Molecular Oncology

FEBS PRESS science publishing by scientists

Cis-regulatory effect of HPV integration is constrained by host chromatin architecture in cervical cancers

Anurag Kumar Singh[1], Kaivalya Walavalkar[1], Daniele Tavernari[2,3,4], Giovanni Ciriello[2,3,5], Dimple Notani[1] and Radhakrishnan Sabarinathan[1]

1 National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bengaluru, India
2 Department of Computational Biology, University of Lausanne (UNIL), Switzerland
3 Swiss Cancer Center Leman, Lausanne, Switzerland
4 Swiss Institute for Experimental Cancer Research (ISREC), EPFL, Lausanne, Switzerland
5 Swiss Institute of Bioinformatics, Lausanne, Switzerland

Keywords
cervical cancer; chromatin structure; gene-regulation; HPV integration; topologically associating domains

Correspondence
R. Sabarinathan, National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bengaluru 560065, India
E-mail: sabari@ncbs.res.in

(Received 10 October 2023, accepted 24 November 2023, available online 9 December 2023)

doi:10.1002/1878-0261.13559

Human papillomavirus (HPV) infections are the primary drivers of cervical cancers, and often HPV DNA gets integrated into the host genome. Although the oncogenic impact of HPV encoded genes is relatively well known, the cis-regulatory effect of integrated HPV DNA on host chromatin structure and gene regulation remains less understood. We investigated genome-wide patterns of HPV integrations and associated host gene expression changes in the context of host chromatin states and topologically associating domains (TADs). HPV integrations were significantly enriched in active chromatin regions and depleted in inactive ones. Interestingly, regardless of chromatin state, genomic regions flanking HPV integrations showed transcriptional upregulation. Nevertheless, upregulation (both local and long-range) was mostly confined to TADs with integration, but not affecting adjacent TADs. Few TADs showed recurrent integrations associated with overexpression of oncogenes within them (e.g. MYC, PVT1, TP63 and ERBB2) regardless of proximity. Hi-C and 4C-seq analyses in cervical cancer cell line (HeLa) demonstrated chromatin looping interactions between integrated HPV and MYC/PVT1 regions (~ 500 kb apart), leading to allele-specific overexpression. Based on these, we propose HPV integrations can trigger multimodal oncogenic activation to promote cancer progression.

1. Introduction

Cervical cancer is the fourth most common cancer type among women worldwide, with the majority being reported from developing and underdeveloped countries [1]. Studies have established that most cervical cancer cases can be attributed to the persistent Human papillomavirus (HPV) infection, particularly the high-risk subtypes such as HPV16 and HPV18 [2,3]. HPV contains a circular double-stranded DNA genome of size ~ 8 kb, and it infects basal epithelial cells in the cervix. During the initial stages of infection, the HPV DNA exists in episomal form. However, over the course of epithelial cell differentiation, proliferation or neoplastic changes, the HPV DNA gets integrated into the host genome. This integration process is likely to occur at the host genomic regions sensitive to DNA strand breaks and share microhomology with the HPV DNA

Abbreviations
APOT, amplification of papillomavirus oncogene transcripts; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; eRNA, enhancer RNA; HPV, human papillomavirus; SE, super enhancer; TAD, topologically associating domain; TCGA, The Cancer Genome Atlas; WGS, whole genome sequencing; WXS, whole exome sequencing.

[4,5]. Large-scale genome study of cervical tumours has shown that 80% of the tumours with HPV had integration in the host genome [6].

Tumours with HPV DNA integration show overexpression of viral oncogenes such as E6 and E7, likely due to the perturbations or DNA breakpoints in the viral regulatory gene E2 (which controls the expression of E6/E7) or increased stability of the viral transcripts upon fusion with the host genes [3]. E6 and E7 proteins are known to interfere with the host p53 and RB pathways respectively, and thus favour cancer cell proliferation by avoiding apoptosis and cell cycle arrest [7,8]. Besides our understanding of the oncogenic roles of these viral proteins (E6/E7), efforts to delineate the cis-regulatory effects of HPV integration on the host chromatin structure and gene regulation have been rather limited.

