A Minimal DNA Methylation Signature in Oral Tongue Squamous Cell Carcinoma Links Altered Methylation with Tumor Attributes

Neeraja M. Krishnan1, Kunal Dhas1, Jayalakshmi Nair1, Vinayak Palve1, Jamir Bagwan1, Gangotri Siddappa2, Amritha Suresh2, Vikram D. Kekatpure3, Moni Abraham Kuriakeose2,3, and Binay Panda1,4

Abstract

Oral tongue squamous cell carcinomas (OTSCC) are a homogenous group of aggressive tumors in the head and neck region that spread early to lymph nodes and have a higher incidence of regional failure. In addition, there is a rising incidence of oral tongue cancer in younger populations. Studies on functional DNA methylation changes linked with altered gene expression are critical for understanding the mechanisms underlying tumor development and metastasis. Such studies also provide important insight into biomarkers linked with viral infection, tumor metastasis, and patient survival in OTSCC. Therefore, we performed genome-wide methylation analysis of tumors (N = 52) and correlated altered methylation with differential gene expression. The minimal tumor-specific DNA 5-methylcytosine signature identified genes near 16 different differentially methylated regions, which were validated using genomic data from The Cancer Genome Atlas cohort. In our cohort, hypermethylation of MIR10B was significantly associated with the differential expression of its target genes NR4A3 and BCL2L11 (P = 0.0125 and P = 0.014, respectively), which was inversely correlated with disease-free survival (P = 9E–15 and P = 2E–15, respectively) in patients. Finally, differential methylation in FUT3, TRIM5, TSPAN7, MAP3K8, RPS6KA2, SLC9A9, and NPAS3 genes was found to be predictive of certain clinical and epidemiologic parameters.

Implications: This study reveals a functional minimal methylation profile in oral tongue tumors with associated risk habits, clinical, and epidemiologic outcomes. In addition, NR4A3 downregulation and correlation with patient survival suggests a potential target for therapeutic intervention in oral tongue tumors. Data from the current study are deposited in the NCBI Geo database (accession number GSE75540).

Mol Cancer Res; 14(9); 805–19. ©2016 AACR.

Introduction

Head and neck squamous cell carcinomas (HNSCC) are a heterogeneous group of tumors with different incidences, mortalities, and prognosis for different subtypes. HNSCCs are the sixth leading cause of cancer worldwide (1) and account for almost 30% of all cancer cases in India (2). Oral cancer is the most common subtype of head and neck cancers with a worldwide incidence of greater than 300,000 cases. The disease is an important cause of morbidity and mortality with a 5-year survival of less than 50% (1, 2). Unlike other subtypes of oral cavity like gingivo-buccal, tumors originating in the anterior part of tongue or oral tongue have an increased association with younger patients (3, 4), spread early to lymph nodes (3), and have a higher regional failure (5). Tobacco and alcohol are common risk factors for this group of cancer. Although the role of human papilloma virus (HPV) is clearly understood in the etiologic agent and/or a prognostic marker in oral tongue squamous cell carcinoma (OTSCC) is not yet established.

In the last 5 years, studies have identified somatic mutations, insertions, deletions and amplifications, copy number alterations, loss of heterozygosity, gene expression, and DNA methylation changes for different tumor subtypes in HNSCC (7–12). In addition, few studies (13, 14) have demonstrated the role of epigenetic modifications and associated pathways in HPV-positive HNSCC patients. Epigenetic changes are known biomarkers for cancer initiation and progression (15). DNA methylation at cytosine residues [5-methylcytosine (5mC)] is one of the best studied epigenetic changes in cancer (16). Cytosine methylation alters gene expression and thus plays an important role in other biologic processes, such as embryonic development, imprinting, and X-chromosome inactivation (17, 18). Cancer development is characterized by two major DNA methylation changes, hypermethylation of CpG islands located in the promoter regions of tumor suppressor genes (making them inactive) and...
global hypomethylation leading to activation of oncogenes and transposons (19). By conducting genome-wide methylation studies, one can identify potential DNA methylation markers for early cancer detection. Previous studies profiling DNA methylation changes in HNSCC primarily used low-through-put assays and identified changes in DCC, DAPK1, SPEN, E-cadherin, CDKN2A, RASSF1, and MGMT genes (20–22). Picking and colleagues (2013) studied methylation in oral squamous cell carcinomas using low-density HumanMethylation27K arrays and found 3,694 loci to be differentially methylated in tumor tissues (9). Oral cancer studies by Shaw and colleagues (2013) used the Illumina GoldenGate assay that comprises 1,305 CpG island loci and found 48 loci to be differentially methylated in tumors (23).

In this study, we profiled tumor-specific 5mC changes using high-density arrays comprising 485,512 probes in 52 pairs of OTSCCs. We identified differentially methylated loci or probes (DMR) between tumors and their matched normals and studied their functional impact by comparing the levels of differential methylation with differential gene expression in tumors. We found varying extents of differential methylation in various subregions. These differentially methylated regions (DMRs) often encompass multiple genes and were associated with differential gene expression, thus representing functional DMRs. We found aberrant methylation in a total of 27,276 hypermethylated and 21,231 hypomethylated loci with an average Δβ (level of differential methylation) of at least 0.2. We discovered a minimal methylation signature comprising 16 different DMRs that harbored ambiguous probes from analyses to be differentially methylated in tumors (23).

