Published in final edited form as:
Cancer Prev Res (Phila). 2011 February ; 4(2): 218–229. doi:10.1158/1940-6207.CAPR-10-0155.

Gene Expression Profiling Predicts the Development of Oral Cancer

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

[1] Department of Thoracic/Head and Neck Medical Oncology, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA [2] Department of Bioinformatics and Computational Biology, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA [3] Department of Pathology, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA [4] Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA [5] University of Maryland Dental School, Baltimore, Maryland

Abstract

Patients with oral preneoplastic lesion (OPL) have 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 develope 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 clinico-pathological risk factors. Based on the gene expression profile data, we also identified 2182 transcripts significantly associated with oral cancer risk associated genes (*P*-value<0.01, single variate Cox proportional hazards model). Functional pathway analysis revealed proteasome machinery, MYC, and ribosomes components as the top gene sets associated with oral cancer risk. In multiple independent datasets, 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.

Keywords

gene expression profiling; oral cancer; oral leukoplakia; biomarker

Requests for reprints: Li Mao, Department of Oncology and Diagnostic Sciences, University of Maryland Dental School, 650 W Baltimore Street, Baltimore, MD 21201, USA; lmao@umaryland.edu.

*P.S. and L.Z. contributed equally to this study and report.

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 diagnose the disease before the cancerous stage and preventing development of invasive cancers.

Leukoplakia and erythroplakia are the most commonly diagnosed oral premalignant lesions (OPLs) with a 17%–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 loss of heterozygosity (LOH) profile (4), polysomy (5), p53 (5), over expression 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 biologic 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 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 pre-specified 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 histological 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 M. D. 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 over-sampled relative to the non-events in order 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. Clinical-pathologic 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 et al (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 M.D. Anderson Cancer Center 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 detailled 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 (13)). Raw data of microarrays were processed using quantile normalization and RMA algorithm (14). Single-variate 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 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 versus dysplasia), podoplanin and deltaNp63 expression. To evaluate the performance of the models, we used the .632+ bootstrap method (17) and prediction error curve estimates (16). The later 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 dataset were dichotomized into short cancer-free and long cancer-free patients based on a follow-up cutoff of 5 years. Seventeen patients with 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, 9/10 of the samples (N=62) were used to (1) identify the most significant 50 transcripts that are associated with OCFS time and (2) to build a DLDA model using the 50 transcripts as the predictors. Then the model is tested in the remaining 1/10 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 > 1 minus the average level of expression of the transcripts that have a hazard ratio < 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 Gene Set Enrichment Analysis (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 two biologic states (e.g., high risk *v* 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 biologic pathways, chromosomal location, or expression profiles in previously generated microarray data sets. The input data for GSEA procedure were the following: (1) a complete table of genes ranked according to the log2 transformed Cox model hazards ratio associated with the development of oral squamous cell carcinoma, (2) a mapping file for identifying transcripts in HG-1.ST platform, and (3) a catalog 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 1436 curated gene sets (canonical pathway gene sets, chemical and genetic perturbations gene sets, BioCarta gene sets, GenMAPP gene sets, 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 0.25 criterion were considered (Supplementary Material 3).

External datasets

Nine independent datasets were used to validate our findings, and downloaded from Gene Expression Omnibus: GSE9844 (20), GSE6791 (21), GSE3524 (22), GSE6631 (23), GSE13601 (24), GSE2379 (25), and GSE686 (9), which compared HNSCC and normal mucosa, GSE10774 that studied normal keratinocyte and various HNSCC cell lines (26), and GSE4115 that 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 National Institutes of Health Gene Expression Omnibus database at www.ncbi.nlm.nih.gov/geo under the accession number GSEXXX [To be available upon acceptance of publication. Reviewers can access the data 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 pathological 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.

Developping 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 transcripts with a non-zero coefficient in the prediction model, an explicit penalty score was added to the fitness function, which is a partial log-likelihood function. In this way, the over parameterization 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, deltaNp63 and podoplanin expression were used as mandatory covariates along with the microarray data. In Model 3, only age, histology, deltaNp63 and podoplanin expression were used as predictors. Table 2a 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 2a showed the frequency of occurrences of the transcripts. Among the 29 transcripts, 21 have frequencies greater than 80%, showing the lists of marker trancripts are mostly consistent bewteen different runs of the algorithm. Similarly, Table 2b shows the transcripts found in Model 2; 15 genes of the 23 in Table 2b have frequencies greater than 80%. There are 9 transcripts shared in Table 2a and Table 2b.