HPV integration in the host genome can be a single or clustered (multiple nearby) event. The latter is often found together with genomic alterations (including amplification, deletion and translocations) in the nearby host regions, likely due to the HPV integration mediated DNA replication and recombination processes [9,10]. Besides, HPV integrations are associated with the upregulation of host genes which are either directly affected by the integration or in its immediate vicinity [6,11–13]. Furthermore, recent studies using cell lines showed that HPV integration can cause long-range effects in *cis* through changes in the host chromatin interactions and subsequent gene dysregulation. For example, HPV16 integration in the W12 (human cervical keratinocyte) cell line was shown to alter chromatin interactions (involving both host:host and host:viral DNA), as well as host gene expression in the nearby regions [14]. Similarly, in the cervical adenocarcinoma cell line HeLa, the HPV18 DNA integration in chromosome 8 was shown to have long-range chromatin interactions with the promoter region of the MYC oncogene (located approximately 500 kb away in the same chromosome) and was associated with its overexpression [15,16]. However, the extent of these long-range chromatin interactions mediated by HPV integrations genome-wide and the associated host gene expression changes are still unexplored in cervical tumours.

Previous studies have shown that the HPV integrations from the cervical tumours and cell lines were enriched in the transcriptionally active open-chromatin regions [17–19]. However, these findings were based on the HPV integrations derived mostly from transcription-based assays (such as RNA-seq and amplification of papillomavirus oncogene transcripts (APOT)), and thus probably have a bias towards transcriptionally active regions. Hence, a whole-genome DNA-based HPV integration detection approach is required to understand the distribution of HPV integrations across the genome and to study their impact on chromatin structure and gene-regulation. Moreover, a recent Pan Cancer Analysis of Whole Genomes consortium study has also demonstrated the need for DNA-based methods to obtain a comprehensive view of viral association with cancers [20].

To address the aforementioned limitations, we explored the genome-wide HPV integration patterns and their impact on host gene expression in the context of chromatin states and topologically associating domains (TADs). For this, we collated genome-wide HPV integrations in cervical cancers detected using DNA-based approaches and compared it with the chromatin state information from cancer and normal cell lines. We found that the HPV integrations are significantly enriched in active chromatin regions and depleted in inactive chromatin regions, as compared to the expected counts. Interestingly, regardless of the host chromatin state, transcriptional upregulation was observed in the immediate vicinity of the HPV integration regions (up to 10 kb). Further investigation of the long-range effects of HPV integration revealed that the TADs with integration have higher gene expression as compared to samples without integration in the same TAD. More importantly, this difference was not observed in the TADs adjacent to the HPV integrated TADs. Moreover, the recurrent HPV integration analysis at the TAD level revealed both the direct and long-range effect of HPV integration on the expression of cancer-related genes (such as MYC, PVT1, TP63 and ERBB2). Additionally, we used Hi-C and 4C-seq analyses to show that the HPV integration in HeLa cells mediates long-range chromatin interactions with the oncogenes MYC and PVT1, and drives their overexpression in an allele-specific manner. Interestingly, these chromatin interactions were also mostly confined to the same TAD and not extending to the neighbouring TADs. Together, our results suggest the cis-regulatory potential of integrated HPV DNA that drives upregulation of host genes through changes in chromatin interactions (but mostly within the same TAD) in cervical cancer. This underscores that the HPV integration can mediate multiple modes of oncogenic activation and thereby acts as a strong driver conferring selective growth advantage to the cancer cells.

2. Materials and methods

2.1. HPV integrations

We collated HPV integrations in cervical cancer patient samples from previous studies (including

TCGA-CESC and others) [4,11]. This dataset consists of HPV integration identified through genome-wide approaches (whole-genome sequencing [WGS] and HPV capture methods) and exome-wide approaches (whole-exome sequencing [WXS] and RNA-seq based integrations). In total, we got 1324 integrations from 326 samples (see Table S1), after removing five samples which had an extreme number of HPV integrations (above the 99th percentile in the respective methods). We categorised the HPV integrations into two main sets: (a) *GW-HPV-int*: containing 617 integrations from 212 samples identified through genome-wide approaches, and (b) *all-HPV-int*: containing 1324 integrations from 326 samples, which includes HPV integrations detected from genome-wide, RNA-seq and exome-based approaches (see Table S1). In the latter set, if in any samples, HPV integrations were identified from both WGS and RNA-seq, we merged the overlapping integrations to avoid redundancy. For the analysis shown in Figs 1 and 2, we used *GW-HPV-int* set whereas for the others (Figs 3 and 4; Fig. S3) we used *all-HPV-int* set. Only samples from TCGA-CESC were used whenever the expression was being plotted together with the HPV integration. The HPV integration (n = 381 from 95 samples) from small cell cervical cancer [21] was treated as an independent dataset to test the TAD based recurrent analysis (Fig. S4A).

