SCIENTIFIC REPORTS
natureresearch

OPEN

Differentially Methylated Super-Enhancers Regulate Target Gene Expression in Human Cancer

Emily L. Flam[1,8], Ludmila Danilova[2,3,8], Dylan Z. Kelley[1], Elena Stavrovskaya[4,5], Theresa Guo[1], Michael Considine[2], Jiang Qian[6], Joseph A. Califano[7], Alexander Favorov[2,3], Elana J. Fertig[2,8*] & Daria A. Gaykalova[1,8*]

Current literature suggests that epigenetically regulated super-enhancers (SEs) are drivers of aberrant gene expression in cancers. Many tumor types are still missing chromatin data to define cancer-specific SEs and their role in carcinogenesis. In this work, we develop a simple pipeline, which can utilize chromatin data from etiologically similar tumors to discover tissue-specific SEs and their target genes using gene expression and DNA methylation data. As an example, we applied our pipeline to human papillomavirus-related oropharyngeal squamous cell carcinoma (HPV + OPSCC). This tumor type is characterized by abundant gene expression changes, which cannot be explained by genetic alterations alone. Chromatin data are still limited for this disease, so we used 3627 SE elements from public domain data for closely related tissues, including normal and tumor lung, and cervical cancer cell lines. We integrated the available DNA methylation and gene expression data for HPV + OPSCC samples to filter the candidate SEs to identify functional SEs and their affected targets, which are essential for cancer development. Overall, we found 159 differentially methylated SEs, including 87 SEs that actively regulate expression of 150 nearby genes (211 SE-gene pairs) in HPV + OPSCC. Of these, 132 SE-gene pairs were validated in a related TCGA cohort. Pathway analysis revealed that the SE-regulated genes were associated with pathways known to regulate nasopharyngeal, breast, melanoma, and bladder carcinogenesis and are regulated by the epigenetic landscape in those cancers. Thus, we propose that gene expression in HPV + OPSCC may be controlled by epigenetic alterations in SE elements, which are common between related tissues. Our pipeline can utilize a diversity of data inputs and can be further adapted to SE analysis of diseased and non-diseased tissues from different organisms.

Super-enhancers (SEs) are tissue- and disease-specific regulatory genomic elements related to chromatin that drive cell-specific gene expression changes in development, differentiation, and disease progression, including cancer[1,2]. SEs are enriched for binding of many transcription factors, as well as Mediator, RNA polymerase II, and BRD4 proteins, which they bring to the promoter regions of *in cis* target genes through the formation of chromatin loops[2–7]. SEs are marked by specific histone modifications, such as H3K27ac and H3K4me1[3,8], suggesting an essential role of the chromatin landscape in SE-mediated gene expression regulation. SEs can cover up to 300 kb regions[9] and influence the expression of genes with transcription start sites (TSS) up to 1.5 Mbp away[3,10]. Genes regulated by SEs are more expressed than those regulated by typical enhancers and are often associated with tissue-specific or disease-specific cell-identity[2,3,11]. Moreover, the chromatin landscape predetermines disease-specific genetic alterations in cancer[12]. SEs can appear *de novo* during carcinogenesis in proximity to their cancer-related gene targets, causing changes in the relative gene expression of multiple genes simultaneously[13,14].

[1]Department of Otolaryngology—Head and Neck Surgery, Johns Hopkins Medical Institutions, Baltimore, Maryland, USA. [2]Division of Oncology Biostatistics and Bioinformatics, Department of Oncology, Johns Hopkins Medical Institutions, Baltimore, Maryland, USA. [3]Laboratory of Systems Biology and Computational Genetics, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia. [4]Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119992, Russia. [5]Institute for Information Transmission Problems, RAS, Moscow, 127994, Russia. [6]Department of Ophthalmology, Johns Hopkins Medical Institutions, Baltimore, Maryland, USA. [7]Department of Surgery, Head and Neck Cancer Center, University of California, San Diego, California, USA. [8]These authors contributed equally: Emily L. Flam and Ludmila Danilova. [9]These authors jointly supervised this work: Elana J. Fertig and Daria A. Gaykalova. *email: ejfertig@jhmi.edu; dgaykal@jhmi.edu

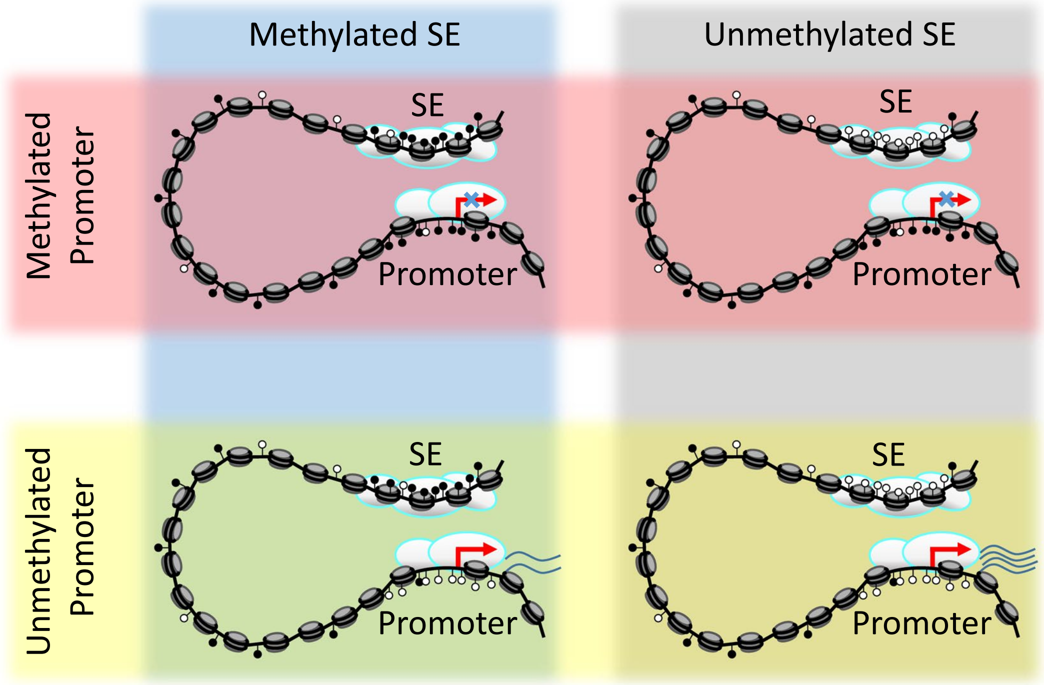


Figure 1. Scheme of how DNA methylation of either promoter or SE can affect target gene expression. Methylation at the promoter region prevents the expression of a gene, regardless of SE methylation. Genes can be minimally expressed with an unmethylated promoter, even with a methylated SE region, but maximal gene expression is reached with both an unmethylated promoter and an unmethylated SE.

Differential methylation has been used as a hallmark for SE detection[8,15–18]. Therefore, it is critical to study the interplay between DNA methylation of SEs and their effects on the regulation of target gene expression. SE methylation can be altered during the carcinogenesis process, independent of global disease-specific methylation changes in cancer cells[8,16]. We propose that hypomethylation of SE regions causes increased expression of nearby oncogenic target genes and that hypermethylation of these SEs causes decreased expression of targets, especially tumor suppressors[8,15,19].