The hazard ratio and Wald-test p-values obtained from using single-variate Cox proportional hazard 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 hazard ratio provided by the CPH model (positive and negative CoxBoost coefficients being associated with a hazard ratio > 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 errors suggest better performance. Following Binder et al. (16), we computed the prediction error curve using boostrap samples and aggregated into .632+ estimates. The .632+ method was invented by Efron et *al.* as an improvement over conventional cross validation schemes for assessing model performance (17). The advantage of .632+ method is that it allows the use of all observations to train the prediction model, but nontheless 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 presviously known factors (i.e., age, histology, deltaNp63 and podoplanin expression). Model 1 and Model 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 used random numbers as predictors, was also shown in the figure 1. As expected, null model has the worst performance. Model 1 and Model 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 K-M survival curves of the two groups and the accompanying curves based on 50 bootstrap samples display marked differences in survival time (log-rank p-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 was 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.

Checking and testing the significant genes

Because overfitting and fragile-inference is a well-recognized concern when microarray data is 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 (Diagnal Linear Discrimination Analysis) 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% Confidence Interval (95CI): 88–93%], specificity was 76% (95CI: 72–78%), positive predictive value was 80% (95CI: 78–82%), and negative predictive value was 89% (95CI: 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 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 $<10^{-16}$ 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 assocaited 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 single variate Cox proportional hazards model to identify the significant transcripts. A *p*-value (Wald test) was computed for each of the transcripts.

We found 2182 significant transcripts that have *p*-value < 0.01 and the false discovery rate (FDR) was estimated to be 11%. Supplemetary Material 8 shows a histogram of the *p*-values, and Supplementary Material 9 provide 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 false discovery rate (FDR) of the significant transcripts, which assumes that distribution of the p values of the non significant transcripts follows a uniform distribution while the distribution of the p values of significant genes follows a beta distribution. Among the 2182 significant transcripts, 1262 were associated with a high risk to develop oral cancer (hazard ratio > 1), and 920 were associated with low

risk (hazard ratio < 1). All but 3 trancripts included in the CoxBoost Model 1 and all but 1 trancripts included in the Model 2 were included in list of 2182 trancripts, with very significant p-values (Table 2). As well, hazard ratios 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 Gene Set Enrichment Analysis (GSEA) agorithm (19), which seeked significant associations between the hazard ratios that 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 ribosomes 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 datasets. We took the 2182 significant transcripts, found them correspond to 1270 gene symbols according to annotation provided by the manufacturer. We then extracted gene expression data that have matched gene symbols from multiple datasets composed of HNSCC, and normal mucosa samples, including one dataset 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 to normal mucosa across 7 independent datasets, and lower in HNSCC cell compared to normal keratinocytes in one cell line. All these differences were highly significant, with few overlapp between cancer and normal samples in most of ths studies. GSE6791 (21) dataset 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 versus 14 normal mucosa; p-value=4e-06) and in the subgroup of HNSCC HPV- (16 HNSCC versus 14 normal mucosa; p-value=1e-13).

Similarly, we found that the gene expression profiles of the 1270 genes we identified, were differentially expressed in normal bronchial epithelial samples from smokers with versus without lung cancer. We used the dataset published by Spira et al. (27), who demonstrated that gene expression profile in histologically normal large-airway epithelial cells obtained at bronchoscopy from 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 dataset and found them to be significantly different between lung cancer and the rest (Figure 4A). Spira *et al.* developed a biomarker score computed from 80-genes that can distinguish lung cancer from normal lung in the dataset. Interestingly, there was a strong correlation between the oral cancer index and the reported biomarker scores for these samples (Figure 4B).

