Analysis Using Random Forest and Data Integration Methods of Non-promoter Methylation and Expression Data Identifies Dysregulated Genes in Central Carbon Metabolism Pathway in Oral Cancer

Srija Mukhopadhyay[1], Sahana Ghosh[1], Debodipta Das[1], P. Arun[2], Bidyut Roy[3], Nidhan K. Biswas[1], Arindam Maitra[1], Partha P. Majumder[1,3]

[1] National Institute of Biomedical Genomics, Kalyani, India
[2] Tata Medical Centre, Kolkata, India
[3] Indian Statistical Institute, Kolkata, India

Address for Correspondence:

Partha P. Majumder
National Institute of Biomedical Genomics
Kalyani 741251, India
ppm1@nibmg.ac.in

Abstract

Background: Studies of epigenomic alterations associated with diseases primarily focus on methylation profiles of promoter regions of genes, but not of other genomic regions. In our past work (Das et al. 2019) on patients suffering from gingivo-buccal oral cancer – the most prevalent form of cancer among males in India – we have also focused on promoter methylation changes and resultant impact on transcription profiles. Here, we have investigated alterations in non-promoter (gene-body) methylation profiles and have carried out an integrative analysis of gene-body methylation and transcriptomic data of oral cancer patients.

Methods: Tumor and adjacent normal tissue samples were collected from 40 patients. Data on methylation in the non-promoter (gene-body) regions of genes and transcriptome profiles were generated and analyzed. Because of high dimensionality and highly correlated nature of these data, we have used Random Forest (RF) and other data-analytical methods.

Results: Our integrative analysis of non-promoter methylation and transcriptome data has revealed significant methylation-driven alterations in some genes that also significantly impact on their transcription levels. These changes result in enrichment of the Central Carbon Metabolism (CCM) pathway, primarily by dysregulation (overexpression) of (a) *NTRK3*, which plays a dual role as an oncogene and a tumor suppressor; (b) *SLC7A5* (*LAT1*) which is a transporter dedicated to essential amino acids, and is overexpressed in cancer cells to meet the increased demand for nutrients that include glucose and essential amino acids; and, (c) *EGFR* which has been earlier implicated in progression, recurrence, and stemness of oral cancer, but we provide evidence of epigenetic impact on overexpression of this gene for the first time.

Conclusions: In rapidly dividing cancer cells, metabolic reprogramming from normal cells takes place to enable enhanced proliferation. In the present analysis, we have identified that among oral cancer patients, genes in the CCM pathway – that plays a

fundamental role in metabolic reprogramming – are significantly dysregulated because of perturbation of methylation in non-promoter regions of the genome. This result compliments our previous result that perturbation of promoter methylation results in significant changes in key genes that regulate the feedback process of DNA methylation for the maintenance of normal cell division.

Keywords: Random Forest; Epigenomic; Transcriptomic; Integrative analysis; Gingivo-buccal oral cancer.

Background

For various cancers both DNA methylation and gene expression data have been analyzed separately and alterations have been found to be associated with susceptibility and outcome [1,2]. It is well known that DNA methylation impacts on gene expression. Therefore, attempts have been made to perform integrative analyses of these two types of data to draw robust inferences [3]. Various methods of data integration have been used [4,5]. Methylation and expression data are high volume, highly correlated data. Further, the number of genes or DNA regions/sites on which data are collected are orders of magnitude higher than the number of patients and controls (the so-called “large p, small n” problem in statistics). Therefore, no method of analysis has been universally accepted that takes into account data volume, correlation, and interaction. Random forest (RF) is a machine learning inferential method that is data-adaptive and tree-based. It handles correlated and large data sets very efficiently and is, therefore particularly appealing for analysis of high-dimensional genome data. Normally, only a small portion of a high-dimensional data is associated with a phenotype. A regression framework does not apply to this scenario. The highly correlated nature of genomic data also makes the application of standard statistical models inappropriate. RF is a non-parametric tree-based approach that is particularly suited for such data-analysis problems. RF can also be used to select and rank variables by taking advantage of variable importance measures. A good review of RF in genomic data analysis can be found in [6].