SEs can be detected through chromatin-focused studies, such as chromatin immunoprecipitation with high throughput sequencing (ChIP-Seq)[3] or assay for transposase-accessible chromatin with high throughput sequencing (ATAC-Seq)[13]. While SEs are currently recognized as one of the main drivers of carcinogenesis, many tumor types are still missing information about SE locations. This lack of information can be explained by the challenges of chromatin analysis procedures on clinical biopsy tissues[20]. Nevertheless, annotation and characterization of the SEs based on H3K27Ac chromatin data have been presented in numerous cancer cell lines[2]. We hypothesized that etiologically similar tumors might share a portion of SE regions to regulate similar genes due to the genetic and epigenetic similarities noticed between certain tumor types[13]. Therefore, it is possible to utilize chromatin-related data available for etiologically-relevant tissues for the discovery of SEs in other tumor types that still lack chromatin data. This type of data is widely available for diverse tumor types through projects like The Cancer Genome Atlas (TCGA) and can be used to navigate through SE candidates from etiologically relevant samples.

In this study, we introduce a pipeline to detect SE regions by using gene expression and DNA methylation for a particular set of samples. This pipeline helps to define the role of SE methylation in target gene expression in human carcinomas. We built our pipeline under the assumption that maximum gene expression occurs under the condition that both the gene promoter region and nearby SE region are hypomethylated, while hypermethylation signal from either the promoter or the SE region could diminish target gene expression (Fig. 1 and refs[8,15,19]).

As an example of the power of our pipeline, we studied high-risk human papillomavirus-related oropharyngeal squamous cell carcinoma (HPV + OPSCC), which has limited available chromatin data. We chose this model because the development of HPV + OPSCC cannot be fully explained by its mutational landscape alone[21–25]. Most mutations in OPSCC are found in tumor suppressors[21–23,25,26] and are coupled with pervasive genome-wide alterations to DNA methylation, but these alterations are still insufficient to explain the widespread gene expression changes observed in HPV + OPSCC[21,27,28]. Mutations in chromatin-related genes have been implicated in head and neck carcinogenesis, including K27 and K36 of the histone 3 tail[28–31]. Given the extensive epigenetic changes in HPV + OPSCC, we hypothesized that methylation of SEs is a critical driver of transcriptional changes in carcinogenesis. We utilized our pipeline to identify actionable SE elements using ChIP-Seq data published for lung and HPV + cervical cell lines[2], which are all closely related to HPV + OPSCC[13,21]. We hypothesized that SE regions are conserved between these tissues and HPV + OPSCC, resulting in transcriptional activation of target genes in HPV + OPSCC.

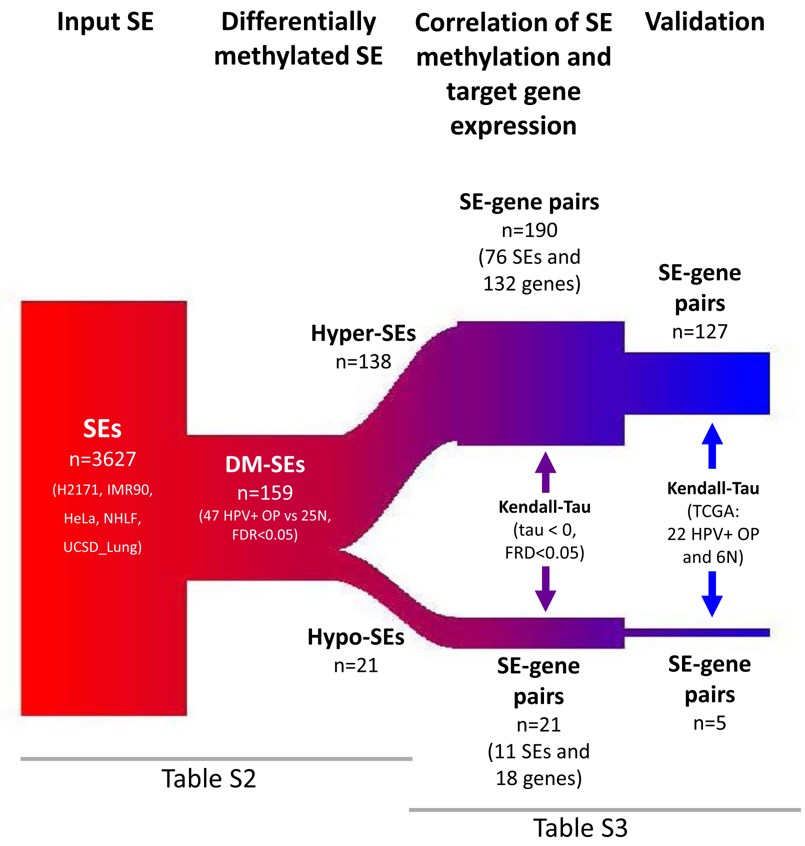


Figure 2. Experimental scheme. Pipeline for analysis of SEs, including SE input and initial filtering steps, detection of differential methylation of SEs, correlation of target gene expression with SE methylation, and validation of results.

Using our pipeline, the candidate SE regions from lung and HPV + cervical samples were filtered through the integrated analysis of gene expression and DNA methylation patterns in our JHU cohort of HPV + OPSCC[32,33]. DNA methylation of gene promoters may also impact gene expression. Therefore, we have also evaluated the methylation status of promoter regions on SE-regulated genes. We identified 211 gene-SE pairs in HPV + OPSCC, of which we have validated n = 132 in the HPV + OPSCC cohort from TCGA project[21]. These critical SE regions suggest epigenetic regulation as a mechanism for HPV + OPSCC carcinogenesis.

We believe our pipeline can be adopted for SE filtering and analysis of other cancer types, non-cancer diseases, and other tissue types for any organism with available data. The pipeline can use SEs for the same tissues or for etiologically-similar samples and can incorporate a range of relevant data input from various platforms.

Results

Pipeline to identify differentially methylated SE that regulate target gene expression.

We developed a pipeline to study the role of SE methylation on target gene expression. The pipeline takes a list of SEs as an input, which can vary depending on the study question, cancer type of interest, data availability, etc. The main steps of the pipeline for the analysis of HPV + OPSCC's SEs are shown in Fig. 2 and include differential methylation analysis of SE and correlation analysis of SE methylation with target gene expression. While we used HPV + OPSCC as an example throughout the manuscript, the pipeline can be applied to any cohort that has DNA methylation and gene expression data available for the same samples. As an output, the pipeline creates SE-gene pairs where methylation of the SE region significantly correlates with target gene expression. This set of pairs can be validated in different cohorts and be supplemented with pathway analysis, functional analysis, correlation analysis, experimental analysis, and so on. The R code for the pipeline is available through this link: https://bitbucket.org/favorov/cervical-lung-se-and-hnscc/downloads/ and in the supplementary information.

Differentially methylated SEs regulate target gene expression in human papillomavirus-related oropharyngeal squamous cell carcinoma.

To demonstrate a possible use for our pipeline, we applied it to find differentially methylated SEs that regulate gene expression in HPV + OPSCC.

Input SEs. As an input list of candidate SEs, we obtained a pooled list of SEs for lung cancer (H2171), non-cancer lung (NHLF and IMR90), and HPV + cervical cancer (HeLa) cell lines and healthy lung tissue (UCSD_Lung) available from a previous study[2] (Table S1) for a total of 3627 candidate SEs (Table S2, Fig. 2). To evaluate whether these candidate SEs were conserved across tissues, we calculated the Jaccard index of 3627 SEs from the five cell lines described above and 54656 SEs from the remaining 81 cell lines available from the same study[2] using the Genometricorr package[34]. We obtained a Jaccard index of 0.17, which means the two sets of SEs had a small intersection, which suggests high conservation between our chosen study input SEs relative to all other tissue types and their tissue- and disease-specificity.

