Published in final edited form as:

Int J Cancer. 2016 January 1; 138(1): 98–109. doi:10.1002/ijc.29691.

Phylogenetic Analysis of Multiple FISH Markers in Oral Tongue Squamous Cell Carcinoma Suggests that a Diverse Distribution of Copy Number Changes Is Associated with Poor Prognosis

Darawalee Wangsa^{1,2}, Salim Akhter Chowdhury^{3,4}, Michael Ryott⁵, E. Michael Gertz⁶, Göran Elmberger⁷, Gert Auer², Elisabeth Åvall Lundqvist^{2,8}, Stefan Küffer⁹, Philipp Ströbel⁹, Alejandro A. Schäffer⁶, Russell Schwartz^{4,10}, Eva Munck-Wikland¹¹, Thomas Ried¹, and Kerstin Heselmeyer-Haddad¹

¹Genetics Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

²Department of Oncology-Pathology, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden

³Joint Carnegie Mellon/University of Pittsburgh Ph.D. Program in Computational Biology, Carnegie Mellon University, Pittsburgh, PA, USA

⁴Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA, USA

⁵Department of Otorhinolaryngology, Sophiahemmet Hospital, Stockholm, Sweden

⁶Computational Biology Branch, National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA

⁷Department of Laboratory Medicine, Pathology, Örebro University Hospital, Örebro, Sweden

⁸Department of Oncology and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden

⁹Institute of Pathology, University Medical Center Göttingen, Göttingen, Germany

¹⁰Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

¹¹Department of Oto-Rhino-Laryngology, Head and Neck Surgery, Karolinska University Hospital and Department of Clinical Science, Intervention and Technology, Karolinska Institutet, Stockholm, Sweden

Abstract

Oral tongue squamous cell carcinoma (OTSCC) is associated with poor prognosis. To improve prognostication, we analyzed four gene probes (*TERC*, *CCND1*, *EGFR*, and *TP53*) and the centromere probe CEP4 as a marker of chromosomal instability, using fluorescence *in situ* hybridization (FISH) in single cells from the tumors of sixty-five OTSCC patients (Stage I, n=15; Stage II, n=30; Stage III, n=7; Stage IV, n=13). Unsupervised hierarchical clustering of the FISH

data distinguished three clusters related to smoking status. Copy number increases of all five markers were found to be correlated to non-smoking habits, while smokers in this cohort had low-level copy number gains. Using the phylogenetic modeling software FISHtrees, we constructed models of tumor progression for each patient based on the four gene probes. Then, we derived test statistics on the models that are significant predictors of disease-free and overall survival, independent of tumor stage and smoking status in multivariate analysis. The patients whose tumors were modeled as progressing by a more diverse distribution of copy number changes across the four genes have poorer prognosis. This is consistent with the view that multiple genetic pathways need to become deregulated in order for cancer to progress.

Keywords

Oral Tongue Cancer; FISH; Genetic Markers; Phylogenetic Modeling; HPV

INTRODUCTION

Oral tongue squamous cell carcinoma (OTSCC) is associated with poor prognosis, with increasing incidence seen among young adults.¹ Known environmental risk factors for OTSCC include tobacco usage, either via smoking² and chewing³, and alcohol consumption⁴. Patients diagnosed at earlier stages (I or II) have a significantly better prognosis. Numerous studies using single-markers attempted to improve disease prognostication could not be validated.⁵

That might be due to the fact that earlier studies on tongue cancer usually combined mobile tongue cancer cases (OTSCC) with base of tongue cancer, other types of oral cancer, or (more broadly) head and neck cancers from other primary sites, despite obvious differences between the subsites. For example, the presence of human papilloma virus (HPV), e.g., is common in base of tongue cancer but not in OTSCC, and is associated with a favorable prognosis in base of tongue cancer.⁶ Therefore, focusing exclusively on the oral (mobile) tongue site is important to obtain a more homogeneous study cohort. Understanding the development and progression of OTSCC could be enhanced by more extensive examination of specific genetic alterations seen in OTSCC, especially by considering multiple markers^{7, 8} and multiple single cells simultaneously in the same tumor.⁹

In this study, we enumerated copy numbers of four genes and one centromere probe in 65 cases of OTSCC with detailed patient data and follow-up including disease-free and overall survival. Our approach of enumerating all probes within the same cells allowed us to analyze intra-tumor heterogeneity and co-occurrence of copy number changes. We analyzed the oncogenes *TERC* on 3q26, *EGFR* on 7p12, *CCND1* on 11q13, and the tumor suppressor *TP53* on 17p13. These genes were selected because they have been frequently implicated in the progression of oral cancer. Specifically, the first three have been suggested to be among the primary targets of copy number gains on chromosome 3q (*TERC*^{10, 11}), 7p (*EGFR*^{11, 12}), and 11q (*CCND1*^{11, 12}), respectively, for review.¹³ Much more detailed rationale for selecting these four genes is provided in Supporting Information A.

A central aim of this study was to evaluate whether the combination of multiple FISH markers can be used to predict prognosis in OTSCC independent of tumor stage. A novel feature of our approach is the use of phylogenetic analysis of tumor progression to infer models of cellular evolution. These models serve as the basis for clustering patients into groups with putatively differential prognoses.

Our approach utilizes multiple probe single-cell FISH data to build tree models of tumor progression using our software FISHTrees.^{14, 15} We then derive summary statistics based on all four genes from the tree models generated and we test for associations between those summary statistics and survival, while taking into account tumor stage and smoking history. In this setting, our use of phylogenetic tools is to build trees for each individual patient, and the taxa in any one tree are gene count patterns of single cells from an individual tumor. The general idea of constructing phylogenetic models from samples of *single patients*, as we do in this study, has been used before (e.g.,¹⁶). Earlier uses of phylogenetics in modeling tumor progression were to build trees or networks where the taxa were either events, such as the gain of one chromosome arm, shared by many samples^{17–19} or entire genomic profiles of patient samples.²⁰

We show that this evolutionary approach combined with test statistics derived from single-tumor phylogenetic models has more predictive power than using the static gene copy number counts for survival analysis. Investigating the relationship between smoking status and the pattern of copy number alterations, we conclude that exposure to tobacco results in different patterns of genomic changes in tongue cancer than those observed in non-smokers.^{2, 21}

METHODS

Patients and samples

Sixty-five formalin-fixed, paraffin-embedded pretreatment biopsy specimens with histopathologically confirmed OTSCC, UICC Stages I–IV, treated at the Department of Oto-Rhino-Laryngology, Karolinska University Hospital (Stockholm, Sweden) from January 2000 to December 2004 were investigated. The present study was carried out with approval from the Regional Ethics Committee in Stockholm and all patients gave informed consent.

Treatment of OTSCC was based mainly on the UICC stage classification and patient treatment response.²² Initially, tumor stage and size were verified with ultrasound guided fine needle aspiration cytology, radiology with CT/MRI, and palpation under general anesthesia. Depending on clinical and tumor characteristics, preoperative radiation therapy was given with a mean total dose of 50–68Gy.

