Title

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

Journal

Cancer Research, 77(23)

ISSN

0008-5472

Authors

Kelley, Dylan Z
Flam, Emily L
Izumchenko, Evgeny
et al.

Publication Date

2017-12-01

DOI

10.1158/0008-5472.can-17-0833

Peer reviewed

Published in final edited form as:

Cancer Res. 2017 December 01; 77(23): 6538–6550. doi:10.1158/0008-5472.CAN-17-0833.

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

Dylan Z. Kelley[1], **Emily L. Flam**[1], **Evgeny Izumchenko**[1], **Ludmila V. Danilova**[2,3], **Hildegard A. Wulf**[1], **Theresa Guo**[1], **Dzov A. Singman**[1], **Bahman Afsari**[2], **Alyza Skaist**[2], **Michael Considine**[2], **Elena Stavrovskaya**[4,5], **Justin A. Bishop**[6], **William H. Westra**[6], **Zubair Khan**[1], **Wayne M. Koch**[1], **David Sidransky**[1], **Sarah Wheelan**[2], **Joseph A. Califano**[7,8], **Alexander V. Favorov**[2,3,9], **Elana J. Fertig**[2,*], and **Daria A. Gaykalova**[1,*]

[1]Department of Otolaryngology-Head and Neck Surgery, Johns Hopkins University School of Medicine, Baltimore, MD USA

[2]Department of Oncology, The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, 1550 Orleans Street, Baltimore, MD 21231

[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 Pathology, Johns Hopkins Medical Institutions, Baltimore, MD USA

[7]Head and Neck Cancer Center, Moores Cancer Center, University of California, San Diego, 3855 Health Sciences Dr., MC 0803, La Jolla, CA, USA

[8]Division of Otolaryngology-Head and Neck Surgery, Department of Surgery, 3855 Health Sciences Dr., MC 0803, La Jolla, CA, USA

[9]Laboratory of Bioinformatics, Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, Russia

Abstract

Chromatin alterations mediate mutations and gene expression changes in cancer. Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) has been utilized to study genome-wide chromatin structure in human cancer cell lines, yet numerous technical challenges limit comparable analyses in primary tumors. Here we have developed a new whole-genome analytical

Corresponding author: Daria A. Gaykalova, PhD, Assistant Professor, Otolaryngology - Head and Neck Surgery, The Johns Hopkins University School of Medicine, 1550 Orleans Street, Rm 5M05C, CRBII, Baltimore, MD 21231, Phone: 410.614.2745, Fax: 410.614.1411, dgaykal1@jhmi.edu.

*These authors contributed equally to this project

Conflict of interest:

The authors declare no potential conflicts of interest.

pipeline to optimize ChIP-Seq protocols on patient-derived xenografts from human papillomavirus-related (HPV+) head and neck squamous cell carcinoma (HNSCC) samples. We further associated chromatin aberrations with gene expression changes from a larger cohort of the tumor and normal samples with RNA-Seq data. We detect differential histone enrichment associated with tumor-specific gene expression variation, sites of HPV integration in the human genome, and HPV-associated histone enrichment sites upstream of cancer driver genes, which play central roles in cancer associated pathways. These comprehensive analyses enable unprecedented characterization of the complex network of molecular changes resulting from chromatin alterations that drive HPV-related tumorigenesis.

Keywords

Chromatin immunoprecipitation; high-throughput data; histone code; HPV integration; head and neck squamous cell carcinoma

Introduction

Aberrations to histone marks and chromatin organization are critical to cancer development and progression (1). Many histone modifications (including H3K4me3 and H3K27ac) are robust cancer biomarkers (1). These alterations induce widespread changes across larger genomic areas than mutations, making them likely regulators of pervasive gene expression changes in cancer (2). Changes in gene expression are considered as a fundamental hallmark of cancer (3). Many of these changes can be explained by the mutational landscape of the disease. However, mutations alone are insufficient to explain vast transcriptomic changes in cancers with lower mutation rates, such as hematologic malignancies or virus-induced cancers that lack actionable genetic alterations (3). We hypothesize that pervasive alterations to chromatin organization can drive functional gene expression changes in virus-induced cancers, such as HPV+ head and neck squamous cell carcinoma (HNSCC). Comprehensive genome-wide analyses of the chromatin structure, gene expression changes, and viral integration sites can shed light on this hypothesis and better elucidate the complex cooperative biological activities occurring in the genome during tumorigenesis. However, such high-dimensional analysis has never been performed in primary cancer samples of this disease and the manner by which chromatin reorganization cooperates with components of host and viral genomes to affect tumor progression remains to be delineated.

Chromatin immunoprecipitation followed by high-resolution whole-genome sequencing (ChIP-Seq) is the gold-standard method for studying the association of modified histones with genomic DNA. However, its use as a modality for large-scale analysis has recognized limitations, such as large sample input requirement, variable antibody binding efficiency, material loss over the purification steps, overall low DNA outcome/output, and complex multilateral quality assessments during the computational procedures (4). Moreover, the chromatin accessibility during fragmentation is not uniform across the genome. Open chromatin regions are amenable to better fragmentation and therefore are preferentially represented in the digested sample. Whereas tightly packed heterochromatin is digested to a lesser extent, thereby confounding weak enrichment of true binding sites for

heterochromatin markers (5). Finally, the chromatin integrity, the rate of its digestion, and strength of DNA-protein binding highly depend on the preservation and processing of the patient's primary cancer tissue (5). Therefore, the majority of ChIP-Seq-generated data in cancer are currently limited to cell line analysis (2) and do not represent the wide heterogeneity of human malignancy that occurs on a population basis (6). Although multiple efforts (such as recently described SimpleChIP protocol) have been made to streamline the workflow and to obtain high-quality, unbiased and reasonable data, to date, there is no single experimental design, which is optimal to evaluate primary tumor samples.

HPV is the second leading cause of HNSCC after cigarette smoking (7). Several high-throughput genomic analyses, aimed at facilitating the development of cancer-related therapy, revealed that HPV+ tumors have fewer genetic alterations than non-HPV-related HNSCCs (3, 8). The virus-related dysfunction of the APOBEC complex in HPV+ HNSCC (9) leads to accumulation of the non-synonymous mutations within the isolated hot-spots, resulting in genetically-homogeneous disease and narrowing down the number of potential targeting candidates. As such, relatively few genetic alterations critical to the development of HPV+ HNSCC are currently recognized. Nonetheless, both HPV+ and HPV-HNSCC subgroups display relevant pervasive gene expression alterations (3). Amplification of the host genome at the site of HPV integration (10, 11) and dysregulation of tumor-suppressor genes by HPV oncoproteins (9) cannot fully describe the genome-wide spectrum of gene expression changes in HPV+ HNSCC (3). These data lead to the hypothesis that epigenetic modifications, such as chromatin reorganization, are central to gene expression dysregulation during the HPV-associated carcinogenesis.

In parallel with acquiring the high-quality ChIP-Seq results, assessing the function of chromatin structure requires integrated bioinformatics analysis with additional genomics data. Recently, mutations in NSD1 were found to define a subtype of HPV-HNSCC in which disruption of H3K36 potentially drives oncogenesis. However, this study lacked chromatin data to evaluate the association between gene expression and chromatin reorganization in head and neck cancers (12). Since modulation in chromatin structure enables transcriptional changes to numerous genes simultaneously, new integrated studies with matched chromatin and high-throughput transcriptomic data are essential to establish the functional relevance of specific chromatin alterations in prevalent types of cancer, such as HPV+ HNSCC.

To determine the role of chromatin structure in HPV-related carcinogenesis, we have performed the first ChIP-Seq characterization of chromatin state in primary HPV+ HNSCC tumor samples. To address technical challenges with ChIP-Seq, we have performed a comprehensive optimization of the current ChIP-Seq methodology aimed at improving its applicability to clinically relevant primary tissue samples (Fig. S1.). We have adjusted various processing parameters and extensively validated this optimized protocol in tumor cell lines, primary patients' samples, and patient-derived xenograft (PDX) models. The chromatin data generated in this study was coupled with matched RNA-Seq data that include samples profiled as part of a larger cohort of 72 HPV+ tumors (13). We performed integrated bioinformatics analysis of the ChIP-Seq and RNA-Seq data to determine the potential functional role of chromatin alterations in HPV+ HNSCC. This integrated analysis was performed with a new Expression Variation Analysis (EVA) algorithm that models inter-

tumor heterogeneity (14) of epigenetic regulation of gene expression. Overall, we have shown a strong disease-specific distribution of H3K4me3 and H3K27ac histone marks, which correlates with differential gene expression of nearby cancer-related genes, and their associated pathways. The analyses further demonstrated a sample-specific association of H3K27ac marks with sites of HPV integration and known HNSCC driver genes. Taken together, this first integrated analysis of chromatin data in primary tumor samples demonstrates the critical role of chromatin distribution in HPV+ HNSCC and is applicable to determining that role in other cancer subtypes.

Materials and Methods

JHU cohort of primary samples

Primary tumor tissue samples were obtained from a cohort of 47 patients with HPV-related oropharyngeal squamous cell carcinoma, as previously described (13). For comparison, healthy oropharynx mucosal tissue from uvulopalatopharyngoplasty (UPPP) surgical specimens were obtained from 25 cancer-unaffected controls (13). All tissue samples were collected from the Johns Hopkins Tissue Core under an approved IRB protocol (NA_00036235) after obtaining the informed written consent from all subjects. This protocol also permitted the usage of the tumor tissue for PDX model development. This study qualified for exemption under the U.S. Department of Health and Human Services policy for protection of human subjects [45 CFR 46.101(b)] (IRB study number is NA_00036235). *Additional details are provided in* Supplementary Materials and Methods.

Cell Lines

Human HPV+ HNSCC cell lines UM-SCC-047 and UPCI-SCC-090 were provided by Dr. Thomas Carey (University of Michigan) and Dr. Susanne Gollin (University of Pittsburgh), respectively. *Additional details are provided in* Supplementary Materials and Methods.

HPV detection

Four independent methodologies were used to validate HPV status in all of our samples: *In situ* hybridization for HR-HPV, ICH staining for p16, qRT-PCR detection for HPV DNA and RNA-Seq based detection of HPV expression. *Additional details are provided in* Supplementary Materials and Methods.