Discussion

In this report, we demonstrate that gene expression profile can significantly improve the prediction of OSCC development over clinical and histological 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 (Figure 1) compared to 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 evaluated the models. We obtained 16% misclassification rate, which is highly statistically significant ($P < 1.0E-16$, compared to null hypothesis) compared with other risk factors alone. These results suggest that the 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 pathological factors. It also provides a list of transcripts that could facilitate future efforts to better understand the disease and intervene in its progress. In order to move this work into a clinical tool, the next step will be (1)-to refine the signature and adapt it on a CLIA-certifiable platform, and (2)-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 where cancers were used. A number of studies have shown a value of gene expression profiles in cancer prognosis. For example, Shedden et al performed a multi-site, 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 heterogenous, 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 histological 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 to 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 et al reported a transformation rate of oral leukoplakia of 24% in never smokers versus 16% and 12% in current and former smokers respectively (2). Einhorn et al and Roed-Petersen et al reported an eightfold risk and five-fold risk for patients never smoker with oral leukoplakia (33). Because the incidence of human papilloma virus 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. Subramnian and Simon identified a number of statistical issues in the design and evluation of the prognostic models in recent studies, which casts some doubts on the readyness 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 Cox proportional hazards 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 *et al.* 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 three other popular methods to compare gene-based versus pathway-based procedures for the identification of prediction models (36). Thus, we thought 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. Eventhough it is typical that RTQ-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 datasets generated from different microarray platforms to test wether 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. Since 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 et al. 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 were 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 et al studied gene expression profiles of normal bronchial samples of smokers (27). The authors developed a multi-gene index that can distinguish smokers with or without lung cancer from non-cancer samples with high sensitivity and specificity. They proposed that this index may also predict lung cancer risk in smokers. Since 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 Spira et al's lung cancer risk index (Figure 4).

Since many of the significant transcripts have been shown altered in cancers, it suggests that gene expression profiles may evolve progressively towards 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). Down regulation 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 (hazard ratio (HR): 4.4, Wald test P = 3.6E-4) (7). Among other very significant genes were 4 of the 5 small integrin-binding ligand N-linked glycoproteins (SIBLINGs), 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 (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 pro-apoptotic proteins and anti-apoptotic 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 et al. reported recently the results of a Phase II randomized, placebo-controlled trial of green tea extract (GTE) in patients with high-risk oral premalignant lesions. 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 two 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 trancript, 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 datasets 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. MicroRNA-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 microRNA 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 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (53). It has been associated in these studieswith the upregulation of cyclooxygenase-2 (*COX2*), and enhancer of zeste homolog 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 microRNA.

The micro-RNAbased 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 pathological 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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

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. [PubMed: 18540529]
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. [PubMed: 6537892]
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. [PubMed: 17944749]
4. Mao L, Lee JS, Fan YH, 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. [PubMed: 8640560]
5. Lee JJ, Hong WK, Hittelman WN, et al. Predicting cancer development in oral leukoplakia: ten years of translational research. Clin Cancer Res. 2000; 6:1702–10. [PubMed: 10815888]
6. Kawaguchi H, El-Naggar AK, Papadimitrakopoulou V, et al. Podoplanin: a novel marker for oral cancer risk in patients with oral premalignancy. J Clin Oncol. 2008; 26:354–60. [PubMed: 18202409]

7. Saintigny P, El-Naggar AK, Papadimitrakopoulou V, 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. [PubMed: 19773378]
8. Taoudi Benchekroun M, Saintigny P, El-Naggar AK, Papadimitrakopoulou V, Ren H, Lang W, Fan YH, Huang J, Feng L, Lee JJ, Kim ES, Hong KW, Johnson FM, Mao L. Epidermal growth factor receptor expression and gene copy number in the risk of oral cancer. Cancer Prev Res.
9. Chung CH, Parker JS, Karaca G, et al. Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. Cancer Cell. 2004; 5:489–500. [PubMed: 15144956]
10. Shedden K, Taylor JM, Enkemann SA, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med. 2008; 14:822–7. [PubMed: 18641660]
11. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002; 347:1999–2009. [PubMed: 12490681]
12. Papadimitrakopoulou VA, Lee JJ, William WN Jr, 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. [PubMed: 19075276]
13. Dudoit S, Gentleman RC, Quackenbush J. Open source software for the analysis of microarray data. Biotechniques. 2003; (Suppl):45–51. [PubMed: 12664684]
14. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003; 4:249–64. [PubMed: 12925520]
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. [PubMed: 12835267]
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. [PubMed: 19244389]
17. Efron BTR. 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, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102:15545–50. [PubMed: 16199517]
20. Ye H, Yu T, Temam S, et al. Transcriptomic dissection of tongue squamous cell carcinoma. BMC Genomics. 2008; 9:69. [PubMed: 18254958]
21. Pyeon D, Newton MA, 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; 67:4605–19. [PubMed: 17510386]
22. Toruner GA, Ulger C, Alkan M, et al. Association between gene expression profile and tumor invasion in oral squamous cell carcinoma. Cancer Genet Cytogenet. 2004; 154:27–35. [PubMed: 15381369]
23. Kuriakose MA, Chen WT, He ZM, et al. Selection and validation of differentially expressed genes in head and neck cancer. Cell Mol Life Sci. 2004; 61:1372–83. [PubMed: 15170515]
24. Estilo CL, POc, Talbot S, et al. Oral tongue cancer gene expression profiling: Identification of novel potential prognosticators by oligonucleotide microarray analysis. BMC Cancer. 2009; 9:11. [PubMed: 19138406]
25. Cromer A, Carles A, Millon R, et al. Identification of genes associated with tumorigenesis and metastatic potential of hypopharyngeal cancer by microarray analysis. Oncogene. 2004; 23:2484–98. [PubMed: 14676830]
26. Lee TL, Yang XP, Yan B, 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. [PubMed: 17908957]
27. Spira A, Beane JE, Shah V, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nat Med. 2007; 13:361–6. [PubMed: 17334370]