Clinical information, including tumor grade, age, treatment modality, treatment response and follow-up were retrieved from medical records (Table 1). Each patient had a minimum follow-up time of at least 36 months after initial cancer diagnosis, unless the patient died earlier than that. Survival and disease-free survival were reported for all patients for up to 73 months (range: 3–73 months). Smoking status was not available for all participants, and is recorded as yes/no rather than in pack-years.

Haematoxylin-eosin (H&E) stained sections (4µm) were cut before and after FISH sections (6µm), to confirm tumor representativity. All sixty-five (Stage I, n=15; Stage II, n=30; Stage III, n=7; Stage IV, n=13) representative formalin fixed, paraffin embedded biopsy specimens were available for FISH.

Sixty-one samples were tested for 16 HPV types (HPV 6, 11, 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66 und 68) using a multiplex based Fluorescence PCR (F-HPV typing™, Genomed Diagnostics AG, Switzerland) and a subsequent fragment analysis using the ABI, 3500 Genetic Analyzer (Life Technologies, CA). A positive control sample (HPV 45) was carried along for each PCR run. The data were analyzed with the GeneMapper 5 Software (Life Technologies, CA).

Sample Preparation and Fluorescence *in situ* hybridization (FISH)

Five-probe FISH was performed using a centromere-specific probe for chromosome 4 (CEP4), and bacterial artificial chromosome (BAC) contigs for the following four probes: *TERC* (3q26), *CCND1* (1q13), *EGFR* (7p12), and *TP53* (17p13) (Figure 1a). Each gene-specific probe consisted of a contig of four to five overlapping BAC clones anchored around the respective gene. Each of the BAC clone contigs were labeled by nick translation as follows: *TERC* and *EGFR* with Spectrum Orange-dUTP (Abbott Molecular Inc., IL), *CCND1* and *TP53* with Rhodamine Green-dUTP (Life Technologies, CA). CEP4 was purchased with a Spectrum Aqua label (Abbott Molecular Inc., IL). Probes were subsequently hybridized to the samples in two panels (Panel 1: CEP4-*TERC*-*CCND1*, Panel 2: *EGFR*-*TP53*) and cells were relocated in the second hybridization, resulting in counts for all five FISH probes within the same nuclei. The hybridization conditions were performed according to previously described protocols.²³ FISH analyses were performed on 65 cases, with 250 interphase cells counted for each case and enumerated as described before.²⁴

DNA cytometry measurement

For quantitative DNA analyses, image cytometry was performed on representative sections with Feulgen staining. The image cytometry methodology of staining, tumor cell selection and internal standardization has been previously described (Steinbeck). The C-value is defined as the nuclear DNA content of a cell (ploidy), with 2C representing a diploid cell. DNA content was measured by the percentage of cells with greater than 2.5C exceeding rates and greater than 5C exceeding rates. Approximately 100 cells were analyzed in each specimen. The average percentage of cells for each category can be found in Table 2.

Unsupervised clustering of patients by mean signal number of five probes

The programs used to produce the heat map included Gene Cluster 3.0 (Laboratory of DNA Information Analysis, University of Tokyo) and Java Tree View 1.1.6r4. The mean signal number per marker in each sample was generated and loaded into Gene Cluster 3.0 with the following options: normalization of gene, centering the genes by the mean, organizing the genes, and clustering via complete linkage. The resulting heat map grouped cases according to the mean signal number per marker in each sample. Each marker per patient is given a color, with green signifying a relatively low mean signal number and red symbolizing high

mean signal. The cluster tree is divided into three groups according to the highest overhang (vertical line), which represents the distinct separation between the groups (see Figure 1b).

One and two-gene survival analysis

Survival tests for one gene, two genes, and ploidy were done using Kaplan-Meier (KM) analysis (survdiff function in R).^{25, 26} Smoking status and stage were not used as covariates. We tested whether gain of *CCND1*, *EGFR* or *TERC* or loss of *TP53* alone was a predictor of survival or disease-free survival. For each gene, we divided the patients into classes, using a threshold of P% or more cells showing a gain for the oncogenes and P% with a loss for the tumor suppressor gene *TP53*. Gains and losses were determined by comparing the copy number of the gene to the copy number of the centromere probe CEP4. For completeness, we tried thresholds of P% = 10%, 20%, 30%, 40%, and 50%. Unlike other studies, we did not draw a distinction between (low-level) “gains” and (high-level) “amplifications” in these tests because that would have created a substantial multiple testing problem if any particular gain vs. amplification criterion had led to significant survival results. For two-gene tests, we compared the group of patients showing above-threshold changes in both genes to all other patients, as done previously.⁸ We also did a test using the cytometry measurement by partitioning the patients into two groups around the median.

Multivariate survival analyses

We used statistics calculated from FISH copy number data and either clustering or phylogenetic trees inferred on these FISH samples to test for association with disease-free survival and overall survival time using KM analysis. We then performed multivariate survival analysis using the Cox proportional hazards (COXPH) model as implemented in R.^{25, 26} The variables used in the multivariate analysis include: tumor stage, smoking status, and various test statistics derived from the phylogenetic analysis, as explained below. The objective of the multivariate analysis is to test whether the new test statistics can predict survival or disease-free survival independent of tumor stage and smoking status.

For cluster-based subgrouping, we followed convention and assigned the cluster associated with longer (shorter) survival of patients as the lower (higher) risk group. For smoking behavior, we assigned smokers (non-smokers) to the higher (lower) risk category based on previous epidemiological results.^{2, 6, 21, 27} Because of these assignments of categories, we expected the hazard ratios (HRs) to be > 1 . Nevertheless, P-values are two-sided.

Clustering of samples by gain/loss patterns

For each cell count pattern in each sample, we estimated whether each gene is amplified or lost by comparing the copy number counts of the centromere probe with the copy number counts of that gene to identify one of three possible gain/loss conditions for each gene: No Change (N), Gain (G) and Loss (L).²⁸ Considering all four genes yields $3^4 = 81$ possible combined gain/loss conditions, each of which we call a “comparator”, for each cell. Although losses of *EGFR* and gains of *TP53* have been found in some cases of oral cancer,^{29, 30} we considered only those comparators that include {gain, no change} for oncogenes and {loss, no change} for the tumor suppressor for a total of $2^4 = 16$ comparator strings for each cell. The comparators represent the genes *TERC*, *CCND1*, *EGFR*, *TP53* in

this order from left to right. We calculated the total number of cells that match each comparator for each sample. Because there is little variation in the number of cells per sample, we counted total cells rather than fractions of cells. Each sample is then represented by an ordered list of 16 integers that are considered “features” (a term of art in machine learning meaning a quantitative measure by which one can distinguish meaningfully different groups of objects). The transformed dataset becomes a 65×16 dimensional matrix with integer-valued count entries.