We have used RF methodology to identify gene-body methylation differences between tumor and adjacent normal tissues in patients with oral squamous cell carcinoma of the gingivo-buccal region (OSCC-GB), the most common form of oral cancer in India [7,8]. We then integrated the knowledge thus obtained with data on levels of transcription of genes, which we use as a proxy for gene-expression levels, to discover methylation-driven alterations in the gene-body regions of the genome that significantly associate with dysregulation of genes in oral cancer.

DNA methylation occurs predominantly on cytosines followed by guanine residues (CpG). This type of methylation is referred to as CpG methylation. We had earlier analyzed data on methylation in CpG sites in the known promoter regions of all genes, but we had ignored gene-body CpG sites; sites that are on the coding regions of genes [4]. By application of modern data-adaptive method (RF) on gene-body methylation data and subsequent integration with gene expression data, we have identified some dysregulated genes and a pathway that were not identified in our earlier [4] analysis of promoter methylation and expression.

Methods

Patient Recruitment and Sample Collection

This study was approved by the Institutional Ethics Committees of the Tata Medical Centre and the National Institute of Biomedical Genomics, India. Patients suffering from oral squamous cell carcinoma of the gingivobuccal region (OSCC-GB) were recruited into this study with written informed consent. From each patient, a sample of tumor tissue and adjacent normal tissue were sampled by one of us (P.A.). The tissue samples were stored appropriately.

DNA Methylation

Methylation data from paired tumour and adjacent normal tissue samples of 40 OSCC-GB patients were generated using the Illumina Infinium MethylationEPIC BeadChip [4]. Using the R package minfi, we estimated for each CpG site, the CpG-specific methylation level (β-value) as the ratio of the intensity of methylated (M) to the combined intensities of both methylated (M) and unmethylated (U) alleles :

$$\beta = \frac{M^*}{M^* + U^* + C}$$

where M^* and U^* denote signal intensities of M and U alleles, respectively, and the constant C set at 100 (as recommended by the BeadChip manufacturer) [4,9,10]. The β-value ranges from 0 (unmethylated) to 1 (methylated). The sites that had a detection p-value ≥ 0.01 and those that mapped to X or Y chromosomes were removed. We further removed from final analysis data on (a) SNP associated probes with minor allele frequency (MAF) > 0.01, (b) probes that mapped to non-coding regions of the genome that are devoid of protein-coding genes, (c) multi-mapped probes, (d) probes that did not map to annotated genes [4,9,11,12], and (e) probes that mapped to 3'UTR region of the genome.

Random Forest Classifier

To analyze the difference between Tumor and Normal samples, a Random Forest (RF) method was used on Methylation data as implemented in the *randomForest* package in R [9,11-16]. The random forest algorithm is an ensemble classifier similar to Classification and Regression Tree (CART) [14]. Each tree in an RF is built by choosing a bootstrap sample of two-third of the total number of individuals; the remaining one-third (Out-Of-Bag [OOB] sample) is utilised for validation. For each node in a tree, a binary splitting rule is used on a sample of CpG sites from the bootstrap sample to find the best split. The variable with the maximum information gain [17] is selected. A parameter *mtry* defines the number of variables randomly selected for each node in a tree, and another parameter *ntree* specifies the number of trees to be built in a forest. Normally, the value of *mtry* is taken to be the square root of the number of variables; this is also the default value in the R package. The output of *randomForest* provides an aggregated misclassification error (OOB error rate), which is estimated from predictions made on the OOB samples, and variable importance, which measures the weighted mean of the improvement in individual trees by each variable [12-14,18]. The most reliable variable importance method is "permutation accuracy importance" or "Mean Decrease Accuracy" (MDA) [18,19]. MDA permutes the data of i^{th} variable in the OOB sample and records the permuted OOB error rate. The difference of the original and permuted OOB error rate averaged over the number the trees gives the importance

score for i^{th} variable (VI_i) in the random forest [16,18-20]. A high value of MDA implies greater importance of the variable [18,19].

