Das *et al.*
European Journal of Medical Research (2024) 29:458
https://doi.org/10.1186/s40001-024-02047-4

European Journal
of Medical Research

RESEARCH Open Access

Universal penalized regression (*Elastic-net*) model with differentially methylated promoters for oral cancer prediction

Shantanab Das[1], Saikat Karuri[1], Joyeeta Chakraborty[1], Baidehi Basu[1], Aditi Chandra[1,2], S. Aravindan[5], Anirvan Chakraborty[3], Debashis Paul[1,4], Jay Gopal Ray[5], Matt Lechner[6], Stephan Beck[6], E. Andrew Teschendorff[6,7] and Raghunath Chatterjee[1*]

Abstract

Background DNA methylation showed notable potential to act as a diagnostic marker in many cancers. Many studies proposed DNA methylation biomarker in OSCC detection, while most of these studies are limited to specific cohorts or geographical location. However, the generalizability of DNA methylation as a diagnostic marker in oral cancer across different geographical locations is yet to be investigated.

Methods We used genome-wide methylation data from 384 oral cavity cancer and normal tissues from TCGA HNSCC and eastern India. The common differentially methylated CpGs in these two cohorts were used to develop an *Elastic-net* model that can be used for the diagnosis of OSCC. The model was validated using 812 HNSCC and normal samples from different anatomical sites of oral cavity from seven countries. Droplet Digital PCR of methyl-sensitive restriction enzyme digested DNA (ddMSRE) was used for quantification of methylation and validation of the model with 22 OSCC and 22 contralateral normal samples. Additionally, pyrosequencing was used to validate the model using 46 OSCC and 25 adjacent normal and 21 contralateral normal tissue samples.

Results With ddMSRE, our model showed 91% sensitivity, 100% specificity, and 95% accuracy in classifying OSCC from the contralateral normal tissues. Validation of the model with pyrosequencing also showed 96% sensitivity, 91% specificity, and 93% accuracy for classifying the OSCC from contralateral normal samples, while in case of adjacent normal samples we found similar sensitivity but with 20% specificity, suggesting the presence of early disease methylation signature at the adjacent normal samples. Methylation array data of HNSCC and normal tissues from different geographical locations and different anatomical sites showed comparable sensitivity, specificity, and accuracy in detecting oral cavity cancer with across. Similar results were also observed for different stages of oral cavity cancer.

Conclusions Our model identified crucial genomic regions affected by DNA methylation in OSCC and showed similar accuracy in detecting oral cancer across different geographical locations. The high specificity of this model in classifying contralateral normal samples from the oral cancer compared to the adjacent normal samples suggested applicability of the model in early detection.

Keywords DNA methylation, Linear regression techniques, HNSCC, Droplet digital PCR, Methylation-sensitive restriction enzyme, Oral squamous cell carcinoma

*Correspondence:
Raghunath Chatterjee
rchatterjee@isical.ac.in
Full list of author information is available at the end of the article

Background

Oral cancer is one of the most common malignancies in Southeast Asia, accounting for up to 30–40% of all malignancies in India [1]. Several oral habits like smoking, chewing tobacco, alcohol, etc., are attributed to the development of Oral squamous cell carcinoma (OSCC). According to the latest GLOBOCAN Statistics, oral cancer is the 2nd most common cancer in India, being the most common among males and 4th most common among females in India [2]. OSCC is one of the most common malignant epithelial neoplasia within the oral cavity [3]. Despite the significant improvements in therapeutic modalities in OSCC, 5 year survival rates are among the lowest of the major cancers and the main reason being the lack of early detection.

The study of DNA methylation has unraveled its role in several fundamental biological processes, like genomic imprinting, activation or silencing of transposons, cell-differentiation, development, etc., signifying the importance of robust regulation of DNA methylation for normal cellular processes. Altered DNA methylation levels have been associated with multiple cancers [4–8].

Aberrant methylations in the promoter region of several genes, including the cell cycle, DNA repair, apoptotic, and tumor suppressor genes, are reported in OSCC [9]. Some recent studies reported differential methylation in OSCC and OPMD (Oral potentially malignant disorder) tissues and evaluated their potential to be used as biomarkers for OSCC detection [10–14]. However, most of them are limited in terms of universality as they are generally not replicated in other cohorts of OSCC patients. The major challenge thus remains in developing a universal set of biomarkers that can be used for detecting OSCC irrespective of the geographical region who might have different etiological factors. In our previous study, we conducted a genome-wide DNA methylation study of OSCC and adjacent normal tissues in patients from India [15]. On comparison of our data with the TCGA HNSC methylation data, we identified a set of hypomethylated and hypermethylated CpGs that were common between these two datasets [15]. These common differentially methylated CpG probes (DMPs) may play a more fundamental role in oral cancer development and should be of primary focus while detecting DNA methylation biomarkers in oral cancer.

Here, we used penalized (regularized) linear regression techniques with these common DMPs and predicted a model using promoter methylation levels, which can be used for detecting oral cancer. Our approach for identifying the common differential promoters relies on a widely used feature selection procedure *Elastic-net* [16]. We selected the tuning parameters for the *Elastic-net* procedure in a data-driven manner by using the principles of cross-validation and aggregation. We validated this model in additional paired OSCC and contralateral normal tissue samples using a method called Droplet Digital PCR amplification of the methyl-sensitive restriction enzyme digested DNA (ddMSRE). To elucidate the potential of these differentially methylated promoters, we used contralateral and adjacent normal samples. Using pyrosequencing, we validated the model for OSCC lesion, adjacent normal, and contralateral normal tissues [17]. Furthermore, our model showed promising results in predicting OSCC with the publicly available methylation array data comprising 812 HNSCC (Head and Neck squamous cell carcinoma) and normal samples, including 131 FFPE samples.

Methods

Illumina infinium humanMethylation450K beadChip array data

For systemic investigation, we have used Illumina Infinium HumanMethylation450K BeadChip array datasets from 982 HNSCC and 214 normal samples. For development of a prediction model, we have used 324 HNSCC and 60 normal samples, and for validation and performance assessment of our predicted model, we have used 658 HNSCC and 154 normal samples (Fig. 1A). For model development, we have used 450 K BeadChip array data of 11 OSCC and 10 adjacent normal samples from our previous study (GSE87053) [15], and 313 HNSCC and 50 normal samples from the TCGA database (https://www.cancer.gov/tcga). For validation, we have downloaded 812 publicly available 450 K BeadChip array data (.IDAT files) of 658 HNSCC and 154 normal samples from the TCGA and 14 GEO databases (GSE123781, GSE79556, GSE38266, GSE67114, GSE75537, GSE41114, GSE97784, GSE52068, GSE204943, GSE178216, GSE178219, GSE38268, GSE136704, and GSE62336) [18–30]. The TCGA and GEO datasets used for validation included tumors from oral cavity, oropharynx, hypopharynx, larynx, and nasopharynx areas from Australia, Germany, the United Kingdom, India, the USA, China, and Brazil (Table 1, Additional file 1).

Patient selection and sample collection

After clinical inspection, patients with a provisional diagnosis of OSCC were recruited for the study with their informed written consent. Histopathologically confirmed 25 well-differentiated squamous cell carcinoma patients were recruited in this study. OSCC and adjacent normal tissues from the 1 cm periphery of the visible tumor border were collected using incisional and 3-mm punch biopsy. A portion of the tumor tissue was collected in the formalin for histopathological analysis. The other portion and adjacent normal tissues were collected in RNA Later