We used K-means clustering (with K=2) as implemented in Matlab to partition the 65 samples into two non-overlapping subgroups. We then performed KM analysis via the survfit function in R²⁵ to compare either the survival time or the disease-free survival time between the two groups obtained from the unsupervised clustering of samples and derive associated two-sided P-values. We derived KM curves derived using (1) two simplified representations of the dataset (“Binarization” and “Thresholding”) and (2) three different distance measures in the K-means clustering algorithm (L1/Rectilinear, Euclidean, Cosine). Details are provided in the Supporting Information section B.

Tumor phylogenetic tree-based statistics and subgrouping of samples

We next investigated whether either of the two different types of statistics computed based on tumor phylogenetic trees inferred from FISH copy number data are associated with disease-free and overall survival time. We first built tumor progression trees for each of the 65 patients using an evolution model incorporating single gene gain/loss (denoted SD) and whole genome duplication (denoted GD), as implemented in the software FISHtrees.¹⁵ The model also allows for gains/losses of entire chromosomes, but that capability is not relevant here because each probe is on a distinct chromosome.

For each gene in each tree, we counted the total number of edges across which that gene’s copy number is increased (SD gene gain or a GD event) and decreased (SD gene lost event) and normalized counts by the total number of SD and GD events, collectively producing an eight-dimensional array of event frequencies for each sample across the four genes. Next, we calculated the Simpson Index (SI) and Shannon Entropy (SE) for each sample. SI and SE are computed by taking sum of squared frequency values and sum of frequency-weighted negative logarithms of frequencies, respectively. For example, if there were only two genes and the (gain, loss) counts for gene 1 and gene 2 were (10,20) and (30,40), then SI of this patient would be: $(10/100)^2 + (20/100)^2 + (30/100)^2 + (40/100)^2 = 0.3$ and SE is: $-0.1 \log(0.1) - 0.2 \log(0.2) - 0.3 \log(0.3) - 0.4 \log(0.4) = 0.56$. Here, we use logarithm base 10 in SE, though it was originally defined using logarithm base 2. A lower SI or a higher SE indicates that the underlying gene count distribution is more diverse (closer to uniform) across the four genes tested. The two measures of diversity are standard in genetics (SI) or information theory (SE), and they have been previously used to study the diversity of tumor cell populations.³¹ We divided the patients into two groups for each measure, with those above versus below the mean SI (meanSE) assigned to different subgroups. Alternative tree-based statistics and subgrouping procedures based on tree topology were also considered and are described in Supporting Information, subsection (D).

RESULTS

HPV Status

HPV status was determined for 61 OTSCC samples. We were unable to perform HPV tests for four samples due to insufficient DNA quantity. Sixty samples tested negative for HPV and one sample tested positive for HPV 16. Interestingly, the one patient that tested positive for HPV 16 was a Stage 4 patient that smoked and was alive at the 60 month follow-up. He was also a complete pathological responder to radiation.

Univariate survival analysis for FISH results and DNA cytometry measurement

FISH detected various degrees of copy number changes for all five markers in all 65 OTSCC biopsy samples (Figure 1a). The mean percentage of cells with elevated FISH signals, more than three copies of each marker is shown in Figure 1c. For all four gene probes, the percentages of cells with a copy number change increases with tumor stage. Nevertheless, survival analysis based on one gene or pairs of genes found no significant predictors of survival independent of tumor stage (Supporting Information Tables 1 and 2). Also DNA cytometry measurement was not a significant predictor of survival (data not shown).

Unsupervised clustering reveals a relationship between smoking habits and copy number changes

Using unsupervised clustering based on all five FISH probes, a heat map for 65 samples was generated and separated into three groups (Figure 1b). Cluster assignment correlated to smoking habits in this patient cohort (Table 2). Further examinations revealed that mean signal numbers for each of the five markers showed moderate to large increases in almost all non-smoking patients (Groups 2 and 3), while low-level changes (Group 1) were observed in >50% of smokers (Figure 1b). These results correlated well with DNA cytometry measurements. Table 2 shows that the increase of mean copy numbers seen by FISH in Group 1–3 in the heat map is reflected by increasing “greater than 2.5C exceeding rates” and “greater than 5C exceeding rates” within these groups.

Two-means clustering and survival analysis based on sample-based statistics

We examined nine combinations of representations of probe count G(ain), N(ormal), L(oss) how data are represented (raw, binarized, or thresholded) and similarity measure (L1/Rectilinear, Euclidean, Cosine). Six of the nine combinations yielded significant ($P < 0.05$) differences in disease-free survival, with all yielding similar patient assignments to high-risk and low-risk clusters (Supporting Information Table 3). Binarizing the data (thresholded so that 60% of values are not 0) with rectilinear distance yielded the most significant difference ($P = 0.0054$) in overall survival between the two patient clusters (Figure 2a) as well as a significant difference for disease-free survival ($P = 0.0107$). The KM curves for the remaining five of the six cluster analyses showing significant survival differences are provided in Supporting Information Figure 1, with additional description of the methodology in Supporting Information section C. Multivariate survival analysis using both cluster assignment and tumor stage showed that cluster assignment is not a significant predictor of survival independent of tumor stage.

To better characterize features predictive of survival, we further clustered the raw G, N, L data to identify short survival (subgroup A) and long survival (subgroup B) clusters. Table 3 provides descriptive statistics about the cluster centers, i.e., the averages of the elements of each cluster. The center of cluster B is heavily biased towards the “No change of any gene” string while cluster A shows bias towards the “GGGN” and “GGGL” strings. In the same table, we show the two representative samples closest to each cluster center. String “NNNN” is most populated across the cluster B representatives and “GGGN” and “GGGL” are most populated in the cluster A representatives. Cluster A thus appears to favor patients with many cells with gain of all three oncogenes while cluster B tends to favor patients with many cells exhibiting no gain or loss of any of the genes.

Supporting Information Table 3 reports the subgroup ID of each sample belonging to two different clusters for the six choices of data representation and similarity measure listed above as Experiments 1–6. The clustering is robust to the methodology, with 50 of the 65 patients assigned to the same clusters for all six experiments, but because the P-values are close to the 0.05 significance threshold, small changes in the clustering assignments can affect whether the KM results are statistically significant. Cluster membership exhibits a clear association with tumor stage, which is unsurprising since stage is known to be associated with survival. For example, in the experiment of Table 3, the long-survival cluster B is strongly associated with tumor stages 1 or 2 ($P = 0.0045$, Fisher’s exact test). The cluster assignment largely predicts tumor stage and thus is not an independent predictor of survival.

We also considered an alternative approach in which we used principal components analysis (PCA, `prcomp` function in R) to represent the patients starting from either the average copy number for the five probes or the 16-item vector of comparator counts. PCA has been previously suggested as a technique for building models of tumor progression (PMID 11319803). The four step analysis of 1) Define initial representation 2) Apply PCA 3) Cluster into two clusters based on the PCA scores and 4) Test the two PCA-derived clusters for association with (disease-free) survival did not yield statistically significant results, even without taking into account tumor stage (data not shown).