$$VI_i = \frac{1}{ntree} \sum_{j=1}^{ntree} (OOBerror_{ij}^{permuted} - OOBerror_{ij})$$

Classification of samples

For efficient computation, only probes with |average Δβ| ≥ 0.2 were considered. A CpG site was considered hypermethylated if average Δβ ≥ 0.2 and hypomethylated if average Δβ ≤ -0.2 [4]. Before implementing the random forest (RF) classifier, ntree and mtry parameters were tuned to generate an accuracy rate [9,13]. The best performing combination of parameters were those for which the OOB error rate stabilised and reached a minimum; i.e., the combination of parameters with the highest accuracy rate. Once the optimum set of parameters was determined, "randomForest" was executed 50 times on the methylation data of 40 paired samples. In each iteration we selected only variables (probes) with MDA-score > 0 [15]. The selected probes were then mapped to their respective genes. A gene was considered for further analyses if it satisfied the following conditions: (a) there were at least two probes in the non-promoter region of the gene, (b) methylation status of all probes in the non-promoter region were unidirectional; either hypermethylated or hypomethylated, and (c) had no probes in the promoter region. The stringency of criteria (a) and (b) were adopted to minimize the chance of false-positive discovery, and the criterion (c) was adopted to make discoveries attributable to gene-body methylation only.

RNA Sequencing

RNA was extracted and RNA sequencing was performed to obtain levels of transcription of genes, on the same set of 40 paired samples. Paired-end libraries were constructed and sequenced using Illumina HiSeq2500 [4,21]. The quality of the RNA-Seq reads was checked by FastQC. *TopHat2* [4,21-23] was then used to align these reads to a hg19 reference transcriptome or genome. Multi-mapped reads and non-concordant reads

were filtered out using *SAMtools* [4] and duplicate reads were removed using *MarkDuplicates* from PICARD [4]. *Cufflinks* [4,21-23] was then used to assemble and reconstruct the transcriptome. Finally, using *Cuffnorm*, normalised FPKM values for each gene were estimated [4]. Only those genes that had non-zero levels of transcription levels in all samples were considered for further analysis. We have used the level of transcription of a gene as a proxy for the level of expression of the gene, and have used transcription and expression levels interchangeably in this report.

Integration of Methylation and Transcription data

Those genes for which there was no promoter probe and with multiple probes in the non-promoter region that were uniformly hyper- or hypo-methylated, and for which the level of transcription/expression change between tumour and normal tissues, averaged over the 40 pairs of samples, was higher than two-fold, were identified to be dysregulated by methylation in non-promoter regions [4]. The genes that had 1st exon and exon boundary probes were removed. Finally, we considered only those genes for mapping on pathways that satisfied the known biological directionality of control; genes with hypermethylation (hypomethylation) in the gene-body region in the tumour tissue should have a significantly higher (lower) level of expression in the tumor tissue [24,25].

Enrichment analysis of pathways

Genes that were so identified by the integration of both methylation and expression data were analyzed for enrichment of biological pathways. We considered pathways in KEGG for this analysis. ClueGo and CluePedia plug-ins of Cytoscape were used. To identify whether a pathway in KEGG was significantly enriched, a right-sided test based on hypergeometric distribution was used. Benjamini-Hochberg correction method was used to correct the *p*-values for multiple testing [4,26].

Results

Identification of genes with abundant methylation in the non-promoter region

A total of 484,420 autosomal probes with detection p-value < 0.01 were associated with 18,688 genes. After removing 3'UTR and unannotated probes, 333,208 probes remained which were associated with 18,684 genes. Of these, 22,711 probes were with |average Δβ| ≥ 0.2 that mapped to 7,027 genes. By fine-tuning [Additional File 1], a stable OOB error rate was obtained with default mtry=150 and ntree=2,000. Random forest was executed 50 times, with these optimal values of the parameters. The MDA scores of each variable and OOB error rate were recorded for 50 iterations. A uniform OOB error rate of 1.25% was observed in each iteration [Additional File 2]. The set of probes with MDA>0 comprised 10,105 probes that mapped to 4,831 genes. Among these, for 433 genes all probes in the non-promoter region were hypermethylated, and for 233 genes all were hypomethylated. We have focused on these 666 unidirectionally methylated genes for drawing further inferences integrated with gene expression patterns in tumor-normal paired tissues.