Differentially methylated SE. We analyzed whole-genome DNA methylation in a previously published cohort of n = 47 HPV + OPSCC samples from tumor patients and n = 25 normal mucosal samples from uvulopalatopharyngoplasty surgery patients (UPPP)[20,35]. In this cohort, 0.07% of the genome was differentially methylated in tumor samples relative to normal controls. To validate the correlation between SE location and differential methylation in HPV + OPSCC, we compared the distribution of differentially methylated 100 bp regions relative to candidate SE sites. Differentially methylated regions were significantly overrepresented in the 3627 SEs (GenometriCorr analysis[34]: observed/expected ratio = 3, p-value $< 10^{-16}$) compared to background DNA. Using the Wilcoxon rank sum test, we found that 159 (4%) of 3627 SE candidates were differentially methylated (DM-SEs) between tumor and normal samples in our cohort (FDR < 0.05, Table S2, Figs 2 and 3). These SEs included 138 (87%) SEs that were hypermethylated (Figs 3A, 4A-B and S1A) and 21 (13%) SEs that were hypomethylated (Figs 3B, 5A,B and S2A).

Correlation of SE methylation and target gene expression. To test whether differential methylation of SEs regulates the expression of at least one *in cis* target gene, we gathered all genes with TSS within 1 Mbp of each DM-SE region. Overall, we detected 2,675 genes to be potentially affected by 159 DM-SEs, suggesting that each DM-SE regulates an average of 28 genes (range: 1–91 genes, Table S2). Target genes were located an average of 471,800 bp (range: 20–1,102,000 bp) from the SE region (Fig. S3). For each of the DM-SE-gene pairs, we correlated the methylation of SEs with the expression of target genes using Kendall's tau test. A statistically significant negative correlation (FDR < 0.05) between gene expression and SE methylation was confirmed for 211 SE-gene pairs (Tables S3). Of these pairs, 190 SE-gene pairs (93%) were linked to hypermethylated SEs and 21 pairs to hypomethylated SEs. The overrepresentation of hypermethylated SE-gene pairs is consistent with OPSCC being a tumor suppressor-dysregulated disease[36].

Hypermethylation of SE regions has been linked to silencing of target genes[15]. Of the 138 DM-SEs, 76 (55%) had at least one target gene with decreased expression in cancer (Figs 2 and S4), totaling 190 SE-gene pairs and 132 individual genes (Table S3). An illustration of one representative hypermethylated SE and its target genes is provided in Fig. 4. For this SE, we found eight potential target genes that were under-expressed in tumors relative to normal samples (Figs 4C and S1B). The representative hypermethylated SE from Fig. 4 is linked to lower expression of *SMAGP* (Small Cell Adhesion Glycoprotein, Fig. S5A-B), which plays a role in epithelial cell-cell contact[37].

Differential hypomethylation of SEs is linked to activation of oncogene expression[8,15,38]. We documented an increased expression of 18 genes linked to 11 (52%) out of 21 hypomethylated SE regions, which formed 21 SE-gene pairs (Fig. 2 and Table S3). An illustration of one representative hypomethylated SEs and its target genes is provided in Fig. 5. For this SE, we found seven potential target genes with overexpression in tumors relative to normal samples (Figs 5C and S2B). Hypomethylation of this SE is strongly linked to overexpression of *GPR107* (G Protein-Coupled Receptor 107, Fig. S5C,D), which is commonly overexpressed in breast cancer patients with worth prognosis[39].

For the genes that had a significant correlation between expression and SE methylation, we also quantified the correlation of its expression with promoter methylation. Most promoters had little to no methylation, and only a small portion of SE-gene pairs showed strong correlation for both promoter and super-enhancer methylation (Figs 4D and 5D). These data emphasize the significant role of super-enhancer methylation on target gene expression.

Validation. To confirm the regulation of gene expression by methylation of the SE regions, we utilized TCGA data from 22 HPV + OPSCC and six normal samples with both DNA methylation and gene expression data[21] to match our JHU cohort. Out of the 211 SE-gene pairs, 132 (63%) had a significant association between SE methylation and gene expression in the TCGA cohort (Table S3 and Figs S1 and S2). Of these 132 validated SE-gene pairs, five were regulated by hypomethylated SEs, and 127 were regulated by hypermethylated SEs (Table S3).

As an additional step of validation, we performed gene set analysis on 132 target genes of hypermethylated SEs. The target genes, whose expression was downregulated in cancer, belong to gene sets such as BREAST CANCER NORMAL LIKE UP, NASOPHARYNGEAL CARCINOMA, DIFFERENTIATING T LYMPHOCYTE, BOUND BY FOXP3, as well as epigenetic related BRAIN HCP WITH H3K4ME3 AND H3K27ME3, ES ICP WITH H3K4ME3, HDAC TARGETS SILENCED BY METHYLATION DN, and many more (Table S4, top). Eighteen upregulated genes linked to hypomethylated SEs were related to BLADDER CANCER WITH LOH IN CHR9Q AND UVEAL MELANOMA UP (Table S4, bottom). This analysis demonstrates that the detected gene targets belong to gene sets that play a significant role in carcinogenesis.

Discussion

In this study, we developed a new bioinformatics pipeline to define and evaluate differentially methylated SEs that regulate target gene expression. Depending on the scientific question, data availability, or disease, the pipeline can be applied to any list of SEs or other defined DNA region, including from other organisms and experimental models. The code for the pipeline is publicly available and can be adapted to use on any cohort of samples that have parallel DNA methylation and gene expression data available. In this paper, we demonstrated how our algorithm integrates methylation and expression data with a list of potential SE sites for the discovery of functional, tissue-specific SEs in OPSCC. The algorithm can use a list of SEs identified from the same study samples or from