Selection of samples for ChIP-Seq analysis

Two HPV+ OPSCC samples from the JHU cohort of primary tumors (13) were used for the preparation of the first generation (F1) PDX models, PDX1 and PDX2, using xenografting procedures described in (6, 15) for ChIP-Seq analysis. RNA-Seq data was collected for these PDXs using the methods and normalization procedures described previously (13). To confirm that the PDX models were similar to the tumor samples from which they were derived, we compared the RNA-Seq gene expression profile to the profile for its corresponding parental tissue. Pearson correlation coefficients were 0.83 for PDX1 and 0.9 for PDX2, and both p-values were below 10^{-16}. This finding was consistent with our previous observations that high-throughput profiles in HNSCC PDX samples were more similar to their parental tumor tissue than to other tumor samples or to cell lines (6). We also

performed ChIP-Seq analysis on 2 HPV+ HNSCC cell lines (UM-SCC-047 and UPCI-SCC-090) and two UPPP samples (UPPP1 and UPPP2), where both UPPP samples were from the same JHU cohort (13). UPPP is the only surgical procedure performed in the oropharyngeal area in healthy individuals, which allows collection of oropharyngeal tissues from non-cancer patients as controls, consistent with previous genomics studies of HNSCC (16–18). Also, the UPPP samples selected for study here had similar gender, race, ethnicity, smoking, and drinking status to that of the HPV+ HNSCC samples selected for ChIP-Seq analysis (Table S1). This matching of tumors and controls enabled inference of tumor-specific differences in chromatin structure, independent of tissue-specific effects on chromatin structure.

Histone marks used in ChIP-Seq analysis

Histone modifications H3K4me3, H3K9ac, H3K9me3, and H3K27ac were chosen for ChIP-Seq analysis. H3K4me3, H3K9ac, and H3K27ac were selected because they were strongly implicated in gene expression regulation (19). The H3K9me3 repressive histone mark was selected as a negative control.

Preservation of samples

Cells were grown to 80% confluence. Each immunoprecipitation (IP) preparation contained 4×10^6 cells. Cell number was verified by Cellometer™ Auto T4 (Nexcelom Bioscience). Viable cells from culture were taken directly to the ChIP experiments. When harvesting the tissue samples, unwanted material such as fat and necrotic material were removed from the sample. Tissue was then snap frozen in liquid nitrogen for later processing. For optimal chromatin yield and ChIP results, we used 25 mg of tissue for each immunoprecipitation to be performed. Frozen tissue was left to thaw on ice and mass was determined by weight.

Protein-DNA cross-linking

ChIP-DNA was prepared using recently developed SimpleChIP Enzymatic Chromatin IP Kit #9005 (Cell Signaling Technology) following manufacturer's protocol with sample-specific adjustments in micrococcal nuclease and sonication steps. 10X phosphate buffered saline (PBS) pH 7.4 from Quality Biological Inc. was used wherever PBS is indicated.

MNase/Sonication

Samples were digested by both micrococcal nuclease and sonication. This process was additionally optimized and followed by gel electrophoresis to ensure uniform shearing of DNA across the genome. *Additional details are provided in* Supplementary Materials and Methods.

Chromatin immunoprecipitation

Equal amounts of chromatin were used per IP step with exceptional performance (XP®) monoclonal antibodies validated for ChIP application (Cell Signaling Technology). Rabbit monoclonal antibodies were added in particular dilution based on an optimized concentration evaluated across a wide variety of commercial monoclonal antibodies. A 1:50 dilution for H3K4me3 (9751), H3K9ac (9649), H9K9me3 (13969) antibodies and a 1:100

dilution for H3K27ac (8173) antibody were used to isolate DNA segments bound by individual histone modification. We used 1:50 diluted total H3 (4620) antibody as a positive control and 1:250 diluted Normal Rabbit IgG (2729) as a negative control. A 3527-5 Incubator Shaker (Lab-Line) was used during elution. ChIP-DNA was purified and measured following the ChIP kit protocol. The 1/50 portion (2%) of the same chromatin for each sample (PDX1, PDX2, UPPP1, UPPP2, UM-SCC-047, UPCI-SCC-090) was used for DNA extraction skipping the antibody enrichment steps and was further used for qRT-PCR and sequencing as an input control.

Quantitative real-time polymerase chain reaction

ChIP-DNA underwent qRT-PCR using a TaqMan® 7900HT Fast Real-Time PCR System (Applied Biosystems) per manufacturer's recommendations. We used Johns Hopkins lab standard 10X PCR Buffer (20), dNTPs (Bioline), FAM (Thermo Fisher Scientific). Primers and probes designed in the promoter region of actively expressed *GAPDH* and *RPL10* genes, and 3′ end of the transcriptionally repressed *ZNF333* gene (19), see Table S2 for details. Each sample was analyzed in triplicate and underwent one cycle of 10min at 95°C, and 50 cycles of 15s 95°C/60s 60°C. Relative fold enrichment of different histones in individual samples was quantified in triplicate relative to the 2% input sample using the 2-ΔΔCT method (21).

ChIP-DNA whole-genome sequencing and normalization

ChIP-DNA for individual sample/antibody and their input controls were sonicated, end-repaired, and ligated to SOLiD P1 and P2 sequencing adaptors lacking 5′ phosphate groups, using the NEBNext DNA Library Prep Set for SOLiD per the manufacturer's recommended protocol (NEB). Libraries were then nick-translated with Platinum Taq. ChIP-DNA was sequenced at the Experimental and Computational Genomics Core (ECGC) at Johns Hopkins University with a target sequencing coverage of approximately 45,000,000x and paired-end reads of 150 bp. Illumina CASAVA 1.8.2 was used to convert BCL files to FASTQ files using default parameters (22). Bowtie 2.2.1 was used to map paired-end reads to the hg19 human reference genome using default parameters and samtools 0.1.19 was used to convert, sort, and index SAM files (23). The count functionality IGVTools package was used to generate a tiled data file using default parameters. MACS (Model-based Analysis of ChIP-Seq algorithm, version 1.4.2) called ChIP-Seq peaks for each mark and each sample using the input DNA in that sample as a control (24). ChIP-Seq peaks were called significant if MACS modeled peak p-values are below a threshold of 10^{-6}, and these peaks were represented as genomic intervals. The cis-regulatory element annotation system (CEAS) was used to associate these genomic intervals with genes (25).

DiffBind Analysis of ChIP-Seq data

To compare the ChIP-Seq peaks for different samples and different modifications, we used the R/Bioconductor package DiffBind (26). MACS bed files for the six samples and their H3K4me3, H3K9ac, H3K27ac, and H3K9me3 histone marks were used as an input using the code in Supplemental File 1. We used DiffBind only to compute pairwise genome-wide correlation coefficients between all possible ChIP-Seq signal pairs (24 × 24 total).

Visualization of Whole Genome ChIP-Seq Enrichments over Genomic Regions

Fold enrichment computed with MACS calls were input to deepTools (27) for visualization. DeepTools heatmap functions were used to visualize ChIP-Seq fold enrichment −1.5 – +1.5 kb region around the transcriptional start sites (TSS) for all known genes. Average profiles for ChIP-Seq enrichment in the same −1.5 – +1.5 kb region around the TSS were also generated with the profiler tool in deepTools.

Identification of disease-specific genes associated with ChIP-Seq peaks

We sought a list of genes with disease-specific coverage from the ChIP-Seq data for each histone mark. To obtain these genes, we compared the CEAS output gene lists for the ChIP-Seq data in each sample. Specifically, we performed set differences to define the lists of genes with ChIP-Seq coverage in $5'$ UTR regions that were specific to either tumor or normal samples for each histone mark (Tables S3–S6). To obtain the normal-specific gene list, we restricted the sets to genes that were shared by both UPPP samples and were not in any cancer cell line or PDX sample. To obtain the tumor-specific gene lists, we restricted the sets to genes that were shared by both cancer cell lines or by both PDXs and were not in any UPPP sample. We created additional annotations for the list of tumor-specific genes in both the PDXs and cell lines and the tumor-specific genes only in the PDXs. These lists of genes were generated using the code in Supplemental File 2 and carried forward for analysis of RNA-Seq data to determine the functional consequences of disease-specific genes.

Gene set analysis

The MSigDB (28) “investigate gene sets” function performed pathway analysis of the disease-specific genes for each tumor- and normal-specific H3K4me3 and H3K27ac histone mark (Tables S3–S4). Gene set analysis in this software was performed with Hallmark gene sets using a hypergeometric test (Tables S7–S10).

Correlation of H3K27ac-enriched genes with other known HPV+ HNSCC gene sets

The R/Bioconductor package GeneOverlap was used to associate the disease-specific gene set for H3K27ac (29, 30). A one-sided Wilcoxon gene set test was further applied to assess the enrichment of the disease-specific H3K27ac gene set with the continuous weights of the gene classifier for HPV+ HNSCC subtypes from (29).

RNA-Seq normalization and analysis

Gene level counts from the RNA-Seq data were obtained from the RSEM V2 pipeline for TCGA (3) as described in (13). Heatmaps of RNA-Seq data for disease-specific genes (listed in Tables S3–S6) were generated for each histone mark. Unsupervised hierarchical clustering in heatmaps used Kendall-Tau dissimilarity distances. Previous work demonstrated that this distance quantified the relative variability of gene expression profiles (14), enabling it to quantify dysregulation of gene expression by enhancers in this study.

Expression Variability Analysis bioinformatics for dysregulation of RNA-Seq in tissue-specific ChIP-Seq peaks

We hypothesized that changes in chromatin structure enabled expression changes in the genes with ChIP-Seq coverage in 5′ UTR regions. However, other epigenetic alterations-based regulatory mechanisms (e.g., transcription factor binding, copy number amplifications, etc.) were still required to alter gene expression. Consequently, the expression changes in genes with tumor-specific ChIP-Seq coverage at 5′ UTR would be more variable than expression changes in genes with normal-specific ChIP-Seq coverage. Consistent with this hypothesis, we used the EVA gene set dysregulation algorithm to quantify the relative dissimilarity measure (e.g. rank proxy of variance) of gene expression profiles in tumor and normal samples using U-theory statistics (14). We applied the EVA algorithm in the R/Bioconductor package GSReg to the RNA-Seq data for the gene sets defined by the disease-specific chromatin modifications (Supplemental File 2).