Integration of Methylation and Gene-Expression

In paired tissues collected from the 40 OSCC-GB patients, non-zero levels of transcription/expression were found for 17,464 genes. Considering the 666 genes that exhibited significant and unidirectional methylation, we found that 132 of these genes showed at least two-fold difference in the level of expression between tumour and normal tissues, averaged over the 40 patients. Of these 132 genes, 8 genes were removed as they had 1st exon and exon boundary probes. However, of these 124 only for 67 (54%) genes, the direction of change of expression level was consistent with that of methylation change [Additional File 3]. That is, genes with hypermethylation (hypomethylation) in the tumour tissue had significantly higher (lower) levels of expression in the tumor tissue.

Enriched pathway

The pathway enrichment analysis using the 67 genes dysregulated by methylation alteration in the gene-body region between tumour and normal tissues, identified enrichment of one significant (corrected p-value = 0.0012) KEGG pathway. This was Central Carbon metabolism in Cancer with three associated genes EGFR, NTRK3, SLC7A5. It has been reported, based on cell line studies, that overexpression of EGFR can impact on the development of solid tumors, including oral cancer [27]. We have found EGFR overexpressed and global hypermethylated.

Discussion

By applying the novel Random Forest data-adaptive method to high-dimensional data (about 500,000 data points per individual) to identify significant alterations in gene-body methylation in gingivo-buccal oral tumor tissue compared to adjacent normal tissue, and subsequent integration with gene expression data we have detected some genes and pathways not earlier inferred to be involved in OSCC-GB only through cell-line studies. The significantly enriched pathway that has been identified using this data-adaptive and data-integrative approach is the Central Carbon Metabolism (CCM) pathway, which is involved in transport and oxidation of main carbon sources inside the cell. Fundamental cellular processes require energy for growth. The catabolic and anabolic reactions in metabolism are finely balanced and tightly regulated. Dysregulation results in cellular transformation and tumor progression. In rapidly dividing cancer cells, metabolic reprogramming from normal cells takes place to enable enhanced proliferation. CCM pathway plays a fundamental role in metabolic reprogramming. Changes in central carbon metabolism of cancer stem cells have also been noted [28]. It is noteworthy that enrichment of the CCM pathway in OSCC-GB takes place by gene-body methylation mediated dysregulation of three key genes, EGFR, NTRK3, SLC7A5.

Significant downregulation of NTRK3 mediated by promoter methylation was noted in our earlier study [4]. NTRK3 is a neurotrophin receptor. It behaves as an oncogene in breast cancer [29,30] and possibly also in hepatocellular carcinoma [31].

However, it also plays a dual function. It acts as a tumor suppressor in colorectal cancer in which it is epigenetically inactivated [32]. In OSCC-GB also, NTRK3 is epigenetically dysregulated and appears to behave as a tumor suppressor.

SLC7A5 – earlier known as LAT1 – is a transporter dedicated to essential amino acids. Cancer cells have an increased demand for nutrients that include glucose and essential amino acids; the so-called “Warburg effect.” Overexpression of SLC7A5, as we have observed here, is explained in part by the presence, in its promoter, of a canonical binding site for the proto-oncogene c-Myc [33] that is known to regulate glucose metabolism [34]. Overexpression of SLC7A5 is also controlled by methylation in the promoter [4] and non-promoter regions (this study).

EGFR has been earlier implicated in progression, recurrence, and stemness of oral cancer [35,36]. EGFR is inappropriately activated in cancer mainly because of amplification and point mutations. Transcriptional upregulation of EGFR due to autocrine/paracrine mechanisms has also been described [37]. Here, we have, for the first time, shown that dysregulation of EGFR takes place by epigenetic mechanisms in oral cancer.