28. Coombes KR, Wang J, Baggerly KA. Microarrays: retracing steps. Nat Med. 2007; 13:1276–7. author reply 7–8. [PubMed: 17987014]
29. Subramanian J, Simon R. Gene Expression-Based Prognostic Signatures in Lung Cancer: Ready for Clinical Use? J Natl Cancer Inst.
30. Tibshirani R. Immune signatures in follicular lymphoma. N Engl J Med. 2005; 352:1496–7. author reply -7. [PubMed: 15814892]
31. van Riggelen J, Yetil A, Felsher DW. MYC as a regulator of ribosome biogenesis and protein synthesis. Nat Rev Cancer. 10:301–9. [PubMed: 20332779]
32. Mueller MM, Fusenig NE. Friends or foes - bipolar effects of the tumour stroma in cancer. Nat Rev Cancer. 2004; 4:839–49. [PubMed: 15516957]
33. Einhorn J, Wersall J. Incidence of oral carcinoma in patients with leukoplakia of the oral mucosa. Cancer. 1967; 20:2184–93. [PubMed: 6073895]
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. [PubMed: 18718395]
35. Dyrskjot L, Zieger K, Real FX, et al. Gene expression signatures predict outcome in non-muscle-invasive bladder carcinoma: a multicenter validation study. Clin Cancer Res. 2007; 13:3545–51. [PubMed: 17575217]
36. Chen X, Wang L, Ishwaran H. An integrative pathway-based clinical genomic model for cancer survival prediction. Statistics and Probability Letters. 2010; 80:1313–9.
37. Bhutani M, Pathak AK, Fan YH, et al. Oral epithelium as a surrogate tissue for assessing smoking-induced molecular alterations in the lungs. Cancer Prev Res (Phila Pa). 2008; 1:39–44.
38. Sidransky D. The oral cavity as a molecular mirror of lung carcinogenesis. Cancer Prev Res (Phila Pa). 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. [PubMed: 15454246]
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. [PubMed: 18292776]
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, 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
43. Wagenblast J, Baghi M, Arnoldner C, et al. Cetuximab enhances the efficacy of bortezomib in squamous cell carcinoma cell lines. J Cancer Res Clin Oncol. 2009; 135:387–93. [PubMed: 18830627]
44. Piva R, Ruggeri B, Williams M, et al. CEP-18770: A novel, orally active proteasome inhibitor with a tumor-selective pharmacologic profile competitive with bortezomib. Blood. 2008; 111:2765–75. [PubMed: 18057228]
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. [PubMed: 16424063]
46. Tsao AS, Liu D, Martin J, 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. [PubMed: 17140695]
48. Nosho K, Shima K, Irahara N, et al. DNMT3B expression might contribute to CpG island methylator phenotype in colorectal cancer. Clin Cancer Res. 2009; 15:3663–71. [PubMed: 19470733]
49. Wang J, Bhutani M, Pathak AK, et al. Delta DNMT3B variants regulate DNA methylation in a promoter-specific manner. Cancer Res. 2007; 67:10647–52. [PubMed: 18006804]

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. [PubMed: 18455294]

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. [PubMed: 14734483]

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 U S A. 2007; 104:15805–10. [PubMed: 17890317]

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. [PubMed: 18780894]

54. Varambally S, Cao Q, Mani RS, et al. Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer. Science. 2008; 322:1695–9. [PubMed: 19008416]