Two-means clustering and survival analysis, taking into account smoking or tumor stage

We next split the samples based on smoker/non-smoker status and performed survival analysis on these two sets of samples separately. 19 patients are non-smokers and 34 are smokers, with the remaining 12 lacking smoking status information and thus not considered in this analysis. Supporting Information Table 4 reports the P-values from the survival analysis for clustering using the choices in Experiments 1–6, and based on subgrouping of smokers/non-smokers. 2-means clustering is associated with survival for the patients belonging to the smoker category, with five of the six P-values < 0.05 and for experiment 2, a P-value of 0.00342.

We then performed COXPH analysis to test the independent predictive power of cluster-based subgrouping, smoking behavior, and tumor stage as explanatory variables. Supporting Information Table 5 reports the results of COXPH analysis using clustering and smoking behavior, showing no significant difference in disease-free (overall) survival for the two

clusters after controlling for smoking behavior and vice versa. None of the six experiments yields a P-value < 0.05 and the confidence intervals include 1 (not significant) except for experiment 2. Supporting Information Table 6 reports results of COXPH analysis on clustering and tumor staging, showing that the cluster-based subgrouping is not significantly associated with disease-free (overall) survival, independent of tumor stage. In every experiment, the HRs using tumor stage exclude 1 in the confidence interval and are significant, while the HRs using cluster subgrouping include 1 and are not significant.

KM survival analysis using tree-based statistics

Based on our previous work with other FISH data sets,^{23, 24} we reasoned that the weakness of the above clustering analysis is that it fails to take into account likely evolutionary relationships between cells with similar combinations of FISH probe counts. Therefore, we used the software FISHtrees^{14, 15} to construct evolutionary models of how each tumor progressed based on the observed FISH patterns of the four genes (not using CEP4) and tabulated the Simpson's Index (SI) and Shannon Entropy (SE) test statistics, as described in Methods.

After computing SI and SE from the tumor phylogenetic trees, we subdivided the patients into two groups based on mean SI, mean SE and computed KM curves based on overall (disease-free) survival time. For SE-based subgrouping, the P-values obtained from the KM analysis for overall (disease-free) survival is 0.0144 (0.0111), and the KM curve is shown in Figure 2b (Figure 2c). Patients with higher SE value, who have similar frequencies of gene gain and loss counts across genes, have shorter overall and disease-free survival time compared to those with preferential gain or loss of specific genes. SI-based subgrouping, shown as Supporting Information Figure 2, yielded comparable results, with P-values from KM analysis of overall (disease-free) survival of 0.0168 (0.0117). Additional KM analysis based on tree topological features, described in Supporting Information section D, failed to yield statistical significance in all but one of the fourteen combinations considered.

Multivariate COXPH survival analysis using tree-based statistics, tumor stage, and smoking

Next, we performed COXPH tests of the relationship between overall (disease-free) survival time and combinations of the SI/SE-based subgrouping, tumor stage, and smoking status as explanatory variables. Supporting Information Table 8(A) and 8(B) show results for SI-based or SE-based subgrouping and tumor stage. For each subgrouping, the combined P-value is statistically significant, showing that the two covariates are independently associated with overall (disease-free) survival time. The HR is higher for either SI-based subgrouping or SE-based subgrouping than tumor stage-based subgrouping, with statistically significant association for the SI-based and SE-based subgroupings, meaning that there is a significant difference in overall (disease-free) survival for the two subgroups after adjusting for tumor stage. SE-based subgrouping shows a higher HR and lower P-values (0.0029, 0.0033) for both disease-free and overall survival in comparison to SI-based subgrouping (0.0036, 0.0045).

Supporting Information Table 9 shows results from comparable experiments taking into account SI/SE-based subgrouping and patients' smoking behavior. The P-values are statistically significant for both SI- and SE-based subgrouping, after adjusting for smoking behavior. The global P-values are statistically significant for the SE-based subgrouping experiment, and the HR confidence intervals are greater than 1 for both SI- and SE-based subgrouping and less than 1 for smoking-pattern-based subgrouping, except for prediction of disease-free survival when considered in combination with SE subgrouping. Therefore, the tree-based statistics provide additional information regarding the overall and disease-free patient survival beyond what can be achieved by looking at smoking behavior alone.

Table 4 shows results of COXPH analysis between all three variables: SI/SE-based subgrouping, smoking behavior and tumor stage. SI- and SE-based subgroupings yield statistically significant results for prediction of both disease-free and overall survival after adjusting for both smoking behavior and tumor stage. The HRs are higher for SI and SE subgrouping compared to the other two.

Supporting Information Table 10 shows the subgroup assignment of the patients based on SI and SE. Low SI and high SE patients are labeled as group "A", and high SI and low SE patients as group "B". The two experiments had the same subgroup assignments for all patients except T27 and T38, who were assigned to the short-surviving group by the SE subgrouping but the long-surviving group in the SI subgrouping.

Order of events

Early papers on tumor phylogenetics and related statistical analysis (reviewed in reference 48) considered the problem of predicting which events were "early events" or "late events". These studies used as input coarse CGH or breakpoint data on a cohort of patients with the same type of cancer. These studies sought to build one, unified model of tumor progression for all patients in the cohort. In contrast, the phylogenetic analysis we do in this study and other studies using our FISHtrees methods, build a separate tree model of progression for each patient or sample and model each copied gained or lost of a gene as a separate event. Despite these inherent differences, we could investigate possible orders of events by combining the 65 patient tree models as follows.

For each edge in each FISHtrees-derived tree that is a gain of *TERC*, gain of *EGFR*, gain of *CCND1*, or loss of *TP53*, we recorded the level of that edge. Level 2 is the level closest to the diploid root and represents the earliest events. For each of 65 patients, we next tabulated the average level for the four events of interest. For each of the four events, we made a box-and-whiskers plot of the 65 per-patient averages. The four box-and-whiskers plots are shown as Supporting Information Figure 3. It is hard to compare *TP53* directly to the three oncogenes because most patients will lose at most two copies of *TP53* (to go from the normal two down to zero), but can gain many more than two copies of the oncogenes. As expected, *TP53* losses are early (at a low mean tree level) in some samples, but the median of the level means of 7.45 for *TP53* losses is surprisingly high. Among the oncogenes, gains of *CCND1* (median of level means 6.52) tend to occur earlier than gains of *EGFR* (7.60) or gains of *TERC* (7.88), but there is substantial overlap in the distributions of level means across trees.

DISCUSSION

We utilized FISH to detect copy number changes of five chromosomal markers (*TERC*, *EGFR*, *CCND1*, *TP53*, and *CEP4*) within the same cells to determine specific genetic alterations in OTSCC and their relation to different clinicopathological parameters. Using these markers allowed discernment of patients according to their smoking habits. Survival analysis using single gene markers or two gene markers simultaneously showed no significant association with prognosis, contrary to some previous studies of the same genes, especially *CCND1*,^{32–42} but consistent with others.^{30, 39, 43–45} One limitation of some of the previous studies claiming associations between some of the four genes and poor prognosis is that tumor stage was not considered as a covariate by these studies.^{32, 35–37, 42, 46} It is important to consider tumor stage as a covariate, as we did here in Cox proportional hazards analysis, if one wants to show that any putative biomarker has predictive value in addition to the predictive value of the tumor stage.