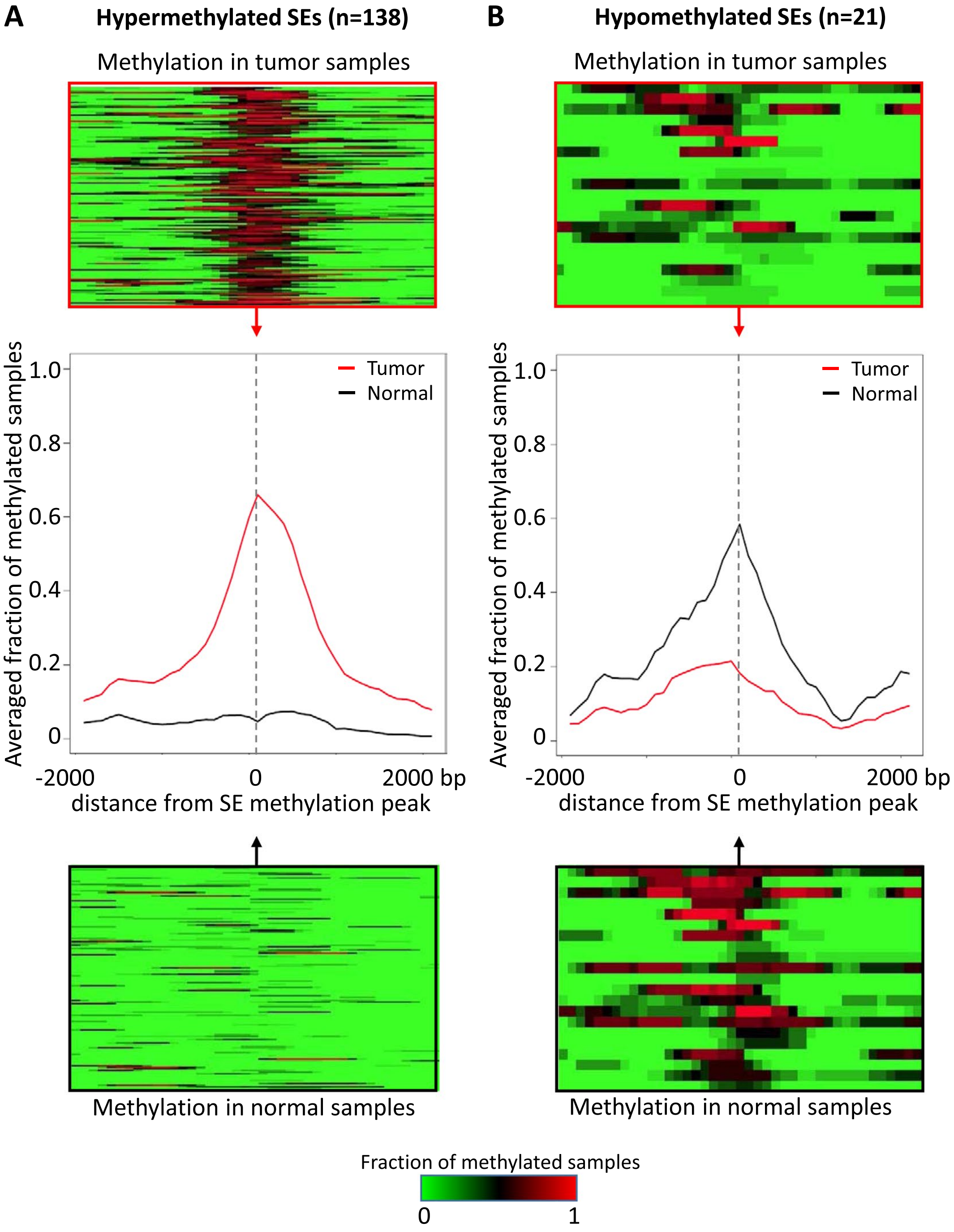
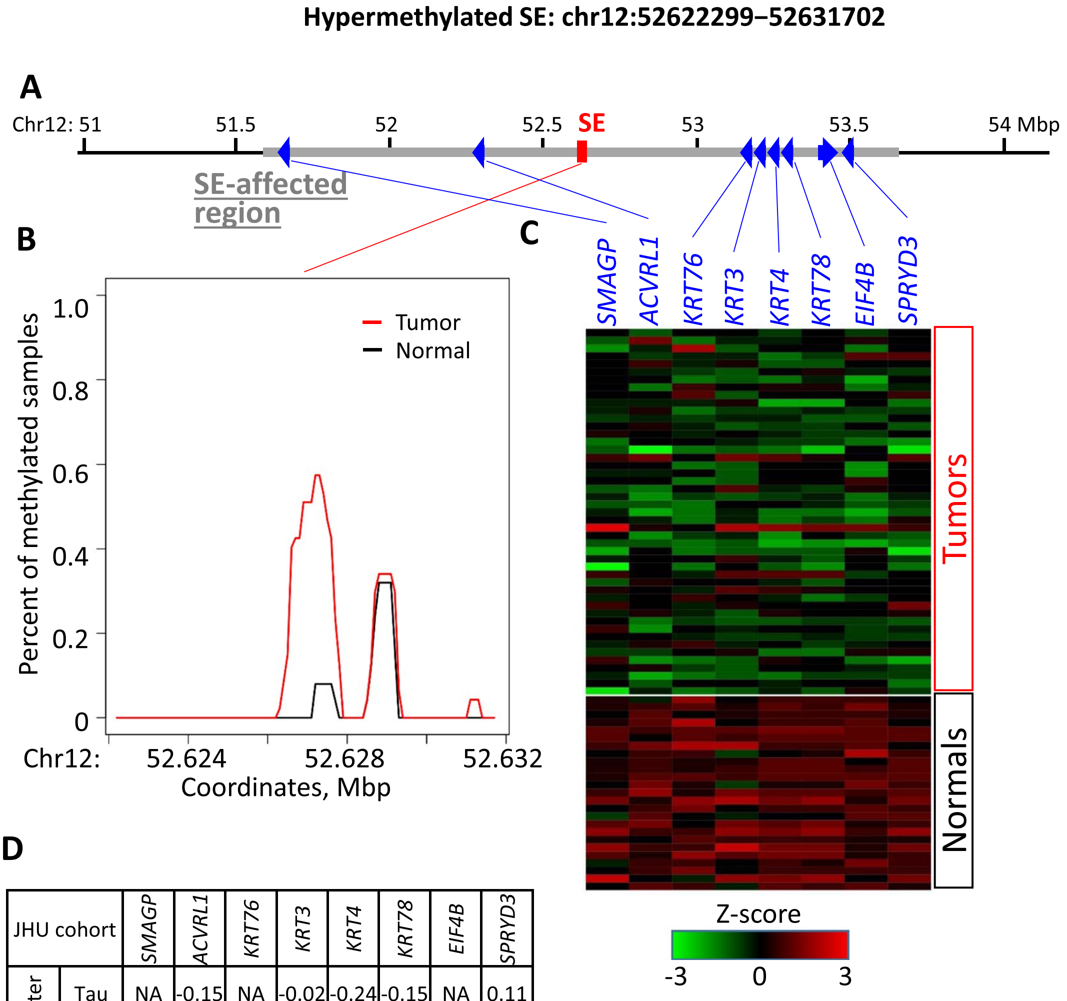


Figure 3. Methylation landscape of DM-SEs. Individual and averaged JHU cohort methylation of the tumor (top, red) and normal (bottom, black) samples across the SE region for (**A**) hypermethylated and (**B**) hypomethylated SEs.

etiologically-similar samples, which is especially valuable for samples that have limited chromatin data. Indeed, our results suggest that biological signatures in one cancer type can be detected and validated using SE data from related cancer types.

The input SE list goes through the stringent sorting process. DNA methylation data from study samples is the primary filter, which helps to remove all non-phenotype-related SE candidates. This assumption is built on recent works demonstrating that actionable SEs in a particular tissue are expected to be differentially methylated[8,16]. Previous studies developed algorithms to predict functional, tissue-specific SEs by applying machine learning algorithms to integrated DNA methylation and SE data[17]. Therefore, the incorporation of such methylation status of SEs regions increases the accuracy and biological relevance of our SE predictions. Our pipeline for this step of the algorithm is based on the well-known MACS peak-calling procedure. Interestingly, in our examples, the full differential signal is provided by local differentially methylated regions (Figs 4, 5, S1, and S2). It is concordant