HPV integration detection by MapSplice

Detection was performed with MapSplice (31), which was run with the option to identify fusions on the RNA-Seq data. The reference for the reads to be mapped was a chimera that was prepared from a joint human and HPV16 genome. In this way, a viral integration site was visible as a fusion of a human chromosome and HPV genome. We considered the viral genome integrated if there were at least three discordant pairs (in which one end of the paired end read mapped to the viral genome, and its mate pair mapped to human genome) and one split read (in which one end of the paired end read spanned the human-viral junction, and its mate pair mapped to either the human or HPV genome). These seven total reads support integration at the same locus, according to our recent analysis (32). *Additional details are provided in* Supplementary Materials and Methods.

Identification of transcriptional enhancers

MACS peaks for H3K27ac were further input to Ranking Of Super Enhancers (ROSE) analysis software (33) to make enhancer calls for each sample. We applied this algorithm to merge H3K27ac peaks from MACS and ranked the resulting merged peaks as enhancers.

Results

High-quality ChIP-Seq data obtained in all tissue types and histone modifications

PDX models, established from primary tumor tissue samples directly implanted into immunodeficient mice, maintain important molecular features of human malignancies and provide sufficient tissue resources for profiling (6). We first used two HPV+ HNSCC patient-derived xenografts, PDX1 and PDX2, to optimize methods for sample preservation and processing required for ChIP-Seq analysis. Since human stromal elements are replaced by murine stroma as the engrafted tumor grows within its new biological niche, only the first passage of PDX tumors was used. These first passage PDX tumors were previously confirmed to have DNA methylation profiles similar to the tumor from which they were derived (6). Chromatin purification was performed using improved ChIP kit protocol, followed by optimized chromatin digestion procedure adjusted for each sample (see

Methods for details). PDX samples were processed in parallel with non-cancer controls: two primary oropharyngeal samples after uvulopalatopharyngoplasty (UPPP1 and UPPP2) and two HPV+ HNSCC cell lines (UPCI-SCC-090 and UM-SCC-047).

ChIP was performed for three well-characterized active histone marks that have been linked to carcinogenesis (H3K4me3, H3K9ac, and H3K27ac) (19) as well as a repressive histone mark (H3K9me3) using exceptional performance (XP®) monoclonal antibodies specifically developed for ChIP applications (see Methods for details). The ChIP-DNA deep sequencing produced an average of 46 million reads per sample (range: 29–88 million), which surpassed the ENCODE recommended guidelines for ChIP-Seq quality (34). Furthermore, over 89% of the reads aligned to the human reference genome (range: 56%–99%). Notably, the read frequency alignment did not differ significantly among samples types (Fig. S2A). PDX samples showed the lowest alignment rates to the human genome (mean: 76%, Fig. S2B) compared to cell lines or UPPP tissues. Nonetheless, the overall mapping rate in PDX samples was higher than that reported in standard mapping protocols (35).

The active H3K9ac histone mark enrichment had the highest number of sequencing reads across all samples analyzed (average: 59M) (Fig. S2C) as well as the highest alignment rate to the genome (average: 93%) (Fig. S2D), followed by the active H3K4me3 and H3K27ac. Due to the nature of chromatin structure, repressive histone marks are known to have lower sequencing quality. Although the repressive H3K9me3 histone mark had both lowest sequencing depth (average: 31M) (Fig S2C) and mapping frequency (average: 82%) (Fig. S2D), the overall sequencing quality was on par with that observed in the active marks.

Technical validation confirmed high concordance of ChIP-Seq and ChIP-based qRT-PCR

We performed a qRT-PCR analysis of the same ChIP-DNA that was used for the deep sequencing experiments for technical validation of the ChIP-Seq data. Three control genomic regions were analyzed: constitutively expressed *GAPDH* and *RPL10*, and transcriptionally repressed *ZNF333* genes (Fig. S3A–H) (2, 19). Furthermore, using RNA-Seq data from the same samples, we confirmed that histone enrichment within each one of the tested genes was associated with their expression level (Fig. S4).

Both ChIP-Seq and qRT-PCR analyses revealed that enrichment of all active histone marks near the transcriptionally repressed *ZNF333* gene was significantly lower than at the transcriptionally active *GAPDH* and *RPL10* genes (Fig. S3A–F). Enrichment of the repressive H3K9me3 mark was minimal or undetectable at actively transcribed *GAPDH* and *RPL10* genes but was significantly enriched near the repressed *ZNF333* gene loci, especially in cell lines and primary non-cancer tissues (Fig. S3G–H). Although the qRT-PCR platform was more sensitive, a strong concordance of histone enrichment was detected by both methodologies (Fig. S3A–H).

Clustering analysis of histone mark enrichment reveals correlation between sample types and histone modifications

The spatial correlation of histone enrichment peaks between all 24 samples (4 histone marks for six specimens) was calculated with DiffBind (26). The 24 × 24 correlation matrix was represented as a heatmap with unsupervised hierarchical clustering. The clustering analysis

segregates samples based on the relative similarity of their genome-wide ChIP-Seq peaks for each histone mark tested (Fig. 1). The analysis produced two distinct clusters independent of tissue type: samples with repressive (H3K9me3) and samples with active (H3K4me3, H3K9ac, H3K27ac) histone marks. The cluster of samples with active histone marks was further segregated by the disease status of the samples, with all non-cancerous UPPP tissues combining into a single cluster. To a lesser degree, a similar disease-specific clustering pattern was observed for samples with the repressive H3K9me3 mark. A strong co-clustering of ChIP-Seq sample-to-sample correlations for active histone marks was observed among normal tissues, suggesting a more homogenous genome-wide ChIP-Seq profile in these samples compared to the corresponding cancerous specimens. The co-clustering of cancerous cell lines and PDX samples further supported this observation. This was also consistent with previous studies that suggest strong tissue- and sample-specificity of H3K27ac, H3K4me3, and H3K9ac (19, 33).

ChIP-Seq analysis detected biologically relevant histone mark distribution around transcription start sites (TSSs)

To validate the biological relevance of the ChIP-Seq peak calls from MACS data, we analyzed the profiles of ChIP-Seq peaks in the vicinity of all TSS for all 24 samples. Increased enrichment of active histone marks was observed at TSSs with a signal drop at the nucleosome-free region (located upstream of the transcription start site) and clear enrichment peaks observed for transcribed nucleosomes (directly downstream of the transcription start site) (Fig. 2A–C). In opposite, repressed histone mark was not enriched around TSS (Fig. 2D). Cell lines showed the highest fold change enrichment of active histone marks, which can be attributed to the increased integrity of chromatin structure due to the absence of snap freezing steps (36). Nonetheless, the profiles of each histone mark were similar across all samples.

The profile of ChIP data relative to ChIP input was lower for the H3K9ac histone mark (Fig. 2B, S5) than either the H3K27ac or H3K4me3 marks in all samples (Fig. 2A, 2C, 2E, S6). On the other hand, the ChIP-Seq data for H3K9ac mark had overall higher MACS signal and a higher rate of alignment to the human genome than other activating marks (Fig. S2C–D), suggesting a wider spread of H3K9ac mark along the genome. The ChIP data for the H3K4me3 mark relative to ChIP input was the highest, and there were MACS peaks present around the TSSs of more than half of the genes used in the profile analysis (Fig. 2E). Peak enrichment for all active marks was highest around 400bp downstream of TSS at third transcribed nucleosome, with peaks from the first four transcribed nucleosomes being the most clearly distinguished. As expected, in all six samples, the repressive H3K9me3 mark did not have any enrichment relative to input control near TSSs (Fig. 2D, S7) (37). Taken together, these results demonstrate that ChIP-Seq data in this study consistently detected biologically relevant histone marks across different sample types and preservation techniques.

Tissue-specific histone enrichment was associated with gene expression in a cohort of primary HPV+ HNSCC and normal samples

The genes with disease-specific histone mark enrichment were identified by comparing the presence of specific mark enrichment at their 5′UTRs between PDX and UPPP samples (Tables S3–S6). We next analyzed the expression of such genes using RNA-Seq data for 47 HPV+ HNSCC tumors and 25 non-cancerous controls (13). This RNA-Seq cohort was inclusive of the two cell lines, and two UPPP samples, as well as two primary tumors from which the PDX models were derived, all of which were used for ChIP-Seq analysis. Unsupervised clustering of the RNA-Seq data was performed separately for the sets of disease-specific genes defined by the ChIP-Seq analysis in each of the individual histone marks (Fig. 3–4, S8–S9).

We detected 948 genes with disease-specific H3K4me3 enrichment: 415 UPPP-specific genes and 533 PDX-specific genes, including 206 genes which were also detected in HPV+ HNSCC cell lines (Table S3). Unsupervised hierarchical clustering of RNA-Seq data revealed a strong separation of the expression of genes that were associated with the disease-specific H3K4me3 histone mark (Fig. 3, columns). Notably, cell lines and PDX models co-clustered with primary tumor samples, whereas most normal controls clustered together (Fig. 3).

Similar to H3K4me3, H3K27ac revealed multiple disease-specific histone enrichment regions and their associated genes (total of 1800 differentially enriched genes: 317 UPPP-specific genes, and 1483 PDX-specific genes, including 519 genes that were also found in cell lines, Table S4). The gene expression analysis of nearby genes also separated tumor samples from healthy controls (Fig. 4).

Active H3K9ac and repressive H3K9me3 histone marks demonstrated comparable distribution patterns across the entire sample cohort, with only a limited number of histone enrichment regions specific to each tissue type (n=27 and n=15 for H3K9ac and H3K9me3 respectively) (Fig. S8–S9, Tables S5–S6). Nonetheless, the expression data for genes that were associated with these tissue-specific regions significantly co-clustered according to the disease state (Table 1). Among all four histone marks analyzed, we observed that the ChIP-Seq data from PDX models reflected the gene expression changes in primary tumor tissues (Fig. 3–4, S8–S9), and noticed broad differences between cell lines and primary tumors (**S10–S13**).