55. Strillacci A, Griffoni C, Sansone P, et al. MiR-101 downregulation is involved in cyclooxygenase-2 overexpression in human colon cancer cells. Exp Cell Res. 2009; 315:1439–47. [PubMed: 19133256]

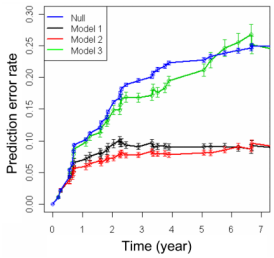

Figure 1.
Model comparison using prediction error curves. Null model used random number data as predictors. Model 1 used only microarray data as predcitors. Model 2 used microarray data as well as age, histology, podoplanin and deltaNp63 expression as predictors. Model 3 used only age, histology, podoplanin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

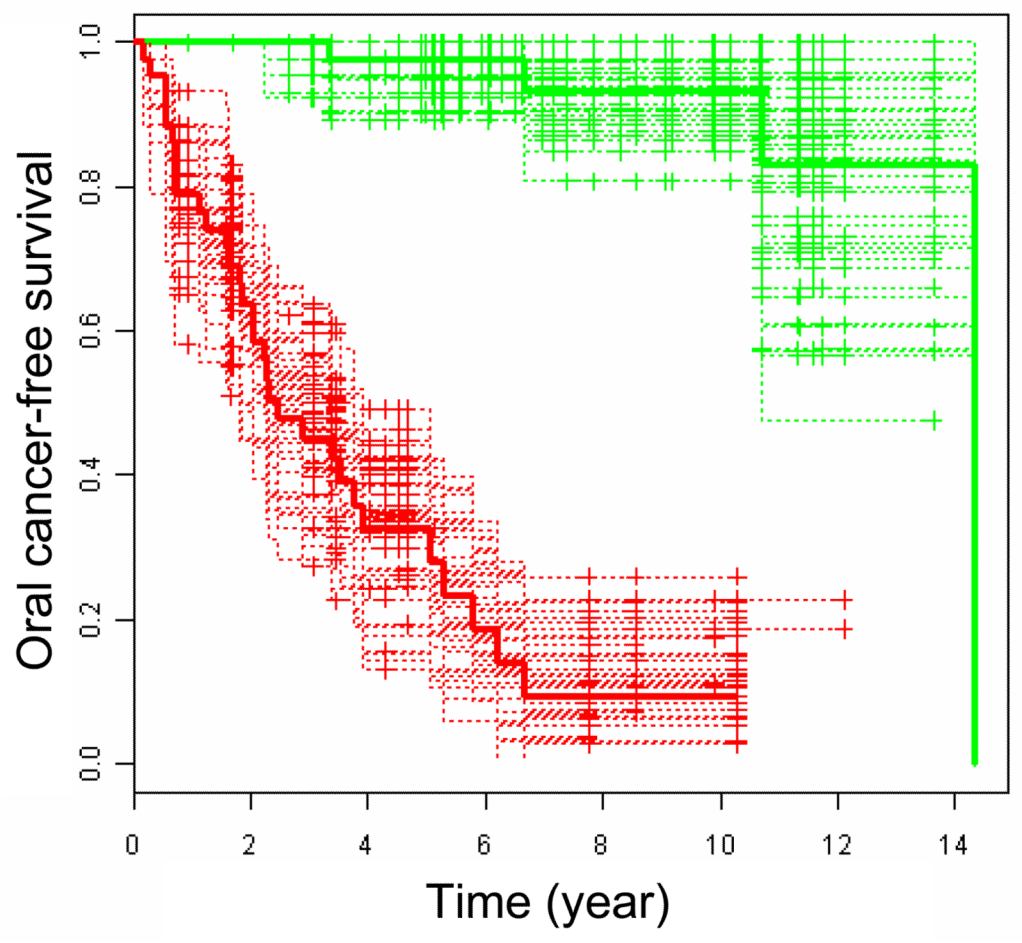

Figure 2.
Oral cancer-free survival dicotomized by oral cancer index. The red solid curve showed the patients with above median cancer risk index (median=−0.42) while the green solid curve showed the patients with below median oral cancer index. The oral cancer risk indices were computed as the hazard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

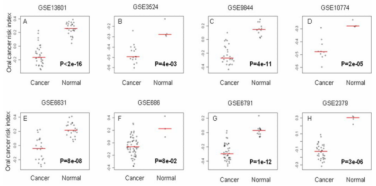

Figure 3.