2.2. Genomic annotations

ChromHMM regions (15 states) for both HeLa and NHEK cell lines were obtained from the Roadmap Consortium [22]. The full-stack ChromHMM annotations representing the universal genome annotations across cell types were obtained from Vu and Ernst [23]. We broadly classified these states into two groups: active state (TssA, TssAFlnk, TxFlnk, Tx, TxWk, Enh, EnhG, ZNF/Rpts) and inactive state (Het, TssBiv, BivFlnk, EnhBiv, ReprPC, ReprPCWk, Quies) according to the Roadmap Consortium [22].

The individual histone marks information (broad peaks) of HeLa and NHEK (used in Fig. 1E) was obtained from ENCODE (https://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/).

The genomic coordinates of non-B forms DNA conformations were obtained from http://nonb.abcc.ncifcrf.gov/.

Super-enhancer coordinates for HeLa were obtained from SEdb (http://www.licpathway.net/sedb) [24]. W12 super-enhancers were obtained from Warburton et al. [4].

HeLa and NHEK TAD coordinates were obtained from Rao et al. [25]. Further, the overlapping TADs were merged using bedtools mergeBed [26] to obtain non-overlapping TADs [27] (Table S2). The size of TADs in HeLa ranges from 130 to 6070 kb (median 290 kb) and TADs in NHEK ranges from 130 to 4240 kb (median 280 kb).

2.3. Enrichment analysis

For genome-wide enrichment analysis of HPV integration (shown in Fig. 1), the genomic coordinates of HPV integration sites were extended (500 bp flanking either sides) to a total length of 1 kb size from the centre (such that all HPV integrations have uniform size distribution and to compute the GC content around HPV integration sites). To compute the expected integrations, we randomly sampled an equal number of regions which were of the same length and similar GC content to the observed HPV integrations. For each feature, we calculated the number of observed HPV integrations overlapping with that and compared it against the expected HPV integration counts using the Chi-squared test. The P-values were subjected to the multiple-hypothesis testing correction using Benjamini–Hochberg method (FDR). For ChromHMM annotations, we considered the centre of HPV integrations (and random regions) for the overlap.

2.4. Expression analysis

To check the expression of genes and non-coding elements in the immediate vicinity (Fig. 2A,C; Fig. S2A,C), we downloaded the pre-computed normalised total RNA expression values (FPKM) in the 10 kb region around the HPV integration regions from Nguyen et al. [11]. These integration regions were defined by merging HPV integrations that were within 10 kb at the sample level (if any, to avoid overlapping biases), and then a 10 kb flanking region was added on both sides to compute the total RNA expression (by considering the normalised expression level of all transcripts that overlapping the genomic region) [11]. In the case of enhancer RNA (eRNA) expression, we followed the above steps, to compute the eRNA expression (from Super-enhancers) around the HPV integration regions. In case of TAD level expression analysis (unique TAD and sample combination), we computed the mean expression using all genes (with expression from TCGA-CESC) in the respective TADs. The eRNA expression from super-enhancer regions and gene expression of TCGA-CESC samples were obtained from the TCeA database (https://bioinformatics.mdanderson.org/public-software/tcea/) and gdc portal (https://portal.gdc.cancer.gov/), respectively.

For the gene level expression comparison with respect to HPV integration status (Fig. 4), we used normalised RSEM values. The expression level represented as *z*-score in Fig. S5 was calculated by using the mean and standard deviation from all the samples with and without integrations (at gene level). For the pathway level enrichment analysis, we computed the single sample GSVA score [28] for each of the 50 hallmark gene sets from MSigDB (http://www.gsea-msigdb.org/gsea/msigdb/). Mann–Whitney *U* test (one-sided) was used to compare the distribution of GSVA scores in samples with and without integration for each pathway. The *P*-values were subjected for multiple-hypothesis testing using Benjamini–Hochberg approach and those pathways significant under FDR < 5% were shown in Fig. S4B,C.

