas M, Kim ES, et al. Randomized trial of 13-cis retinoic acid compared with retinyl palmitate with or without beta-carotene in oral premalignancy. J Clin Oncol 2009;27: 599–604.
13. Dudoit S, Gentleman RC, Quackenbush J. Open source software for the analysis of microarray data. Biotechniques 2003;Suppl:45–51.
14. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249–64.
15. Pounds S, Morris SW. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. Bioinformatics 2003;19:1236–42.
16. Binder H, Allignol A, Schumacher M, Beyersmann J. Boosting for high-dimensional time-to-event data with competing risks. Bioinformatics 2009;25:890–6.

17. Efron B TR. Improvements on cross-validation: the.632+ bootstrap method. J Am Stat Assoc 1997;92:548–60.
18. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97:77–87.
19. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102:15545–50.
20. Ye H, Yu T, Temam S, Ziober BL, Wang J, Schwartz JL, et al. Transcriptomic dissection of tongue squamous cell carcinoma. BMC Genomics 2008;9:69.
21. 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 Res 2007;67:4605–19.
22. Toruner GA, Ulger C, Alkan M, Galante AT, Rinaggio J, Wilk R, et al. Association between gene expression profile and tumor invasion in oral squamous cell carcinoma. Cancer Genet Cytogenet 2004;154:27–35.
23. Kuriakose MA, Chen WT, He ZM, Sikora AG, Zhang P, Zhang ZY, et al. Selection and validation of differentially expressed genes in head and neck cancer. Cell Mol Life Sci 2004;61:1372–83.
24. Estilo CL, Oc P, Talbot S, Socci ND, Carlson DL, Ghossein R, et al. Oral tongue cancer gene expression profiling: identification of novel potential prognosticators by oligonucleotide microarray analysis. BMC Cancer 2009;9:11.
25. Cromer A, Carles A, Millon R, Ganguli G, Chalmel F, Lemaire F, et al. Identification of genes associated with tumorigenesis and metastatic potential of hypopharyngeal cancer by microarray analysis. Oncogene 2004;23:2484–98.
26. Lee TL, Yang XP, Yan B, Friedman J, Duggal P, Bagain L, et al. A novel nuclear factor-kappaB gene signature is differentially expressed in head and neck squamous cell carcinomas in association with TP53 status. Clin Cancer Res 2007;13:5680–91.
27. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nat Med 2007;13:361–6.
28. Coombes KR, Wang J, Baggerly KA. Microarrays: retracing steps. Nat Med 2007;13:1276–7; author reply 7–8.
29. Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use?J Natl Cancer Inst 2010;102:464–74.
30. Tibshirani R. Immune signatures in follicular lymphoma. N Engl J Med 2005;352:1496–7; author reply 7.
31. van Riggelen J, Yetil A, Felsher DW. MYC as a regulator of ribosome biogenesis and protein synthesis. Nat Rev Cancer 2010;10:301–9.
32. Mueller MM, Fusenig NE. Friends or foes – bipolar effects of the tumour stroma in cancer. Nat Rev Cancer 2004;4:839–49.
33. Einhorn J, Wersall J. Incidence of oral carcinoma in patients with leukoplakia of the oral mucosa. Cancer 1967;20:2184–93.
34. Liang XH, Lewis J, Foote R, Smith D, Kademani D. Prevalence and significance of human papillomavirus in oral tongue cancer: the Mayo Clinic experience. J Oral Maxillofac Surg 2008;66:1875–80.
35. Dyrskjot L, Zieger K, Real FX, Malats N, Carrato A, Hurst C, et al. Gene expression signatures predict outcome in non-muscle-invasive bladder carcinoma: a multicenter validation study. Clin Cancer Res 2007;13:3545–51.
36. Chen X, Wang L, Ishwaran H. An integrative pathway-based clinical-genomic model for cancer survival prediction. Stat Probab Lett 2010;80:1313–9.
37. Bhutani M, Pathak AK, Fan YH, Liu DD, Lee JJ, Tang H, et al. Oral epithelium as a surrogate tissue for assessing smoking-induced molecular alterations in the lungs. Cancer Prev Res 2008;1:39–44.