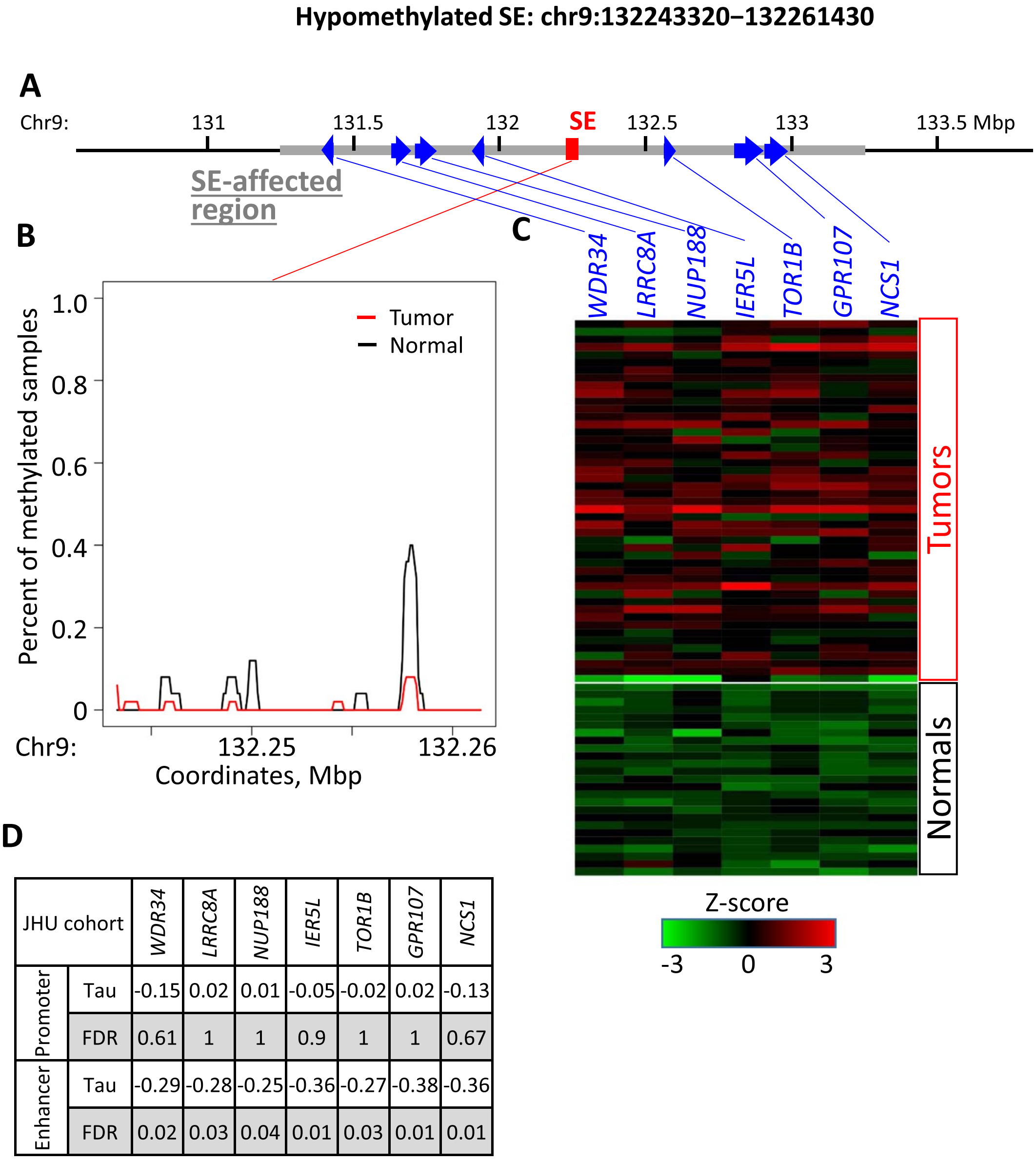


JHU cohort		*SMAGP*	*ACVRL1*	*KRT76*	*KRT3*	*KRT4*	*KRT78*	*EIF4B*	*SPRYD3*
Promoter	Tau	NA	-0.15	NA	-0.02	-0.24	-0.15	NA	0.11
	FDR	NA	0.6	NA	1	0.13	0.53	NA	0.66
Enhancer	Tau	-0.23	-0.25	-0.25	-0.24	-0.23	-0.29	-0.25	-0.23
	FDR	0.04	0.03	0.03	0.04	0.04	0.01	0.03	0.04

Figure 4. Methylation and genetic landscape of the representative hypermethylated SE: chr12:52622299−52631702 in JHU cohort. (**A**) Genomic landscape of the SE region and potential target genes within one Mbp of the SE. (**B**) Relative average methylation coverage across the SE region (red – tumor, black – normal). (**C**) Log-transformed RNA expression of the potential target genes (z-score). (**D**) Kendall-tau values and corresponding FDR for correlation of promoter methylation with gene expression, as well as SE region methylation with gene expression of target genes.

with the work of others[40] that SEs and their parts (individual typical enhancers) both carry the regulatory signal. Still, we do not explicitly relate the unsupervised uniformed probe regions we use with specific SEs. Furthermore, in the future, we are going to compare with modern algorithms, such as MBDDiff[41].

The application of our pipeline allowed us to define tissue-specific SE candidates in HPV + OPSCC, their methylation status, and its correlation with target gene expression. This tissue type has minimal chromatin data available, and SEs for this disease are not yet described. Although many SEs are disease- and tissue-specific[1,2,5,38], similar genetic profiles between head and neck, cervical, and lung cancers[21] suggest conservation of SE regions between these diseases and tissues. We observed many examples of analogous SEs from two or more different cell lines, suggesting the presence of pan-SEs that are common across tissues types and disease status. Our findings were consistent with recent data, which suggest that a subset of multiple myeloma-specific SE candidates was differentially methylated in six head and neck primary samples (three HPV + and three HPV-)[42]. The current study presented a high confidence list of candidate head and neck SEs that were detected in cervical and lung tissue and were differentially methylated in HPV + OPSCC samples. These SE had statistically significant links to gene expression of their *in cis* targets, which participate in cancer-relevant pathways. Such work provides the groundwork for the future discovery of novel, HPV + OPSCC-specific SEs.



JHU cohort		*WDR34*	*LRRC8A*	*NUP188*	*IER5L*	*TOR1B*	*GPR107*	*NCS1*
Promoter	Tau	-0.15	0.02	0.01	-0.05	-0.02	0.02	-0.13
	FDR	0.61	1	1	0.9	1	1	0.67
Enhancer	Tau	-0.29	-0.28	-0.25	-0.36	-0.27	-0.38	-0.36
	FDR	0.02	0.03	0.04	0.01	0.03	0.01	0.01

Figure 5. Methylation and genetic landscape of the representative hypomethylated SE: chr9:132243320−132261430 in JHU cohort. (**A**) Genomic landscape of the SE region and potential target genes within one Mbp of the SE. (**B**) Relative average methylation coverage across the SE region. (**C**) Log-transformed RNA expression of the potential target genes. (**D**) Kendall-tau values and corresponding FDR for correlation of promoter methylation with gene expression, as well as SE region methylation with gene expression of target genes.

We identified 21 differentially hypomethylated SEs and 138 differentially hypermethylated SEs that were linked to the expression of *in cis* target genes in HPV + OPSCC. The prevalence of hypermethylated SEs suggests that hypermethylation of SEs leads to downregulation of normal cellular homeostasis. For example, the observed down-regulated genes are important in breast cancer (BREAST CANCER NORMAL LIKE UP; BREAST CANCER LUMINAL B DN). The significance of those SE-gene links was validated using one of the largest HPV + OPSCC cohorts with in-parallel gene expression and DNA methylation data available for the same samples from TCGA[21,32,33].