Gene expression changes in disease-specific histone enrichment sets had significantly more inter-sample variability in tumor than in normal samples

Heatmaps of gene expression data for genes associated with each histone mark visually showed greater inter-sample heterogeneity in tumor than normal samples (Fig. 3–4). We applied Expression Variation Analysis (EVA) algorithm (14) to the RNA-Seq data for sets of genes that are specifically enriched by individual marks (Fig. S14). This algorithm quantified their variability in tumor samples relative to normal specimens, by reporting p-values testing the null hypothesis that there is no difference in variation for genes in the disease-specific sets for each histone mark. The EVA analysis of the gene expression data

found significant dysregulation of the disease-specific gene sets for all histone marks (Table 1).

Genes with nearby tumor-specific enrichment of H3K4me3 and H3K27ac were associated with established cancer-related pathways

H3K27ac and H3K4me3 had the largest sets of genes with disease-specific histone modification enrichment (Tables S3–S4). The PDX-specific gene sets include numerous functional cancer-related genes that have been implicated in HPV+ HNSCC, including E2F transcription factors (*E2F1, E2F4* etc), growth factors and their receptors (*EGFR, FGF1* etc), forkhead box proteins (*FOXE1, FOXE3*), RAS oncogenes (*RAB23, RAB35*, etc), SRY-box proteins (*SOX2, SOX15,* etc), tumor necrosis factors (*TNFAIP1, TNFSF15,* etc), and NOTCH pathway proteins (*NOTCH3, JAG1, DVL3*), as well as *TP63, P53AIP1, TK1, ANO1, BIRC5,* and *SNAI1* (Tables S3–S4). Many of these genes are involved in cancer-related KRAS, NOTCH, p53, NFκB, IL/STAT, MYC, G2M checkpoint, glycosis, spermatogenesis and UV response pathways (Tables S7 and S9). Notably, the normal-specific gene sets for H3K4me3 and H3K27ac are enriched for pathways associated with allograft rejection, apoptosis, inflammatory response and IFNγ response pathways (Tables S8 and S10). These data suggest that histone marks can control the disease-specific gene expression.

H3K27ac enrichment segregates tumor samples by their HPV integration status, and H3K27ac enrichment regions correlate with sites of HPV integration

The clustering analysis of RNA-Seq data in gene sets associated with the H3K27ac mark enrichment segregated the tumor samples into two main subgroups (Clusters 1 and 2, Fig. 4), which was not observed for other histone marks (Fig. 3, S8–S9). Consistent with our findings, recent publications have established two subtypes of HPV+ HNSCC from gene expression profiles (29, 30), called HPV-KRT and HPV-IMU in (30) or mesenchymal and classical in (29). The HPV-KRT was associated with keratinocyte differentiation and episomal HPV, while HPV-IMU was associated with strong immune response, mesenchymal differentiation and HPV viral integration within the DNA (30). Moreover, that study also found that the episomal HPV infection correlated with the HPV-positive mesenchymal subtype from (29). The subtypes defined in these studies are correlated, and also associated with HPV-integration. In our cohort, HPV integration was detected only in one tumor specimen of ten in Cluster 1, whereas samples within the Cluster 2 were enriched for positive HPV integration (13 out of 29, Fisher Exact Test p-value of 0.064). Gene set enrichment analysis showed that genes with tumor-specific H3K27ac-associated enrichment were significantly associated with HPV-KRT (30) (Table S11) and with the classical subtypes (29) (Fig. S15A–D), confirming the relationship of H3K27ac enrichment to HPV integrated HNSCC subtypes.

Because of these observations, we hypothesized that H3K27ac enrichment is associated with HPV-integration. To test this hypothesis, we investigated whether the H3K27ac marks themselves co-localized with the sites of HPV integration in each sample profiled with ChIP-Seq. Of the four cancer samples with ChIP-Seq data, three had HPV integrated into the host genome. The HPV genome was integrated into narrow genomic regions that were

unique to each sample (PDX2 - 9q34.3, UPCI-SCC-090 – 9q22.33, UM-SCC-047 – 3q28) (Fig. 5). In all three cases, HPV integration sites co-localized with H3K27ac enhancers (38) in a sample-specific manner (Fig. 5). Notably, HPV integration and H3K27ac histone enrichment co-localized upstream of *TP63*, *FOXE1*, *NOTCH1*, and *EGFL7*, which all have been implicated in HNSCC tumorigenesis. Moreover, HPV integration-specific histone enrichment at locus 9q22.33 is proximal to the FRA3C DNA fragile site (10), which further confirmed chromatin's multifaceted role in HPV integration and carcinogenesis.

Discussion

The primary novelty of our study was a generation of the first ChIP-Seq data for human oropharyngeal samples, composed of two primary healthy tissues, two HPV+ HNSCC cell lines, and two HPV+ HNSCC PDXs. This study also presented the first chromatin-based analysis of HNSCC tumors, defining high disease-specificity of H3K4me3 and H3K27ac histone marks. We demonstrated that these histone marks were associated with tumor-specific transcriptional changes in their target genes. These chromatin-regulated genesets included well-characterized HNSCC-driving genes, such as *EGFR*, *FGFR1*, and *FOXA1* (Tables S3–S4). Our analysis also correlated tumor-specific dysregulation of the known cancer-related pathways, such as NOTCH and NFκB (Table S7 and S9) with chromatin re-organization. Moreover, this was the first HNSCC-based study to describe the relationship between chromatin structure and HPV integration status of HPV+ HNSCC subtypes.

ChIP-Seq requires a large amount of input DNA. Due to this limitation, large consortia, such as ENCODE, limited ChIP-Seq analysis for cancer samples to cell lines (2). Many recent high-profile publications utilizing ChIP-Seq to study tumorigenesis had similar difficulties and limited their ChIP-Seq analysis to a minimal number of samples, often only to a single specimen for each tissue type (38). The limitations of input specimen size posed a particular challenge for studying primary HPV+ HNSCC tumors, which can be relatively small. To overcome this limitation and enlarge the tissue volume available for ChIP-Seq analysis, we have performed primary tumor xenografting. We acknowledge that minor changes associated with the xenografting procedure may occur due to tumor evolution and altered tumor environment. To address this concern, we limited ChIP-Seq studies for first generation PDXs. We found that a single xenografting step preserved the gene expression profile of the primary tissues, similar to previous observations that early stage xenografts preserved DNA methylation profiles of the parental neoplasm (6).

One limitation of our study is a modest sample size of 2 HPV+ HNSCC PDXs and 2 HPV+ HNSCC cell lines. In addition to being small, HPV+ HNSCC tumors are often not surgically excised, limiting primary tissue availability. In fact, there are fewer than 10 HPV+ HNSCC cell lines worldwide (10), and a limited number of HPV+ HNSCC PDXs are currently described (39). These limitations are reflected in relatively small cohorts of HPV+ HNSCC samples in large consortia for HNSCC genomics profiling, such as TCGA (3). Thus, our dataset is still a unique resource to characterize the chromatin landscape of these virally associated tumors.

To evaluate this small cohort of ChIP-Seq samples, we selected DiffBind, a bioinformatics algorithm that allowed for appropriately powered genome-wide correlation analyses using integration with RNA-Seq data from a larger cohort of HPV+ HNSCC samples (47 tumors and 25 normals). Specifically, DiffBind (Fig. 1) considered each gene as an event and was thus sufficiently powered to perform sample-to-sample comparisons for the genomic regions defined by the ChIP-Seq profile.

In RNA-Seq analysis with available data for a larger HPV+ HNSCC cohort, we observed much larger inter-sample heterogeneity of gene expression in tumor samples within tumor-specific ChIP-Seq peaks than that of normal samples within normal-specific ChIP-Seq peaks (Fig. 3–4). Within individual tumor samples, changes in gene expression may arise from histone modifications, but may also be influenced by mutations, DNA methylation, and copy number variations. Therefore, to tease out chromatin structure related changes, it is essential to relate the tissue-type-specific ChIP-Seq peaks analysis with the inter-tissue-type heterogeneity of gene expression, rather than directly evaluating gene expressional levels. Recently, the EVA algorithm was developed to compare the variability of expression profiles for gene sets in tumor samples relative to that of normal specimens (14). This reliable quantification of the relative variability in sample phenotypes can indicate significant pathway dysregulation in one phenotype relative to another. Therefore, the EVA algorithm was uniquely suited to determine gene expression dysregulation associated with tissue-specific chromatin structure in our study (Table 1). Future work is needed to adapt EVA to account for sample-specific marks for integrated analysis of large cohorts with matched ChIP-Seq and RNA-Seq data.

Both the correlation-based ChIP-Seq and RNA-Seq analyses demonstrated higher tissue-type specificity in active H3K4me3 and H3K27ac histone marks than the active mark H3K9ac or repressive mark H3K9me3. Both of these histone marks have been found to be pervasive in actively transcribed genes, as well as enhancer regions reported in previous studies (38). Enhancers are known as tissue-specific regulators of gene expression during cell differentiation and cancer development (33, 38, 40, 41). Enhancers are commonly identified as genomic elements enriched by histone modification (H3K27ac and different H3K4me isoforms), predominantly hypomethylated (42, 43), and occupied by various transcription factors, BRD4 (BET bromodomain protein, an activator of RNA polymerase II - PolII), MED1 (PolII transcription subunit) proteins, and PolII itself (33). H3K27ac enrichment at enhancer regions is recognized by BRD4, whose inhibition leads to a dysregulation in gene expression of oncogenes such as *MYC, MYB, MMP9, BCL2,* and *CCND1* (40, 41, 44, 45). This is in a strong concordance with our observation that the regions with differential enrichment of the H3K27ac histone mark were found near known HNSCC associated genes: *EGFR, CEBPD, TP63, FOXE1, NOTCH1*, *GATA6, SOX2*, and *EGFL7* (Table S4), whose expression may also be regulated through BRD4 and its inhibitors (40, 41, 44–46). This suggests that BRD4 may play a role in expression regulation of those genes (Table S4) and their associated HNSCC-related pathways (Table S7), which merits further investigated.