Oral cancer risk index in head and neck cancers. Samples under comparison are head and neck squamous cell carcinoma (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 hazard ratio > 1 minus the average level of expression of the transcripts that have a hazard ratio < 1. Panels A–C compared HNSCC and normal mucosa from the oral cavity. Panels E–G compared HNSCC and normal mucosa from various anatomic locations. Panel H compared hypopharynx squamous cell carcinoma from normal hypopharynx mucosa. The microarray datasets were downloaded from Gene Expression Omnibus (GEO). The GEO asscession numbers associated with the datasets were shown at the top of each panel.

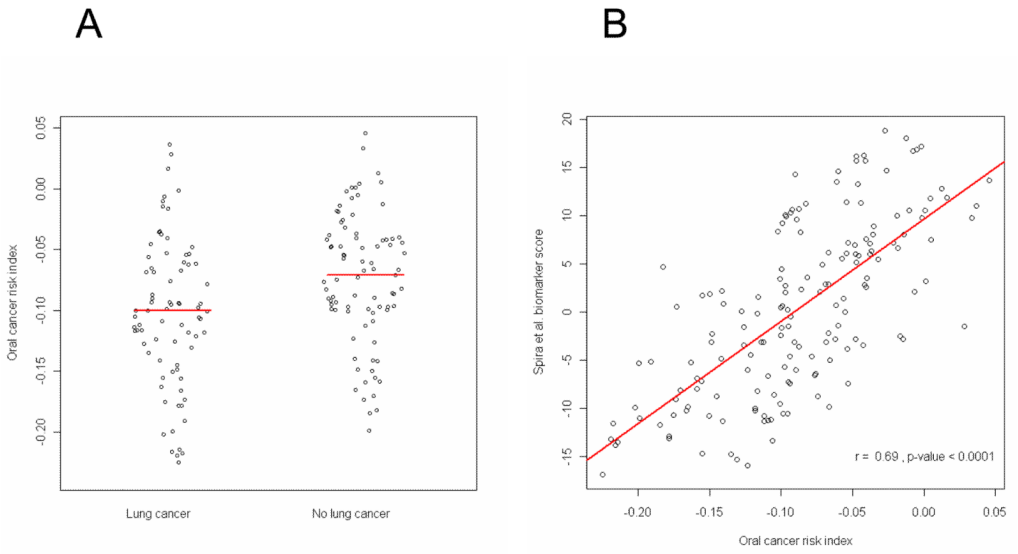

Figure 4.

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

Table 1

Characteristics of the 86 patients included in the gene expression study, and the whole population of the trial that included 162 patients (BC: β-carotene; RP: retinyl palmitate; 13cRA: 13-cis-retinoic acid; NA: not available)

Variable	Whole population	Patients included	Patients not	*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(years)	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)	

Variable	Whole population	Patients included	Patients not	*P* -value
Podoplanin				0.0009
Low	94 (63)	41 (51)	53 (77)	
High	56 (37)	40 (49)	16 (23)	

Table 2

Models generated by the CoxBoost approach; model 1 includes microarray data only (29 transcripts), whereas model 2 includes microarray data (23 transcripts) as well as age, histology at baseline, deltaNp63, and podoplanin expression at baseline (clinical and pathological covariates were mandatory); *P*-values and hazard ratio (HR) are from the single-variate 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.

Model 1 includes 29 transcripts					
Probeset ID	**Gene symbol[§]**	**CoxBoost coefficient**	**Frequency**	**Cox p-value**	**Cox hazard ratio**
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

Model 1 includes 29 transcripts					
Probeset ID	**Gene symbol[§]**	**CoxBoost coefficient**	**Frequency**	**Cox p-value**	**Cox hazard ratio**
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 includes 23 transcripts and age, histology at baseline, podoplanin and deltaNp63 expression at baseline					
Probeset_ID	**Gene symbol[§]**	**CoxBoost coefficient**	**Frequency**	**Cox p-value**	**Cox hazard ratio**
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

Cancer Prev Res (Phila). Author manuscript; available in PMC 2012 February 1.

Model 2 includes 23 transcripts and age, histology at baseline, podoplanin and deltaNp63 expression at baseline

Probeset_ID	Gene symbol[§]	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

Clinical and Pathological covariates	CoxBoost coefficient	Frequency	Cox p-value	Cox hazard ratio
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