We note several limitations of our study. First, the clinical characteristics between tumor and non-tumor groups are not matched in the JHU patient cohort[32,33] due to the demographics of UPPP and OPSCC populations[43–46], with differences in age and smoking status. Nonetheless, a similar UPPP population has helped to reveal strong cancer-specific signatures of OPSCC in previous studies[43–46]. Moreover, the employment of TCGA's control population with matched clinical characteristics validated our original discovery of tissue-specific SEs in HPV + OPSCC. Utilization of DNA methylation array data (Illumina Infinium HumanMethylation450 BeadChip) in TCGA restricted our validation of methylation-expression correlations for both SEs and promoters in this cohort due to the limited number of DNA methylation probes. The availability of probes weakens the correlation, as we are only able to link expression to methylation of one or a few coordinates in the SE region, as opposed to the entire region. The comprehensive discovery of SE candidates is possible only with MBD-Seq whole-genome data, as is employed in our JHU HPV + OPSCC cohort[32,33], which is the only HPV + OPSCC cohort with MBD-Seq data available in parallel with RNA-Seq data for the same samples. The observed differences between the JHU and TCGA cohorts are expected, as normal TCGA samples are not from healthy patients, but are adjacent to the tumor sites of oral cavity tumor patients. These adjacent tissues are known to carry genetic and epigenetic alterations characteristic of OPSCC, skewing analyses[47]. Moreover, we could not define a correlation with survival because only three patients in the discovery cohort recurred in the last five years. Lastly, none of the SE candidates were functionally evaluated within the scope of this study, but they can be assessed in future work, which will also define the OPSCC-specific transcription factors associated with the SE regions.

In conclusion, we developed a new bioinformatics pipeline to define and evaluate SE activity in individual tissue types, which can be further adapted for a wide range of tissues to facilitate analysis of epigenetic mediators. Our pipeline and the SE candidates presented here provide an important next step in developing novel epigenetic therapies and biomarkers for detection of a variety of diseases.

Materials and Methods

JHU study cohort. The employed cohort of study samples is composed of 47 primary HPV + OPSCC tissue specimens and 25 control normal mucosal samples from uvulopalatopharyngoplasty (UPPP) surgeries of non-cancer-affected patients[32,33]. The clinical differences between tumor and control populations in this cohort were previously identified and acknowledged[32,33]. All tissue samples were obtained from the Johns Hopkins Tissue Core, as a part of the Head and Neck Cancer Specialized Program of Research Excellence (HNC-SPORE). These samples were acquired under the Internal Review Board-approved research protocol #NA_00036235. Informed consent was obtained from all patients recruited under this protocol prior to participation in the study. All methods for processing the high-throughput data for these samples were performed in accordance with the relevant guidelines and regulations.

TCGA study cohort. The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/cancersselected/headandneck) recently finished high throughput analyses of head and neck squamous cell carcinoma samples (HNSCC)[21]. The study analyzed data on 279 HNSCC tumors, including 35 HPV + tumors and 50 total adjacent non-cancer control tissues from the same HNSCC patients. Of the 35 HPV + HNSCC tumors, 22 samples were from the oropharynx, similar to JHU cohort, which were selected for cross-study validation to avoid introducing any cohort-specific biases in the analysis. Of 50 controls, we analyzed six samples confirmed as squamous epithelium tissues, while the other 44 TCGA samples designated as controls belong to muscle, salivary gland, and other tissues[21]. After the employed JHU cohort, TCGA is the largest HPV + OPSCC cohort with in-parallel analysis of DNA methylation and gene expression. All methods for processing the high-throughput data for these samples were performed in accordance with the relevant guidelines and regulations.

Candidate SE regions. SEs were previously reported for NHLF (human lung fibroblast), IMR90 (human fetal lung), H2171 (small cell lung carcinoma), HeLa (HPV + cervical adenocarcinoma) cell lines, and UCSD_Lung (healthy lung tissue)[2] (Table S1). Due to their genetic similarity to HPV + OPSCC[21], this pooled list of SEs ($n = 3627$) was used as an input in our pipeline to look for SEs that are specific for HPV + OPSCC (Fig. 2, Table S2).

Gene expression data and processing. RNA-Seq data was obtained for JHU[32,33] and TCGA[21] cohorts. JHU stranded RNA-Seq libraries from ribosomal RNA depleted total RNA were prepared using the Illumina TruSeq stranded total RNA Seq Gold kit and sequenced on the HiSeq 2500 (JHU) or HiSeq 2000 (TCGA) platform sequencer (Illumina) and the TruSeq Cluster Kit. RNA sequencing data from both cohorts were normalized based on the version 2 protocols developed by TCGA[21]. Gene expression values were quantified from RNA sequencing data using RSEM version 1.2.9 and upper quartile normalization according to the TCGA RSEM v2 normalization pipeline[21,32].

DNA methylation data and processing. *MBD-Seq DNA methylation analysis for JHU samples.* Genome-wide DNA methylation analysis was carried out using MBD-Seq (Methyl-CpG binding domain protein sequencing), as previously described[48,49] using the NEBNext DNA Library Prep Set for the SOLiD sequencer. Methylated regions were identified as positional peaks of the population of aligned sequencing reads in the MBD-enriched data compared with the total input fraction using the MACS v1.4 software[50,51]. MACS builds an HMM model to identify peaks and indirectly takes the CpG density into account. This algorithm identifies peaks after accounting for both global and local biases using the enriched-to-input fraction. MACS p-value cut-off ($p < 10^{-6}$) was used to identify reliable methylation peaks. Using the distribution of MACS-called DNA methylation regions, we performed a whole-genome identification of differentially methylated DNA regions between 47 HPV + OPSCC and 25 UPPP samples.

Illumina infinium humanmethylation450 array analysis of TCGA samples. Methylation array data were collected from the TCGA database. This platform includes probes for more than 480,000 CpG sites, spanning 99% of RefSeq. In total, 96% of CpG islands and 92% of CpG shores are represented by at least one probe. Beta values (percent methylation) were estimated from unmethylated (U) and methylated (M) measurements on a probe level basis: $\beta = M/(M + U)$[43,45].

Differential methylation analysis preparation. The methylation calculation was done by a function for intersection length calculations for a set of intervals (SEs, promoters) in multiple samples, which was provided by the "differential.coverage" R package[52]. Annotation functions, e.g., the enumeration of all genes with transcription start site around a SE, was also provided by this package[52]. The package was also used to prepare 100 bp probe intervals inside SE regions and to calculate their methylation for visualization (Figs 4 and 5).

Whole genome differential methylation by 100 bp regions. To identify differentially methylated genome regions using MACS-processed MBD-Seq data for 47 OPSCC and 25 UPPP (normal) samples, we separately tested 30,975,368 nonoverlapping regions of 100 bp each, which together completely cover the human genome. The methylation status of each 100 bp segment for each sample in the discovery cohort was determined as the presence of any intersection of the segment with regions of DNA methylation, as identified by MACS peak calling[50] for that sample. The differentially methylated probes between diseased and normal phenotypes were identified by exact Fisher test for association of the probe methylation status with the sample status, followed by FDR correction.

GenometriCorr analysis. We used overlap statistics provided by the GenometriCorr package[34] to compare the genome-wide distribution of differentially methylated 100 bp regions relative to candidate SE sites and to test the conservation of the input list of 3627 SEs.

SE differential methylation detection. Methylation of a SE region in each sample was identified as the length of the intersection of the SE region with methylation peaks provided by the MACS software for each sample[50,51]. MACS identifies peaks after accounting for both global and local biases using the enriched-to-input fraction. Using MACS-processed MBD-Seq data, we calculate the net length of methylated regions that overlap with each of the enhancers in each of the samples. For each SE, we calculated the Wilcoxon p-values for the difference of the methylation in the SE region between sample types of cases (n = 47) and controls (n = 25), followed by FDR correction. All the SE regions with FDR p-value < 0.05 were considered as differentially methylated. Of the initial 3627 SEs, 159 were differentially methylated (DM) between 47 tumors and 25 normal samples in JHU HPV + HNSCC cohort. The SEs with higher methylation in tumors relative to normal (138 SEs) were referred to as hypermethylated, while the remaining (21 SEs) with higher methylation in normal were referred to as hypomethylated.

