Journal of
Cancer Science & Therapy

Lu et al., J Cancer Sci Ther 2011, S5
DOI: 10.4172/1948-5956.S5-001

Commentary Open Access

Distinct Gene Profiles for Tumor and Non-Tumor Tissue in the Head and Neck: An Analytical Approach

Mei Lu[1]*, Josena K Stephen[2], Kang Mei Chen[2], Shaleta Havard[2] and Maria J. Worsham[2]

[1]*Department of Public Health Sciences, Henry Ford Health System, Detroit, MI 48202, USA*
[2]*Department of Otolaryngology/Head and Neck Surgery, Henry Ford Health System, Detroit, MI 48202, USA*

Abstract

In a study of genetic alterations, the Multiplex Ligation-dependent Probe Amplification (MLPA) assay was used to measure gain or loss of 113 gene-probes in tumor and non-tumor tissue samples collected from each of the 220 patients with squamous head and neck cancer (HNSCC). Conditional and marginal models were available; both models account for correlated data but have different aspects. The conditional logistic regression model was proposed to estimate the subject-specific risk of tumor based on the paired tumor and non-tumor data collection, which was in contrast with the marginal model to estimate population-average risk.

The modeling process included rigorous variable selection, an initial multivariable model, a final model selection, and model validation. Genes with individual effect ($p<0.01$) were considered as candidates for the initial multivariable model for tumor. The final model included gene-probes with $p<0.01$ and estimations of odds ratios (OR) 95% Confidence Intervals (CIs) and the model's predictive ability, measured by the receiver operating characteristic curve (ROC). A 10-fold cross-validation was performed to validate the model. Of 113 gene-probes, using the conditional approach, 16 genes in 7 chromosomes, remained in the final multivariable model with $p<0.01$ and an ROC score of 0.94. The cross-validation showed ROC mean (SD) score of 0.96(0.04). The marginal model, in contrast, ended with 8 gene-probes and had an observed ROC of 0.81.

Conclusion: The conditional approach appears to be the model of choice when assessing gene-probe risks of subjects with paired data collection and fewer missing covariates, compared to the marginal approach. This multiple gene model demonstrated excellent ability to discriminate tumor from non-tumor, and supports its contribution to the pathogenesis of HNSCC as well as their potential utility for further markers of early tumor detection.

Abbreviations: AA: African American; CA: Caucasian American; CIS: Carcinoma *in situ*; *CTNNB1*: Catenin Beta-1; FGFs: Fibroblast growth factors; HNSCC: Squamous Head and Neck Cancer; MLPA: Multiplex Ligation-dependent Probe Amplification assay; *PRKDC*: Protein Kinase, DNA-activated, Catalytic Polypeptide; PTPs: Prenylated Protein Tyrosine Phosphatases; ROC: Receiver Operating Characteristic; *STCH*: Stress 70 Protein Chaperone; TFF1: Trefoil Factor-1; TRAIL: Tumor Necrosis Factor-related Apoptosis-inducing Ligand

Introduction

The overwhelming majority of mucosal head and neck cancers are squamous cell carcinomas (HNSCC), affecting more than 500,000 people worldwide each year, accounting for 5% of all malignancies [1]. In the United States, approximately 52,140 new cases are expected in 2011 with an estimated 11,460 deaths for HNC of the oral cavity, pharynx, and larynx [2]. According to the SEER data, between 1995 and 2001, the five-year relative survival rate for patients with localized disease (with no evidence of spread) was 82%; patients diagnosed with regional disease (with spread to nearby lymph nodes and other organs) had a higher five-year survival rate (51%) compared to those who had distant disease (with spread to distant organs and lymph nodes) (27.6%).

Cancer is the result of transformation from a normal to a malignant cell that results from accumulated mutations. Knowledge of the genetic mechanisms that drive cancer growth and development are important in understanding the pathogenesis of malignancy and provide insights into the tumorigenesis process. Acquisition of a fully malignant phenotype in colon cancer is thought to occur from alterations of growth-promoting oncogenes and growth-inhibiting cancer suppressor genes in a step-wise manner. In HNSCC, the evolution in transformation from a normal squamous epithelial cell to a cancer cell is likewise assumed to require several steps, some defined by genetic alterations. The underlying hypothesis is that a malignant phenotype is characterized by specific genetic alternations.

Genetic alterations provide means of identifying tumor cells as well as defining changes that presumably determine biological differences from their normal counterparts. Chromosome aberrations have served as landmarks to identify cancer genes in many tumor types, however, individual gene loci altered in tumors cannot be deduced solely from the type of chromosome rearrangement [3]. Historically, the molecular pathogenesis of cancer has been teased out one gene at a time. Recent high-throughput genome-wide candidate strategies such as the Multiplex Ligation-dependent Probe Amplification (MLPA) assay [4] to identify specific genes for gain and loss concurred with

***Corresponding author:** Mei Lu, PhD, Department of Public Health Sciences, 1 Ford Place, 3E, Detroit, MI 48202, USA, Tel: 313-874-6413; Fax: 313-874-6730; E-mail: mlu1@hfhs.org

Received October 21, 2011; **Accepted** December 05, 2011; **Published** December 07, 2011

Citation: Lu M, Stephen JK, Chen KM, Havard S, Worsham MJ (2011) Distinct Gene Profiles for Tumor and Non-Tumor Tissue in the Head and Neck: An Analytical Approach. J Cancer Sci Ther S5:001. doi:10.4172/1948-5956.S5-001

chromosomal aberrations, and provide a novel index to estimate the extent of genomic abnormality with disease progression [3].