In the study, we profiled tumor-specific 5mC changes using high-density arrays comprising 485,512 probes in 52 pairs of OTSCCs. We identified differentially methylated loci or probes (DMR) between tumors and their matched normals and studied their functional impact by comparing the levels of differential methylation with differential gene expression in tumors. We found varying extents of differential methylation in various subregions. These differentially methylated regions (DMRs) often encompass multiple genes and were associated with differential gene expression, thus representing functional DMRs. We found aberrant methylation in a total of 27,276 hypermethylated and 21,231 hypomethylated loci with an average Δβ (level of differential methylation) of at least 0.2. We discovered a minimal methylation signature comprising 16 different DMRs that harbored ambiguous probes from analyses to be differentially methylated in tumors (23).

Only those patients whose histologic sections confirmed the presence of squamous cell carcinoma with at least 70% tumor cells in the specimen were used in this study. At the time of admission, patients were asked about their habits (smoking, chewing, and/or alcohol consumption) and other risk factors. Fifty-two treatment-naive patients who underwent staging according to AJCC criteria and curative intent treatment as per NCCN guidelines involving surgery with or without postoperative adjuvant radiation or chemotherapy at the Mazumdar Shaw Medical Centre were accrued for the study (Supplementary Table S1).

Whole-genome methylation assay
Genomic DNA was extracted from tissues using DNeasy Blood and Tissue Kit (Qiagen) as per the manufacturer’s specifications. Genomic DNA quality was assessed by agarose gel electrophoresis, and samples with intact, high molecular weight DNA bands were used for bisulfite conversion. Five hundred nanograms of genomic DNA was used for bisulfite conversion using the Zymo EZ DNA Methylation Kit (Zymo Research Corp.) following the manufacturer’s instructions and eluted in 12 µL of elution buffer. Four microliters of the bisulfite-converted DNA was used as template for target preparation to hybridize on the BeadChip and was processed as per the manufacturer’s instructions following Infinium HD and Infinium protocol for HumanMethylation450K BeadChip (Illumina), respectively. Data were collected using the HiScan (Illumina) reader and analyzed with GenomeStudio V2011.1 methylation module 1.1.1 (Illumina).

Methylation450 BeadChip probe alignment and removal of ambiguous probes from analyses
The probe sequences were aligned using bowtie2 (25), while allowing up to 2 mismatches to the human reference hg19 sequence to detect and exclude probes mapping ambiguously to multiple locations in the reference genome (Supplementary Text S1). We found 17,649 probes (3.6% of the total number of probes) to be ambiguously mapping in this manner. These probes mapped to multiple locations, ranging from 2 to 6265 (Supplementary Table S2).

Whole-genome gene expression assay
Whole-genome gene expression profiling was carried out with Illumina HumanHT-12 v4 expression BeadChip (Illumina) with 21 tumors and their matched adjacent normal tissues. Total RNA was extracted using PureLink RNA Mini Kit (Invitrogen), and RNA quality was checked on the bioanalyzer chip (Agilent). RNA samples with poor RNA integrity numbers (RIN; ≤7) on the bioanalyzer chip, indicating partial degradation of RNA, were processed using Illumina WG64SL assay and the ones with good RIN numbers (≥7) were labeled using Illumina TotalPrep RNA Amplification Kit (Ambion) as per the manufacturer’s recommendations. Targets were used to hybridize arrays, and arrays were processed according to the manufacturer’s recommendations. Arrays were scanned using HiScan, Illumina, and the data collected were analyzed with GenomeStudio V2011.1 Gene Expression module 1.9.0 (Illumina) to check for the assay quality control probes. Raw signal intensities were exported from GenomeStudio for transformation, normalization, and differential expression analyses using R.

Data preprocessing and differential methylation analyses
Raw data for 52 tumor:matched control sample pairs obtained from the Illumina Infinium Methylation450 BeadChip scanned using the HiScan system were analyzed in GenomeStudio Methylation Module v1.9.0 (Illumina) for assay controls, before exporting the β values for analysis using R. For each CpG site, ratio of fluorescent signal was measured by that of a methylated
Using the formula mentioned below:

\[
\beta = \frac{P_m}{P_m + P_{um}}
\]

where \( P_m \) is the signal intensity of the methylation detection probe and \( P_{um} \) is the signal intensity of the nonmethylation detection probe. Infinium methylation assay has built-in sample-dependent and -independent controls, which were used to test the overall efficiency of the assay for each sample. Raw signal intensities were exported as idat files, preprocessed (transformed and normalized), and analyzed using R Bioconductor packages, watermelon, minfi, and ComBat (26–28,29), to estimate differential methylation. The Illumina 450K data were preprocessed using a functional normalization procedure implemented by the "preprocessFunnorm" function within the R Bioconductor package minfi (Supplementary Text S2). This algorithm was recently developed and found to outperform other 450K data normalization and batch correction methods (29). After functional normalization, we identified DMRs using "bumphunter" function in minfi, retaining significant ones with \( P \leq 0.05 \) and \( \text{FDR} \leq 0.05 \). Dasen normalization was also performed using watermelon to obtain differentially methylated loci using minfi (Supplementary Text S3). A clear difference before and after normalization was seen in the density plots (Supplementary Fig. S1). The normalized data were further batch corrected using ComBat (28) using empirical Bayes methods. Dasen-normalized DMPs corresponding to these DMRs were extracted. Postprocessing of the differentially methylated data for visualization was performed as per the method provided in the Supplementary Material (Supplementary Text S4–S6).