Identification of in cis SE targets. Recent data suggest that SE effects can reach *in cis* targets up to one Mbp away through chromatin loop formation[3–5,8,10,14,42]. Therefore, all genes with transcription start sites within one Mbp from the SE region were considered potential targets of the SE[2,3]. For each differentially methylated SE (n = 159), a list of all potential targets was assembled. SE candidates that covered the same genomic coordinates due to the utilization of four different cell lines were treated individually.

Promoter methylation detection. We defined the gene's promoter region as the genomic interval 1500 bp upstream and 500 bp downstream of the transcription start site. The methylation of each promoter was assessed in the same way as the SEs (see *SE differential methylation detection*).

Methylation to expression correlation analysis. *Correlation between target gene expression and SE methylation.* For each pair of a differentially methylated super-enhancer and a gene with TSS within 1 Mbp of the SE region, we tested the hypothesis that the SE methylation regulated the gene expression. This hypothesis was tested by calculation of correlation (negative concordance) between the RSEM-estimated expression of the gene in each sample and DNA methylation of the SE region in the same sample for all samples in the JHU cohort[32,33]. We applied Kendall's tau test using the Kendall package for R, version 2.2[53]. We used rank-based statistics to compare gene expression (measured by RNA-Seq) and DNA methylation (quantified by MBD-Seq or Illumina 450 k array). These data types produce different values and cannot be compared directly, only through their ranks. FDR-corrected Kendall's tau test p-value < 0.05 was considered significant to link SE methylation and gene expression. Then, we filtered out all gene-SE pairs with positive correlation as artifacts. The remaining pairs were then considered separately for hypo- and hyper-SEs.

Correlation between target gene expression and promoter methylation. For all the target genes (with TSS in 1 Mbp from a DM-SE) of a SE, we also estimated the Kendall's tau rank correlation between promoter methylation and gene expression to test whether the gene expression is regulated by promoter methylation. Similar to SEs, we used the Kendall package for R[53] and FDR-corrected Kendall's tau test p-values < 0.05 were considered significant.

TCGA validation. We used 22 HPV + OPSCC and 6 normal samples from TCGA[21] that have both Illumina 450 k DNA methylation and RNA-Seq expression data to validated gene-SE pairs. For every pair, we found Illumina probes in a SE region and applied Kendall's tau test[53] to methylation beta values of these probes and the gene expression values of potential target genes. For every SE-gene pair for which the SE region has at least one Illumina probe, we found the probe with the minimum correlation coefficient with the gene expression value. A pair was considered as validated if Kendall's test p-value for the probe was less than 0.05.

Overrepresentation gene set analysis. Overrepresentation gene set analysis was done by computing overlaps with annotated Hallmark gene sets in MSigDB v6.1[54].

Received: 17 August 2017; Accepted: 16 September 2019;
Published online: 21 October 2019

References

1. Drier, Y. *et al*. An oncogenic MYB feedback loop drives alternate cell fates in adenoid cystic carcinoma. *Nature genetics* **48**, 265–272, https://doi.org/10.1038/ng.3502 (2016).
2. Hnisz, D. *et al*. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947, https://doi.org/10.1016/j.cell.2013.09.053 (2013).
3. Whyte, W. A. *et al*. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319, https://doi.org/10.1016/j.cell.2013.03.035 (2013).
4. Dixon, J. R. *et al*. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380, https://doi.org/10.1038/nature11082 (2012).
5. Heintzman, N. D. *et al*. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112, https://doi.org/10.1038/nature07829 (2009).
6. Kulaeva, O. I., Nizovtseva, E. V., Polikanov, Y. S., Ulianov, S. V. & Studitsky, V. M. Distant activation of transcription: mechanisms of enhancer action. *Molecular and cellular biology* **32**, 4892–4897, https://doi.org/10.1128/MCB.01127-12 (2012).
7. Shen, Y. *et al*. A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116–120, https://doi.org/10.1038/nature11243 (2012).
8. Heyn, H. *et al*. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol* **17**, 11, https://doi.org/10.1186/s13059-016-0879-2 (2016).
9. Hnisz, D. *et al*. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol Cell* **58**, 362–370, https://doi.org/10.1016/j.molcel.2015.02.014 (2015).
10. Herranz, D. *et al*. A NOTCH1-driven MYC enhancer promotes T cell development, transformation and acute lymphoblastic leukemia. *Nat Med* **20**, 1130–1137, https://doi.org/10.1038/nm.3665 (2014).
11. Loven, J. *et al*. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334, https://doi.org/10.1016/j.cell.2013.03.036 (2013).
12. Polak, P. *et al*. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364, https://doi.org/10.1038/nature14221 (2015).
13. Corces, M. R. *et al*. The chromatin accessibility landscape of primary human cancers. *Science* **362**, https://doi.org/10.1126/science.aav1898 (2018).
14. Hnisz, D. *et al*. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458, https://doi.org/10.1126/science.aad9024 (2016).
15. Charlet, J. *et al*. Bivalent Regions of Cytosine Methylation and H3K27 Acetylation Suggest an Active Role for DNA Methylation at Enhancers. *Mol Cell* **62**, 422–431, https://doi.org/10.1016/j.molcel.2016.03.033 (2016).
16. Bell, R. E. *et al*. Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. *Genome Res* **26**, 601–611, https://doi.org/10.1101/gr.197194.115 (2016).
17. He, Y. *et al*. Improved regulatory element prediction based on tissue-specific local epigenomic signatures. *Proc Natl Acad Sci USA* **114**, E1633–E1640, https://doi.org/10.1073/pnas.1618353114 (2017).
18. Hon, G. C. *et al*. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nature genetics* **45**, 1198–1206, https://doi.org/10.1038/ng.2746 (2013).
19. Baylin, S. B. *et al*. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Hum Mol Genet* **10**, 687–692 (2001).
20. Kelley, D. Z. *et al*. Integrated Analysis of Whole-Genome ChIP-Seq and RNA-Seq Data of Primary Head and Neck Tumor Samples Associates HPV Integration Sites with Open Chromatin Marks. *Cancer Res* **77**, 6538–6550, https://doi.org/10.1158/0008-5472.CAN-17-0833 (2017).
21. Cancer Genome Atlas, N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582, https://doi.org/10.1038/nature14129 (2015).
22. Agrawal, N. *et al*. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* **333**, 1154–1157, https://doi.org/10.1126/science.1206923 (2011).
23. Gaykalova, D. A. *et al*. Novel insight into mutational landscape of head and neck squamous cell carcinoma. *PLoS One* **9**, e93102, https://doi.org/10.1371/journal.pone.0093102 (2014).
24. Hayes, D. N., Van Waes, C. & Seiwert, T. Y. Genetic Landscape of Human Papillomavirus-Associated Head and Neck Cancer and Comparison to Tobacco-Related Tumors. *J Clin Oncol* **33**, 3227–3234, https://doi.org/10.1200/JCO.2015.62.1086 (2015).
25. Seiwert, T. Y. *et al*. Integrative and comparative genomic analysis of HPV-positive and HPV-negative head and neck squamous cell carcinomas. *Clin Cancer Res* **21**, 632–641, https://doi.org/10.1158/1078-0432.CCR-13-3310 (2015).
26. Stransky, N. *et al*. The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157–1160, https://doi.org/10.1126/science.1208130 (2011).
27. Lee, T. L. *et al*. A novel nuclear factor-kappaB gene signature is differentially expressed in head and neck squamous cell carcinomas in association with TP53 status. *Clin Cancer Res* **13**, 5680–5691, https://doi.org/10.1158/1078-0432.CCR-07-0670 (2007).
28. Papillon-Cavanagh, S. *et al*. Impaired H3K36 methylation defines a subset of head and neck squamous cell carcinomas. *Nature genetics* **49**, 180–185, https://doi.org/10.1038/ng.3757 (2017).
29. Chan, K. M. *et al*. The histone H3.3K27M mutation in pediatric glioma reprograms H3K27 methylation and gene expression. *Genes Dev* **27**, 985–990, https://doi.org/10.1101/gad.217778.113 (2013).
30. Khuong-Quang, D. A. *et al*. K27M mutation in histone H3.3 defines clinically and biologically distinct subgroups of pediatric diffuse intrinsic pontine gliomas. *Acta Neuropathol* **124**, 439–447, https://doi.org/10.1007/s00401-012-0998-0 (2012).
31. Solomon, D. A. *et al*. Diffuse Midline Gliomas with Histone H3-K27M Mutation: A Series of 47 Cases Assessing the Spectrum of Morphologic Variation and Associated Genetic Alterations. *Brain Pathol* **26**, 569–580, https://doi.org/10.1111/bpa.12336 (2016).
32. Guo, T. *et al*. Characterization of functionally active gene fusions in human papillomavirus related oropharyngeal squamous cell carcinoma. *Int J Cancer* **139**, 373–382, https://doi.org/10.1002/ijc.30081 (2016).
33. Guo, T. *et al*. A Novel Functional Splice Variant of AKT3 Defined by Analysis of Alternative Splice Expression in HPV-Positive Oropharyngeal Cancers. *Cancer Res* **77**, 5248–5258, https://doi.org/10.1158/0008-5472.CAN-16-3106 (2017).
34. Favorov, A. *et al*. Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS Comput Biol* **8**, e1002529, https://doi.org/10.1371/journal.pcbi.1002529 (2012).
35. Ren, S. *et al*. Discovery and development of differentially methylated regions in human papillomavirus-related oropharyngeal squamous cell carcinoma. *Int J Cancer* **143**, 2425–2436, https://doi.org/10.1002/ijc.31778 (2018).
36. Irizarry, R. A. *et al*. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics* **41**, 178–186, https://doi.org/10.1038/ng.298 (2009).
37. Tarbe, N. G., Rio, M. C. & Weidle, U. H. SMAGP, a new small trans-membrane glycoprotein altered in cancer. *Oncogene* **23**, 3395–3403, https://doi.org/10.1038/sj.onc.1207469 (2004).
38. Sur, I. & Taipale, J. The role of enhancers in cancer. *Nat Rev Cancer* **16**, 483–493, https://doi.org/10.1038/nrc.2016.62 (2016).