MLPA is a cost effective approach to detect gene alterations of loss and gain in a single tube, providing tumor profiles based on multiple genes, with relevance in understanding the molecular pathogenesis of HNSCC. Because MLPA requires only small amounts of DNA, it is ideally suited for DNA from formalin fixed paraffin embedded material. Currently, MLPA has been used for validation of array-based comparative genomic hybridization (array-CGH) and SNP arrays [5-8]. More recently, several modifications of the original technique have been implemented. MLPA has a potential major role in the analysis of common copy number variation in genome-wide association analyses [6].

In this study, we hypothesized that MLPA genetic profiles characterized tumor from non-tumor in HNSCC patients. Both tumor and non-tumor tissues were available from each patient for gene-probe measurements using MLPA. A multivariable model is often used when multiple genes are involved in prediction(of tumor in this study). This approach estimates the joint effect of multiple gene-probes on tumor. Available approaches included the conditional logistic regression (CLR) and marginal logistic regression (MLR) models [9]. The marginal model (or so called population-average) is often contrasted with the conditional model, which is a subject-specific model [10]. We proposed the conditional model to study genetic profile differences between tumor and non-tumor for individual subjects, and describe the similarity and differences between marginal and conditional models.

Materials and Methods

Patients

This is a prospective study derived from an ongoing cohort of patients with primary HNSCC. Patients (n=1000) were included in the study cohort if they were 18 years and older and had a biopsy-proven primary HNSCC diagnosis at Henry Ford Health System from 1986–2006. The use of formalin-fixed paraffin embedded tissue blocks from patients with both tumor and non-tumor records within the same biopsy and the collection of related patient information was approved by the Henry Ford Health System Institutional Review Board (IRB) Committee.

Each histopathology tissue record underwent the process of subsequent lesion microdissection, and DNA extraction for MLPA genetic profiles. A subgroup of patients (n=220) with both tumor and non-tumor tissue records were included in this study.

Histopathological evaluation of tumor versus non-tumor

Pathology review of paraffin embedded tissue sections captured all types of lesions in a biopsy to include normal squamous epithelium, benign (mild, moderate dysplasia), preneoplastic lesions (severe dysplasia/carcinoma *in situ*) and tumor. Only 3 severe dysplasia and 1 carcinoma *in situ* lesions were identified and classified as malignant outcomes.

Processing lesion specimens for molecular analysis

DNA was obtained from either whole 5 micron tissue sections/blocks or from microdissected tissues sections as previously described [11]. The microdissected lesions were mounted on glass slides using a single-use-disposable scalpel blade under a dissecting microscope. This procedure minimizes mixing of normal and tumor subpopulations, and yields lesion and tumor samples estimated to be at least 90% free from contamination with normal cells [11,12].

Multiplex ligation-dependent probe amplification (MLPA) processing

MLPA is a high throughput assay allowing simultaneous interrogation of 41 genes using minute amounts (20 ng) of DNA. Validated using real-time PCR, it is ideally suited for DNA from formalin-fixed paraffin embedded tissues [3,13-15]. Gene gain and loss by MLPA concur with chromosomal aberrations, and provide a novel index to estimate the extent of genomic abnormality with disease progression [11]. Three gene-probe panels (www.mlpa.com) comprising 113 unique genes were examined. The panel primarily detects oncogenes and tumor suppressor genes located at chromosomal segments, which have been implicated in cancer, including HNSCC, and distributed throughout the genome [3,13-15].

For each gene (probe), the area under the peak is expressed as a percent of the total surface area of all peaks of a sample in an assay run. Relative copy number for each probe is obtained as a ratio of the normalized value for each locus (peak) of the sample to that of the normal control, and in general copy numbers in the range of 0.75 to 1.3 is regarded as normal, <0.75 as loss and >1.3 as gain [16], adjusting for gender because of chromosome differences.

Data collection

Demographic information (e.g., age, gender and race) for 220 patients were collected at time of HNSCC diagnosis. A total of 1076 tissue samples were derived from 932 tissue blocks, which had pathological classification of tumor and non-tumor. Both tumor and non-tumor tissue samples were collected from the same subject, and the number of samples varied in a range of 1 to 7 per subject and per tissue type (tumor or non-tumor). For the majority of patients (86%) tumor and non-tumor was collected from separated tissue blocks (37% blocks has tumor tissue alone, and 39% of blocks had non-tumor tissue alone). Only 14% of tissue blocks had both tumor and non-tumor tissues obtained from the same block. Tumor and non-tumor were outcomes of interest and they were correlated or clustered within the subject.