In this study, we demonstrated the correlation of H3K27ac enrichment and HPV integration status. However, currently, there is no gold-standard for evaluation of HPV integration.

Whole genome sequencing (WGS) is a preferred but expensive methodology (32). Some researchers agree that the detection of multiple reads across the human-viral fusions defines the HPV integration into the genome, whereas no detection of human-viral reads represents the episomal virus (32). Nonetheless, in many cases such as samples with lower sequencing depth, low HPV copy number, utilization of whole exome sequencing (WES) or RNA-Seq evaluation, the desired number of the human-viral junctions may not be reached or detected. Such results could produce false negative detection of HPV integration. Parfenov and colleagues developed a protocol with stringent criteria and detected 71% of HPV integration in HPV+ TCGA population using three independent high-throughput methodologies: RNA-Seq, WGS, and WES (32). Using only RNA-Seq analysis, a lower rate (34%) of HPV integration was detected, and the high false-negative rate of integration detection may be because only transcriptomic data was evaluated. Nonetheless, analysis of gene expression from this RNA-Seq data in genes associated with ChIP-Seq marks for H3K27ac distinguished clusters of HPV+ HNSCC samples with different numbers of HPV-integration calls.

The role of HPV-integration in HNSCC is an area of active research. A recent study supported the hypothesis that HPV is only transiently integrated into the host genome and then subsequently excised out of the human genome together with human sequences to form human-viral chimeric episomes with multiple copies of HPV (32, 47). If this hypothesis is true, the human-viral fusion sites, that we called integration sites in this study, could instead represent chimeric human-viral episomes described in other studies (32, 47). The fact that the episomal and integrated HPV can co-exist in the same tissue or even the same cell (30, 32, 47, 48), and both can be transcribed, adds another twist to the complex picture of HPV infection in human cancers. Southern blots or FISH-based technologies followed by high-resolution microscopy might help to resolve this scientific dispute (32, 49), and should be evaluated in conjunction with H3K27ac binding sites in future studies.

Although the role of HPV-integration is controversial, recent studies have demonstrated that HPV+ HNSCC samples with episomal and integrated HPV have unique signatures of gene expression, and can be separated based on them (22, 30, 32, 47–49). Clustering of gene expression data described two HPV+ HNSCC subtypes: mesenchymal and classical (29). A later study independently found two subtypes: HPV-KRT (integrated HPV) and HPV-IMU (episomal HPV) (30). The episomal HPV infection also correlated with the HPV-positive mesenchymal subtype reported in (29). Our analysis of both gene expression signatures confirmed the predominant association of tumor-specific H3K27ac-associated genes with HPV-KRT (30) and classical subtype (29), confirming the strong correlation between HPV integration and H3K27ac enrichment (Fig. S15A–D and Table S11). This association was further supported by co-localization of HPV-integration sites with H3K27ac peaks (Fig. 5).

Published data suggest that integration of different viruses correlates with an “open chromatin” landscape (50, 51). The integration of the HPV genome is crucial to HPV-related carcinogenesis and progression (52). Indeed, 34–71% of HPV+ HNSCC tumors have the virus integrated into the host genome (10, 48). Furthermore, recent data confirm that integration of oncogenic viruses, including HPV, into the host genome, is not random (32). The murine leukemia virus (MLV) tends to integrate into DNase I hyper-sensitive sites (50),

most likely accommodated by the interaction of MLV integrase with BRD proteins, known to bind “open chromatin” mark, H3K27ac (46). Integration of gamma-retrovirus also correlates with H3K27Ac modification (51). Our results further indicated that H3K27ac marks distinguish tumor samples with and without HPV integration (Fig. 4 and S15A-D, Table S11). HPV integration sites also strongly co-localize with H3K27ac marks in each sample (Fig. 5). HPV integration in the UM-SCC-047 cell line was located in close proximity (5.5Mb) to the fragile FRA3C site, which is known to have a high double-stranded break rate (10, 11). Additionally, sites of HPV integration were near known cancer-related genes in our samples: *TP63* (18) (047 cell line); *FOXE1* (53) (090 cell line); *NOTCH1* (8) (PDX2); and *EGFL7* (54) (PDX2). This suggests that HPV integration occurs near genes that play a central role in HNSCC development (Tables S3–S6).

The correlation of the HPV integration status and tumor stage is still an active area of research. In our study cohort, only 3 out of 47 HPV+ HNSCC patients were TNM stage III, whereas the other 44 were TNM stage IV tumors. This is reflective of the clinical presentation of HPV+ HNSCC, which is predominantly diagnosed with the large regional nodal disease. This homogenous tumor population limited any conclusions regarding the correlation of HPV integration and TNM stage. A larger number of early stage samples will be required for the further detailed investigation to elucidate the correlation of the tumor stage and HPV integration status in HNSCC.

This study has several limitations, primarily associated with the small sample size. First, the small ChIP-Seq cohort limited statistical analyses of high-throughput data between samples. Also, the simple detection of HPV integration did not allow us to draw any conclusions regarding HPV excision during carcinogenesis. Homogeneous distribution of TNM stages within HPV+ HNSCC patients prevented drawing a conclusion between clinical characteristics and HPV integration status.

A large amount of tumor volume required for ChIP-Seq analysis, prohibited the use of primary tumor tissue, limiting our study to cell lines and xenografts. We recognize that changes occur in the tumor microenvironment during xenografting, and in combination with small sample size, some limited conclusions could be drawn. However, utilizing bioinformatic methods such as DiffBind, an adequate statistical power was achieved. The small size of our cohort also limited any definitive conclusions about the genetic and epigenetic difference between samples with and without HPV integration into the host genome.

Our confirmation of the correlation of H3K27ac mark and HPV integration is a critical first step to delineating the relationship between enhancers and viral HNSCC carcinogenesis. Further functional studies among larger cohorts are necessary to establish the role of chromatin structure in mediating HPV integration and alterations to cancer-related genes in HPV+ HNSCC. The presented optimized ChIP protocol for primary tumor samples and integration of the ChIP-Seq results with RNA-Seq data has wide applicability and can be expanded to better understand the interplay between chromatin structure changes and their downstream effects on gene expression in various types of cancer.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support:

- D.A. Gaykalova received R21DE025398
- D.A. Gaykalova, E.J. Fertig, D. Sidransky, W.M. Koch, and E. Izumchenko received 5P50DE019032
- E.J. Fertig received R01CA177669
- E.J. Fertig received P30CA006973
- J.A. Califano received R01DE023347
- E. Izumchenko received ALCF-IASLC Award.

We thank SKCCC Next Generation Sequencing Center and JHMI Deep Sequencing & Microarray Core on performing and providing advice on ChIP-Seq and RNA-Seq data, respectively; L. Kagohara, G. Stein-O'Brien, T. Ou, F. Zamuner, and K. Zambo for critical comments and feedback during the preparation of the manuscript.

References

1. Lutz L, Fitzner IC, Ahrens T, Geissler AL, Makowiec F, Hopt UT, et al. Histone modifiers and marks define heterogeneous groups of colorectal carcinomas and affect responses to HDAC inhibitors in vitro. Am J Cancer Res. 2016; 6:664–76. [PubMed: 27152243]
2. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
3. Cancer Genome Atlas N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015; 517:576–82. [PubMed: 25631445]
4. Zwart W, Koornstra R, Wesseling J, Rutgers E, Linn S, Carroll JS. A carrier-assisted ChIP-seq method for estrogen receptor-chromatin interactions from breast cancer core needle biopsy samples. BMC Genomics. 2013; 14:232. [PubMed: 23565824]
5. Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Brief Bioinform. 2017; 18:279–90. [PubMed: 26979602]
6. Hennessey PT, Ochs MF, Mydlarz WW, Hsueh W, Cope L, Yu W, et al. Promoter methylation in head and neck squamous cell carcinoma cell lines is significantly different than methylation in primary tumors and xenografts. PLoS One. 2011; 6:e20584. [PubMed: 21637785]
7. Pulte D, Brenner H. Changes in survival in head and neck cancers in the late 20th and early 21st century: a period analysis. Oncologist. 2010; 15:994–1001. [PubMed: 20798198]
8. Gaykalova DA, Mambo E, Choudhary A, Houghton J, Buddavarapu K, Sanford T, et al. Novel insight into mutational landscape of head and neck squamous cell carcinoma. PLoS One. 2014; 9:e93102. [PubMed: 24667986]
9. Hayes DN, Van Waes C, Seiwert TY. Genetic Landscape of Human Papillomavirus-Associated Head and Neck Cancer and Comparison to Tobacco-Related Tumors. J Clin Oncol. 2015; 33:3227–34. [PubMed: 26351353]
10. Olthof NC, Huebbers CU, Kolligs J, Henfling M, Ramaekers FC, Cornet I, et al. Viral load, gene expression and mapping of viral integration sites in HPV16-associated HNSCC cell lines. Int J Cancer. 2015; 136:E207–18. [PubMed: 25082736]
11. Richards RI. Fragile and unstable chromosomes in cancer: causes and consequences. Trends Genet. 2001; 17:339–45. [PubMed: 11377796]
12. Papillon-Cavanagh S, Lu C, Gayden T, Mikael LG, Bechet D, Karamboulas C, et al. Impaired H3K36 methylation defines a subset of head and neck squamous cell carcinomas. Nat Genet. 2017; 49:180–5. [PubMed: 28067913]