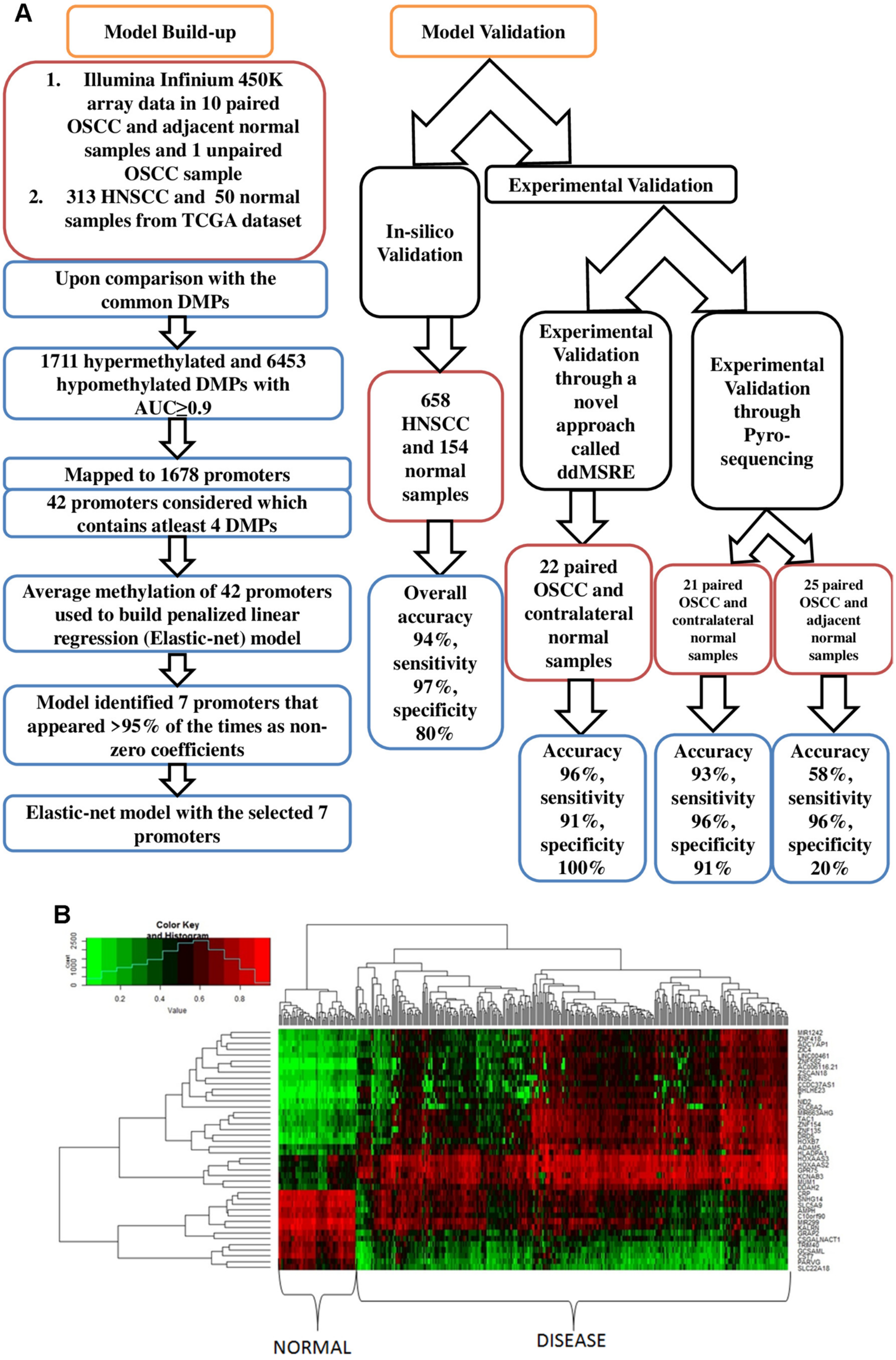
A
Model Build-up
1. Illumina Infinium 450K array data in 10 paired OSCC and adjacent normal samples and 1 unpaired OSCC sample
2. 313 HNSCC and 50 normal samples from TCGA dataset
Upon comparison with the common DMPs
1711 hypermethylated and 6453 hypomethylated DMPs with AUC≥0.9
Mapped to 1678 promoters
42 promoters considered which contains atleast 4 DMPs
Average methylation of 42 promoters used to build penalized linear regression (Elastic-net) model
Model identified 7 promoters that appeared >95% of the times as non-zero coefficients
Elastic-net model with the selected 7 promoters
Model Validation
In-silico Validation
Experimental Validation
658 HNSCC and 154 normal samples
Overall accuracy 94%, sensitivity 97%, specificity 80%
Experimental Validation through a novel approach called ddMSRE
Experimental Validation through Pyro-sequencing
22 paired OSCC and contralateral normal samples
21 paired OSCC and contralateral normal samples
25 paired OSCC and adjacent normal samples
Accuracy 96%, sensitivity 91%, specificity 100%
Accuracy 93%, sensitivity 96%, specificity 91%
Accuracy 58%, sensitivity 96%, specificity 20%
B
Color Key and Histogram
Count
0
1000
2500
0.2
0.4
0.6
0.8
Value
MIR1242
ZNF418
ADCYAP1
ZIC4
LINC00461
ZNF582
AC006116.21
ZSCAN18
INSC
CCDC37AS1
BHLHE23
T
NID2
SLC6A2
MIR663AHG
TAC1
ZNF154
ZNF135
DRD5
HOXB7
ADAM5
HLADPA1
HOXAAS3
HOXAAS2
GPR75
KCNAB3
MUM1
DDAH2
CRP
SNHG14
SLC5A9
AMPH
C10orf90
MIR299
KALRN
GRAP2
CSGALNACT1
TRIM40
GCSAML
CST7
PARVG
SLC22A18
NORMAL
DISEASE

Fig. 1 **A** A schematic representation of the study design. **B** Heatmap showing the average methylation (*β*-value) of 42 promoters among 324 HNSCC and 60 normal samples. Hierarchical clustering shows classification of differentially methylated CpG probes (DMPs) for oral cancer and normal tissues

Table 1 Validation of the model using publicly available HNSCC methylation data

Site	Sample Status		Predicted								
	Disease	Normal	TP*	FP*	TN*	FN*	PLR*	NLR*	SN*	SP*	ACC*
Oral cavity	314	87	308	20	67	6	4.27	0.02	98.09	77.01	93.52
Oropharynx	167	18	155	6	12	12	2.78	0.11	92.81	66.67	90.27
Larynx	118	NA	117	NA	NA	1	NA	NA	99.15	NA	NA
Hypopharynx	10	NA	10	NA	NA	0	NA	NA	100	NA	NA
Nasopharynx	49	49	48	40	9	1	1.2	0.11	97.96	18.37	58.16

(Thermo Fisher) and stored at − 80 C until processing. Apart from the paired OSCC and adjacent normal tissues, contralateral normal samples were also used for the experimental validation of the model. DNAs of the 43 OSCC and contralateral normal samples were obtained from the samples reported in our previous study [15]. This study was approved by the Institutional Ethics Committee of the Indian Statistical Institute, Kolkata, India.

HumanMethylation450K BeadChip array data analysis

The array data (.IDAT files) downloaded from the TCGA and GEO databases were analyzed using RnBeads [31]. The β values, which are the ratio of the methylated probe intensity to the overall probe intensity (sum of methylated and unmethylated probe intensities plus constant α, where $\alpha=100$), for each DMP represent the methylation status of the corresponding CpG. Receiver operating curve (ROC) analysis was performed for common DMPs with SPSS Statistics v18.00 (IBM). Bedtools were used to identify the DMPs that overlapped with the promoters (− 2000 bp to TSS) of all genes annotated in the human genome (hg19) [32]. The average methylation of 42 promoters was calculated for all 812 validation samples (Additional file 2: Table S2). Additionally, the model was validated using two experimental approaches with completely different patient cohorts: first, using a newly developed method ddMSRE with 22 paired OSCC and contralateral normal tissue DNA samples [15], and using pyrosequencing with 46 paired OSCC samples. Among these 46 paired samples, we used 21 paired OSCC and contralateral normal samples, whereas the rest of the 25 paired samples consisted of OSCC and adjacent normal tissues. The demographic details of all these samples are presented in Additional file 3: Tables S3a, S3b, and S3c.

Feature selection and building predictive model using *Elastic-net*

Regularization techniques use penalties to control the complexity of a model. L1 norm and L2 norm are used to control the least square or least absolute errors between the observed and predicted target values. LASSO regression model [33] uses the L1 regularization technique, i.e., it imposes a penalty on the sum of the absolute values of the regression coefficients [34]. In contrast, ridge regression uses L2 regularization techniques and adds a penalty on the sum of the squares of the coefficients [35]. The L1 penalization encourages a few coefficients to be non-zero in the fitted model, thereby attaining the variable selection along with estimation. In contrast, L2 penalization is used to obtain better control on the risk characteristics of the resulting estimator. A hybrid method, such as *Elastic-net*, is a regularized regression model that uses a linear combination of the L1 penalty term from LASSO and the L2 penalty term from ridge regression [16, 35]. The *Elastic-net* procedure was proposed with the aim of obtaining a better bias-variance trade-off than what would be possible through either L1 or L2 penalized regression, while also achieving a good variable selection performance. Assuming that there are p features or predictors, the corresponding optimization problem in the regression setting can be expressed as follows:

$$min_{\beta_0,\beta_1}\left[\sum_{i=1}^{N}\left(y_i-\beta_0-X_i^T\beta\right)^2+\lambda P_\alpha(\beta)\right],$$

where y_i corresponds to the disease outcome for the *i-th* patient. X_i is a vector of features (average promoter methylation) for the *i-th* patient. The βs are regression coefficients that we estimate. When the response variable y_i is binary (1 or 0 corresponding to OSCC and normal, in our setting), there is a version of *Elastic-net* that replaces the squared error loss in the above equation with the negative log-likelihood loss under a logistic regression model, assuming that the logit transformation of the conditional probability of the diseases outcome logit[Pr($y=1$|features)] is a linear function of the features [35].

The tuning parameter λ is the weight of the regularization term and is chosen to minimize the mean square error. The regularization term $\mathrm{P}_\alpha(\beta)$ is given by the following:

$$P_\alpha(\beta) = \sum_{j=1}^{|x_i|} \left(\frac{1-\alpha}{2}\beta_j^2 + \alpha\left|\beta_j\right| \right)$$

Here, α is a number between 0 and 1 with $\alpha=0$ corresponding to ridge regression, $\alpha=1$ corresponding to LASSO, and $\alpha=0.5$ corresponding to *Elastic-net.* The parameter λ is non-negative and controls the level of the penalty, with larger λ leading to sparser (in the sense of having fewer features with non-zero estimated coefficients) models, but also larger bias in the estimates. Thus, there is a need for specifying this parameter in a data-driven manner so that the resulting regression model has a satisfactory predictive performance.

For our current problem, *Elastic-net* regression was used to concurrently select features and produce a linear regression model for predicting the model using average promoter methylation of 42 genes. Following the principle of cross-validation [36], we randomly sampled 80% of the data as a training dataset, while the remaining 20% data were used as the validation dataset. Model fitting and selection of the tuning parameter λ was done based on the training dataset. To determine the value of λ, we performed a tenfold cross-validation using the training dataset for λ value that gives the minimum mean cross-validation error. Cross-validation folds of the training set were randomly assigned with respect to the balance of classes within the training set and were repeated 100 times. The minimum value of the selected λ across 100 searches was subsequently defined as the best λ. With this best λ, the model was fitted using the training dataset and features with non-zero coefficient were recorded. The remaining 20% validation samples were only used for calculating the mean error rate and recording the non-zero coefficient. The above process (involving the 80–20 random splitting of the data) was run 500 times and the features that were selected (i.e., with non-zero coefficient) more than 95% of the times were used as potential biomarkers for classifying the disease and normal tissues. This aggregation of the results, based on sub-sampling of the data, ensures that the feature selection procedure is robust to random fluctuations of the data. This resulted in the retention of 7 features which were used as the predictors in the final model. We used a cross-validated *Elastic-net* procedure with these 7 features to fit the final model with the full dataset.