A total of 113 gene-probes, with known to be associated in HNSCC and others cancers, were interrogated in the paraffin tissue DNA using MLPA. The underlying hypothesis was that a set of MLPA gene-probes could discriminate tumor from non-tumor.

Statistical methods

Two analytical approaches, conditional logistic regression (CLR) and marginal logistic regression (MLR) models [9] were considered to address the correlated or cluster data.

Conditional model versus marginal model

We consider observation (Y_i, X_i), for $i = 1,2, \cdots N$ (the number of subjects), where, vector $Y_i' = (y_{ij})$ for $j = 1,2, \cdots J_i$ represents tissue record J status with response of value 1 (tumor), or 0 (non-tumor) at patient i and $X_i' = (x_{ijk})$ represents k^{th} covariate (e.g., Gene-probe) for $k = 1,2, \cdots K$ some integer K. Notices that the "paired" of tumor versus non-tumor responses were collected with unequal numbers of tissue records in each subject (cluster) and they are correlated.

Suppose η_{ij} is the link function of y_{ij} and for binary response, the logit link is considered and can be expressed as

$$\eta_{ij} = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = E\left(\log\left(\frac{pr(y_{ij}=1)}{pr(y_{ij}=0)}\right)\right) \quad (1)$$

In practice, two types of statistical models are widely used to model binary data while accounting for correlation of the binary measurements into the analysis. The first one is conditional/ subject-specific model and the second is the marginal/ population-averaged model.

Mathematically, the conditional model has a form as described

$$\eta_{ij} = \upsilon_i + \beta_0 + \sum_{k=1}^{K} x_{ijk}\,\beta_k \quad (2)$$

where, υ_i is a random effect and normal distributed as ~ $N(0,\lambda^2)$ for patient i and β_k for k = 1,2, ··· K are coefficient parameters of the covariates, and the marginal model can be expressed as

$$\eta_{ij} = \beta_0 + \sum_{k=1}^{K} x_{ijk}\,\beta_k \quad (3)$$

The conditional/random-effect model for binary outcome in Eq (2) can be implemented in SAS using PROC LOGISTIC with STRAUM as the cluster (subject ID) and the maximum likelihood estimation [17].

The random effect only uses data from the individual with discordant responses and the covariates. Thus, the concordant responses of the individual contribute no information to the likelihood. Since conditional likelihood is unaffected by the sampling scheme (e.g., retrospective versus prospective sampling) [18], the method can be used in a matched case-control study.

Use of the random-effect approach in modeling the tumor versus-non tumor data is preferred given the prospective nested cohort data collection. The random-effect model assumes that the logit varies from one individual to the next by υ_i. This assumption is reasonable because each individual has its own unique genetic profile. Thus, this variability reflects natural heterogeneity due to unmeasured genetic factors among individuals. This heterogeneity is represented by υ_i for i = 1,2, ··· N assuming that υ_i is normally distributed with mean zero and variance of λ^2.

An advantage of the random-effect models (conditional regression models) is that they allow conditional inferences in addition to marginal inferences [19]. With random effect model in Eq(2), we can obtain not only a conditional estimation of the odds, but also the marginal estimate of the odds, see Eq(3). The conditional model takes the variation or heterogeneities within individual subject into account, while the marginal model considers a population average assuming the estimation of odds obtained by integrating out individuals' heterogeneities. Because individuals' deviations have been eliminated in Eq(3) by integration, the estimation of the odds ratio does not involve any individual. The two models are equivalent if and only if the individuals in the study can be regarded as a random sample from a population (behaviors the same), which may not be practical.

For this "paired" tumor versus non-tumor study, we preferred conditional/ random-effect model, given the above discussion. However, the marginal approach is still considered to be a valid model to handle the "paired" data [20,21]. The marginal model has an advantage of accommodating missing covariates or response variables, while the conditional model, in contrast, cannot accommodate any missing observations.

Modeling

The modeling process included rigorous variable selection, an initial multivariable model, a final model selection and model validation. Given the large number of gene-probe covariates involved, the variable selection for inclusion in the initial multivariable model is a necessary step to avoid the model over-fitting [22]. The univariate analysis approach was used for the variable selection. Genes with individual risks ($p<0.01$ based on the univariate analysis) were then considered as candidates for the initial multivariable model. The cut-off $p<0.01$ was considered based on the number of gene-probes (m=113) and number of subjects (n=220). Neither p-value as 0.2 (testing the risks of clinical factors) nor 10^{-7} (exploring risks of micro-arrays to simultaneously measure the mRNA abundance of thousands of genes) was considered. Prior to multivariable modeling, genes were evaluated for their correlation and missing values. Highly correlated genes (correlation coefficient, r, >0.7) or genes with larger missing values (>5%) were fitted separately along with other uncorrelated ($r<0.7$) genes.