HPV status is an important factor to consider when studying head and neck cancers since it has been shown that many sub-sites are HPV-positive.^{1, 2} The presence of HPV is very rare in OTSCC as compared to other sub-sites with one study showing an incidence of 2.4% positive cases in 85 tumors.⁶ After testing for 16 of the major HPV types, we concur that HPV positivity is infrequent in OTSCC. This finding reiterates the importance of investigating tumors of different sub-sites separately, e.g., base of tongue versus OTSCC.

We showed that various applications of K-means clustering could group our 65 cases into two clusters that had significantly different disease-free survival. However, in most of these analyses, the assignment to clusters was statistically associated with tumor stage, and the association with survival was not statistically significant in multivariate COXPH analysis that combined cluster assignment and tumor stage. Thus, while these markers may be associated with OTSCC survival, they appear not to provide significant independent predictive value.

We built evolutionary models of tumor progression for each patient using the gene probes for *TERC*, *EGFR*, *CCND1*, and *TP53*. The concept that cancer is an evolutionary process and can be modeled using phylogenetic algorithms has been repeatedly demonstrated.⁴⁷ Patient-centric analysis combining all data for one patient at a time, rather than analyzing all patients one gene at a time, has advantages for the analysis of somatic mutation data from solid tumors.⁴⁸ Using the tumor progression trees, we derived test statistics that are significantly associated with disease-free and overall survival in multivariate analysis, even after taking into account tumor stage and smoking.

We have developed a general analysis framework for FISH data sets using several (4–8) gene markers on hundreds (here ~250 cells per tumor) of single cells from numerous samples (here 65 samples). Previously, we used part of this phylogenetic framework to analyze the progression from ductal carcinoma in situ (DCIS) to invasive ductal carcinoma (IDC) of the breast²⁴ and from patients with prostate cancer²³ with different study designs. In the breast cancer study, one objective was to derive a test statistic that could distinguish DCIS from IDC.²⁴ The prostate cancer study compared non-paired progressing and non-

progressing tumors, and one analysis objective was to derive a test statistic that could distinguish between these two categories. In both those studies, the patients were selected to have similar clinical parameters (e.g., almost identical Gleason scores in the prostate study), so that when testing for associations with prognosis, the clinical covariates would not be confounding factors.

In contrast, in the present study we analyzed 65 oral tongue cancer samples with known tumor stage, tobacco usage, and survival. The tumor stages ranged from 1 to 4 and the study included both smokers and non-smokers. Therefore, a principal objective of the data analysis was to derive a test statistic T that would (i) partition the cases into two or more categories and (ii) be associated with significantly different (disease-free) survival *after taking into account tumor stage and smoking status*. Due to the different study designs, it is surprising that the same phylogenetic analysis framework has been effective in all three studies, but not surprising that the test statistics in the present dataset differ from the previously published ones, as explained below.

Using the trees computed by FISHtrees, we can obtain a variety of test statistics. For the breast and prostate cancer studies, test statistics characterizing overall tree topology, specifically concerning the average depth of tree nodes, were most effective. In the present study, the best test statistics were the Simpson's Index (SI) or Shannon Entropy (SE) of the distribution of copy number changes per gene, measures of the overall diversity of variation in the tree. Patients whose trees showed a diversity of changes in all four genes had worse disease-free survival, suggesting that each of the four selected genes is relevant to OTSCC tumor progression, but making any one of them a biomarker may be too reductionist.

The concept that a series of genomic changes in *different* genes or pathways can transform a healthy cell to a tumor cell was demonstrated *in vitro*.⁴⁹ Cancer studies focused on the analysis of somatic point mutations have identified that there are at least 12 pathways that may be dysregulated in solid tumors.⁵⁰ For each tumor type, there may be some pathways that are more frequently mutated, but generally a breadth of gene changes are needed to cause cancer.⁵⁰ The same principle should apply to all types of genomic changes, including gene copy number changes. This recurrent observation can be used to find new pathways using the principles of coverage (many patients should have genes in the same pathway mutated) and mutual exclusivity (one mutated gene in a pathway can be sufficient to dysregulate the pathway).⁵¹ Moreover, these principles can effectively be used to find pathways when mutation and copy number data are combined.⁵² Here, we showed that the general paradigm of assessing breadth of genetic changes extends to single-cell FISH data.

In a similar analysis of oral cancer cases, Smeets et al.⁵³ used unsupervised clustering to partition samples into two clusters, such that one cluster of patients had significantly better prognosis than the other. Their results are especially compelling because they were replicated on a second data set from Snijders et al.⁵⁴ Indeed, one weakness of our study is that by focusing on OTSCC instead of all oral cancer, we could not collect an independent validation set. However, a particular strength of our study in comparison is the focus on a very defined disease entity.

One novel aspect of our study is the simultaneous enumeration of multiple FISH markers. A second novel aspect is that we collected data on hundreds of cells per tumor and made explicit progression models for how the *cells evolve*, rather than using rules to determine whether each gene is gained or lost *in an entire sample* (see Uzawa et al.⁷ for example). We designed the study to focus exclusively on OTSCC to avoid the inter-tumor heterogeneity observed in other studies that combined tumors from multiple sites in the oral cavity. This is also the first study, as far as we know, to demonstrate correlations between an increase in copy numbers of multiple genes and smoking habits in OTSCC. In our data, patients with low copy number changes were mostly smokers while those with higher copy number changes for all markers were more likely to be non-smokers, which is consistent with prior observations of Liloglu et al.⁵⁵ in head and neck cancers. The increasing 5C levels, as determined by DNA cytometry, from Group 1 to Group 3 of our hierarchical clustering, is a reflection of increased genomic instability (Table 2). Group 3, with the highest rate of genomic instability, was consistent with samples that revealed the highest copy number changes, suggesting that tumor aneuploidy may be more extensive in non-smokers as compared to smokers. Development of OTSCC in individuals exposed to tobacco may involve mechanisms distinct from those that drive carcinogenesis in patients that were not exposed.²

In summary, our results provide a better understanding of the genetic alterations in OTSCC, the phylogeny of tumor progression, and their relation to prognosis. Further delineation of the genetic basis of tongue cancer may aid in the development of individualized treatment options.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Cancer Institute, National Library of Medicine, NIH Extramural grants (1R01CA140214 and 1R01AI076318), the Swedish Cancer Society (Cancerfonden), the Cancer Society of Stockholm (Cancerföreningen), Laryngfonden, and the Karolinska Institutet. We thank Ann Olsson and Margaretha Waern for technical assistance. We thank Quang-Tri Nguyen for technical discussions and suggestions regarding the heat map clustering analysis.