The final model was validated using 812 publicly available 450 K BeadChip array data of 658 HNSCC and 154 normal samples. Sensitivity (SN), specificity (SP), accuracy (AC), Positive Likelihood ratio (PLR), and Negative Likelihood ratio (NLR) of the validation set were calculated for each anatomical site using true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions by the predicted *Elastic-net* model.

Droplet digital PCR amplification of the methyl-sensitive restriction enzyme digested DNA (ddMSRE)

To establish the reliability of the prediction model, we used an approach for exact quantification of methylation level and named it as Droplet Digital PCR amplification of the methyl-sensitive restriction enzyme digested DNA (ddMSRE). To establish and validate ddMSRE as a reliable tool for quantification of methylation, we generated methylation standards of different methylation levels. We cloned a 296 bp human DNA (chr8: 19540274–19540569; promoter of *CSGALNACT1* gene), having 49% GC-content into a pCpGL vector (InvivoGen, USA; Catalogue No. pcpgf-promlc) where the vector backbone is devoid of any CG sites. The DNA region was PCR amplified with primers containing BamHI and HindIII restriction enzyme cut sites at their 5' ends (Additional File 4: Table S4). Both the insert and the vector backbone were digested with BamHI and HindIII (NEB) followed by ligation using T4 DNA Ligase (NEB) and transformation into *E.coli*-GT115 *pir* strain (InvivoGen, USA). Zeocin-resistant-positive colonies were verified by colony-PCR and were cultured for plasmid isolation. The cloning was further confirmed by Sanger sequencing in 3100 Genetic Analyzer (Applied Biosystems, California––USA). Around 1 µg of construct containing 296 bp CSGALNACT1 promoter (CSP-pCpGL) was subjected to in vitro methylation using CpG methyltransferase M.SssI (NEB) in the presence of S-adenosylmethionine (SAM) (NEB). To verify the completion of methylation, the methylated constructs were digested with methylation-sensitive restriction enzymes and run on 1% Agarose gel. The absence of any digested fragment ensured the complete methylation (Additional File 8: Figure S1). Finally, methylated and unmethylated CSP-pCpGL were mixed in different proportions according to the copy number to achieve methylation levels of 0%, 20%, 40%, 60%, 80%, and 100%, respectively.

Methylation-sensitive restriction enzymes (MSRE) digestion of genomic DNA and vectors

CSP-pCpGL with different methylation proportions and 1 µg of genomic DNA samples from the paired tissue samples were digested using 10 units of four methylation-sensitive restriction enzymes (MSREs), namely, AciI, HpaII, HpyCH4IV, and Hinp1I (NEB) in a 30 µl reaction volume for 1 h at 37 C. MSRE digested CSP-pCpGL along with the digested genomic DNA was further evaluated through ddPCR to validate the consistency of this newly developed method and then to determine the methylation status of 7 promoters.

Droplet digital PCR of MSRE digested DNA

Droplet Digital PCR enables absolute quantification of nucleic acid target sequences by counting the nucleic acid molecules that are encapsulated in discrete water-in-oil droplets. Using the QX200 Droplet Digital PCR (Bio-Rad), at first, we validated its reliability as a potential novel tool for quantifying the exact methylation status and secondly, we evaluated the methylation status of selected promoters in paired tissue samples of OSCC and normal individuals. The ddPCR reaction mixture consisted of template DNA, 1X QX200 ddPCR EvaGreen super-mix (Bio-Rad), 500 nM of each primer (Additional File 4: Table S4), and nuclease-free water up to a final volume of 20 µl. For genomic DNA, we used DNA that has ~ 8000 copies (30 ng), and for plasmids, we used ~ 10,000 copies (0.00003 ng) as template DNA. The 20 µl reaction mixture and the 70 µl of QX200 Droplet Generation Oil were loaded gently into the DG8 Droplet Generator Cartridge. The loaded cartridge was covered with a DG8 Gasket and placed in the QX200 Droplet Generator. The droplet generator uses microfluidics and partitions around 20,000 nanoliter-sized water-in-oil droplets where each droplet represents a single polymerase chain reaction (PCR) with or without a single template DNA molecule. Around 40 µl of nanoliter-sized water-in-oil droplets was transferred to a 96-well plate and sealed with PX1 PCR Plate Sealer (Bio-Rad) followed by PCR amplification using a T100 Thermal Cycler (Bio-Rad). Depending on the fluorescence amplitude of the droplets, distinct positive fractions were defined, and quantification of nucleic acids was performed using Poisson distribution [37]. The droplets were read by QX200 Droplet Reader following the manufacturer's instructions. Quantasoft Software (version 1.7) was used to analyze the data. The percent methylation for each promoter was calculated by taking the ratio of positive droplets obtained from the ddPCR reaction of DNA digested with methylation-sensitive restriction enzymes versus DNA without enzymatic digestion. The percent methylation of CSP-pCpGL methylation standards and for each of the 7 promoters from 22 paired OSCC and adjacent normal tissues were plotted using GraphPad Prism software and a paired t-test was performed to determine the level of significance.

Bisulfite conversion and bisulfite sequencing PCR

For all OSCC and normal samples, 1 µg of isolated genomic DNA was treated with sodium bisulfite using EZ DNA Methylation Gold Kit (Zymo-Research Corp) following the manufacturer's instructions. Converted DNA samples were used for bisulfite sequencing PCR to amplify selected promoter regions. The reverse primer, used for bisulfite sequencing PCR, was biotinylated in the 5'-end and the forward primer was used as the sequencing primer in order to cover all the CpGs in the PCR amplified region. The primer details are given in Additional File 4: Table S4.

DNA methylation quantification using pyrosequencing

After bisulfite sequencing PCR, the PCR amplified product was purified using a MinElute PCR Purification Kit (Qiagen, Germany) and subjected to pyrosequencing (PyroMark Q48 Autoprep, Qiagen). The methylation percentages for each CpG were calculated using PyroMark Q48 Autoprep software (Qiagen, Germany).

Survival data analysis

Apart from the model-based validation of the seven selected features, to elucidate any clinical significance of those selected features, we analyzed methylation values of those seven promoters with clinical follow-up data available at TCGA. For each of those seven promoters, top and bottom 25 percentile of methylation values was considered as high and low groups, respectively. Survival analyses of these seven features with the methylation value were done using the Kaplan–Meier survival analysis in SPSS. Mantel–Cox log-rank P-value ≤ 0.05 was considered significant. Additionally, gene expression data for those seven genes were retrieved from the TCGA database and Kaplan–Meier survival analysis was performed with top and bottom 25 percentile of gene expression (TPM) values.

Statistical analysis and code availability

All the statistical analyses apart from survival analysis were performed in the R programming language (version 4.1.0). An adjusted p-value < 0.05 was considered to be the significance level. Heatmap was generated using gplots package in R. The *Elastic-net* regression model was developed in the R environment and the necessary codes are available upon request.

Results

Identification of potential candidates for universal set of markers

The comparison of genome-wide DNA methylation data of OSCC patients in India with the TCGA HNSCC data from anatomical sites like oral cavity, oral tongue, floor of mouth, base of tongue, buccal mucosa, and hard palate was used to identify 20,645 common differentially methylated CpG probes (DMPs)(5670 hypermethylated and 14,975 hypomethylated CpGs) between these two datasets[15]. We intended to evaluate the potential of these common DMPs as epigenetic markers for OSCC detection. A brief schematic representation of our study design is shown in Fig. 1A. To identify the usable features for OSCC prediction, receiver operating curve (ROC) analysis was performed using these common DMPs. We identified 6453 hypo and 1711 hypermethylated CpGs that classified the normal and OSCC tissues with an area under the curve (AUC) ≥ 0.9 (Additional File 5: Table S5). These DMPs overlapped with 1,678 promoters, among which 42 promoters contained at least 4 DMPs (Additional File 6: Table S6). Average methylation of these 42 promoters distinctly classified the normal and disease samples, indicating the significance of these promoters in predicting the disease (Fig. 1B).

Final model development using elastic-net/development of prediction model using elastic-net

The average methylation values of 42 promoters in 324 HNSCC and 60 normal tissues were used to build a penalized (regularized) linear regression model. Randomly sampled 80% of the data were used as training dataset, and the remaining 20% of data were kept for validation of the model. Model fitting and selection of the tuning parameter λ was done based on the training dataset. The best λ was determined using a tenfold cross-validation from the training dataset for which the model produced the minimum mean cross-validation error. With this best λ, the model was fitted using the training dataset and features with non-zero coefficient were recorded. To obtain the relevant features and estimate the mean error rate, this process was repeated 500 times. Box plot of the coefficients of the 42 promoters in 500 runs appeared among the top contributing factors in the model (Fig. 2A). The mean error rate, calculated using the remaining 20% validation data, was found to be 0.02% with a range of 0 to 6% (Fig. 2B). The model predicted 7 features that appeared > 95% of the times as non-zero coefficients in the model, thereby indicating them as promising candidates for potential DNA methylation markers of OSCC (Fig. 2C). Among these, *CSGALNACT1, SLC5A9,* and *HOXAAS2* were non-zero coefficient in 100% of the runs, while coefficient for *ZNF154, SNHG14, GPR75,* and *HOXAAS3* was non-zero in 99.8%, 99.2%, 97%, and 94% of the 500 runs, respectively. The final *Elastic-net* model with these 7 promoters is given by the following equation:

$$y = 3.95 - 7.97 \times M_{CSGALNACT1} - 6.36 \times M_{SLC5A9} - 2.30 \times M_{SNHG14} + 2.52 \times M_{GPR75} + 5.43 \times M_{HOXAAS2} + 2.14 \times M_{HOXAAS3} + 3.26 \times M_{ZNF154},$$

where M indicates the average promoter methylation for the genes that are presented as the subscript.