The stepwise model selection approach was considered. The final model included genes with $p<0.01$ along with odds ratios for loss or gain as risk predictors and ROC (the receiver operating characteristic curve) was calculated to measure the model predictive ability [23]. A 10-fold cross-validation was performed. The cohort of 220 patients was randomly partitioned into 10 experiments. For each k, k = 1,2, ··· 10, 9 experiments except k^{th} experiment were used as training set, and the k^{th} experiment was used as test set. We retained genes from scratch (113 genes) with a training set to build a multivariable model as described above and then estimated the ROC score with the test set [22]. The true ROC estimation was obtained as the average of ROC values from 10 test sets.

Results

The 220 HNSCC patient cohort contributed both tumor and non-tumor (control) samples with a total of 1076 tissue samples. Race distribution was 50% Caucasian American (CA), 38% African American (AA) and 12% other or unknown. Thirty-four percent (34%) were males. A patient served as his/her own controls and, therefore, the clinical variables were balanced given 'paired" data collection under the conditional approach.

Of 1076 tissue samples, 513 were non-tumor tissue records (495: normal squamous epithelium, 9: benign lesions, 6: mild dysplasia, and 3: moderate dysplasia), and 563 tumor tissue records (559: tumor, 1: severe dysplasia, 3: carcinoma *in situ*). One-hundred-thirteen (113) unique gene-probes were examined for each tissue record for gene copy number outcomes of gain, normal, or loss. Gene missing values ranged from 0% to 4.2%. Using the conditional model approach, fifty-three (53) genes were identified with individual effects ($p<0.01$) and correlation among gene-probes was moderate ($r<0.50$), therefore, those genes were considered as the candidate genes for the initial multivariable model.

After the stepwise model selection process, 16 genes remained in the multivariable model with $p<0.01$, estimation of odds ratio (OR) and its 95% confidence interval (CI) shown in Tables 1, 2 and 3. The model had an excellent predictive ability with an ROC of 0.94. The 16 genes in the final model with alterations of loss and/or gain accounted for loci along 7 chromosomes: 3, 4, 6, 8, 9, 11, and 21(Tables 1,2 and 3). Of these, 50% were altered in both tumor and non-tumor, with loss or gain

Effect	Chromosome	Odds Ratio Estimate	Lower CL*	Upper CL
TFF1 Loss vs. Normal	21q22.3	3.019	1.514	6.02
TFF1 Gain vs. Normal		0.08	0.024	0.268
PRKDC Loss vs. Normal	8q11	0.276	0.11	0.692
PRKDC Gain vs. Normal		5.449	2.09	14.206
MYC Loss vs. Normal	8q24.12	0.221	0.097	0.503
MYC Gain vs. Normal		2.218	1.136	4.332
LTA Loss vs. Normal	6p21.3	2.156	1.172	3.965
LTA Gain vs. Normal		0.266	0.108	0.655
IL2 Loss vs. Normal	4q26	3.697	1.774	7.705
IL2 Gain vs. Normal		0.149	0.055	0.407
FGFR1 Loss vs. Normal	8p21	0.275	0.126	0.598
FGFR1 Gain vs. Normal		5.555	1.689	18.267
CTNNB1 Loss vs. Normal	3p22	2.682	1.394	5.162
CTNNB1 Gain vs. Normal		0.323	0.147	0.71
CDKN2A Loss vs. Normal	9p21	1.845	1.013	3.362
CDKN2A Gain vs. Normal		0.14	0.056	0.35

*CL: Confidence Limit

Table 1: Genes with corresponding loss and gain that predict tumor and non-tumor.

Effect	Chromosome	Odds Ratio Estimate	Lower CL*	Upper CL
PTP4A3 Loss vs. Normal	8q24.3	0.493	0.214	1.135
***PTP4A3* Gain vs. Normal**		**12.158**	**3.461**	**42.71**
***LMO2* Loss vs. Normal**	11p13	**4.977**	**2.16**	**11.466**
LMO2 Gain vs. Normal		0.573	0.205	1.607
FGF3 Loss vs. Normal	11q13	0.882	0.447	1.741
***FGF3* Gain vs. Normal**		**7.819**	**3.286**	**18.604**
***CDKN2B* Loss vs. Normal**	9p21	**3.256**	**1.676**	**6.325**
CDKN2B Gain vs. Normal		1.168	0.442	3.087
BCL6 Loss vs. Normal	3q27	0.55	0.27	1.12
***BCL6* Gain vs. Normal**		**8.989**	**3.155**	**25.612**

*CL: Confidence Limit

Table 2: Genes with loss or gain that predict tumor (highlighted in bold).