Data preprocessing and differential expression analyses
Whole-genome gene expression data were analyzed using final reports exported from Genome Studio, using the R Bioconductor packages lumi (30). The VST transformation and LOESS normalization methods in lumi performed best in preprocessing data from the DASL and WG assays. They were chosen based on their highest mean IACs (interrarray correlation coefficients), among various combinations of three types of transformations (VST, log2, and cubicRoot) and six types of normalizations (SSN, quantile, RSN, VSN, RankInvarariant, and LOESS). The preprocessed data was further batch corrected using ComBat, as the experiments had been performed in multiple batches. The batch-corrected gene expression data were then analyzed for differential expression using limma (Supplementary Text S7). This method of transformation, normalization, and batch correction was performed as described previously (31).

Correlation of methylation and expression data
All probes were annotated using the Ensemble v75 database and classified on the basis of their locations to genic subregions, such as 5' and 3' untranslated regions (UTR), north and south shelves and shores, TSS1500, promoters, CpG islands, gene bodies, and first exons. The relationship between differential expression of genes and differential methylation of genic subregions was visualized (Supplementary Text S8).

Minimal differentially methylated loci using machine-learning algorithm
Sample-wise dasen-normalized and batch-corrected intensities of probes corresponding to DMRs, identified postfunctional normalization using minfi, along with the tumor-matched control tissue type for each sample was used as a training set (training set #1) for the random forest analyses.

Variable selection
The method employed here follows the method described before (32). Briefly, the algorithm performs both backward elimination of variables and selection based on the importance spectrum. The Bioconductor package VarSelRF (33) was used to perform this function. About 20% of the least important variables are eliminated iteratively until the current out-of-bag (OOB) error rate becomes larger than the initial OOB error rate or the OOB error rate in the previous iteration. We computed a metric by multiplying the prediction accuracy (sensitivity and specificity) and repeatability, and dividing the product by the number of parameters contained within the predicted set. The parameter set with the highest value for this metric. Five hundred of such iterations were performed, and each of the iterations was permuted across 3,000 random forest trees (Supplementary Text S9).

A score was computed for predictive parameter sets across all iterations, for each sample, which factored in the prediction accuracy of the tissue type of that sample, the repeatability (number of instances out of 500) of the predictive parameter set, presence of other predictive parameter sets, and the parameter set size. The first two weighed the score positively, and the latter two, negatively. The R commands used in our method are provided below:

```r
library(randomForest)
library(varSelRF)
DS.read.table("/home/neeraja/RF/OSCC/TminN.NoFlanks.forRF",header=T,na.strings="NA")
DS<-na.omit(DS)
for (i in 1:500) {
  set.seed(i)
  DS.rf.vsf=varSelRF(DS[,-1],DS[,1],ntree=3000,
nlooerat=2000,vars.drop.frac=0.2,whole.range = FALSE,keep.
  forest = TRUE)
  print(DS.rf.vsf$selected.vars)
  print(predict(DS.rf.vsf$rf$model,subset(DS[,-1],select=DS.rf.vsf
$selected.vars)))
}
```

For all iterations of all RF analyses, we confirmed that the rankings of variable importances remained highly correlated (Supplementary Table S3) before and after correcting for multiple hypotheses comparisons using pre- and post- Benjamin–Hochberg FDR-corrected \( P \) values (Supplementary Text S10). The R commands used to recompute importance values after multiple comparisons testing for the six somatic mutation category RF analyses are provided below:

```r
# Selected variables from varSelRF analyses
DS.rf.vsf$selected.vars
d.s.rf = randomForest(DS[,1] ~ ., DS, ntree = 3000, norm.votes =
FALSE, importance = TRUE)
#Importance values before FDR correction
d.s.rf$importance[-1,3]
```
The random forest algorithm was also implemented to predict tumor-specific minimal methylation signature, this time training with a categorical input for the probes, instead of the actual intensity values (training set #2). This training set was created on the basis of the normalized $\Delta F$ (tumor-matched control $B$) values falling into predefined categories. Following the $\Delta F$ density distributions across probes (Supplementary Fig. S2), for $-0.2 < \Delta F < 0.2$, bins were determined at an interval of 0.032, and for $\Delta F \geq 0.2$, the binning frequency was reduced to 0.08. For example, $\Delta F = 0.02$ for a probe corresponding to a tumor:matched control sample pair results in categories A for the tumor and B for the matched control, whereas $\Delta F = -0.02$ would result in reversal of the categories, that is, B for the tumor and A for the matched control. In addition to the above two DMP-based training sets, we created another categorical set (training set #3), based on hypo- or hypermethylation of DMRs housing only CpG island, TSS1500, or promoter-associated probes. In this scenario, DMRs housing heterogeneously methylated probes (i.e., a mixture of hyper- and hypomethylated probes) were discarded, and only the homogeneous DMRs (with all hypo- or all hypermethylated probes) were considered.