Among these 7 promoters, *CSGALNACT1, SLC5A9,* and *SNHG14* were hypomethylated, while the remaining 4 promoters, *GPR75, HOXAAS2, HOXAAS3,* and *ZNF154,* were hypermethylated in OSCC tissues compared with the adjacent normal tissues. To evaluate the clinical significance of these 7 promoters, we performed survival analysis with the methylation values (β-value) of these promoters. Among the hypomethylated promoters, *CSGALNACT1* (P-value = 0.010) and *SNHG14* (P-value = 0.043) showed significant prognostic relevance, while for the hypermethylated promoters, the survival analysis failed to show any significant association (Additional File 8: Figure S2). We also performed survival analysis based on the gene expression values (TPM) of these respective genes. *ZNF154* and *HOXAAS3* were hypermethylated in OSCC samples and showed better survival (P-value = 0.021 and P = 0.027) with higher gene expression (Additional File 8: Figure S3).

Validation of the predicted model using publicly available HNSCC data

The prediction accuracy of the model was determined with publicly available genome-wide methylation data of 812 (658 HNSCC and 154 normal) HNSCC samples from different anatomical sites, and seven different countries (Table 1, Additional File 1: Table S1). The data have been classified broadly according to the site of occurrence, country of origin, tumor staging, and specific dataset (Table 1, Additional File 7: Table S7, S8, S9), and sensitivity (SN), specificity (SP), and accuracy (AC) of the model were calculated. The accuracy of our model is the best for patients with tumors on oral cavity, which showed an overall accuracy of 93.52% with 98.09% sensitivity and 77.01% specificity (Table 1). For the oropharynx, larynx, and hypopharynx, the sensitivities are 92 to 100%. The positive likelihood ratio (i.e., true positive rate/false positive rate) of oral cavity (4.27) was higher compared to oropharynx (2.78). The tumor on the nasopharynx showed 98% sensitivity, but the specificity of this anatomical site is very low (18%) (Table 1), suggesting its limitation in predicting tumors at nasopharynx. Methylation data generated from FFPE tissues (GSE38266 and GSE79556) also predicted 100% sensitivity in cancer of the oral cavity and 98% sensitivity in cancers of the oropharynx. The efficiency of the model in predicting the disease was also comparable across studies conducted in different geographical locations, except the study reported from China, which mainly included tumors at nasopharynx, depicting the universality of our predictive model in oral cavity cancer (Additional File 7: Table S7). Our model showed outstanding sensitivity

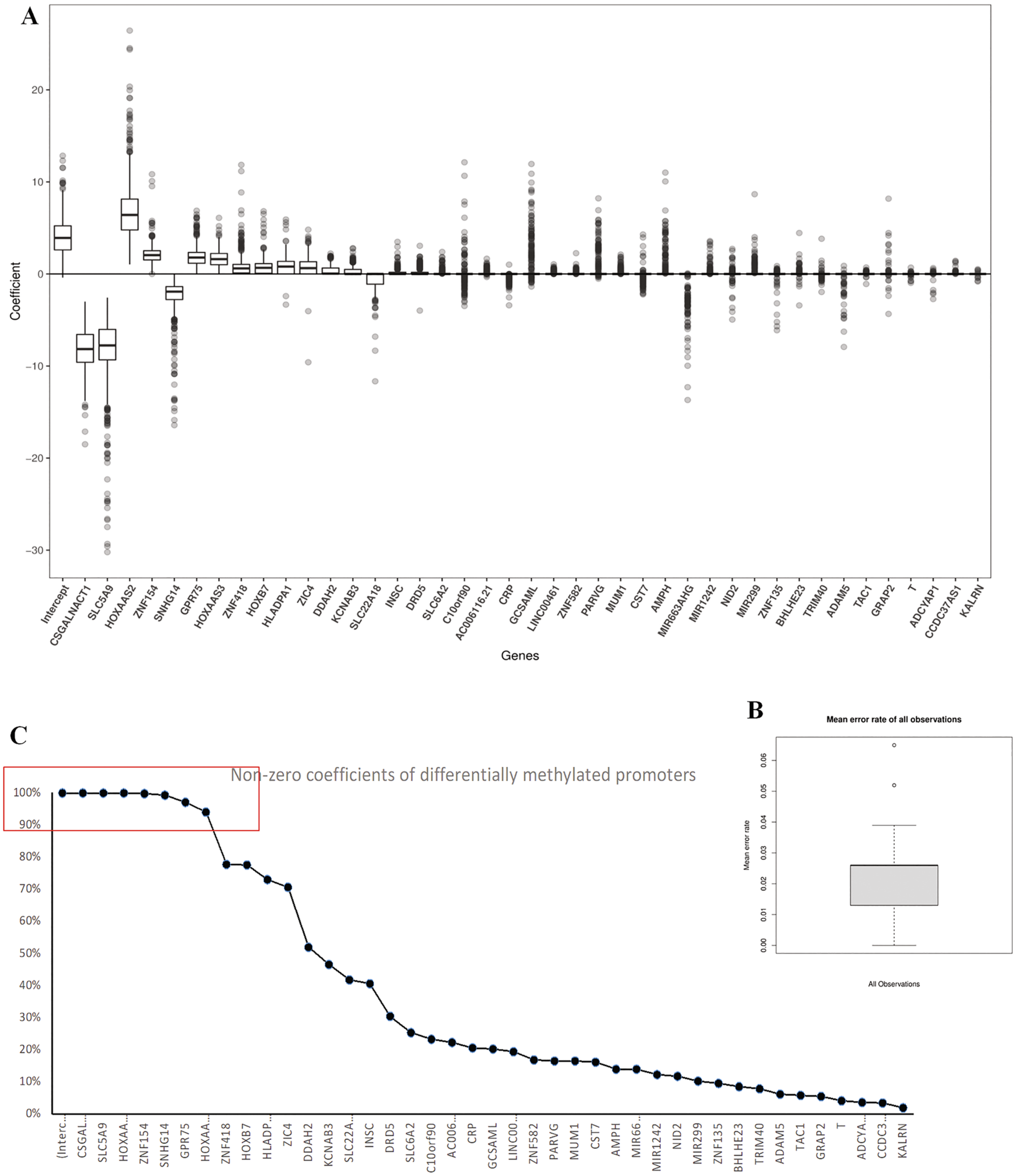

Fig. 2 **A** Box plot of *Elastic-net* coefficients for 42 promoters in 500 runs. **B** Mean error rate of the model as measured using the validation samples. **C** Percentage of occurrences of differentially methylated promoters as non-zero coefficients in the predicted model

for different tumor stages. We found almost similar sensitivity for both early-stage (Stages I and II) and late-stage (Stages III and IV) tumors (98.93 and 98.67%, respectively) (Additional File 7: Table S8). Study-specific classification also showed the excellent performance of the model for predicting oral cavity cancer (Additional

File 7: Table S9). Even for the TCGA samples from Alveolar Ridge, Hypopharynx, Larynx, Oropharynx, and tonsil ($n=200$), the model showed 99.5% sensitivity.

Validation with droplet digital PCR of MSRE digested DNA (ddMSRE)

We next attempted to validate the predicted model using methylation data generated from a completely different experimental approach. To calculate the average methylation, we developed a novel method using methylation-sensitive restriction enzymes (MSRE) digestion followed by amplification with the Droplet Digital PCR (ddPCR) using primers for methylated CpGs. Here, MSREs were chosen so that only unmethylated DNA gets digested, while methylated DNA remains intact and is available for amplification. Droplets receiving undigested methylated DNA will amplify and be counted as positive droplets and those without DNA or with digested DNA will not amplify and be counted as negative droplets. With the number of positive droplets obtained from each sample, we can calculate the exact copies of undigested methylated DNA (Fig. 3A). The copy number ratio between the digested and undigested samples was used to calculate the percent methylation of a DNA segment (Additional File 8: Figure S4). A strong positive correlation between the known and observed methylation values indicated the robustness of our proposed method (Correlation coefficient = 0.99 and *P*-value < 0.0001) (Fig. 3B). The