Effect	Chromosome	Odds Ratio Estimate	Lower CL*	Upper CL
STCH Loss vs. Normal	21q11.1	1.788	0.833	3.839
***STCH* Gain vs. Normal**		**0.124**	**0.043**	**0.359**
***CCND1* Loss vs. Normal**	**11q13**	**0.403**	**0.22**	**0.736**
CCND1 Gain vs. Normal		1.239	0.634	2.421
***BAK1* Loss vs. Normal**	**6p21.3**	**0.262**	**0.103**	**0.666**
BAK1 Gain vs. Normal		0.438	0.192	0.999

*CL: Confidence Limit

Table 3: Genes with loss or gain that predict non-tumor (highlighted in bold).

reflective of chromosomal aneuploidy. This copy number instability favored loss of *CDKN2A* (9p21), *CTNNB1* (3p21), *IL2* (4q26), *LTA* (6p21.3), and *TFF1* (21q22.3) in tumor, with corresponding gain in non-tumor lesions, and gain of *FGFR1* (8p21), *c-MYC* (8q24.12), and *PRKDC* (8q11) in tumor, with corresponding loss in non-tumor (Figure 1,Table 1). Loss of *CDKN2B* (9p21) and *LMO2* (11p13), and gain of *BCL6* (3q27), *FGF3* (11q13), and *PTP4A3* (8q24.3) predicted tumor (Figure 2,Table 2). Loss of *BAK1* (6p21.3) and *CCND1* (11q13), and gain of *STCH* (21q11.1) predicted non-tumor (Figure 3,Table 3). In addition, the model remained unchanged after excluding the 9 benign lesions (6- mild dysplasia, 3-moderate dysplasia) from the non-tumor group. The cross-validation showed mean (SD) ROC as 0.96 (0.04). Model validation based on 53 candidate genes showed an ROC score of 0.94 (0.07).

Including the above mentioned 16 gene-probes into a marginal model, 9 gene-probes were significant ($p<0.01$) after adjusting for other variables. The ROC for the MLR 16 gene model was 0.84. Univariate analysis, followed by multivariable modeling resulted in forty-three (43) candidate genes with individual effects ($p<0.01$), and 8 genes, *AR*, *BCL6* (3q27), *CDKN2B* (9p21), *FGF3* (11q13), *IL2* (4q26), *c-MYC*

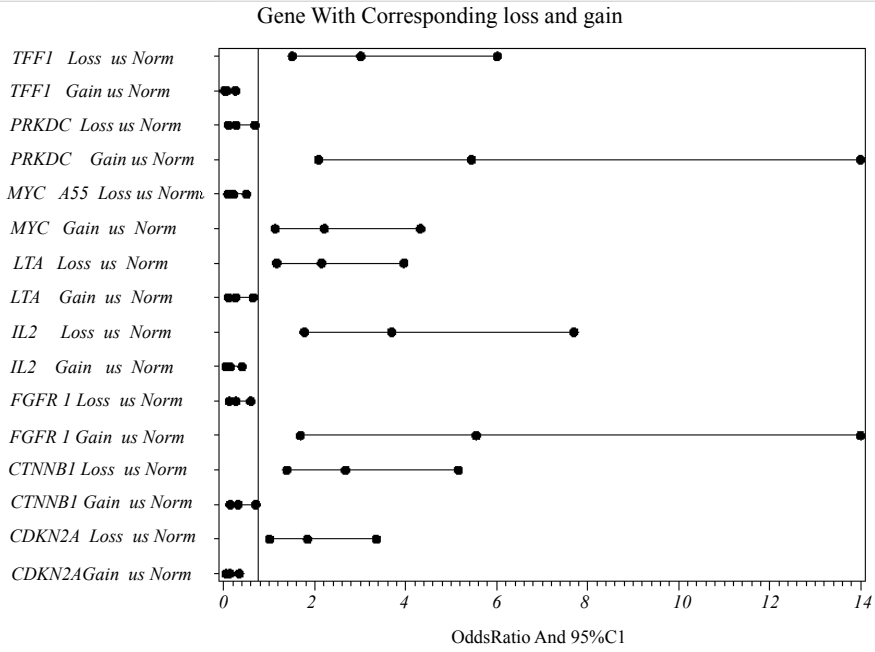

Figure 1: Gene alterations of corresponding loss and gain for the 8 genes in the final model that predict tumor (odds ratio > 1) and non-tumor (odds ratio < 1) (Table 1**)** and 95% CI (Confidence Interval). Norm: normal copy number.