Validation of minimal methylation signature in TCGA HNSCC data

The best-scoring signatures identified from the random forest analyses on the OTSCC data, using three training sets, were examined in the 50 tumor:matched control pairs (various subsites in head and neck region) from the TCGA cohort. TCGA data were preprocessed and normalized exactly the same way as described above for the OTSCC data.

Prediction of minimal differential methylation signature in OTSCC for various clinical parameters

Various epidemiologic and clinical parameters were used for predictive analyses using the importance-based variable elimination by the random forest approach (24), where the $\Delta F$ of DMPs corresponding to DMRs were used as the training sets. Minimal differential methylation signatures with the best scores were visualized for patients with different states for the selected epidemiologic and clinical parameters.

Validation of NR4A3 and BCL2L11 expression using qPCR

Total RNA was isolated using RNeasy Mini Kits (Qiagen) following the manufacturers’ instructions with on column DNase treatment (Qiagen) to remove contaminating genomic DNA. The RNA integrity was analyzed by electrophoresis, and samples were preserved at $-80^\circ$C until further use. Total RNA (500 ng) was reverse transcribed using SuperScript III First-Strand Synthesis Kit, following the manufacturer's instructions (Thermo Fisher Scientific). qPCR was performed using KAPPA SYBR Fast qPCR Master Mix (Kapa Biosystems) on NR4A3 using primers: forward 5’-GGGTACGCAGCGTTCTTCTCAAA-3’, reverse 5’-TCGGTTTCGAGCTCTTGGTT-3’ and on BCL2L11 using primers: forward 5’-CAACACAAACCCAACTGTTCA-3’, reverse 5’-TCGTGGCGCCAATCAATCTGTC-3’. GAPDH was amplified in parallel using primers: forward 5’-TCGACAGCTGCGGCTATCCTT-3’, reverse 5’-GCGCAATACCCAAATCCGGTGA-3’ as an internal housekeeping control. qPCR was performed using initial denaturation at 95°C for 5 minutes, followed by 40 cycles of 95°C for 10 seconds, 60°C for 30 seconds. The amplification was followed by dissociation curve analysis using Stratagene MX3000P instrument (Agilent Technologies) and were analyzed by Agilent Mx3000P software (Agilent Technologies). All amplification reactions were done in triplicates, using nuclease-free water as negative controls. The analysis of gene expression data was performed using the comparative $C_t$ method of relative quantification as described by Schmittgen and Livak, 2008 (38).

Validation of miR-10B methylation using quantitative methylation-specific PCR

The genomic DNA was isolated from paired oral tongue tumor and the adjacent normal tissue samples using DNase treatment (Qiagen) following the manufacturer’s instructions. Genomic DNA (total 500 ng) was bisulfite converted using EZ DNA Methylation Kit (Zymo Research) following the manufacturer’s guidelines. For the specific amplification of CpG Islands of MIR10B promoter, methylation-specific primers (forward, 5’-GTGAGTATTTTCCGAGGATGC-3’ and reverse, 5’-CGAAC-TAACCTATAACCCGGAGC-3’) were designed using MethPrimer software (http://www.urogene.org/methprimer/). Quantitative methylation-specific PCR (MSP) was carried out using KAPA SYBR Fast qPCR Master Mix (Kapa Biosystems). Quantitative methylation analyses were carried out using the UCSC Genome Browser.

Visualization of variants

We visualized all the variants and the corresponding genes using Circos v0.66. We integrated information on LOHs, somatic SNPs and indels, and copy number insertions and deletions from 48 patients described previously (13) with $\geq10\%$ frequency of patients bearing the variants, genes undergoing significant expression, and methylation changes and visualized the changes using the circular genomic representation. The DNAseI hypersensitivity tracks, methylation, histone and chromatin modifications, and TFBs tracks were loaded and visualized using the UCSC Genome Browser. Genome-wide profiles of differential expression and differential methylation of DMRs housing island/TSS1500/promoter probes were visualized using GenomeGraphs, a tool in the UCSC Genome Browser.
methylation in each sample was normalized using the methylated primers for ACTB (forward, 5'-AACCATAAAACCCCTTCCTCCCTTAA-3' and reverse, 5'-TCCCTGGACGGGAGGTTTAGTAAGT-3'). The reaction mix was denatured initially at 95°C for 5 minutes and amplified for 40 cycles at 95°C for 10 seconds, 53°C for 30 seconds (59°C for ACTB primer), followed by extension at 72°C for 1 minute. The amplification was followed by dissociation curve analysis. PCR reactions were performed using Stratagene MX3000P (Agilent Technologies) and were analyzed by MxPro Mx-3000p software (Agilent Technologies). Human HCT116 DKO methylated and nonmethylated DNA set standards (Zymo Research) were used as positive and negative assay controls, respectively, for qMSP.