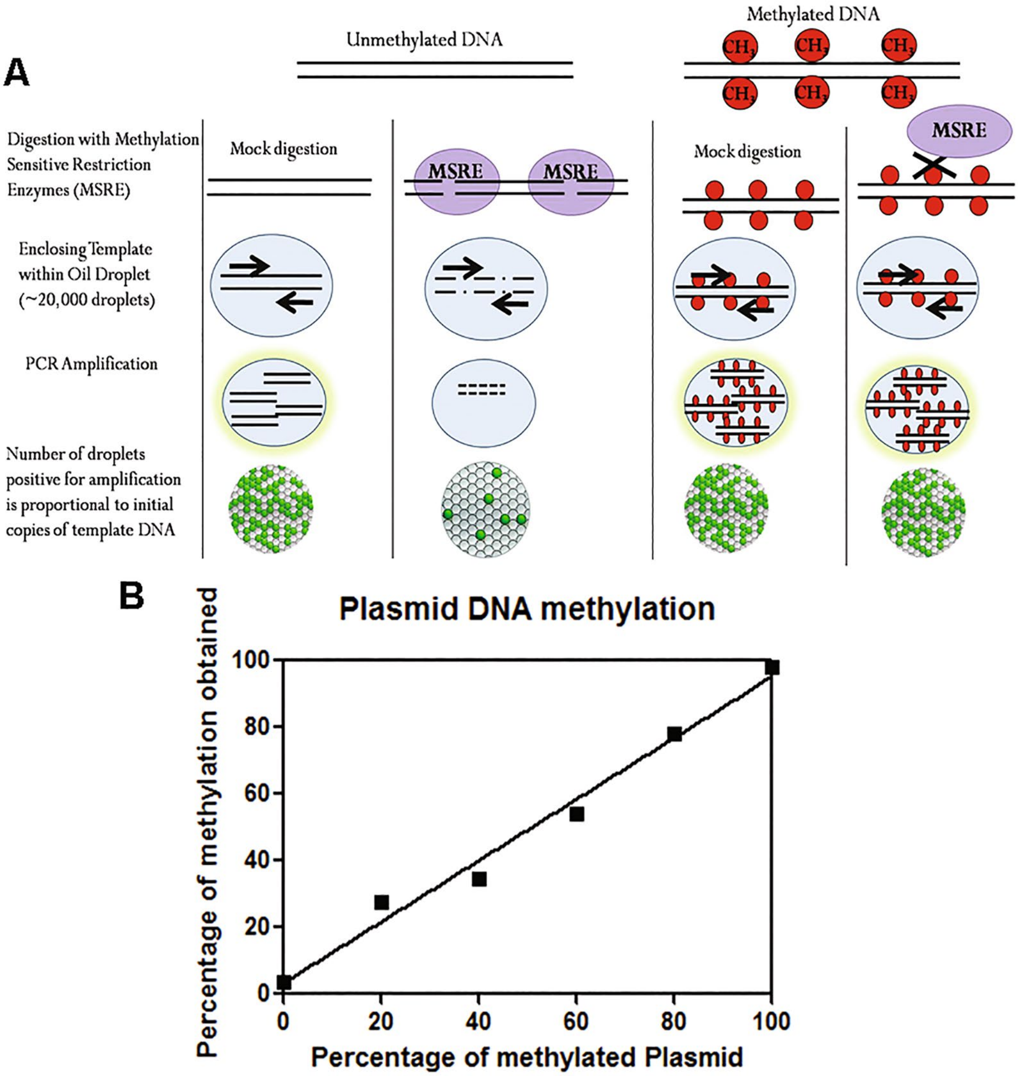
A
Unmethylated DNA
Methylated DNA
CH_3
Digestion with Methylation Sensitive Restriction Enzymes (MSRE)
Mock digestion
MSRE
MSRE
Mock digestion
MSRE
Enclosing Template within Oil Droplet (~20,000 droplets)
PCR Amplification
Number of droplets positive for amplification is proportional to initial copies of template DNA
B
Plasmid DNA methylation
Percentage of methylation obtained
Percentage of methylated Plasmid
0
20
40
60
80
100

Fig. 3 **A** The schematic presentation of the proposed methodology for quantifying DNA methylation using Methylation-sensitive restriction enzymes followed by droplet digital PCR (ddMSRE). **B** Line graph showing the correlation between the actual and observed percent methylation using ddMSRE

proportional increase in the positive droplets was also in concordance with the increasing methylation percentage (Additional File 8: Figure S4). To validate the reliability of our predicted *Elastic-net* model, we determined the methylation level of 7 promoters in an additional set of 22 paired OSCC and contralateral normal tissues using ddMSRE (Additional File 7: Table S10). DNA isolated from paired tissue samples was digested with MSRE and subjected to ddPCR amplification (Additional File 8: Figure S5). We observed a significant hypomethylation for the promoters of*CSGALNACT1* (*P*-value = 0.0047), *SLC5A9* (P-value = 0.0002) and *SNHG14* (P-value = 0.0033), and significant hypermethylated (*P*-value < 0.0001) for *ZNF154, HOXAAS2, HOXAAS3,* and *GPR75* in OSCC compared to adjacent normal tissues (Fig. 4). The methylation status of these seven promoters was used in our predicted *Elastic-net* model. Among the 22 OSCC samples, our model predicted 20 as the disease, while among 22 contralateral normal samples, it could predict all 22 as the normal samples. The sensitivity, specificity, and accuracy of our independent validation are 91%, 100%, and 96%, respectively (Table 2).

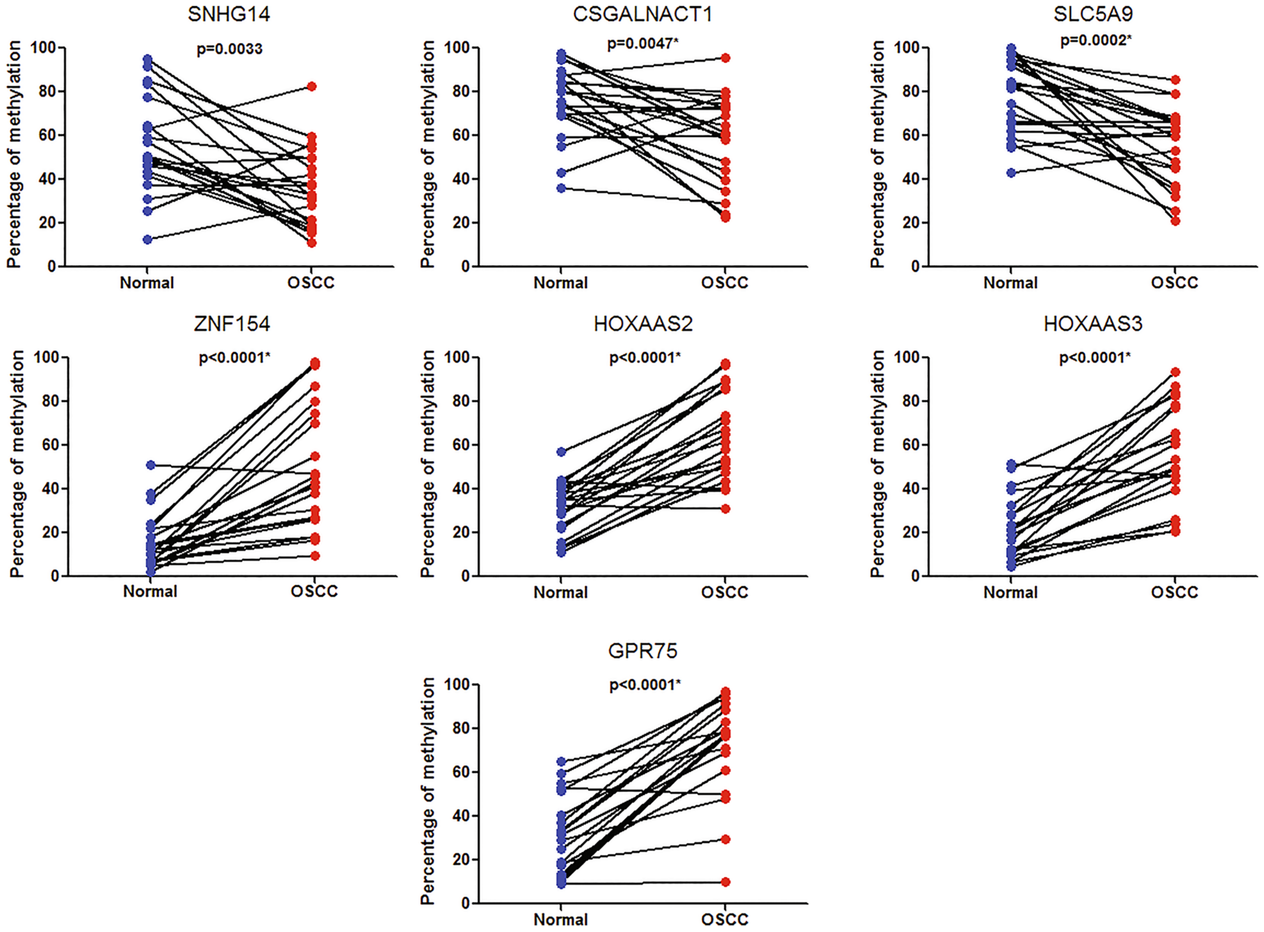

Fig. 4 Methylation level of 7 promoters of 22 paired OSCC and contralateral normal tissue samples using ddMSRE. *P*-values were calculated using paired t-test for each promoter

Table 2 Validation of the model using ddMSRE

	Sample Status		Predicted								
Sample type	Disease	Normal	TP*	FP*	TN*	FN*	PLR*	NLR*	SN*	SP*	ACC*
OSCC (Indian patient cohort)	22	22	20	0	22	2	–	0.09	90.91	100	95.45