13. Guo T, Gaykalova DA, Considine M, Wheelan S, Pallavajjala A, Bishop JA, et al. Characterization of functionally active gene fusions in human papillomavirus related oropharyngeal squamous cell carcinoma. Int J Cancer. 2016; 139:373–82. [PubMed: 26949921]
14. Afsari B, Geman D, Fertig EJ. Learning dysregulated pathways in cancers from differential variability analysis. Cancer Inform. 2014; 13:61–7. [PubMed: 25392694]
15. Stebbing J, Paz K, Schwartz GK, Wexler LH, Maki R, Pollock RE, et al. Patient-derived xenografts for individualized care in advanced sarcoma. Cancer. 2014; 120:2006–15. [PubMed: 24705963]
16. Gaykalova DA, Zizkova V, Guo T, Tiscareno I, Wei Y, Vatapalli R, et al. Integrative computational analysis of transcriptional and epigenetic alterations implicates DTX1 as a putative tumor suppressor gene in HNSCC. Oncotarget. 2017
17. Gaykalova DA, Vatapalli R, Wei Y, Tsai HL, Wang H, Zhang C, et al. Outlier Analysis Defines Zinc Finger Gene Family DNA Methylation in Tumors and Saliva of Head and Neck Cancer Patients. PLoS One. 2015; 10:e0142148. [PubMed: 26544568]
18. Li R, Ochs MF, Ahn SM, Hennessey P, Tan M, Soudry E, et al. Expression microarray analysis reveals alternative splicing of LAMA3 and DST genes in head and neck squamous cell carcinoma. PLoS One. 2014; 9:e91263. [PubMed: 24675808]
19. Kimura H, Hayashi-Takanaka Y, Goto Y, Takizawa N, Nozaki N. The organization of histone H3 modifications as revealed by a panel of specific monoclonal antibodies. Cell Struct Funct. 2008; 33:61–73. [PubMed: 18227620]
20. Sidransky D, Von Eschenbach A, Tsai YC, Jones P, Summerhayes I, Marshall F, et al. Identification of p53 gene mutations in bladder cancers and urine samples. Science. 1991; 252:706–9. [PubMed: 2024123]
21. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001; 25:402–8. [PubMed: 11846609]
22. Castera L, Krieger S, Rousselin A, Legros A, Baumann JJ, Bruet O, et al. Next-generation sequencing for the diagnosis of hereditary breast and ovarian cancer using genomic capture targeting multiple candidate genes. Eur J Hum Genet. 2014; 22:1305–13. [PubMed: 24549055]
23. Kumar P, Al-Shafai M, Al Muftah WA, Chalhoub N, Elsaid MF, Aleem AA, et al. Evaluation of SNP calling using single and multiple-sample calling algorithms by validation against array base genotyping and Mendelian inheritance. BMC Res Notes. 2014; 7:747. [PubMed: 25339461]
24. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nature protocols. 2012; 7:1728–40. [PubMed: 22936215]
25. Shin H, Liu T, Manrai AK, Liu XS. CEAS: cis-regulatory element annotation system. Bioinformatics. 2009; 25:2605–6. [PubMed: 19689956]
26. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature. 2012; 481:389–93. [PubMed: 22217937]
27. Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014; 42:W187–91. [PubMed: 24799436]
28. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102:15545–50. [PubMed: 16199517]
29. Keck MK, Zuo Z, Khattri A, Stricker TP, Brown CD, Imanguli M, et al. Integrative analysis of head and neck cancer identifies two biologically distinct HPV and three non-HPV subtypes. Clin Cancer Res. 2015; 21:870–81. [PubMed: 25492084]
30. Zhang Y, Koneva LA, Virani S, Arthur AE, Virani A, Hall PB, et al. Subtypes of HPV-Positive Head and Neck Cancers Are Associated with HPV Characteristics, Copy Number Alterations, PIK3CA Mutation, and Pathway Signatures. Clin Cancer Res. 2016; 22:4735–45. [PubMed: 27091409]
31. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010; 38:e178. [PubMed: 20802226]

32. Parfenov M, Pedamallu CS, Gehlenborg N, Freeman SS, Danilova L, Bristow CA, et al. Characterization of HPV and host genome interactions in primary head and neck cancers. Proc Natl Acad Sci U S A. 2014; 111:15544–9. [PubMed: 25313082]
33. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013; 153:307–19. [PubMed: 23582322]
34. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014; 15:121–32. [PubMed: 24434847]
35. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol. 2013; 9:e1003326. [PubMed: 24244136]
36. Milani P, Escalante-Chong R, Shelley BC, Patel-Murray NL, Xin X, Adam M, et al. Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells. Sci Rep. 2016; 6:25474. [PubMed: 27146274]
37. Barth TK, Imhof A. Fast signals and slow marks: the dynamics of histone modifications. Trends Biochem Sci. 2010; 35:618–26. [PubMed: 20685123]
38. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013; 155:934–47. [PubMed: 24119843]
39. Facompre ND, Sahu V, Montone KT, Harmeyer KM, Nakagawa H, Rustgi AK, et al. Barriers to generating PDX models of HPV-related head and neck cancer. Laryngoscope. 2017
40. Drier Y, Cotton MJ, Williamson KE, Gillespie SM, Ryan RJ, Kluk MJ, et al. An oncogenic MYB feedback loop drives alternate cell fates in adenoid cystic carcinoma. Nat Genet. 2016; 48:265–72. [PubMed: 26829750]
41. Liu F, Hon GC, Villa GR, Turner KM, Ikegami S, Yang H, et al. EGFR Mutation Promotes Glioblastoma through Epigenome and Transcription Factor Network Remodeling. Mol Cell. 2015; 60:307–18. [PubMed: 26455392]
42. Charlet J, Duymich CE, Lay FD, Mundbjerg K, Dalsgaard Sorensen K, Liang G, et al. Bivalent Regions of Cytosine Methylation and H3K27 Acetylation Suggest an Active Role for DNA Methylation at Enhancers. Mol Cell. 2016; 62:422–31. [PubMed: 27153539]
43. Heyn H, Vidal E, Ferreira HJ, Vizoso M, Sayols S, Gomez A, et al. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. Genome Biol. 2016; 17:11. [PubMed: 26813288]
44. Wu SY, Nin DS, Lee AY, Simanski S, Kodadek T, Chiang CM. BRD4 Phosphorylation Regulates HPV E2-Mediated Viral Transcription, Origin Replication, and Cellular MMP-9 Expression. Cell Rep. 2016; 16:1733–48. [PubMed: 27477287]
45. Dooley KE, Warburton A, McBride AA. Tandemly Integrated HPV16 Can Form a Brd4-Dependent Super-Enhancer-Like Element That Drives Transcription of Viral Oncogenes. MBio. 2016:7.
46. Zhou H, Schmidt SC, Jiang S, Willox B, Bernhardt K, Liang J, et al. Epstein-Barr virus oncoprotein super-enhancers control B cell growth. Cell Host Microbe. 2015; 17:205–16. [PubMed: 25639793]
47. Nulton TJ, Olex AL, Dozmorov M, Morgan IM, Windle B. Analysis of The Cancer Genome Atlas sequencing data reveals novel properties of the human papillomavirus 16 genome in head and neck squamous cell carcinoma. Oncotarget. 2017; 8:17684–99. [PubMed: 28187443]
48. Olthof NC, Speel EJ, Kolligs J, Haesevoets A, Henfling M, Ramaekers FC, et al. Comprehensive analysis of HPV16 integration in OSCC reveals no significant impact of physical status on viral oncogene and virally disrupted human gene expression. PLoS One. 2014; 9:e88718. [PubMed: 24586376]
49. Vojtechova Z, Sabol I, Salakova M, Turek L, Grega M, Smahelova J, et al. Analysis of the integration of human papillomaviruses in head and neck tumours in relation to patients' prognosis. Int J Cancer. 2016; 138:386–95. [PubMed: 26239888]
50. Wu X, Li Y, Crise B, Burgess SM. Transcription start regions in the human genome are favored targets for MLV integration. Science. 2003; 300:1749–51. [PubMed: 12805549]

51. Gilroy KL, Terry A, Naseer A, de Ridder J, Allahyar A, Wang W, et al. Gamma-Retrovirus Integration Marks Cell Type-Specific Cancer Genes: A Novel Profiling Tool in Cancer Genomics. PLoS One. 2016; 11:e0154070. [PubMed: 27097319]

52. Wentzensen N, Vinokurova S, von Knebel Doeberitz M. Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. Cancer Res. 2004; 64:3878–84. [PubMed: 15172997]

53. Pereira JS, da Silva JG, Tomaz RA, Pinto AE, Bugalho MJ, Leite V, et al. Identification of a novel germline FOXE1 variant in patients with familial non-medullary thyroid carcinoma (FNMTC). Endocrine. 2015; 49:204–14. [PubMed: 25381600]

54. Diaz R, Silva J, Garcia JM, Lorenzo Y, Garcia V, Pena C, et al. Deregulated expression of miR-106a predicts survival in human colon cancer patients. Genes Chromosomes Cancer. 2008; 47:794–802. [PubMed: 18521848]