The relative level of methylated DNA was calculated for each tumor versus the adjacent normal using the method described by Schmitzgen and Livak, 2008 (38).

Differential expression analysis of NR4A3 in TCGA cohort

Level 3 RNA-seq expression data on TCGA HNSCC (N = 42 for all subsites and N = 12 for oral tongue, where expression data were available for tumor:normal pairs), along with survival and living status, were downloaded from the UCSC Cancer Browser (https://genome-cancer.ucsc.edu/proj/site/hgHeatmap/). Tumor-specific ratios of the normalized expression values for NR4A3 gene were log transformed before further analyses.

Survival analysis

Event-free survival probability distributions were estimated using the Kaplan–Meier method, and significance was assessed using the log-rank test. We calculated event-free survival for patients as the time from diagnosis to the event (recurrence, disease progression, or death) or last examination if the patient had no event. Significance was calculated as two-tailed P values, on association with event-free survival (months), of habits alone, NR4A3 expression status, and habits and NR4A3 expression together. The R commands used for Kaplan–Meier analyses are provided in Supplementary Text S12.

Results

Analysis workflow and differentially methylated loci in OTSCC

The analytic pipeline for genome-wide methylation and gene expression data analyses along with linking methylation with gene expression, machine-learning approach to discover methylation signature, and follow-up validation is provided in Fig. 1. The methylation data were normalized by multiple methods (see Materials and Methods), batch corrected, and then used for the identification of DMPs and regions encompassing multiple probes (DMRs) in the tumors. Scanning of the preprocessed data in tumor samples before normalization, batch correction, and pairing with corresponding normal tissues revealed DMPs in seven genes, UBE4B, CCDC13, LRPSL, BCL3, MIR4260, FOX2, and COL18A1, to be hypermethylated, and one in gene GSTM2, to be hypomethylated in nearly 95% of the tumors. We found more DMRs to be hypomethylated than hypermethylated (Fig. 2A), which is consistent with the data from the TCGA cohort, including when data from the oral tongue subsite alone was used (Fig. 2A). The fraction of hypermethylated probes located within the CpG islands, TSS1500, and promoter subregions, combined, were higher, whereas hypomethylated probes were more frequent in the gene bodies for OTSCC (Fig. 2B). The average Δβ of the island DMRs had a narrow distribution for the OTSCC and HNSCC (all and oral tongue subsites from the TCGA cohort), with median values around 0.30 and 0.40, respectively, indicative of greater hypermethylation events (Fig. 2C). The interquartile ranges for average Δβ distributions of TSS1500 DMRs spanned both hyper- and hypomethylation ranges, for the OTSCC and HNSCC cancer types, respectively. The median, however, was more indicative of hypermethylation in the OTSCC and hypomethylation in the HNSCC. The average Δβ distributions were identical for island DMRs between the OTSCC and the TCGA HNSCC cohort for early- and late-stage tumors, with their medians suggestive of hypermethylation (Fig. 2D). The Δβ distribution for TSS1500 DMRs in OTSCC was much wider in the late-stage patient samples, as compared with those in early stages of the disease and their medians (both ~0.3–0.4) were representative of hypermethylation. Overall, the
intraregion differences (within a dataset and across tumor stages) for the Δβ distributions were significant (P < 0.0001; unpaired t test) over the interregion differences across datasets for the same subregions (TSS1500, P = 0.0751 or CpG island, P = 0.2758). A comprehensive representation of all 5mC changes, along with previously identified somatic variants in OTSCC (37), is presented as a Circos graph (Supplementary Fig. S3; Supplementary Table S4).

**Correlation between expression and methylation**

We wanted to identify the functional relevance of 5mC changes by correlating methylation changes with changes in gene expression. To do this, we extracted gene level information using probes spanning hyper- or hypomethylated regions that were up- or downregulated, respectively, for the same patient. First, we separated probes in gene bodies, CpG islands, 5' and 3' UTRs, north and south shelves and shores, promoters, 1.5 kb flanking the transcription start sites (TSS1500), and first exons. For gene expression analysis, whole-genome expression data were processed posttransformation; normalization and batch correction (see Materials and Methods) and tumor-specific up- and downregulated genes were considered for comparison. We found a maximum of 27 genes satisfying these criteria in the gene bodies, closely followed by 26 genes each, in the CpG islands and 5' UTR and 22 genes in the first exon (Supplementary Fig. S4). Finally, we obtained 12 genes across all the regions where the differential methylation across all tumors was significantly linked (2-tailed t test, P < 0.05) with gene expression changes in the same tumor:normal paired samples (Fig. 3). The magnitude of correlation between expression and methylation varied widely across genes for various subregions (Fig. 3 and Supplementary Fig. S4), with a significant negative correlation (P < 0.05, two-tailed t test) observed for most subregions: 5' UTR, R = −0.51; promoter, R = −0.5; TSS1500, R = −0.56; body, R = −0.35; S shelf, R = −0.39; first exon, R = −0.46; S shore, −0.66; and N shelf, R = −0.36. The island region harbored the largest fraction of hypermethylated and downregulated genes, whereas the promoter and TSS1500...
Figure 3.
Region-wise correlation between expression and methylation. The relationship between preprocessed differential methylation (tumor MINUS normal) values for probes from specific regions, such as gene bodies, CpG islands, TSS1500, N shore, S shore, N shelf, S shelf, 5' UTR, and first exon, and the log2 fold change (tumor/normal fold change) of the corresponding gene expression for the same patient. Blue and orange colors, hypo- and hypermethylation, respectively. Filled circles indicate a correlation in the expected direction, that is, hypermethylation and downregulation or hypomethylation and upregulation. Empty circles indicate a counterintuitive correlation, namely, hypermethylation and up-regulation or hypomethylation and downregulation. The sizes of the circles indicate the magnitude of gene expression. Only genes with significant associations (P < 0.05) across all tumors are shown.
subregions displayed a strong inverse relationship between differential expression and differential methylation (Supplementary Fig. S4). This observation formed a basis for considering DMRs housed in these three subregions as one of our training sets to identify a minimal methylation signature.