Validation of the model using pyrosequencing

To establish the rigor and robustness of the model, we further validated the model in an additional set of paired tissue samples through pyrosequencing. Owing to the specificity of the model for contralateral normal samples, we attempted to explore contralateral and adjacent normal samples for the validation of the model. We used 21 paired OSCC and contralateral normal samples as well as 25 paired OSCC and adjacent normal samples (Additional File 7: Tables S11, S12). In case of 21 paired OSCC and contralateral normal samples, we observed significant hypomethylation for *CSGALNACT1, SNHG14,* and *SLC5A9*, and significant hypermethylation for *ZNF154, HOXAAS2, HOXAAS3,* and *GPR75* in OSCC compared to contralateral normal tissues (Fig. 5). However, we did not observe any significant changes in methylation for *SNHG14, SLC5A9, HOXAAS3,* and *GPR75* in 25 paired OSCC and adjacent normal samples (Additional File 8: Figure S6), possibly due to hyperplasia or dysplasia in the phenotypically normal adjacent normal samples. The average methylation values of seven promoters from both sample sets were used to

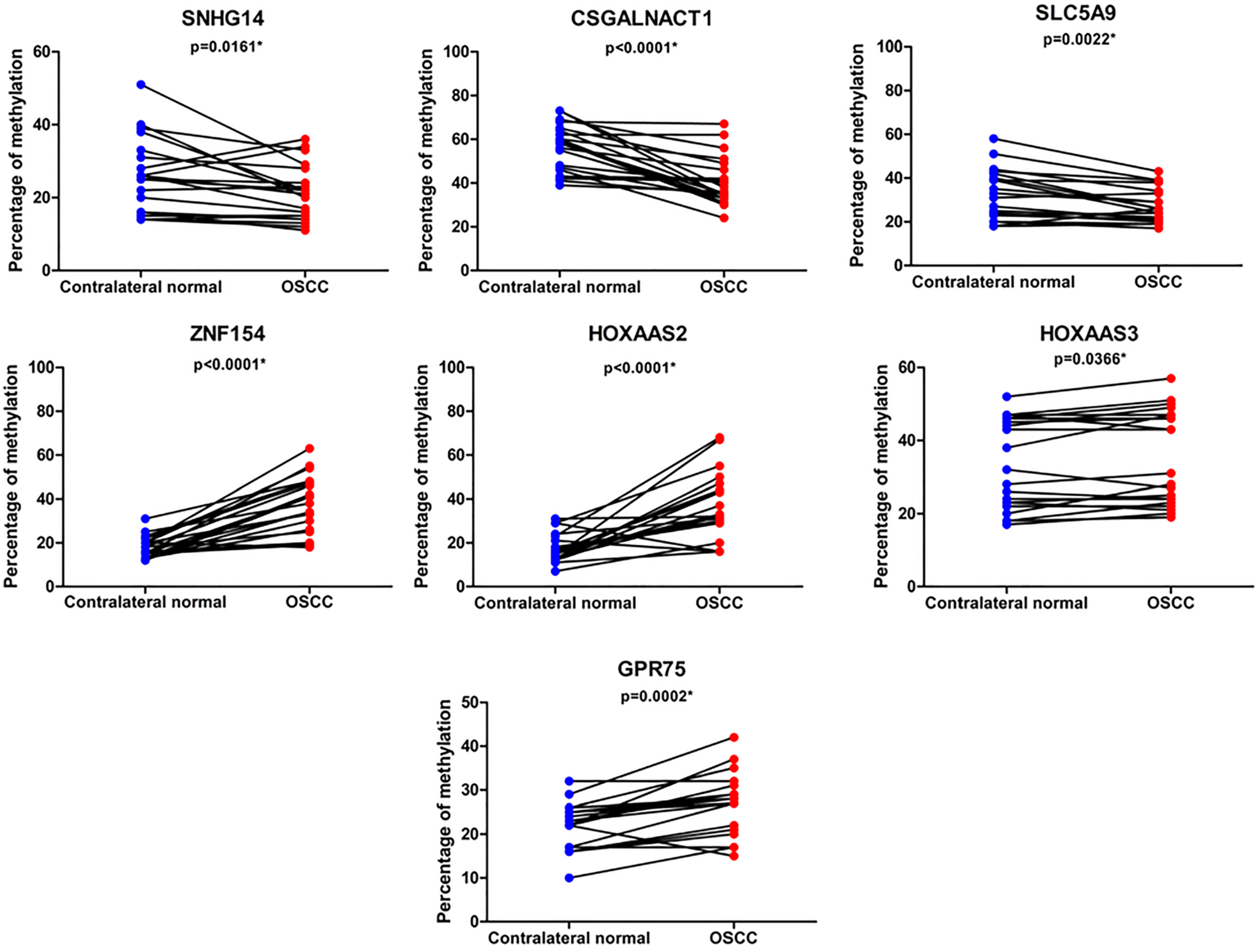

Fig. 5 Methylation level of 7 promoters of 21 paired OSCC and contralateral normal tissue samples using pyrosequencing. *P*-values were calculated using paired *t*-test for each promoter

Table 3 Validation of the model of paired contralateral normal samples using pyrosequencing

Sample type	Sample Status		Predicted								
	Disease	Normal	TP*	FP*	TN*	FN*	PLR*	NLR*	SN*	SP*	ACC*
OSCC (Indian patient cohort)	21	21	20	2	19	1	10	0.05	95.24	90.48	92.86

validate the model. For 21 paired OSCC and contralateral normal samples, our model showed 95.24% sensitivity, 90.48% specificity, and 92.86% accuracy (Table 3). But in case of 25 paired OSCC and adjacent normal, the model showed 96% sensitivity, 20% specificity, and 58% accuracy (Table 4), indicating potential early detection of this predictive model.

Discussion

Aberrant DNA methylation is one of the hallmarks of cancer including oral squamous cell carcinoma. Potential of CpG methylation as a biomarker for OSCC detection has been evaluated, but most of the time it was not replicated in patient cohorts. In this study, we proposed a model for detecting oral cancer in patients irrespective of their anatomical origin, oral habits, etiology, and geographical locations of the study. Since DNA methylation is known to be an early molecular event appearing prior to phenotypic changes, we intended to construct a universal set of DNA methylation markers that can distinguish the disease state from normal. In our previous study, we reported a set of common differentially methylated CpGs between OSCC patients of India and any of the four stages (S1–S4) of the TCGA methylation data. From these common differentially methylated CpGs, usable features were selected to develop a regularized linear regression *Elastic-net* model. Among different machine learning algorithms, such as Lasso, Ridge, Random Forest, and KNN, the *Elastic-net* showed the best performance in terms of sensitivity, specificity, and accuracy of the model (data not shown). The proposed *Elastic-net* model uses average methylation values of 7 promoters in classifying oral cancer and normal oral tissues. We evaluated the applicability and universality of the predicted model using 812 HNSCC 450 k methylation array and EPIC array data (Disease = 658, Normal = 154) available in TCGA HNSCC and 14 GEO databases reported from Australia, China, Germany, Brazil, India, the UK, and the USA. We further classified the data based on the anatomical sites of disease occurrence. The model showed a classification accuracy of 93.52% with 98.09% sensitivity in identifying tumor and normal tissues of the oral cavity. However, the model was developed using methylation data excluding oropharyngeal, laryngeal, hypo-pharyngeal, and nasopharyngeal carcinoma samples, but we obtained 92% to 100% sensitivity for oropharyngeal, laryngeal, and hypo-pharyngeal samples. We could not determine the specificity of our model for laryngeal and hypo-pharyngeal samples due to lack of data for corresponding normal samples. Interestingly, our model worked well for the methylation data generated from FFPE tissues with sensitivity ranging from 98 to 100%. Our model showed 98% sensitivity for both early and late-stage samples, suggesting the applicability of the model in predicting very early-stage cancer. The model was also evaluated according to the country of origin and it showed almost consistent sensitivity and accuracy across Germany, India, the UK, the USA, and Australia. However, one of the limitations of this study is the low sample size for some of the countries. A multi-institutional study with sufficient number of samples may further substantiate the prediction of this model.

In addition to the validation of our model using 450 k methylation array data, we explored its applicability using a ddMSRE. DNA methylation of 7 promoters was quantified using ddMSRE and used to assess the performance of the model in classifying OSCC and contralateral normal tissues. For validation of the DNA methylation quantified using ddMSRE, we determined the methylation status of known methylation standards. Comparison of the observed methylation with the known methylation values showed a strong positive correlation, suggesting the robustness of the method. Our predicted model was further validated using an additional set of 22 paired OSCC and contralateral normal samples by ddMSRE (Additional File 8: Figure S5). The predicted model identified the normal samples with 100% precision and classified the disease and normal samples with 95% accuracy. All 3 hypomethylated (*SNHG14, CSGALNACT1*, and *SLC5A9*) and 4 hypermethylated (*ZNF154, HOXAAS2, HOXAAS3,* and *GPR75*) promoters showed significant differential methylation in OSCC compared to adjacent normal samples. As ddMSRE showed 100% precision for contralateral normal samples, we also preferred to choose a different set of clinically normal samples to evaluate the precision of the model. We selected 21 paired OSCC with

Table 4 Validation of the model of paired adjacent normal samples using pyrosequencing

Sample type	Sample status		Predicted								
	Disease	Normal	TP*	FP*	TN*	FN*	PLR*	NLR*	SN*	SP*	ACC*
OSCC (Indian patient cohort)	25	25	24	20	5	1	1.2	0.2	96	20	58

NA# Not calculated due to lack of normal samples

* *TP* true positive, *FP* false positive, *TN* true negative, *FN* false negative, *PLR* positive likelihood ratio, *NLR* negative likelihood ratio, *SN* Sensitivity, *SP* Specificity, *ACC* Accuracy.

contralateral normal and 25 paired OSCC with adjacent normal samples. Interestingly, all the 7 promoters were significantly differentially methylated among 21 paired OSCC and contralateral normal samples, but only 3 promoters showed significant differential methylation for 25 paired OSCC and adjacent normal samples. Interestingly, for both sample sets, the model showed 96% sensitivity which suggests the rigor of the model in disease prediction. For contralateral normal samples, the model showed 91% specificity but the specificity drops down to 20% for adjacent normal samples. This observation suggests the occurrence of early molecular events in the vicinity of tumor samples and those early molecular events in the adjacent clinically normal samples can also be captured through the methylation status of these 7 promoters. Our *Elastic-net* model is rigorous as it predicted the disease phenotype based on the early molecular events. Certain morphological alterations known as malignancy-associated changes (MACs) develop in adjacent histologically normal cells due to the close proximity of the tumor [38]. Assays based on those molecular events also found significant differences between adjacent normal and contralateral normal samples [39]. These observations broadly suggest two important aspects, first, the methylation value from any qualified in vitro methylation determining assay can be used in our model and second, the model with these seven promoters distinguishes oral cancer from the normal samples.