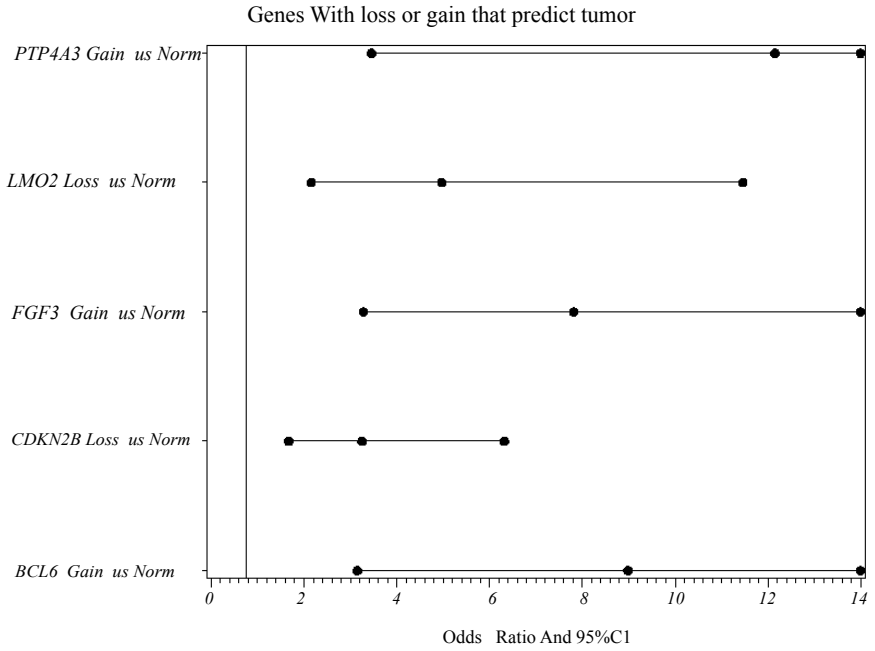

Figure 2: Gene alterations (Table 2) that predict tumor with odds ratio >1 and 95% CI (Confidence Interval). Norm: normal copy number.

(8q24.12), *PTP4A3* (8q24.3), and *NTF,* that remained in the final multivariable model with p<0.01 respectively with an ROC of 0.81. Of those 8 gene-probes in the final MLR model, 6 probes, except *AR* and *NTF*, were also retained in the final CLR model.

Discussion

Our study illustrates that gene copy number alterations can distinguish malignant tumor from the non-tumor tissue samples (pathologically defined normal tissues). Excluding the benign lesions from non-tumor, the model remained the same. A limitation of this study is that it was conducted on tumor and non-tumor tissue from a cancer cohort. Regardless, multivariate models clearly detected significant differences between tumor and non-tumor tissue samples, keeping in mind that "normal" tissues in a malignant environment may be contaminated. The availability of 47% of tumor samples and 39% non-tumor samples from separate tissue blocks, with only 14% of tumor and non-tumor samples from the same tissue block was an important factor in minimizing this contamination. The resultant gene-

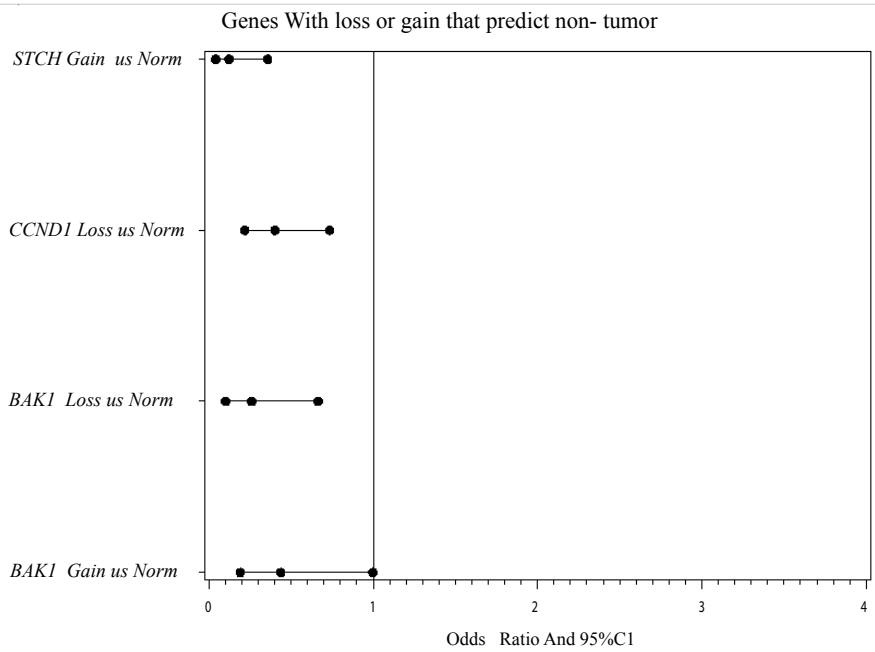

Figure 3: Gene alterations (Table 3) that predict non-tumor with odds ratio <1 and 95% CI (Confidence Interval). Norm: normal copy number.

based model is therefore more representative of the clinical setting with potential to determine tumor-specific risk profiles in biopsies of patients without a clinical diagnosis of HNSCC.

The multiple 16 gene-probe CLR model demonstrated excellent predictive ability to discriminate tumor from non-tumor (ROC=0.94 based on observed data and 0.96 based on 10-fold validation), and is a significant improvement when compared to a single gene model such as *BAK1*, *CCD1* or *STCH* (with observed ROC as 0.67, 0.68, or 0.68, respectively). The multiple gene-probes model undoubtedly requires validation in an independent cohort with similar patient characteristics.