Conclusions

Three key genes NTRK3, SLC7A5 (LAT1) and EGFR were overexpressed in the CCM pathway. Of these, NTRK3 [4,38] and SLC7A5 [4] were earlier identified to be associated with oral cancer. However, we provide the first evidence of epigenetic impact on overexpression of EGFR in oral cancer. To enable enhanced proliferation of cells in a cancer tissue, metabolic reprogramming from normal cells usually takes place. In the present analysis, we have identified that among oral cancer patients, genes in the CCM pathway – that plays a fundamental role in metabolic reprogramming – are significantly dysregulated because of perturbation of methylation in non-promoter regions of the genome. This result compliments our previous result that perturbation of promoter methylation results in significant changes in key genes that regulate the feedback

process of DNA methylation for the maintenance of normal cell division. Taken together, it is evident that in oral cancer methylation driven alterations in both promoter and non-promoter genomic regions result in disruption of normal cell division accompanied by metabolic reprogramming to enable rapid cell proliferation.

Declaration

Abbreviations

RF: Random Forest; OSCC-GB: Oral squamous cell carcinoma of the gingivobuccal region; CCM: Central Carbon Metabolism pathway; MAF: Minor allele frequency; CART: Classification and Regression trees; OOB: Out-of-Bag sample; MDA: Mean decrease accuracy; VI: Variable importance; FPKM: Fragments Per Kilobase of transcript per Million mapped reads; KEGG: Kyoto Encyclopedia of Genes and Genomes.

Acknowledgements

We are grateful to all participating members of Systems Medicine Cluster (SyMeC) and International Cancer Genome Consortium (ICGC) India project for their guidance and advice during the course of this study.

Authors' Contributions

PPM and SM conceived of the study and designed the analysis of data. SM carried out data analyses. PA coordinated patient recruitment and sample collection. AM, NKB, DD and SG coordinated data generation and data collection. PPM and BR contributed towards the fine-tuning of the method. PPM and SM wrote the manuscript. PPM, DD, SG, PA, BR, AM and NKB edited the draft manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the J.C. Bose National Fellowship to PPM. The present work was supported by a grant from the Department of Biotechnology (DBT), Govt. of India, through the SyMeC project (BT/Med-II/NIBMG/SyMeC/2014/Vol. II).

Competing Interests

The authors declare that they have no competing interest.

Ethics approval and consent to participate

The study was approved by the Institutional Ethics Committees of Dr. R. Ahmed Dental College & Hospital (RADCH), Kolkata, Chittaranjan National Cancer Institute (CNCI), Kolkata, National Institute of Biomedical Genomics (NIBMG), Kalyani, and Indian Statistical Institute (ISI), Kolkata. Prior written informed consent was obtained from each study participant.

Consent for publication

All authors have agreed to publish in BMC Cancer.

Availability of data and materials

Raw IDAT files of 40 samples generated using Illumina Infinium methylation array were deposited under EGAS00001003896 EGA study ID and aligned bam files for transcriptome data of 40 samples were deposited under EGAS00001003893 EGA study ID. Biospecimens may be shared on request, if not exhausted. Dispatch of biospecimens requires prior approval from the Government of India.

References

1. Jiao, Y., Widschwendter, M., Teschendorff, A., E. A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control.

Bioinformatics. 2014; 30:2360–2366. https://doi.org/10.1093/bioinformatics/btu316.