Higher expression of long non-coding RNA (lncRNA) *SNHG14* is already attributed to colorectal cancer metastasis, cervical cancer, bladder cancer, non-small cell lung carcinoma, hepatocellular carcinoma, ovarian cancer, retinoblastoma, pancreatic cancer, colorectal cancer, and endometrial cancer progression [40–42]. Promoter hypomethylation usually leads to overexpression of genes, and hence, *SNHG14* promoter hypomethylation in OSCC might also play a significant role in OSCC prognosis. Investigation on copy number alterations in oral cancer identified *CSGALNACT1* gene to be one of the most frequent gene losses, but no epigenetic regulation has been reported yet [43]. In case of *SLC5A9*, better survival was attributed to low expression for renal cancer [44], but no epigenetic regulation or expression-based studies are reported for oral cancer. Promoter hypermethylation of tumor suppressor *ZNF154* is reported in many cancer types including triple-negative breast cancer, clear cell renal cell carcinomas, and nasopharyngeal carcinoma [45–49]. Overexpression of *ZNF154* in a gastric cancer cell line (MGC-803) showed reduced cell proliferation, migration, and invasion and enhanced apoptosis. Restoration of the gene expression with treatment of 5-aza-2-deoxycytidine showed higher expression of *ZNF154* followed by inhibition of cell-migration and invasion in nasopharyngeal carcinoma cells [45]. Moreover, the robust hypermethylation of *ZNF154* as a multi-cancer signature made this to be a promising blood-based diagnostic marker for cancer [48, 49].

The oncogenic role of another lncRNA, *HOXAAS2*, is also well established in the literature [50–53]. The knockdown of *HOXAAS2* showed a significant reduction in cell proliferation while promoting apoptosis in colorectal cancer [53]. Knockdown of this lncRNA showed inhibition in cell viability, migration, and invasion in osteosarcoma cells [54]. Higher expression of *HOXAAS3* was also reported in lung adenocarcinoma, while shRNA-mediated knockdown of this lncRNA showed inhibited tumorigenesis in ng adenocarcinoma cells [55]. Similarly, *GPR75* was also reported to be methylation-driven gene and proposed to be one of the independent prognostic biomarkers in lung squamous cell carcinoma [56]. In order to decipher the clinical significance of seven promoters examined in our study, we performed survival analysis using the average promoter methylation values in the TCGA HNSCC dataset. Average promoter methylation of *SNHG14* and *CSGALNACT1* showed a significant association, but the remaining 5 promoters did not show any significant association with the survival of HNSCC patients (Additional File 8: Figure S2). Survival analysis with gene expression values of these genes also showed that higher expression of *ZNF154* and *HOXAAS3* leads to significant association of better survival.

Conclusions

To conclude, we identified a set of DNA methylation markers that may have the potential to be used as biomarkers in detecting oral cavity cancer irrespective of geographical locations and indigeneity, having different oral habits or etiological factors. We used *Elastic-net* regression to develop the predictive model with seven promoters. Validation of the model showed excellent performance in classifying oral cavity cancer from the contralateral normal samples, while the performance of the model is relatively low in classifying the oral cavity cancer from the adjacent normal sample, suggesting possible molecular involvement in the adjacent but clinically normal regions near the tumor lesion.

Abbreviations

OSCC	Oral squamous cell carcinoma
OPMD	Oral potentially malignant disorder
DMPs	Differentially methylated CpG probes
ddPCR	Droplet digital PCR
ddMSRE	Droplet digital PCR amplification of the methyl-sensitive restriction enzyme digested DNA
HNSCC	Head and Neck squamous cell carcinoma
ROC	Receiver operating curve
LASSO	Least absolute Shrinkage and selection operator
CSP-pCpGL	*CSGALNACT1* Promoter cloned pCpGL vector
MSRE	Methylation-sensitive restriction enzymes

AUC	Area under the curve
CSGALNACT1	Chondroitin Sulfate N-Acetylgalactosaminyltransferase 1
SLC5A9	Solute carrier family 5 (sodium/glucose cotransporter), member 9
SNHG14	Small nucleolar RNA host gene 14
ZNF154	Zinc finger protein 154
HOXAAS2	HOXA cluster antisense RNA 2
HOXAAS3	HOXA cluster antisense RNA 3
GPR75	G protein-coupled receptor 75

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40001-024-02047-4.

Additional file 1

Additional file 2

Additional file 3

Additional file 4

Additional file 5

Additional file 6

Additional file 7

Additional file 8

Acknowledgements

SD was supported by the SERB grant (CRG/2020/003837). SK was supported by the DBT-Junior Research Fellowship. BB and AC were supported by the CSIR fellowship. JC was supported by the DST-INSPIRE fellowship. The authors want to acknowledge all the patients who participated in this study.

Author contributions

SD, SK, AC and BB conducted the experiments. JC, AC, DP, AET and RC conducted the analysis. AS and JG recruited the patients and collected samples. DP, AC and RC conceptualize the study. ML, SB and AET provided the clinical information from UK cohort. SD, AC, DP and RC wrote the manuscript. ML, SB and AET edited the manuscript. All the authors have reviewed the manuscript.

Funding

The research is funded by the DST, Govt. of India for partial support through the Technology Innovation Hub (TIH) on Data Science, Big Data Analytics, and Data Curation at ISI [grant number NMICPS/006/MD/2020-21].

Availability of data and materials

All data used in the manuscript were downloaded from the publicly available database. Illumina Infinium HumanMethylation450BeadChip array data for 21 OSCC and adjacent normal tissues from OSCC patients in India are available in the GEO database (GSE87053).

Declarations

Ethics approval and consent to participate

This study was conducted after obtaining ethical approval from the "Review Committee for Protection of Research Risks to Humans" of the Indian Statistical Institute and obtaining consent from each participant.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Human Genetics Unit, Indian Statistical Institute, 203 B T Road, Kolkata 700 108, India. [2]Univeristy of Pennsylvania, Philadelphia 19104, USA. [3]Department of Mathematical Sciences, IISER Kolkata, Kalyani, India. [4]Department of Statistics, U C Davis, 4222 Mathematical Sciences Building, Davis, CA 95616, USA. [5]Department of Oral Pathology, Dr. R. Ahmed Dental College & Hospital, Kolkata, India. [6]University College London Cancer Institute, University College London, 72 Huntley St, London WC1E 6DD, UK. [7]CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China.

Received: 23 May 2024 Accepted: 1 September 2024
Published online: 11 September 2024

References

1. Llewellyn CD, Johnson NW, Warnakulasuriya KA. Risk factors for squamous cell carcinoma of the oral cavity in young people–a comprehensive literature review. Oral Oncol. 2001;37(5):401–18.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.
3. Lingen MW, Kalmar JR, Karrison T, Speight PM. Critical evaluation of diagnostic aids for the detection of oral cancer. Oral Oncol. 2008;44(1):10–22.
4. Baylin SB, Jones PA. A decade of exploring the cancer epigenome—biological and translational implications. Nat Rev Cancer. 2011;11(10):726–34.
5. Irimie AI, Ciocan C, Gulei D, Mehterov N, Atanasov AG, Dudea D, et al. Current insights into oral cancer epigenetics. Int J Mol Sci. 2018;19(3):670.
6. Bakhtiar SM, Ali A, Barh D. Epigenetics in head and neck cancer. Methods Mol Biol. 2015;1238:751–69.
7. Esteller M. Cancer epigenetics for the 21st century: what's next? Genes Cancer. 2011;2(6):604–6.
8. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003;33(Suppl):245–54.
9. Shaw RJ, Liloglou T, Rogers SN, Brown JS, Vaughan ED, Lowe D, et al. Promoter methylation of P16, RARbeta, E-cadherin, cyclin A1 and cytoglobin in oral cancer: quantitative evaluation using pyrosequencing. Br J Cancer. 2006;94(4):561–8.
10. Huang YK, Peng BY, Wu CY, Su CT, Wang HC, Lai HC. DNA methylation of PAX1 as a biomarker for oral squamous cell carcinoma. Clin Oral Investig. 2014;18(3):801–8.
11. Cheng SJ, Chang CF, Ko HH, Lee JJ, Chen HM, Wang HJ, et al. Hypermethylated ZNF582 and PAX1 genes in mouth rinse samples as biomarkers for oral dysplasia and oral cancer detection. Head Neck. 2018;40(2):355–68.
12. Cheng SJ, Chang CF, Ko HH, Liu YC, Peng HH, Wang HJ, et al. Hypermethylated ZNF582 and PAX1 genes in oral scrapings collected from cancer-adjacent normal oral mucosal sites are associated with aggressive progression and poor prognosis of oral cancer. Oral Oncol. 2017;75:169–77.
13. Li YF, Hsiao YH, Lai YH, Chen YC, Chen YJ, Chou JL, et al. DNA methylation profiles and biomarkers of oral squamous cell carcinoma. Epigenetics. 2015;10(3):229–36.
14. Foy JP, Pickering CR, Papadimitrakopoulou VA, Jelinek J, Lin SH, William WN Jr, et al. New DNA methylation markers and global DNA hypomethylation are associated with oral cancer development. Cancer Prev Res. 2015;8(11):1027–35.
15. Basu B, Chakraborty J, Chandra A, Katarkar A, Baldevbhai JRK, Dhar Chowdhury D, et al. Genome-wide DNA methylation profile identified a unique set of differentially methylated immune genes in oral squamous cell carcinoma patients in India. Clin Epigenet. 2017;9:13.
16. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Statist Soc Series B Statist Methodol. 2005;67(2):301–20.
17. Poulin M, Zhou JY, Yan L, Shioda T. Pyrosequencing methylation analysis. Methods Mol Biol. 1856;2018:283–96.
18. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.
19. Nemeth CG, Rocken C, Siebert R, Wiltfang J, Ammerpohl O, Gassling V. Recurrent chromosomal and epigenetic alterations in oral squamous cell carcinoma and its putative premalignant condition oral lichen planus. PLoS ONE. 2019;14(4): e0215055.