38. Sidransky D. The oral cavity as a molecular mirror of lung carcinogenesis. Cancer Prev Res 2008;1:12–4.
39. Burger AM, Seth AK. The ubiquitin-mediated protein degradation pathway in cancer: therapeutic implications. Eur J Cancer 2004;40:2217–29.
40. Bellahcene A, Castronovo V, Ogbureke KU, Fisher LW, Fedarko NS. Small integrin-binding ligand N-linked glycoproteins (SIBLINGs): multifunctional proteins in cancer. Nat Rev Cancer 2008;8:212–26.
41. Yang H, Zonder JA, Dou QP. Clinical development of novel proteasome inhibitors for cancer treatment. Expert Opin Investig Drugs 2009;18:957–71.
42. Li C, Zang Y, Sen M, Leeman-Neill RJ, Man DS, Grandis JR, et al. Bortezomib up-regulates activated signal transducer and activator of transcription-3 and synergizes with inhibitors of signal transducer and activator of transcription-3 to promote head and neck squamous cell carcinoma cell death. Mol Cancer Ther 2009;8:2211–20.
43. Wagenblast J, Baghi M, Arnoldner C, Bisdas S, Gstöttner W, Ackermann H, et al. Cetuximab enhances the efficacy of bortezomib in squamous cell carcinoma cell lines. J Cancer Res Clin Oncol 2009;135:387–93.
44. Piva R, Ruggeri B, Williams M, Costa G, Tamagno I, Ferrero D, et al. CEP-18770: a novel, orally active proteasome inhibitor with a tumor-selective pharmacologic profile competitive with bortezomib. Blood 2008;111:2765–75.
45. Bettuzzi S, Brausi M, Rizzi F, Castagnetti G, Peracchia G, Corti A. Chemoprevention of human prostate cancer by oral administration of green tea catechins in volunteers with high-grade prostate intraepithelial neoplasia: a preliminary report from a one-year proof-of-principle study. Cancer Res 2006;66:1234–40.
46. Tsao AS, Liu D, Martin J, Tang XM, Lee JJ, El-Naggar AK, et al. Phase II randomized, placebo-controlled trial of green tea extract in patients with high-risk oral premalignant lesions. Cancer Prev Res (Phila Pa) 2009;2:931–41.
47. Lin RK, Hsu HS, Chang JW, Chen CY, Chen JT, Wang YC. Alteration of DNA methyltransferases contributes to 5'CpG methylation and poor prognosis in lung cancer. Lung Cancer 2007;55:205–13.
48. Nosho K, Shima K, Irahara N, Kure S, Baba Y, Kirkner GJ, et al. DNMT3B expression might contribute to CpG island methylator phenotype in colorectal cancer. Clin Cancer Res 2009;15:3663–71.
49. Wang J, Bhutani M, Pathak AK, Lang W, Ren H, Jelinek J, et al. Delta DNMT3B variants regulate DNA methylation in a promoter-specific manner. Cancer Res 2007;67:10647–52.
50. Liu Z, Wang L, Wang LE, Sturgis EM, Wei Q. Polymorphisms of the DNMT3B gene and risk of squamous cell carcinoma of the head and neck: a case-control study. Cancer Lett 2008;268:158–65.
51. Tang XH, Knudsen B, Bemis D, Tickoo S, Gudas LJ. Oral cavity and esophageal carcinogenesis modeled in carcinogen-treated mice. Clin Cancer Res 2004;10:301–13.
52. Fabbri M, Garzon R, Cimmino A, et al. MicroRNA-29 family reverts aberrant methylation in lung cancer by targeting DNA methyltransferases 3A and 3B. Proc Natl Acad Sci USA 2007;104:15805–10.
53. Kalscheuer S, Zhang X, Zeng Y, Upadhyaya P. Differential expression of microRNAs in early-stage neoplastic transformation in the lungs of F344 rats chronically treated with the tobacco carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone. Carcinogenesis 2008;29:2394–9.
54. Varambally S, Cao Q, Mani RS, Shankar S, Wang X, Ateeq B, et al. Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer. Science 2008;322:1695–9.
55. Strillacci A, Griffoni C, Sansone P, Paterini P, Piazzi G, Lazzarini G, et al. MiR-101 downregulation is involved in cyclooxygenase-2 overexpression in human colon cancer cells. Exp Cell Res 2009;315:1439–47.