When compared to the marginal model, the conditional model is the preferred analytical approach to deal with the paired case-control data. Conditional modeling takes individual variation into account, while marginal modeling addresses averages. The model accuracy and predictive ability were excellent using the conditional approach (ROC=0.94), compared to the marginal approach (ROC=0.84 based on 16 gene-probes or 0.81 based on the final 8 gene-probe model). The predictive ability was improved 12% to 16% by using the CLR model. For the "paired" case-control data collected for each subject, the subject served as his/her own control using the conditional approach. This implies that the known clinical risk factors (e.g., the tumor site, race, age, gender, and tumor stage) or unknown factors would be balanced within each subject, and therefore, suggests that the "paired" case-control data collected at each subject level has utility to identify solely a gene model. The marginal approach is still considered a valid analytical approach although it is less powered, compared to the conditional approach. Also, the marginal approach is significantly different (See Eq 2 and Eq 3) and can be used to address scientific question (e.g., the average risk or effect).

Nevertheless, the conditional approach may be less sufficient or even misleading when cases and controls are not paired or the data has a large amount of missing observations (>5%). The random effect only uses data from the individual with discordant responses and the covariates. Thus, the concordant responses of the individual contribute no information to the likelihood. The SAS Proc LOGISTIC with statement STRATUM (subject ID) will eliminate any unpaired rerecords so that the model would be generated based on subpopulation rather than the whole population. In contrast to CRL, marginal approach (MRL) implemented using Proc SURVEYLOGISTIC, with statement STRATUM/CLUSTER (subject ID), accounts for correlations between the pairs throughout the maximum likelihood estimation. It models marginal distribution of the tumor and treats the correlates data as though it were unpaired, and would count for all records. Therefore, the analysis is analogous to an unmatched analysis.

Generalized Estimation Equations (GEE) [24,25] using Proc GENMOD in SAS, models the marginal model and accounts for correlation throughout the Quasi-likelihood estimation. The latter is less restrictive on distribution function with more robust results [24,25]. This has been used in many studies including the study of paired discordant responses, as well as combination concordances and discordances [20,21,26]. However, concerns were noted when correlations within a cluster were negative [27]. This was the case for our data with discordant pairs. Therefore, if a study is interested in the average risks within a population and if the clustered data has both concordant and discordant responses, the marginal model can be considered. GEE is a cautionary approach when discordances dominate or negative corrections are observed [27].

The model's discriminatory abilities (c-index/ROC of 0.93) support molecular distinctiveness of malignant versus non-malignant tissue with significant predictive power. Genetic alterations at 16 chromosomal loci underscore the association of already known genes as well as newer gene targets in HNSCC pathogenesis. The sixteen gene predictors spanning loci along 7 chromosomes cover an array of essential functions that ensure normal homeostasis to include DNA repair (*PRKDC*), initiation of carcinogenesis (*TFF1*), immune surveillance (*IL2*, *LTA*), cell cycle regulators (*CDKN2A*, *CDKN2B*), apoptosis (*BAK1*, *STCH*), regulation of cell proliferation

and differentiation (*CCND1*, *FGF3*, *MYC*), transcription factors (*BCL6*), stem cell hematopoesis (*LMO2*), adhesion, invasion and metastasis (*CTNNB1*, *FGFR1*), and acquisition of metastatic potential of tumor cells (*PTP4A3*), implicating these genes as key players in the tumorigenesis continuum.

Our data support distinct genetic profiles for tumor and non-tumor. One of the hallmarks of malignant transformation is genomic instability, which promotes a wide range of mutations, including chromosome deletions, gene amplifications, translocations and polyploidy [28]. In this study, genomic instability evident from the directional loss and gain for several genes, underscored the contribution of aneuploidy in early HNSCC tumorigenesis. Analogous to studies, where prognostic information is augmented when gene outcomes of amplifications and deletions (e.g,, *CCND1* and/or *MYC* amplification, in combined with *CDKN2A* deletion) as compared to analysis of single genetic aberrations [29]. The 16 gene alteration compendium in this study likely reflects finely choreographed genomic instability events to achieve biological distinctiveness.

Competing Interests

The authors have no potential conflicts of interest.

Acknowledgements

This research was supported by NIH R01 DE 15990 (MJW).