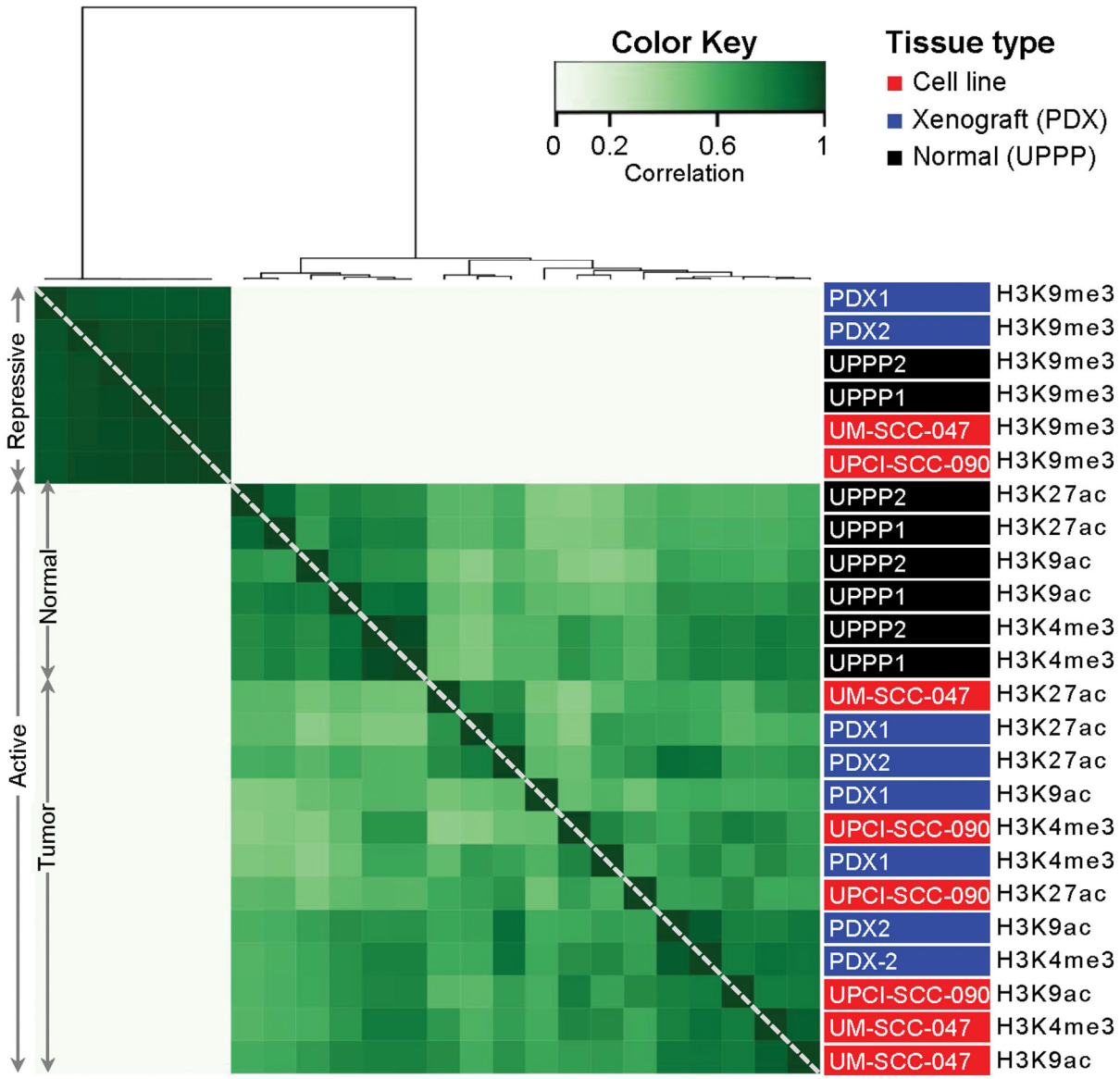

Figure 1. Non-ordered DiffBind analysis revealed a spatial correlation between histone marks in different sample types

The histone mark enrichment distribution was analyzed by the DiffBind algorithm, which calculates the spatial correlations between genomic distributions of histone marks in two different samples. The correlations range from 0 (no correlation, white) to 1 (strong correlation, dark green). Overall, 24 samples were used in the analysis to build 24 × 24 matrix: two of each tissue type (cell lines – red; xenografts - blue; normal controls – back) and four histone marks (active: H3K4me3, H3K9ac, and H3K27ac, as well as repressive: H3K9me3). Two main patterns were revealed: independent clustering of repressive mark regardless of tissue type and the cluster of normal controls regardless of active histone mark nature.

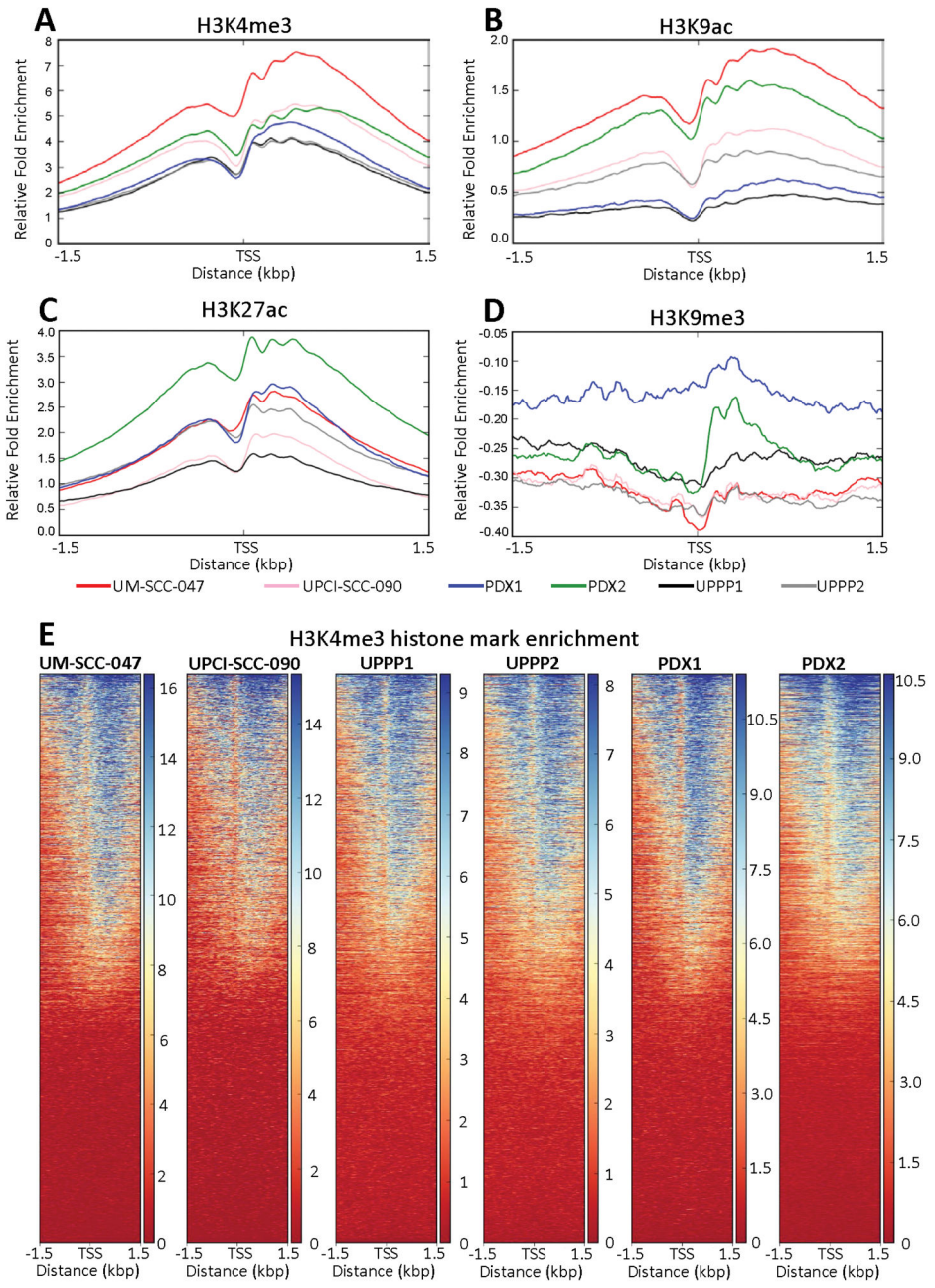

Figure 2. Histone mark enrichment distribution near transcription start sites

The average genome-wide histone enrichment calculated by MACS near transcriptions start sites (TSS, ±1.5 kbp) was calculated for all known genes for each individual histone modification: H3K4me3 (A), H3K9ac (B), H3K27ac (C), and H3K9me3 (D) and shown for individual samples (cell lines: UM-SCC-047 [red] and UPCI-SCC-090 [pink]; xenografts: PDX1 [blue] and PDX2 [green]; normal controls: UPPP1 [black] and UPPP2 [gray]). The relative fold enrichment was calculated by MACS algorithm, which accounted for the background signal by comparing the ChIP peaks within an individual study sample to its own 2% input DNA control, via looking for read orientation and mapping density that

indicates histone binding. Notably, histone mark enrichment near the TSS was detected only for active, but not the repressive modifications. (E) H3K4me3 histone mark enrichment calculated by MACS near individual TSS (±1.5 kbp) genome wide was ranked by the overall fold enrichment for individual samples. The scale of fold enrichment distribution for individual samples is on the right of each enrichment matrix. Notably, half of all known genes had H3K4me3 enrichment near TSSs.

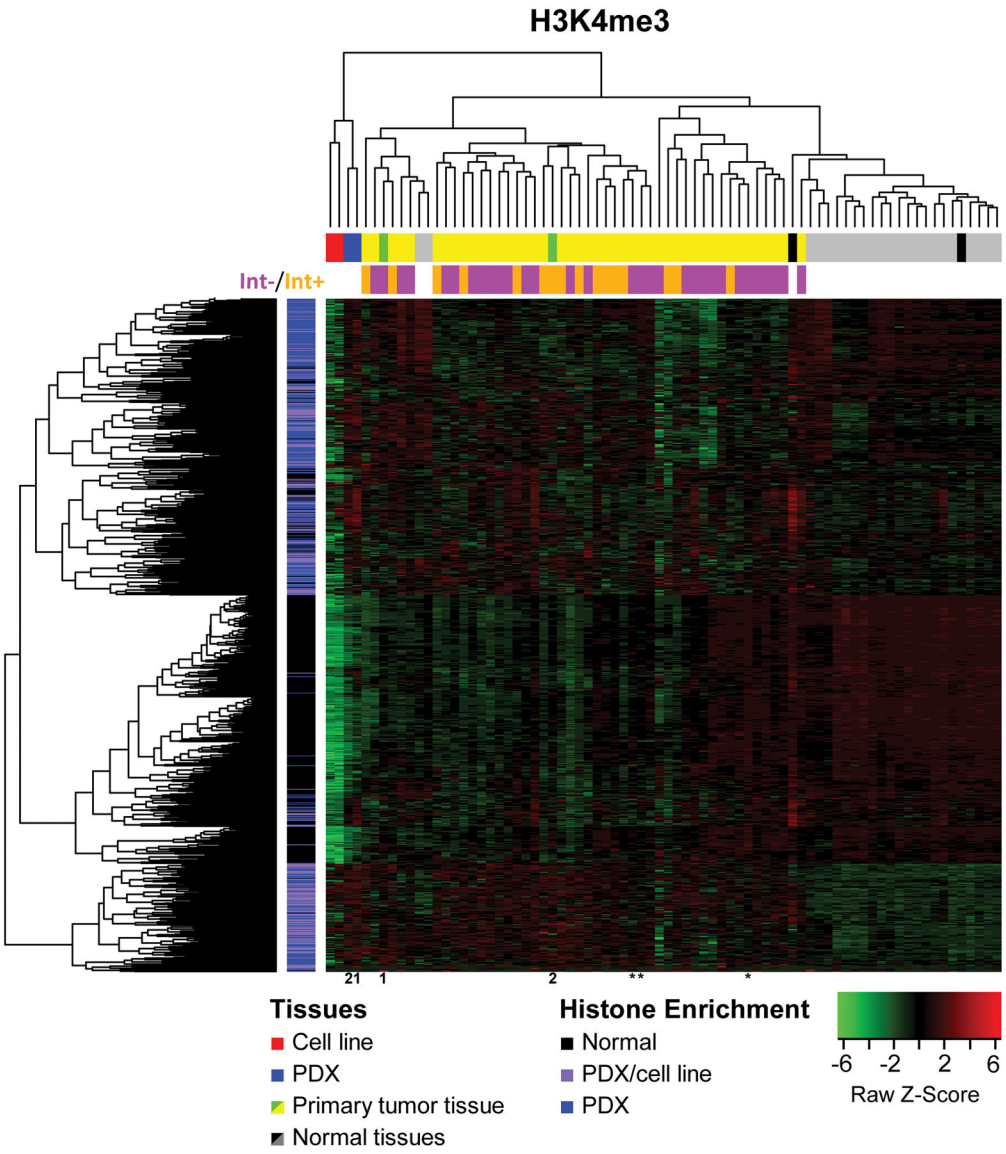

Figure 3. Expression variation analysis revealed the strong correlation of tissue specific H3K4me3 histone peaks with an expression of nearby genes