39. Uhlen, M. *et al*. Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* **28**, 1248–1250, https://doi.org/10.1038/nbt1210-1248 (2010).
40. Moorthy, S. D. *et al*. Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Res* **27**, 246–258, https://doi.org/10.1101/gr.210930.116 (2017).
41. Liu, Y., Wilson, D., Leach, R. J. & Chen, Y. MBDDiff: an R package designed specifically for processing MBDcap-seq datasets. *BMC Genomics* **17**(Suppl 4), 432, https://doi.org/10.1186/s12864-016-2794-z (2016).
42. Wilson, G. A. *et al*. Integrated virus-host methylome analysis in head and neck squamous cell carcinoma. *Epigenetics* **8**, 953–961, https://doi.org/10.4161/epi.25614 (2013).
43. Fertig, E. J. *et al*. Preferential activation of the hedgehog pathway by epigenetic modulations in HPV negative HNSCC identified with meta-pathway analysis. *PLoS One* **8**, e78127, https://doi.org/10.1371/journal.pone.0078127 (2013).
44. Li, R. *et al*. Expression microarray analysis reveals alternative splicing of LAMA3 and DST genes in head and neck squamous cell carcinoma. *PLoS One* **9**, e91263, https://doi.org/10.1371/journal.pone.0091263 (2014).
45. Rathi, K. S., Gaykalova, D. A., Hennessey, P., Califano, J. A. & Ochs, M. F. Correcting transcription factor gene sets for copy number and promoter methylation variations. *Drug development research* **75**, 343–347, https://doi.org/10.1002/ddr.21220 (2014).
46. Sun, W. *et al*. Activation of the NOTCH pathway in head and neck cancer. *Cancer Res* **74**, 1091–1104, https://doi.org/10.1158/0008-5472.CAN-13-1259 (2014).
47. Braakhuis, B. J. *et al*. Second primary tumors and field cancerization in oral and oropharyngeal cancer: molecular techniques provide new insights and definitions. *Head Neck* **24**, 198–206 (2002).
48. Sinclair, S. H., Yegnasubramanian, S. & Dumler, J. S. Global DNA methylation changes and differential gene expression in Anaplasma phagocytophilum-infected human neutrophils. *Clin Epigenetics* **7**, 77, https://doi.org/10.1186/s13148-015-0105-1 (2015).
49. Yegnasubramanian, S. *et al*. Chromosome-wide mapping of DNA methylation patterns in normal and malignant prostate cells reveals pervasive methylation of gene-associated and conserved intergenic sequences. *BMC Genomics* **12**, 313, https://doi.org/10.1186/1471-2164-12-313 (2011).
50. Zhang, Y. *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, https://doi.org/10.1186/gb-2008-9-9-r137 (2008).
51. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nature protocols* **7**, 1728–1740, https://doi.org/10.1038/nprot.2012.101 (2012).
52. Favorov, A. differential.coverage. R package, https://github.com/favorov/differential.coverage. (2015).
53. McLeod, A. I. Kendall: Kendall rank correlation and Mann-Kendall trend test. *2.2* R package (2011).
54. Subramanian, A. *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102**, 15545–15550, https://doi.org/10.1073/pnas.0506580102 (2005).

Acknowledgements

This work was supported by National Institutes of Health (NIH) Grants R21DE025398 and R01DE027809 (D.A.G.), R01CA177669 (E.J.F.), P30 CA006973 (E.J.F., L.D. and A.F.), R01DE023347 (J.A.C.), and Specialized Programs of Research Excellence in Human Cancers (SPORE) P50DE019032 (D.A.G. and E.J.F.).

Author contributions

E.L.F., L.D., D.Z.K., A.F., E.J.F. and D.A.G. conceived the study. D.A.G. and J.A.C. provided the high-throughput data for analysis. E.L.F., L.D., D.Z.K., E.S., T.G., M.C., A.F. and E.J.F. analyzed and processed the data. D.A.G. and J.Q. overviewed the biological significance of SEs and their targets analyzed in the manuscript. E.L.F., L.D., E.J.F. and D.A.G. prepared the figures. E.L.F., L.D., D.Z.K., T.G., A.F., E.J.F. and D.A.G. wrote the manuscript. E.J.F. and D.A.G. co-led the study. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-51018-x.

Correspondence and requests for materials should be addressed to E.J.F. or D.A.G.

Reprints and permissions information is available at www.nature.com/reprints.

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