Abbreviations used

OTSCC	oral tongue squamous cell carcinoma
FISH	fluorescence in situ hybridization
CGH	comparative genomic hybridization
SNP	single nucleotide polymorphism
BAC	bacterial artificial chromosome
KM	Kaplan-Meier

G	gain
N	normal/no change
L	loss
SI	Simpson's Index
HR	hazard ratio
COXPH	Cox proportional hazards
HPV	Human Papilloma Virus

References

- Shiboski CH, Schmidt BL, Jordan RC. Tongue and tonsil carcinoma: increasing trends in the U.S. population ages 20–44 years. *Cancer*. 2005; 103:1843–9. [PubMed: 15772957]
- Li R, Faden DL, Fakhry C, Langelier C, Jiao Y, Wang Y, Wilkerson MD, Peadarallu CS, Old M, Lang J, Loyo M, Ahn SM, et al. Clinical, genomic, and metagenomic characterization of oral tongue squamous cell carcinoma in patients who do not smoke. *Head Neck*. 2014
- Boffetta P, Hecht S, Gray N, Gupta P, Straif K. Smokeless tobacco and cancer. *Lancet Oncol*. 2008; 9:667–75. [PubMed: 18598931]
- Maasland DH, van den Brandt PA, Kremer B, Goldbohm RA, Schouten LJ. Alcohol consumption, cigarette smoking and the risk of subtypes of head-neck cancer: results from the Netherlands Cohort Study. *BMC Cancer*. 2014; 14:187. [PubMed: 24629046]
- Ambatipudi S, Gerstung M, Pandey M, Samant T, Patil A, Kane S, Desai RS, Schaffer AA, Beerenwinkel N, Mahimkar MB. Genome-wide expression and copy number analysis identifies driver genes in gingivobuccal cancers. *Genes Chromosomes Cancer*. 2012; 51:161–73. [PubMed: 22072328]
- Dahlgren L, Dahlstrand HM, Lindquist D, Hogmo A, Bjornestal L, Lindholm J, Lundberg B, Dalianis T, Munck-Wikland E. Human papillomavirus is more common in base of tongue than in mobile tongue cancer and is a favorable prognostic factor in base of tongue cancer patients. *Int J Cancer*. 2004; 112:1015–9. [PubMed: 15386365]
- Uzawa N, Sonoda I, Myo K, Takahashi K, Miyamoto R, Amagasa T. Fluorescence in situ hybridization for detecting genomic alterations of cyclin D1 and p16 in oral squamous cell carcinomas. *Cancer*. 2007; 110:2230–9. [PubMed: 17893905]
- Bhattacharya A, Roy R, Snijders AM, Hamilton G, Paquette J, Tokuyasu T, Bengtsson H, Jordan RC, Olshen AB, Pinkel D, Schmidt BL, Albertson DG. Two distinct routes to oral cancer differing in genome instability and risk for cervical node metastasis. *Clin Cancer Res*. 2011; 17:7024–34. [PubMed: 22068658]
- Viet CT, Schmidt BL. Understanding oral cancer in the genome era. *Head Neck*. 2010; 32:1246–68. [PubMed: 20187013]
- Soder AI, Hoare SF, Muir S, Going JJ, Parkinson EK, Keith WN. Amplification, increased dosage and in situ expression of the telomerase RNA gene in human cancer. *Oncogene*. 1997; 14:1013–21. [PubMed: 9070649]
- Bockmuhl U, Schluns K, Schmidt S, Matthias S, Petersen I. Chromosomal alterations during metastasis formation of head and neck squamous cell carcinoma. *Genes Chromosomes Cancer*. 2002; 33:29–35. [PubMed: 11746985]
- Garnis C, Campbell J, Zhang L, Rosin MP, Lam WL. OCGR array: an oral cancer genomic regional array for comparative genomic hybridization analysis. *Oral Oncol*. 2004; 40:511–9. [PubMed: 15006624]
- Gollin SM. Cytogenetic alterations and their molecular genetic correlates in head and neck squamous cell carcinoma: A next generation window to the biology of disease. *Genes Chromosomes Cancer*. 2014; 53:972–90. [PubMed: 25183546]

14. Chowdhury SA, Shackney SE, Heselmeyer-Haddad K, Ried T, Schaffer AA, Schwartz R. Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations. *Bioinformatics*. 2013; 29:i189–98. [PubMed: 23812984]
15. Chowdhury SA, Shackney SE, Heselmeyer-Haddad K, Ried T, Schaffer AA, Schwartz R. Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics. *PLoS Comput Biol*. 2014; 10:e1003740. [PubMed: 25078894]
16. Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, Kendall J, Riggs M, Eberling Y, Troge J, Grubor V, Levy D, Lundin P, et al. Inferring tumor progression from genomic heterogeneity. *Genome Res*. 2010; 20:68–80. [PubMed: 19903760]
17. Desper R, Jiang F, Kallioniemi OP, Moch H, Papadimitriou CH, Schaffer AA. Inferring tree models for oncogenesis from comparative genome hybridization data. *J Comput Biol*. 1999; 6:37–51. [PubMed: 10223663]
18. Beerenwinkel N, Rahnenfuhrer J, Daumer M, Hoffmann D, Kaiser R, Selbig J, Lengauer T. Learning multiple evolutionary pathways from cross-sectional data. *J Comput Biol*. 2005; 12:584–98. [PubMed: 16108705]
19. Cheng YK, Beroukhi R, Levine RL, Mellinghoff IK, Holland EC, Michor F. A mathematical methodology for determining the temporal order of pathway alterations arising during gliomagenesis. *PLoS Comput Biol*. 2012; 8:e1002337. [PubMed: 22241976]
20. Desper R, Jiang F, Kallioniemi OP, Moch H, Papadimitriou CH, Schaffer AA. Distance-based reconstruction of tree models for oncogenesis. *J Comput Biol*. 2000; 7:789–803. [PubMed: 11382362]
21. Brennan JA, Boyle JO, Koch WM, Goodman SN, Hruban RH, Eby YJ, Couch MJ, Forastiere AA, Sidransky D. Association between cigarette smoking and mutation of the p53 gene in squamous-cell carcinoma of the head and neck. *N Engl J Med*. 1995; 332:712–7. [PubMed: 7854378]
22. Greene, FL. American Joint Committee on Cancer., American Cancer Society. *AJCC cancer staging manual*. 6. New York: Springer; 2002. p. xivp. 421
23. Heselmeyer-Haddad KM, Berroa Garcia LY, Bradley A, Hernandez L, Hu Y, Habermann JK, Dumke C, Thorns C, Perner S, Pestova E, Burke C, Chowdhury SA, et al. Single-Cell Genetic Analysis Reveals Insights into Clonal Development of Prostate Cancers and Indicates Loss of PTEN as a Marker of Poor Prognosis. *Am J Pathol*. 2014; 184:2671–86. [PubMed: 25131421]
24. Heselmeyer-Haddad K, Berroa Garcia LY, Bradley A, Ortiz-Melendez C, Lee WJ, Christensen R, Prindiville SA, Calzone KA, Soballe PW, Hu Y, Chowdhury SA, Schwartz R, et al. Single-cell genetic analysis of ductal carcinoma in situ and invasive breast cancer reveals enormous tumor heterogeneity yet conserved genomic imbalances and gain of MYC during progression. *Am J Pathol*. 2012; 181:1807–22. [PubMed: 23062488]
25. Therneau, TM.; Grambsch, PM. *Modeling survival data : extending the Cox model*. New York: Springer; 2000. p. xiiip. 350
26. Tableman, M.; Kim, JS. *Survival Analysis Using S : Analysis of Time-To-Event Data*. New York, N.Y: Chapman and Hall/CRC; 2004.
27. Nordfors C, Vlastos A, Du J, Ahrlund-Richter A, Tertipis N, Grun N, Romanitan M, Haeggbloom L, Roosaar A, Dahllof G, Dona MG, Benevolo M, et al. Human papillomavirus prevalence is high in oral samples of patients with tonsillar and base of tongue cancer. *Oral Oncol*. 2014; 50:491–7. [PubMed: 24613649]
28. Wangsa D, Heselmeyer-Haddad K, Ried P, Eriksson E, Schaffer AA, Morrison LE, Luo J, Auer G, Munck-Wikland E, Ried T, Lundqvist EA. Fluorescence in situ hybridization markers for prediction of cervical lymph node metastases. *Am J Pathol*. 2009; 175:2637–45. [PubMed: 19893027]
29. Heah KG, Hassan MI, Huat SC. p53 Expression as a marker of microinvasion in oral squamous cell carcinoma. *Asian Pac J Cancer Prev*. 2011; 12:1017–22. [PubMed: 21790244]
30. Werkmeister R, Brandt B, Joos U. Clinical relevance of erbB-1 and -2 oncogenes in oral carcinomas. *Oral Oncol*. 2000; 36:100–5. [PubMed: 10889928]
31. Park SY, Gonen M, Kim HJ, Michor F, Polyak K. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *J Clin Invest*. 2010; 120:636–44. [PubMed: 20101094]