Tumor-specific DMPs and DMRs

We used three different training sets (Materials and Methods section) to discover minimal methylation signature using variable selection by elimination analyses implemented by the random forest approach. The scores for the discovered signatures are presented in Supplementary Table S5. We found two sets of two probes in four different DMRs (encompassing GPER1 and OR2T6, TTLL8 and RHPN1 genes), each to be the best diagnostic predictor for tumors (Fig. 4A, top) with training set #1 (see Materials and Methods). These probes along with other loci housed within the same DMRs showed a hypomethylation signature for most tumor samples when compared with the normal samples (Fig. 4A). The signature was strongest for the predicted probe (DMP) within each DMR. Using the categorical training set based on binning Δβ of paired tumor and matched control samples (training set #2; see Materials and Methods), we further identified three more DMRs that encompassed MIR10B, ENAH, and EMX2OS genes (Fig. 4A). For MIR10B and EMX2OS, probes were distinctly hypermethylated for the majority of tumors (94% and 88%, respectively), while ENAH exhibited a mixed trend. Analyzing homogeneous DMRs in CpG islands/TSS1500 and promoters, alone (training set #3; see Materials and Methods), we obtained DMRs encompassing 10 genes (BTK, C11orf53, CENPVIL1, COL9A1, DEFA1, ODF3, RARRES2, RHPN1, SHF, and SP6) with the highest sensitivity and specificity (Fig. 4A). We found RHPN1 gene showed up in multiple training sets. The DMRs corresponding to BTK, C11orf53, COL9A1, DEFA1, ODF3 along with RHPN1 were hypomethylated, while the remaining were relatively hypermethylated. The results from our 0.632+ bootstrapping errors were low, with a median error of approximately 0.025, for predictions of normal, tumor, or either tissues (Supplementary Fig. S5) indicative of a high overall prediction accuracy.

We further validated the tumor-specific DMRs discovered in our RF analysis in the TCGA cohort. As presented in Fig. 4B (bottom), the differential methylation trends observed in our OTSCC cohort were also observed in both oral tongue and other subsites in the TCGA HNSCC cohort, except in the case of ENAH gene, where the majority of tumors were hypomethylated for the particular probe in the TCGA cohort (Fig. 4B; training set #2). In addition, we also looked at the cell line data from ENCODE consortium that revealed DNAse I hypersensitivity and/or CpG methylation in these genes, including specific locations within the DMRs housing the tissue type predictive

Figure 4.

Discovery of a minimal DNA methylation signature. A, minimal differential methylation profiles of methylation loci from random forest analyses using three different training sets (see Materials and Methods). B, validation of specific DMPs and DMRs using the same training sets in the TCGA HNSCC data.
probes (Supplementary Fig. S6), possibly indicating the role of these CpG sites in regulation of gene expression.

**Hypermethylation of MIR10B and downregulation of its target gene expression**

We scanned upstream and downstream of the DMR that housed MIR10B gene in an attempt to understand its differential methylation relative to its neighborhood. We found a DMR of nearly 1.31 MB to be differentially methylated in all the tumor samples (Fig. S5A). We found that the sizes of such functional DMRs discovered for the MIR10B gene to be exaggerated relative to the stringent DMR boundaries determined by miniFL, also for ENAH and EMX2OS, codiscovered with MIR10B, as the minimal set (Supplementary Fig. S7).

We observed negative correlations between the differential methylation of MIR10B gene and the differential expression of two of its validated gene targets, BCL2L11 and NR4A3 ($P = 0.0125$ and $P = 0.014$, respectively; Fig. 5B and C). NR4A3 expression was downregulated in all except one tumor studied (>95% of the tumors). In addition to mir-10b hypermethylation in 60% of the tumors, and very frequent downregulation of NR4A3, we found that 50% to 60% of the tumors have hypermethylation in the NR4A3 gene itself, near the first exon and first intron, spanning a CpG island (Supplementary Fig. S8). Similarly, about 60% of the tumors had hypermethylation in the promoter region of BCL2L11 gene, also spanning a CpG island (Supplementary Fig. S8). We found the extent of NR4A3 downregulation was correlated, although weakly, in patients with no risk habits than those patients with at least one of the risk habits, smoking, alcohol consumption, and chewing tobacco. The tumor-specific differential expression of BCL2L11 and NR4A3 genes was inversely correlated with disease-free survival ($P = 9E−15$ and $P = 2E−15$, respectively; Fig. 5D and E), with absence of risk habits associating more with improved disease-free survival.