Two normal and two xenograft tissues were compared by histone enrichment distribution at 5′UTR of all known genes to detect disease specific histone enrichment peaks (vertical bar beside heat map: normal – black, cancer – blue and purple), with all overlapping histone enrichment peaks removed from the analysis (Table S3). Additionally, purple regions were detected both in xenografts and cell lines. The expression of the associated gene to each differentially enriched region was evaluated by RNA-Seq for all six ChIP-Seq study samples (horizontal bar above heat map: cell lines – red, xenografts – blue [1 and 2], and normal controls – black) as well as an extended cohort of 47 HPV+ tumors (yellow, including PDX-parental tissues – green [1 and 2]) and 25 non-cancer controls (gray/black). The expression of the nearest gene was calculated as Z-score ranging from −6 (underexpression) to 6 (overexpression). Both samples- (columns) and tissue-specific histone enrichment regions (rows) were hierarchically clustered without supervision, which revealed segregation of

samples and histone enrichment by disease status. The p-values for disease-specific samples segregation is listed in Table 1. The HPV status of HNSCC samples was indicated as pink (Int-, episomal HPV genome with no detected HPV integration into the host genome by MapSplice) or orange (Int+, integration of HPV into host genome detected by MapSplice). Three tumor samples with TNM stage III are indicated by asterisks.

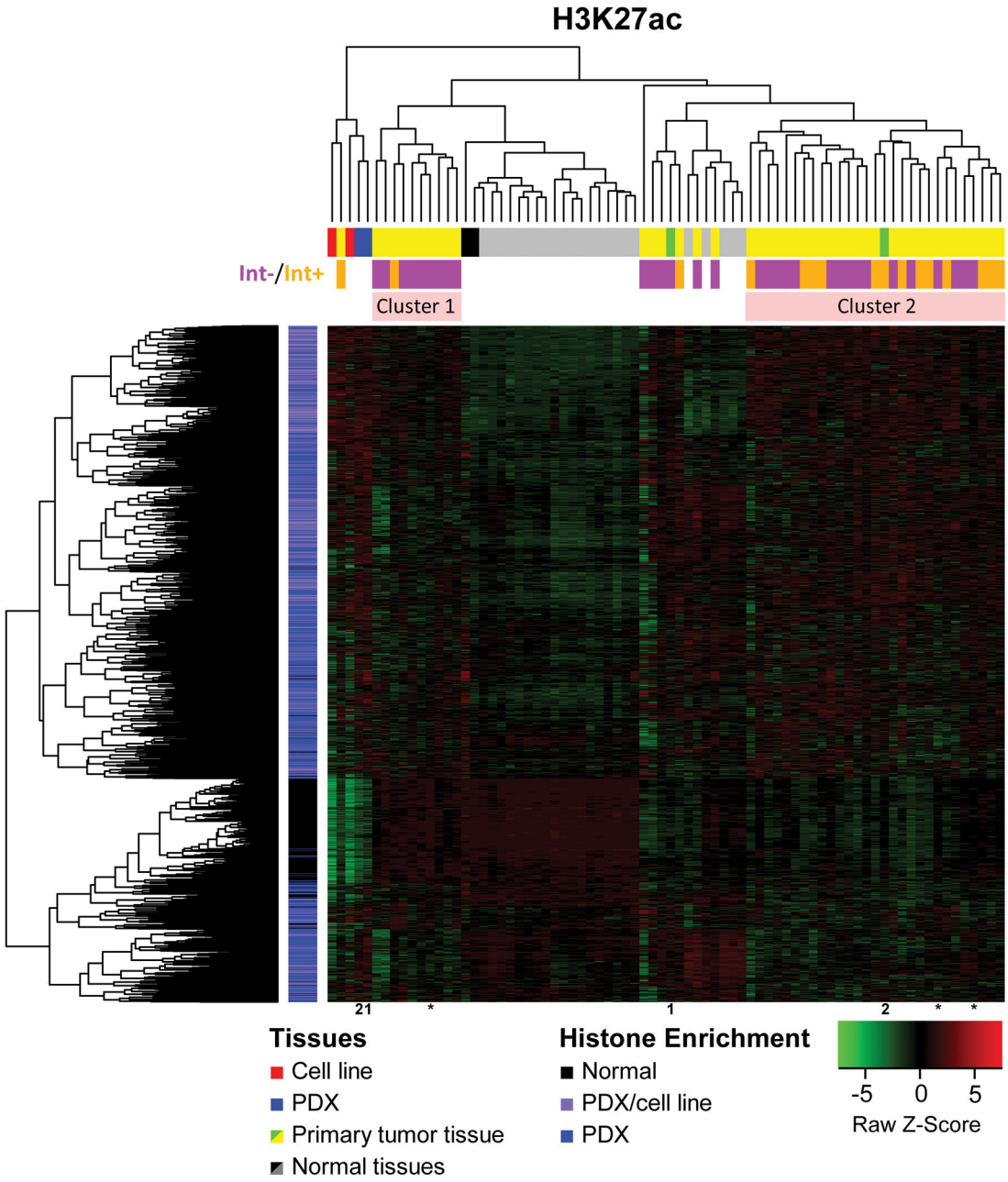

Figure 4. Expression of genes next to tumor-specific H3K27ac enrichment signatures correlates with HPV integration status of HNSCC samples

The analysis was performed similarly to that shown in Fig. 3. Two normal and two xenograft tissues were compared by histone enrichment distribution near 5′UTR of known genes to detect disease-specific histone enrichment peaks (vertical bar beside heat map: normal – black, cancer – blue and purple, see Table S3 for gene list). The expression of the closest gene to each differentially enriched region was evaluated by RNA-Seq for six ChIP-Seq study samples (horizontal bar above the heat map: cell lines – red, xenografts – blue [1 and 2], and normal controls – black) as well as extended HPV+ cohort of 47 tumors (yellow or green [1 and 2]) and 25 non-cancer controls (gray/black). The expression of the associated

gene was calculated as Z-score ranging from −5 (underexpression) to 5 (overexpression). The HPV status of HNSCC samples was indicated as pink (Int-, episomal HPV genome with no detected HPV integration into the host genome by MapSplice) or orange (Int+, integration of HPV into host genome detected by MapSplice). Segregation of samples by gene expression of H3K27ac-enriched genes revealed two dominate tumor clusters (salmon, Clusters 1 and 2) with a different distribution of Int+ HPV HNSCC samples. Three tumor samples with TNM stage III are indicated by asterisks.

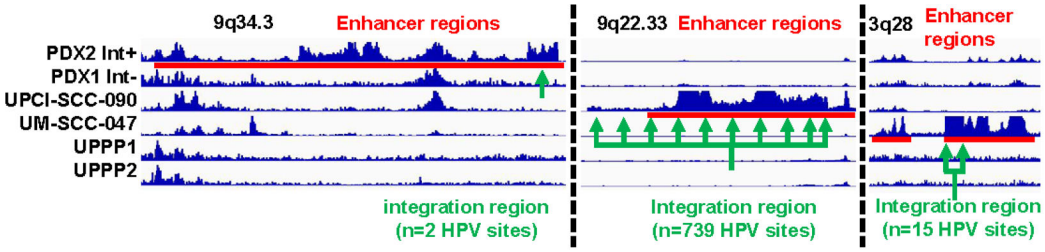

Figure 5. Detection of enhancer at the sites of HPV integration

Integrated genome viewer (IGV) visualization of H3K27ac-ChIP-Seq peaks for six study samples are shown (xenografts: PDX1 without detected HPV integration into host genome [Int-], and PDX2 with integrated HPV [Int+]; HPV+ HNSCC cell lines: UPCI-SCC-090 and UM-SCC-047 both with integrated HPV genome and normal controls: UPPP1 and UPPP2, both HPV-). Three genomic regions with detected sites of HPV integration for the three Int+ HPV+ HNSCC samples are shown: 9q34.3 (HPV integration for PDX2 sample); 9q22.3 (HPV integration for UPCI-SCC-090 sample); and 3q28 (HPV integration for UM-SCC-047 sample). HPV integration sites in each sample were detected by MapSplice of RNA-Seq data and are shown in green. Enhancer regions (red) were defined by ROSE analysis.

Table 1

EVA analysis defines tissue specificity of the differentially enriched histone peaks

	PDX-specific histone enrichment regions	UPPP-specific histone enrichment regions
H3K4me3	$<1\times10^{-10}$	9.35×10^{-10}
H3K9ac	2.59×10^{-4}	2.46×10^{-3}
H3K9me3	$<1\times10^{-10}$	6.14×10^{-4}
H3K27ac	1.23×10^{-9}	3.78×10^{-8}

The list of genes next to the tissue-specific histone enrichment peaks (Tables S3–S6) were used for EVA algorithm analysis (14). PDX-specific regions and their genes (column 1) and UPPP-specific regions and their genes (column 2) were used for p-value calculations.