32. Akervall JA, Michalides RJ, Mineta H, Balm A, Borg A, Dictor MR, Jin Y, Loftus B, Mertens F, Wennerberg JP. Amplification of cyclin D1 in squamous cell carcinoma of the head and neck and the prognostic value of chromosomal abnormalities and cyclin D1 overexpression. *Cancer*. 1997; 79:380–9. [PubMed: 9010112]
33. Bellacosa A, Almadori G, Cavallo S, Cadoni G, Galli J, Ferrandina G, Scambia G, Neri G. Cyclin D1 gene amplification in human laryngeal squamous cell carcinomas: prognostic significance and clinical implications. *Clin Cancer Res*. 1996; 2:175–80. [PubMed: 9816104]
34. Bova RJ, Quinn DI, Nankervis JS, Cole IE, Sheridan BF, Jensen MJ, Morgan GJ, Hughes CJ, Sutherland RL. Cyclin D1 and p16INK4A expression predict reduced survival in carcinoma of the anterior tongue. *Clin Cancer Res*. 1999; 5:2810–9. [PubMed: 10537346]
35. Fujii M, Ishiguro R, Yamashita T, Tashiro M. Cyclin D1 amplification correlates with early recurrence of squamous cell carcinoma of the tongue. *Cancer Lett*. 2001; 172:187–92. [PubMed: 11566495]
36. Hassan NM, Tada M, Hamada J, Kashiwazaki H, Kameyama T, Akhter R, Yamazaki Y, Yano M, Inoue N, Moriuchi T. Presence of dominant negative mutation of TP53 is a risk of early recurrence in oral cancer. *Cancer Lett*. 2008; 270:108–19. [PubMed: 18555592]
37. Kaminagakura E, Werneck da Cunha I, Soares FA, Nishimoto IN, Kowalski LP. CCND1 amplification and protein overexpression in oral squamous cell carcinoma of young patients. *Head Neck*. 2011; 33:1413–9. [PubMed: 21254295]
38. Kyomoto R, Kumazawa H, Toda Y, Sakaida N, Okamura A, Iwanaga M, Shintaku M, Yamashita T, Hiai H, Fukumoto M. Cyclin-D1-gene amplification is a more potent prognostic factor than its protein over-expression in human head-and-neck squamous-cell carcinoma. *Int J Cancer*. 1997; 74:576–81. [PubMed: 9421351]
39. Michalides RJ, van Veelen NM, Kristel PM, Hart AA, Loftus BM, Hilgers FJ, Balm AJ. Overexpression of cyclin D1 indicates a poor prognosis in squamous cell carcinoma of the head and neck. *Arch Otolaryngol Head Neck Surg*. 1997; 123:497–502. [PubMed: 9158396]
40. Miyamoto R, Uzawa N, Nagaoka S, Nakakuki K, Hirata Y, Amagasa T. Potential marker of oral squamous cell carcinoma aggressiveness detected by fluorescence in situ hybridization in fine-needle aspiration biopsies. *Cancer*. 2002; 95:2152–9. [PubMed: 12412169]
41. Myo K, Uzawa N, Miyamoto R, Sonoda I, Yuki Y, Amagasa T. Cyclin D1 gene numerical aberration is a predictive marker for occult cervical lymph node metastasis in TNM Stage I and II squamous cell carcinoma of the oral cavity. *Cancer*. 2005; 104:2709–16. [PubMed: 16265665]
42. Nakata Y, Uzawa N, Takahashi K, Sumino J, Michikawa C, Sato H, Sonoda I, Ohyama Y, Okada N, Amagasa T. EGFR gene copy number alteration is a better prognostic indicator than protein overexpression in oral tongue squamous cell carcinomas. *Eur J Cancer*. 2011; 47:2364–72. [PubMed: 21852109]
43. Ogmundsdottir HM, Bjornsson J, Holbrook WP. Role of TP53 in the progression of pre-malignant and malignant oral mucosal lesions. A follow-up study of 144 patients. *J Oral Pathol Med*. 2009; 38:565–71. [PubMed: 19473450]
44. Sathyan KM, Sailasree R, Jayasurya R, Lakshminarayanan K, Abraham T, Nalinakumari KR, Abraham EK, Kannan S. Carcinoma of tongue and the buccal mucosa represent different biological subentities of the oral carcinoma. *J Cancer Res Clin Oncol*. 2006; 132:601–9. [PubMed: 16767431]
45. Wang L, Liu T, Nishioka M, Aguirre RL, Win SS, Okada N. Activation of ERK1/2 and cyclin D1 expression in oral tongue squamous cell carcinomas: relationship between clinicopathological appearances and cell proliferation. *Oral Oncol*. 2006; 42:625–31. [PubMed: 16483832]
46. Gebhart E, Ries J, Wiltfang J, Liehr T, Efferth T. Genomic gain of the epidermal growth factor receptor harboring band 7p12 is part of a complex pattern of genomic imbalances in oral squamous cell carcinomas. *Arch Med Res*. 2004; 35:385–94. [PubMed: 15610907]
47. Attolini CS, Michor F. Evolutionary theory of cancer. *Ann N Y Acad Sci*. 2009; 1168:23–51. [PubMed: 19566702]
48. Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. Patient-oriented gene set analysis for cancer mutation data. *Genome Biol*. 2010; 11:R112. [PubMed: 21092299]