We validated the inverse association between differential methylation of MIR10B and differential expression in its downstream targets, NR4A3 ($R = 0.66; P = 0.04$) and BCL2L11 ($R = 0.78; P = 0.01$), experimentally by performing qMSP for MIR10B and qPCR for BCL2L11 and NR4A3 (see Materials and Methods), in 12 additional tumor:matched control pairs (Supplementary Fig. S9). We found that MIR10B is hypermethylated in nearly 60% of the samples, with 83% and 42% of the tumors showing downregulation for NR4A3 and BCL2L11, respectively (Supplementary Fig. S9), validating the findings from the arrays.

**NR4A3 expression and patient survival**

We studied the effect of MIR10B target gene expression on patient survival and found that the extent of NR4A3 downregulation was correlated, although weakly, in patients with no risk habits than those patients with at least one of the risk habits, smoking, alcohol consumption, and chewing tobacco. The tumor-specific differential expression of BCL2L11 and NR4A3 genes was inversely correlated with disease-free survival ($P = 9E−15$ and $P = 2E−15$, respectively; Fig. 5D and E), with absence of risk habits associating more with improved disease-free survival. In our discovery set, we found 95% of the tumors with NR4A3 downregulation that we validated in 83% and 42% of the tumors in an independent validation set using qPCR and by using data from TCGA oral tongue cohort (Supplementary Fig. S10). We extended our earlier observations by performing Kaplan–Meier survival analyses on the effects of habits alone, NR4A3 downregulation alone, or their combined effects on survival. We observed that the habits alone do not have significant impact on patient survival (Fig. 6A, $P = 0.494$; Fig. 6D, $P = 0.283$), whereas NR4A3 expression does (Fig. 6B, $P = 0.034$; Fig. 6E, $P = 0.019$). We further observed that the combination of no risk habits and NR4A3 expression has a significant effect on survival (Fig. 6C, $P = 0.046$ and 6F, $P = 0.021$). Overall, patients without habits (habits negative) expressed NR4A3 at a lower level compared with those with habits (Fig. 5C).

**DMRs predicting various epidemiologic and clinical parameters**

Differential methylation (Δβ) for all tumors used as training sets along with five parameters, namely HPV infection, nodal status, risk habits, tumor stage, and prognosis, resulted in best-scoring DMPs corresponding to TSPAN2, FUT3, TRIM5, SLC9A9, NPSA3, RPS6KA2, MAP3K8, TIMM8A, and RNF113A genes. The set of predictive DMPs displayed contrasting patterns of differential methylation among the various categories of each parameter (Fig. 7).

**Discussion**

Epigenetic changes like DNA methylation play important roles in cancer initiation and progression. One of the first steps in studying DNA methylation is to catalog the loci where methylation changes take place and link those to the disease. Previous studies in head and neck cancer were primarily aimed at identifying 5mC changes in a handful of loci in the genome or using low-throughput methylation arrays. In the current study, we have used >482,000 probes to profile genome-wide changes in 5mC patterns in oral tongue tumors (OTSCC) and defined functional regions of importance by comparing DNA methylation changes with those in gene expression. We term these regions of the genome as functional DMRs. We also linked global DNA methylation changes with various clinical and epidemiologic parameters in OTSCC. Our observation of the presence of genome-wide hypomethylated regions alongside promoter hypermethylation (Fig. 2) corroborates previous study in other cancers (19). Although the role of global hypomethylation is unclear, several animal experiments using DNA methylation inhibitors (39, 40) are indicative of its involvement in oncogenesis and tumor progression. The hypermethylation of CpG islands observed in our study (Fig. 2B–D) was supportive of the theory that CpGs located within islands in normal cells are relatively unmethylated, whereas those located outside of the islands are more methylated, with the reverse pattern occurring in tumor cells (41). We observed a shift of the global hypermethylation in TSS1500 toward a relative hypomethylation, from early to advanced stages of the disease (Fig. 2D), a trend also found in hepatocellular carcinoma (42). We found some uniquely differentially methylated loci in our study; different from other studies so far, a possible indication of specific epigenetic marks linked with our cohort and patient geography as the epigenome is known to be dynamic and highly dependent on environmental and dietary factors (43).