2. Udali, S., Guarini, P., Ruzzenente, A., Ferrarini, A, Guglielmi, A., Lotto, V., Tononi, P., Pattini, P., Moruzzi, S., Campagnaro, T., Conci, S., Olivieri, O., Corrocher, R., Delledonne, M., Choi, S.-W., Friso, S. DNA methylation and gene expression profiles show novel regulatory pathways in hepatocellular carcinoma. *Clin Epigenet*. 2015; 7:43. https://doi.org/10.1186/s13148-015-0077-1.
3. Li, M., Balch, C., Montgomery, J., S., Jeong, M., Chung, J., H., Yan, P., Huang, T., H., M., Kim, S., Nephew, K., P. Integrated analysis of DNA methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer. *BMC Med Genomics*. 2009; 2:34. https://doi.org/10.1186/1755-8794-2-34.
4. Das, D., Ghosh, S., Maitra, A., Biswas, N., K., Panda, C., K., Roy, B., Sarin, R., Majumder, P., P. Epigenomic dysregulation-mediated alterations of key biological pathways and tumor immune evasion are hallmarks of gingivo-buccal oral cancer. *Clin Epigenet*. 2019; 11(1):178. https://doi.org/10.1186/s13148-019-0782-2.
5. Ma, X., Liu, Z., Zhang, Z., Huang, X., Tang, W. Multiple network algorithm for epigenetic modules via the integration of genome-wide DNA methylation and gene expression data. *BMC Bioinformatics*. 2017; 18:72. https://doi.org/10.1186/s12859-017-1490-6.
6. Chen, X., Ishwaran, H. Random forests for genomic data analysis. *Genomics*. 2012; 99:323-329. https://doi.org/10.1016/j.ygeno.2012.04.003.
7. Muttagi, S., S., Patil, B., R., Godhi, A., S., Arora, D., K., Hallikerimath, S., R., Kale, A., D. Clinico-pathological factors affecting lymph node yield in Indian patients with locally advanced squamous cell carcinoma of mandibular Gingivo-Buccal sulcus. *Indian J Cancer*. 2016; 53:239–243. https://doi.org/10.4103/0019-509X.197724.
8. Pathak, K., A., Gupta, S., Talole, S., Khanna, V., Chaturvedi, P., Deshpande, M., S., Pai, P., S., Chaukar, D., A., D'Cruz, A., K. Advanced squamous cell

carcinoma of lower gingivobuccal complex: patterns of spread and failure. *Head Neck*. 2005; 27:597–602. https://doi.org/10.1002/hed.20195.

9. Zhang, W., Spector, T., D., Deloukas, P., Bell, J., T., Engelhardt, B., E. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol*. 2015; 16:14. https://doi.org/10.1186/s13059-015-0581-9.
10. Ma, X., Wang, Y.-W., Zhang, M. Q., & Gazdar, A. F. DNA methylation data analysis and its application to cancer research. *Epigenomics.* 2013; 5(3):301–316. https://doi.org/10.2217/epi.13.26.
11. Everson, T., M., Lyons, G., Zhang, H., Soto-Ramírez, N., Lockett, G., A., V., K., Patil, Merid, S., K., Söderhäll, C., Melén, E., Holloway, J., W., Arshad, S., H., Karmaus, W. DNA methylation loci associated with atopy and high serum IgE: a genome-wide application of recursive Random Forest feature selection. *Genome Med.* 2015; 7:89. https://doi.org/10.1186/s13073-015-0213-8.
12. Naue, J., Hoefsloot, H. C. J., Mook, O. R. F., Rijlaarsdam-Hoekstra, L., van der Zwalm, M. C. H., Henneman, P., Kloosterman, A. D., Verschure, P. J. Chronological age prediction based on DNA methylation: Massive parallel sequencing and random forest regression. *Forensic Science International: Genetics.* 2017; 31: 19–28. https://doi.org/10.1016/j.fsigen.2017.07.015.
13. Houseman, E., A., Christensen, B., C., Yeh, R.-F., Marsit, C., J., Karagas, M., R., Wrensch, M., Nelson, H., H., Wiemels, J., Zheng, S., Wiencke, J., K., Kelsey, K., T. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics.* 2018; 9:365. https://doi.org/10.1186/1471-2105-9-365.
14. Christensen, B., C., Houseman, E., A., Marsit, C., J., Zheng, S., Wrensch, M., R., Wiemels, J., L., Nelson, H., H., Karagas, M., R., Padbury, J., F., Bueno, R., Sugarbaker, D., J., Yeh, R.-F., Wiencke, J., K., Kelsey, K., T. Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context. *PLoS Genet*. 2009; 5(8): e1000602. https://doi.org/10.1371/journal.pgen.1000602.