49. Hahn WC, Counter CM, Lundberg AS, Beijersbergen RL, Brooks MW, Weinberg RA. Creation of human tumour cells with defined genetic elements. *Nature*. 1999; 400:464–8. [PubMed: 10440377]
50. Gerstung M, Eriksson N, Lin J, Vogelstein B, Beerenwinkel N. The temporal order of genetic and pathway alterations in tumorigenesis. *PLoS One*. 2011; 6:e27136. [PubMed: 22069497]
51. Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. *Genome Res*. 2012; 22:375–85. [PubMed: 21653252]
52. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. *PLoS One*. 2010; 5:e8918. [PubMed: 20169195]
53. Smeets SJ, Brakenhoff RH, Ylstra B, van Wieringen WN, van de Wiel MA, Leemans CR, Braakhuis BJ. Genetic classification of oral and oropharyngeal carcinomas identifies subgroups with a different prognosis. *Cell Oncol*. 2009; 31:291–300. [PubMed: 19633365]
54. Snijders AM, Schmidt BL, Fridlyand J, Dekker N, Pinkel D, Jordan RC, Albertson DG. Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. *Oncogene*. 2005; 24:4232–42. [PubMed: 15824737]
55. Liloglou T, Scholes AG, Spandidos DA, Vaughan ED, Jones AS, Field JK. p53 mutations in squamous cell carcinoma of the head and neck predominate in a subgroup of former and present smokers with a low frequency of genetic instability. *Cancer Res*. 1997; 57:4070–4. [PubMed: 9307295]

Novelty and Impact statements

We present the largest genetic marker study on oral tongue squamous cell carcinoma (OTSCC), a rare entity of head and neck cancers. We used a novel approach utilizing multiple FISH probes on single cells to build phylogenetic tree models of tumor progression using the software FISHTrees. We were able to separate patients according to overall and disease-free survival, independent of tumor stage and smoking status, in multivariate analysis.

Figure 1.

(a) Representative image of fluorescence in situ hybridization; (b) Heat map depicting average cluster analysis of 65 cases. The heat map is generated by calculating the mean signal number per marker in each patient. The mean signals are assigned a color with green depicting low mean signal number compared to red which represents high mean signal number. The heat map is divided into three groups (Group 1, Group 2 and Group 3); (c) The mean percentage of cells with more than three signals per marker by stage.

Figure 2.

(a) KM curve for test of association between overall survival time and subgrouping of patients based on binarized data and L1 distance based K-means clustering. (b) KM curve for the test of association between overall survival time and SE-based subgrouping of patients. (c) KM curve for the test of association between disease-free survival time and SE-based subgrouping of patients. All the P-values in (a), (b) and (c) are two-sided.

Table 1

Patient and Tumor Characteristics According to Stage

Characteristics	Stage I ¹ n=15	Stage II ¹ n=30	Stage III ¹ n=7	Stage IV ¹ n=13	Total n=65 (%)
Age at Diagnosis					
20–39 years	2	5	1	1	9 (13)
40–59 years	6	13	3	2	24 (41)
60 years	7	12	3	10	32 (46)
Gender					
Male	4	22	4	7	37 (58)
Female	11	8	3	6	28 (42)
Smoking Habit					
Smoking	9	14	5	6	34 (51)
Non-smoker	4	12	2	2	20 (31)
No information	2	4	0	5	11 (18)
Histopathologic Grade²					
Well differentiated	7	4	1	3	15 (23)
Moderately differentiated	7	19	6	7	39 (62)
Poorly differentiated	1	7	0	3	11 (15)
Pre-operative Radiation					
Response					
pCR	0	5	1	0	6
non-pCR	0	19	4	3	26
Survival					
Alive	12	15	1	3	31 (48)
Dead	3	15	6	10	34 (52)
Recurrence					
Locoregional	4	12			
Systemic	0	1			
No Recurrence	9	17	N/A	N/A	N/A
Secondary Primary	2	0			

Abbreviations: pCR (complete pathological remission); non-pCR (incomplete pathological remission)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TNM stage, tumor stage according to UICC

Tumor differentiation grade according to the World Health Organization International Histological Classification of Tumors

Table 2

Heat map grouping according to stage and smoking habits

Smoking Status	Heat Map Clusters			P-value ^I
	Group 1	Group 2	Group 3	
Stage 1				
Non-Smoking	0	2	1	0.018
Smoking	8	1	0	
Stage 2				
Non-Smoking	1	3	8	0.029
Smoking	6	4	4	
Stage 1 and 2				
Non-Smoking	1	5	9	0.001
Smoking	14	5	4	
Stage 3 and 4				
Non-Smoking	1	2	1	0.845
Smoking	4	4	3	
All Stages				
Non-Smoking	2	7	10	0.002
Smoking	18	9	7	
DNA Cytometry				
> 2.5 exceeding rates	61.9%	83.6%	88.7%	
> 5c exceeding rates	12.5%	17.7%	32.9%	

^I Mantel-Haenszel Chi-square test

Table 3

Cluster centers for subgrouping of patients from two-means clustering using the original FISH data (no binarization or thresholding) and rectilinear distance.

String	Cluster Center		Nearest Sample	
	Cluster B	Cluster A	Cluster B	Cluster A
NNNN	11	2	9	2
NNNL	2	1	0	1
NNGN	2	1	2	1
NNGL	1	1	1	1
NGNN	3	2	2	3
NGNL	1	3	0	3
NGGN	3	7	5	8
NGGL	1	4	1	3
GNNN	4	1	4	0
GNNL	1	1	0	0
GNGN	2	2	1	2
GNGL	1	2	0	2
GGNN	4	5.5	3	9
GGNL	1	4	1	3
GGGN	6	18.5	5	16
GGGL	2	11	3	11

Table 4
Results of COXPH analysis using SI (A) and SE (B) based subgrouping of patients, smoking behavior and tumor stage as explanatory variable

A	Disease-free survival				Overall survival			
	Global P-value		95% CI		HR		95% CI	
	HR	P-value	HR	P-value	HR	P-value	HR	P-value
SI Subgrouping	2.633	1.093–6.343	3.09E-02		2.705	1.124–6.508	2.63E-02	
Tumor stage	9.55E-03	1.567	1.090–2.252	1.53E-02	1.843	1.274–2.665	1.16E-03	
Smoking behavior	1.029	0.454–2.333	9.45E-01		0.899	0.395–2.049	8.01E-01	

B	Disease-free survival				Overall survival			
	Global P-value		95% CI		HR		95% CI	
	HR	P-value	HR	P-value	HR	P-value	HR	P-value
SE Subgrouping	2.881	1.143–7.264	2.49E-02		3.022	1.199–7.611	1.90E-02	
Tumor stage	7.12E-03	1.587	1.102–2.285	1.31E-02	1.868	1.289–2.707	9.59E-04	
Smoking behavior	1.052	0.464–2.386	9.03E-01		0.918	0.403–2.089	8.38E-01	