References

1. Brockstein B, Haraf DJ, Rademaker AW, Kies MS, Stenson KM, et al. (2004) Patterns of failure, prognostic factors and survival in locoregionally advanced head and neck cancer treated with concomitant chemoradiotherapy: a 9-year, 337-patient, multi-institutional experience. Ann Oncol 15:1179-1186.
2. Cancer Facts & Figures (2011) American Cancer Society.
3. Worsham MJ, Pals G, Schouten JP, Van Spaendonk RM, Concus A, et al. (2003) Delineating genetic pathways of disease progression in head and neck squamous cell carcinoma. Arch Otolaryngol Head Neck Surg 129:702-708.
4. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, et al. (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res 30: e57.
5. Eijk-Van Os PG, Schouten JP (2011) Multiplex Ligation-dependent Probe Amplification (MLPA®) for the detection of copy number variation in genomic sequences. Methods Mol Biol 688: 97-126.
6. Kozlowski P, Jasinska AJ, Kwiatkowski DJ (2008) New applications and developments in the use of multiplex ligation-dependent probe amplification. Electrophoresis 29: 4627-4636.
7. Sørensen KM, Andersen PS, Larsen LA, Schwartz M, Schouten JP, et al. (2008) Multiplex ligation-dependent probe amplification technique for copy number analysis on small amounts of DNA material. Anal Chem 80: 9363-9368.
8. van Eijk R, Eilers PH, Natté R, Cleton-Jansen AM, Morreau H, et al. (2010) MLPAinter for MLPA interpretation: an integrated approach for the analysis, visualisation and data management of Multiplex Ligation-dependent Probe Amplification. BMC Bioinformatics 11: 67.
9. Nelder JA, Wedderburn RW (1972) Generalized linear models. J R Statist Soc A 135: 370-384.
10. Lee Y, Nelder JA (2004) Conditional and Marginal Models: Another View. Statist Sci 19: 219-228.
11. Raju U, Mei L, Seema S, Hina Q, Wolman SR, et al. (2006) Molecular classification of breast carcinoma in situ. Curr Genomics 7: 523-532.
12. Stephen JK, Vaught LE, Chen KM, Shah V, Schweitzer VG, et al. (2007) An epigenetically derived monoclonal origin for recurrent respiratory papillomatosis. Arch Otolaryngol Head Neck Surg 133: 684-692.
13. Worsham MJ, Chen KM, Tiwari N, Pals G, Schouten JP, et al. (2006) Fine-mapping loss of gene architecture at the CDKN2B (p15INK4b), CDKN2A (p14ARF, p16INK4a), and MTAP genes in head and neck squamous cell carcinoma. Arch Otolaryngol Head Neck Surg 132: 409-415.
14. Worsham MJ, Pals G, Schouten JP, Miller F, Tiwari N, et al. (2006) High-resolution mapping of molecular events associated with immortalization, transformation, and progression to breast cancer in the MCF10 model. Breast Cancer Res Treat 96: 177-186.
15. Kunjoonju JP, Raitanen M, Grénman S, Tiwari N, Worsham MJ (2005) Identification of individual genes altered in squamous cell carcinoma of the vulva. Genes Chromosomes Cancer 44: 185-193.
16. Bremmer JF, Braakhuis BJ, Ruijter-Schippers HJ, Brink A, Duarte HM, et al. (2005) A noninvasive genetic screening test to detect oral preneoplastic lesions. Lab Invest 85: 1481-1488.
17. McFadden D (1974) Conditional Logit Analysis of Qualitative Choice Behaviour. Handbook of transport modeling. New York: Academic Press.
18. Neuhaus JM, Jewell NP (1990) The effect of retrospective sampling on binary regression models for clustered data. Biometrics 46: 977-990.
19. Robinson GK (1991) That BLUP is a good thing: The estimation of random effects (with discussion). Statist Sci 6: 15-51.
20. Hu FB, Goldberg J, Hedeker D, Henderson WG (1998) Modelling ordinal responses from co-twin control studies. Stat Med 17: 957-970.
21. Katz J, Zeger S, Liang KY (1994) Appropriate statistical methods to account for similarities in binary outcomes between fellow eyes. Invest Ophthalmol Vis Sci 35: 2461-2465.
22. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, (2nd Edn), Springer.
23. Hanley JA, McNeil BJ (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148: 839-843.
24. Liang K, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 72: 13-22.
25. Zeger SL, Liang K (1986) Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42: 121-130.
26. Hanley JA, Negassa A, Edwardes MD, Forrester JE (2003) Statistical analysis of correlated data using generalized estimating equations: an orientation. Am J Epidemiol 157: 364-375.
27. Hanley JA, Negassa A, Edwardes MD (2000) GEE analysis of negatively correlated binary responses: a caution. Stat Med 19: 715-722.
28. Someya M, Sakata KI, Matsumoto Y, Kamdar RP, Kai M, et al. (2011) The association of DNA-dependent protein kinase activity of peripheral blood lymphocytes with prognosis of cancer. Br J Cancer 104: 1724-1729.
29. Akervall J, Bockmühl U, Petersen I, Yang K, Carey TE, et al. (2003) The gene ratios c-MYC:cyclin-dependent kinase (CDK)N2A and CCND1:CDK2A correlate with poor prognosis in squamous cell carcinoma of the head and neck. Clin Cancer Res 9: 1750-1755.

This article was originally published in a special issue, **Cancer Research: Clinical & Experimental** handled by Editor(s). Dr. Richard D. Finkelman, AstraZeneca LP Clinical, USA; Dr. Jimmy Thomas Efird, University of North Carolina, USA; Dr. Yanming Wang, Case Western Reserve University, USA