15. Yang, Y., Nephew, K., Kim, S. A novel k-mer mixture logistic regression for methylation susceptibility modeling of CpG dinucleotides in human gene promoters. *BMC Bioinformatics*. 2012; 13:S15. https://doi.org/10.1186/1471-2105-13-S3-S15.
16. Archer, K. J., Kimes, R. V. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis.* 2008; 52(4):2249–2260. https://doi.org/10.1016/j.csda.2007.08.015.
17. Deng, H., Runger, G. Gene selection with guided regularized random forest. *Pattern Recognition*. 2013; 46:3483-3489. http://dx.doi.org/10.1016/j.patcog.2013.05.018.
18. Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., Zeileis, A. Conditional variable importance for random forests. *BMC Bioinformatics.* 2008; 9:307. https://doi.org/10.1186/1471-2105-9-307.
19. Grömping, U. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician*. 2009; 63(4):308–319. https://doi.org/10.1198/tast.2009.08199.
20. Strobl, C., Boulesteix, A., Zeileis, A., Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*. 2007; 8:25. https://doi.org/10.1186/1471-2105-8-25.
21. Ghosh, S., & Chan, C.-K. K. Analysis of RNA-Seq Data Using TopHat and Cufflinks. *Methods in Molecular Biology.* 2016; 1374:339–361. https://doi.org/10.1007/978-1-4939-3167-5_18.
22. Chu, Y., & Corey, D. R. RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Therapeutics.* 2012; 22(4):271–274. https://doi.org/10.1089/nat.2012.0367.
23. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D., R., Pimentel, H., Salzberg, S., L., Rinn, J., L., Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012; 7:562–578. https://doi.org/10.1038/nprot.2012.016.

24. Jjingo, D., Conley, A., B., Yi, S., V., Lunyak, V., V., Jordan, I. On the presence and role of human gene-body DNA methylation. *Oncotarget.* 2012; 3:462-474. https://doi.org/10.18632/oncotarget.497.
25. Yang, X., Han, H., Carvalho, D., D, D., Lay, F., D., Jones, P., A., Liang G. Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. *Cancer Cell*. 2014; 26:1-14. http://dx.doi.org/10.1016/j.ccr.2014.07.028.
26. Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.-H., Pages, F., Trajanoski, Z., Galon, J. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. 2009; 25(8):1091–1093. https://doi.org/10.1093/bioinformatics/btp101.
27. Huang, C.-Y., Chan, C.-Y., Chou, I.-T., Lien, C.-H., Hung, H.-C., & Lee, M.-F. Quercetin induces growth arrest through activation of FOXO1 transcription factor in EGFR-overexpressing oral cancer cells. *The Journal of Nutritional Biochemistry*. 2013; 24(9):1596–1603. https://doi.org/10.1016/j.jnutbio.2013.01.010.
28. Wong, T., L., Che, N., Ma, S. Reprogramming of central carbon metabolism in cancer stem cells. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease.* 2017; 1863:1728-1738. https://doi.org/10.1016/j.bbadis.2017.05.012.
29. Li, Z., Tognon, C., E., Godinho, F., J., Yasaitis, L., Hock, H., Herschkowitz, J., I., Lannon, C., L., Cho, E., Kim, S.-J., Bronson, R., T., Perou, C., M., Sorensen, P., H., Orkin, S., H. ETV6-NTRK3 Fusion Oncogene Initiates Breast Cancer from Committed Mammary Progenitors via Activation of AP1 Complex. *Cancer Cell.* 2007; 12:542-558. https://doi.org/10.1016/j.ccr.2007.11.012.
30. Tognon, C., Knezevich, S., R., Huntsman, D., Roskelley, C., D., Melnyk, N., Mathers, J., A., Becker, L., Carneiro, F., MacPherson, N., Horsman, D., Poremba, C., Sorensen, P., H., B. Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. *Cancer Cell.* 2002; 2:367-376. https://doi.org/10.1016/S1535-6108(02)00180-0.
31. Xiong, D., Sheng, Y., Ding, S., Chen, J., Tan, X., Zeng, T., Qin, D., Zhu, L., Huang, A., Tang, H. LINC00052 regulates the expression of NTRK3 by miR-128

and miR-485-3p to strengthen HCC cells invasion and migration. *Oncotarget*. 2016; 7(30): 47593–47608. https://doi.org/10.18632/oncotarget.10250.