In our study, we define the importance of functional and differentially methylated regions or DMRs in tumors based on
two factors. First, we identified genes that were encompassed in DMRs by correlating differential methylation in a number of probes within specific genomic loci, especially in the Cpg islands, promoter, and TSS1500, with the differential expression in these genes, between the same tumor:matched normal pairs, and estimating the significance of the association (2-tailed t-test, \( P < 0.05 \)) between those two events (Fig. 3). Second, we identified the entire stretch of DMPs and assigned importance to all the genes in that region. It is vital to assign function to the entire DMR, rather than a single probe or a gene housed within it, as the probe hybridization patterns vary widely, even within a given subregion, and methylation changes are not necessarily localized single-gene events. Both these criteria led us to identify functional DMRs. A genome-wide representation of such functional DMRs is shown in Supplementary Fig. S11. We found that methylation changes at certain regions do not go hand-in-hand with gene expression changes in the expected direction (i.e., hyper- and hypomethylated loci giving rise to up- and downregulation of gene expression, respectively. Supplementary Fig. S4). Even though such phenomena were previously reported (44), their significance is currently not known.

Our analyses using an ensemble machine-learning algorithm followed by estimation of errors picked several genes predictive of clinical and epidemiologic factors (Fig. 7). Several of those are shown in the past to have a role in cancer. For example, TSPAN7, a tetraspan family member upregulated in multiple myeloma patients (45) in human epithelial malignancies (46), FUT3 in gastric cancer (47), SLC9A9, overexpressed in most cancers (http://www.proteinatlas.org/ENSG00000181804-SLC9A9/cancer), NPA3, a neuronal PAS transcription factor family member as a tumor suppressor in astrocyomal progression (48), RP56K2A, a ribosomal kinase and as a putative tumor suppressor in sporadic epithelial ovarian cancer (49) and a modifier of the EGFR signaling pathway in pancreatic cancer (50), and MAP3K8 as a proto-oncogene that plays a role in lung (51), prostate (52), and colorectal (53) cancers.

We identified a signature containing differentially methylated loci encompassing 16 genes using a machine-learning algorithm using three different training sets (Fig. 4A) that was further validated using data from TCGA cohort (Fig 4B). Although the machine-learning algorithm identified DMPs, we could identify the functional DMRs encompassing these DMPs (Fig. 5A and Supplementary Fig. S12). Many of the genes within those DMRs are cancer-associated genes. Among those, GPER1 (G-protein–coupled estrogen receptor 1) is known to act as a tumor suppressor in ovarian cancer (54) and has low expression in breast cancer (55), RHPS1 (rhophilin, Rho GTPase-binding protein 1) is reported to be highly expressed in most cancers (http://www.proteinatlas.org/ENSG00000158106-RHPS1/cancer) and CpG islands adjacent to the RHPS1 gene is reported to be linked with survival of colorectal cancer patients (56). In addition, TITL8 (tubulin tyrosine ligase-like family, member B) gene is differentially methylated in diseases such as myelodysplastic syndrome (57), and OR2T6 shows copy number gains in many cancers, including HNSCC (http://ow.ly/RniAf).

**Figure 5.**
Differential methylation of **MIR10B**, linking it to the expression of its downstream target genes and disease-free survival. **A**, median differential methylation of chromosomal regions flanking MIR10B gene. The actual scatter is in gray, exponential smoothing of the scatter in red, and MIR10B is highlighted in blue. **B** and **C**, correlation between methylation (MIR10B) and expression (NR4A3 and BCL2L11). **D**, fold change. **E** and **F**, correlation of NR4A3 and BCL2L11 expression with disease-free survival. Green filled circles, patients without any risk habits; red filled circles, patients with at least one risk habit (smoking, chewing tobacco, or alcohol).
associate survival with the expression of these genes linked with MIR10B hypermethylation. Therefore, we combined the discovery and validation sets to accurately ascertain such a relationship.

We found that NR4A3 was downregulated in a large fraction of tumors, both in the discovery set as well as an independent validation cohort assayed using qPCR and in TCGA sample cohort (Supplementary Fig. S10) and that the downregulation of this gene was significantly linked with better survival (Fig. 6).

Kaplan–Meier survival analyses revealed that the downregulation of NR4A3 expression alone (Fig. 6B and E), but not the absence of habits (Fig. 6A and D), is sufficient for providing...
survival advantage to the patients. It is possible that habits upregulate NR4A3 expression (as most habits positive patients have lowered extent of NR4A3 downregulation; Fig. 5C), thereby working as a negative factor for survival (Fig. 6C and F). However, future work is needed to define mechanistic details of the mode of action and the implications of the key genes like NR4A3 in OTSCC biology.

In summary, we have identified a set of differential methylation signatures with functional relevance in oral tongue tumors. The association between the differentially
methylated signatures with clinical and epidemiologic parameters points to their potential application in diagnosis and disease stratification. In addition, our data on NRAS present potential avenues of exploiting the molecule as a target for therapeutic intervention in head and neck cancer.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors’ Contributions

Conception and design: N.M. Krishnan, A. Suresh, M.A. Kurikakose, B. Panda
Development of methodology: N.M. Krishnan, B. Panda
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): K. Dhas, J. Nair, V. Palve, J. Bagawan, G. Siddappa, V.D. Kekatpure, M.A. Kurikakose
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): N.M. Krishnan, B. Panda

References


A Minimal DNA Methylation Signature in Oral Tongue Squamous Cell Carcinoma Links Altered Methylation with Tumor Attributes

Neeraja M. Krishnan, Kunal Dhas, Jayalakshmi Nair, et al.