20. Lim AM, Wong NC, Pidsley R, Zotenko E, Corry J, Dobrovic A, et al. Genome-scale methylation assessment did not identify prognostic biomarkers in oral tongue carcinomas. Clin Epigenetics. 2016;8:74.
21. Lechner M, Fenton T, West J, Wilson G, Feber A, Henderson S, et al. Identification and functional validation of HPV-mediated hypermethylation in head and neck squamous cell carcinoma. Genome Med. 2013;5(2):15.
22. Worsham MJ, Chen KM, Datta I, Stephen JK, Chitale D, Gothard A, et al. The biological significance of methylome differences in human papilloma virus associated head and neck cancer. Oncol Lett. 2016;12(6):4949–56.
23. Krishnan NM, Dhas K, Nair J, Palve V, Bagwan J, Siddappa G, et al. A minimal DNA methylation signature in oral tongue squamous cell carcinoma links altered methylation with tumor attributes. Mol Cancer Res. 2016;14(9):805–19.
24. Pickering CR, Zhang J, Yoo SY, Bengtsson L, Moorthy S, Neskey DM, et al. Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. Cancer Discov. 2013;3(7):770–81.
25. Khongsti S, Lamare FA, Shunyu NB, Ghosh S, Maitra A. Whole genome DNA methylation profiling of oral cancer in ethnic population of Meghalaya, North East India reveals novel genes. Genomics. 2018;110(2):112–23.
26. Jiang W, Liu N, Chen XZ, Sun Y, Li B, Ren XY, et al. Genome-wide identification of a methylation gene panel as a prognostic biomarker in nasopharyngeal carcinoma. Mol Cancer Ther. 2015;14(12):2864–73.
27. Dai W, Cheung AK, Ko JM, Cheng Y, Zheng H, Ngan RK, et al. Comparative methylome analysis in solid tumors reveals aberrant methylation at chromosome 6p in nasopharyngeal carcinoma. Cancer Med. 2015;4(7):1079–90.
28. Inchanalkar M, Srivatsa S, Ambatipudi S, Bhosale PG, Patil A, Schaffer AA, et al. Genome-wide DNA methylation profiling of HPV-negative leukoplakia and gingivobuccal complex cancers. Clin Epigenet. 2023;15(1):93.
29. Soares-Lima SC, Mehanna H, Camuzi D, de Souza-Santos PT, Simao TA, Nicolau-Neto P, et al. Upper aerodigestive tract squamous cell carcinomas show distinct overall DNA methylation profiles and different molecular mechanisms behind WNT signaling disruption. Cancers. 2021;13(12):3014.
30. Marthong L, Ghosh S, Palodhi A, Imran M, Shunyu NB, Maitra A, et al. Whole genome DNA methylation and gene expression profiling of oropharyngeal cancer patients in North-Eastern india: identification of epigenetically altered gene expression reveals potential biomarkers. Front Genet. 2020;11:986.
31. Assenov Y, Muller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. Nat Methods. 2014;11(11):1138–40.
32. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
33. Tibshirani R. Regression shrinkage and selection via the Lasso. J Roy Stat Soc: Ser B. 1996;58(1):267–88.
34. Kennard RW. Ridge regression: biased. Estimation Nonorthogonal Problems Technometr. 1970;12(1):55–67.
35. Bunea F. Honest variable selection in linear and logistic regression models via ℓ1 and ℓ1+ℓ2 penalization. Electron J Statist. 2008;2:1153–94.
36. Friedman J, Hastie T. The elements of statistical learning data mining inference, and prediction. Berlin: Springer; 2009.
37. Hindson BJ, Ness KD, Masquelier DA, Belgrader P, Heredia NJ, Makarewicz AJ, et al. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal Chem. 2011;83(22):8604–10.
38. Palcic B. Nuclear texture: can it be used as a surrogate endpoint biomarker? J Cell Biochem Suppl. 1994;19:40–6.
39. Jabalee J, Carraro A, Ng T, Prisman E, Garnis C, Guillaud M. Identification of malignancy-associated changes in histologically normal tumor-adjacent epithelium of patients with HPV-positive oropharyngeal cancer. Anal Cell Pathol. 2018;2018:1607814.
40. Di W, Weinan X, Xin L, Zhiwei Y, Xinyue G, Jinxue T, et al. Long noncoding RNA SNHG14 facilitates colorectal cancer metastasis through targeting EZH2-regulated EPHA7. Cell Death Dis. 2019;10(7):514.
41. Zhang YY, Li M, Xu YD, Shang J. LncRNA SNHG14 promotes the development of cervical cancer and predicts poor prognosis. Eur Rev Med Pharmacol Sci. 2019;23(9):3664–71.
42. Shen S, Wang Y, Zhang Y, Dong Z, Xing J. Long non-coding RNA small nucleolar RNA host gene 14, a promising biomarker and therapeutic target in malignancy. Front Cell Dev Biol. 2021;9: 746714.
43. Yong ZW, Zaini ZM, Kallarakkal TG, Karen-Ng LP, Rahman ZA, Ismail SM, et al. Genetic alterations of chromosome 8 genes in oral cancer. Sci Rep. 2014;4:6073.
44. Gatto F, Ferreira R, Nielsen J. Pan-cancer analysis of the metabolic reaction network. Metab Eng. 2020;57:51–62.
45. Hu Y, Qi MF, Xu QL, Kong XY, Cai R, Chen QQ, et al. Candidate tumor suppressor ZNF154 suppresses invasion and metastasis in NPC by inhibiting the EMT via Wnt/beta-catenin signalling. Oncotarget. 2017;8(49):85749–58.
46. Stirzaker C, Zotenko E, Song JZ, Qu W, Nair SS, Locke WJ, et al. Methylome sequencing in triple-negative breast cancer reveals distinct methylation clusters with prognostic value. Nat Commun. 2015;6:5899.
47. Arai E, Chiku S, Mori T, Gotoh M, Nakagawa T, Fujimoto H, et al. Single-CpG-resolution methylome analysis identifies clinicopathologically aggressive CpG island methylator phenotype clear cell renal cell carcinomas. Carcinogenesis. 2012;33(8):1487–93.
48. Margolin G, Petrykowska HM, Jameel N, Bell DW, Young AC, Elnitski L. Robust detection of DNA hypermethylation of ZNF154 as a pan-cancer locus with in silico modeling for blood-based diagnostic development. J Mol Diagn. 2016;18(2):283–98.
49. Miller BF, Petrykowska HM, Elnitski L. Assessing ZNF154 methylation in patient plasma as a multicancer marker in liquid biopsies from colon, liver, ovarian and pancreatic cancer patients. Sci Rep. 2021;11(1):221.
50. Feng Y, Hu S, Li L, Peng X, Chen F. Long noncoding RNA HOXA-AS2 functions as an oncogene by binding to EZH2 and suppressing LATS2 in acute myeloid leukemia (AML). Cell Death Dis. 2020;11(12):1025.
51. Xiao S, Song B. LncRNA HOXA-AS2 promotes the progression of prostate cancer via targeting miR-509–3p/PBX3 axis. 2020. Biosci Rep. https://doi.org/10.1042/BSR20193287.
52. Song N, Zhang Y, Kong F, Yang H, Ma X. HOXA-AS2 promotes type I endometrial carcinoma via miRNA-302c-3p-mediated regulation of ZFX. Cancer Cell Int. 2020;20:359.
53. Tong G, Wu X, Cheng B, Li L, Li X, Li Z, et al. Knockdown of HOXA-AS2 suppresses proliferation and induces apoptosis in colorectal cancer. Am J Transl Res. 2017;9(10):4545–52.
54. Wang L, Zhang X. Knockdown of lncRNA HOXA-AS2 inhibits viability, migration and invasion of osteosarcoma cells by miR-124-3p/E2F3. Onco Targets Ther. 2019;12:10851–61.
55. Zhang H, Liu Y, Yan L, Zhang M, Yu X, Du W, et al. Increased levels of the long noncoding RNA, HOXA-AS3, promote proliferation of A549 cells. Cell Death Dis. 2018;9(6):707.
56. Han P, Liu Q, Xiang J. Monitoring methylation-driven genes as prognostic biomarkers in patients with lung squamous cell cancer. Oncol Lett. 2020;19(1):707–16.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.