32. Luo, Y., Kaz, A., M., Kanngurn, S., Welsch, P., Morris, S., M., Wang, J., Lutterbaugh, J., D., Markowitz, S., D., Grady, W., M. NTRK3 Is a Potential Tumor Suppressor Gene Commonly Inactivated by Epigenetic Mechanisms in Colorectal Cancer. *PLoS Genet*. 2013; 9(7): e1003552. https://doi.org/10.1371/journal.pgen.1003552.
33. Hayashi, K., Jutabha, P., Endou, H., Anzai, N. c-Myc is crucial for the expression of LAT1 in MIA Paca-2 human pancreatic cancer cells. *Oncology Reports.* 2012; 28(3):862-866. https://doi.org/10.3892/or.2012.1878.
34. Kim, J., W., Zeller, K., I., Wang, Y., Jegga, A., G., Aronow, B., J., O'Donnell, K., A., Dang, C., V. Evaluation of myc E-box phylogenetic footprints in glycolytic genes by chromatin immunoprecipitation assays. *Mol. Cell. Biol.* 2004; 24:5923–5936. https://doi.org/10.1128/MCB.24.13.5923-5936.2004.
35. Mirza, Y., Ali, S., M., A., Awan, M., S., Idress, R., Naeem, S., Zahid, N., Qadeer, U. Overexpression of EGFR in Oral Premalignant Lesions and OSCC and Its Impact on Survival and Recurrence. *Oncomedicine*. 2018; 3:28-36. https://doi.org/10.7150/oncm.22614.
36. Lv, X.-X., Zheng, X.-Y., Yu, J.-J., Ma, H.-R., Hua, C., Gao., R.-T. EGFR enhances the stemness and progression of oral cancer through inhibiting autophagic degradation of SOX2. *Cancer Med*. 2019; 00:1–10. https://doi.org/10.1002/cam4.2772.
37. Wilson, K., J., Mill, C., Lambert, S., Buchman, J., Wilson, T., R., Hernandez-Gordillo, V., Gallo, R., M., Ades, L., M., C., Settleman, J., Riese II, D., J. EGFR ligands exhibit functional differences in models of paracrine and autocrine signaling. *Growth Factors*. 2012; 30(2):107-116. https://doi.org/10.3109/08977194.2011.649918.
38. Campbell, P., J., Getz, G., Korbel, J., O. *et al*. Pan-cancer analysis of whole genomes. *Nature,* 2020; 578: 82–93. https://doi.org/10.1038/s41586-020-1969-6.

Supplementary Information

Additional File 1.

Title: Fine-tuning *randomForest* parameters *mtry* and *ntree*.

Legend: (a) Tuning *ntree* using R package "*tuneNTREE*" with values starting from 500 to 10,000. Each *ntree* value was iterated 5 times and obtained a 1.25% error rate. Error bar is displaying the standard deviation of the OOB error rate for each *ntree*. (b) Tuning *mtry* using R package "*tuneRF*" with mtry values ½√p, √p (default value=150) and 2√p, where p is the number of variables = 22,711 and ntree value 2000 for each 50 iterations. The error rate for ½√p, √p and 2√p for each iteration was 1.25%. Error bar is displaying the standard deviation of OOB error rate for different *mtry* in each iteration.

Additional File 2.

Title: Results of 50 iterations of tuned Random Forest classifier showing the number of variables selected and the OOB error rate.

Additional File 3.

Title: Results of 72 genes that showed significant relationship between methylation in the non-promoter region and gene expression.