2.5. Allele-specific analysis

For allele-specific expression/tf-binding analysis GATK ASEReadCounter (v4.1.9.0) [29] was used. Only those reads with minimum base quality of 10 and minimum read mapping quality of 20 were used. Also minimum read depth of 8 reads (5 for GRO-seq and TF ChIP-seq datasets) at each heterozygous SNPs was used as a cutoff. Only those features which were supported by at least 15 reads at the heterozygous SNP positions in total were further used. For copy number correction, we used read depth from WGS. Binomial test followed by Bonferroni correction was used to detect the significance of allele specific expression/tf-binding with respect to copy number at each feature level from haplotype A.

For the haplotype-specific binding analysis of RNA polymerase 2 (Pol2) in HeLa, the bam files were obtained from ENCODE (encodeDCC/wgEncodeSydhTfbs/). HeLa RNA-seq data and haplotype specific heterozygous SNP positions from HeLa genome [16] were obtained from the dbGaP repository phs000640.v1.p1. The GRO-seq data for HeLa was obtained from GSE63872 [30].

2.6. 4C-seq experiment and data analysis

The 4C-seq experiment was done following the protocol mentioned in Farooq et al. [31]. The primary digestion was performed with DpnII (NEB) and the secondary digestion was done with NlaIII enzyme (NEB). The primer used for the HPV18 viewpoint in HeLa and the number of reads in each replicate is mentioned in Table S3. 4C-ker package [32] was used to analyse the 4C-seq data and the hg19 genome was used as reference.

2.7. Hi-C analysis

Hi-C data from HeLa was obtained from ENCODE (https://www.encodeproject.org/experiments/ENCSR693GXU/). A hybrid hg19-HPV18 genome, considering the HPV18 genome as an additional 'chromosome', was constructed (human genome version: GRCh37.75, HPV18 version: NC_001357.1; the HPV18 genome orientation was reversed, as shown in Adey et al. [16]). The hybrid hg19-HPV18 genome was indexed with BWA v0.7.17 'index' mode and SAMTOOLS v1.6 'faidx'. Hi-C reads from each of the replicates were mapped separately to the hybrid genome using BWA (v0.7.17, 'mem' mode, parameters: -t 20 -E 50 -L 0 -v 0). Filtering of reads was done based on mapping parameters (min mapq = 1, samtools view -Sb -q 1 -F 256). After intersecting reads with HINDIII intervals and removing self ligation and duplicate pairs, replicates were pooled together. HICEXPLORER (v2.1.1) was used to obtain the contact list and the contact matrices at 10 kb resolution. In order to jointly normalise HPV18 and human chr8 Hi-C contacts, the interaction profile involving HPV18 and each 10 kb locus of chr8 were inserted in the integration site discovered by Adey et al. [16] (approximately chr8:128230000–128240000 in hg19 coordinates). Practically, this consisted in replacing, in the intra-chr8 Hi-C contact matrix, the intra-chr8 contact row (and column) of that 10 kb locus with the HPV18-chr8 contact profile. The matrix thus constructed was then normalised with the algorithm Iterative Correction of Hi-C data (ICE) [33] implemented in the function 'normICE' of the R package 'HiTC' [34]. Finally, the distance-matched interaction analyses were performed in the following way: (a) we took all within-TAD3189 HPV18-chr8 normalised interaction values and the genomic distances between loci; (b) we extracted all HPV18-chr8 normalised interaction values at the same genomic distances in (a) but occurring outside TAD3189; (c) for each of the distances in (a), we randomly sampled 1000 chr8-chr8 pairs of loci and extracted their normalised interaction values.

2.8. Promoter capture Hi-C analysis

Raw sequencing data from promoter capture Hi-C in HeLa (upon Cohesin/CTCF depletion) was downloaded from GEO (accession code: GSE145736) [35]. FASTQ files were processed as described in Section 2.7. Contacts between HPV18 and MYC/PVT1 were extracted by subsetting paired-end reads in which one mate fell in HPV18 and the other mate fell in the promoter regions of MYC and PVT1 (namely, in their transcription start sites ±10 kb). Fisher's Exact tests (implemented with R base function fisher.test) were performed between the number of HPV18-chr8 contacts within and outside X promoter in condition of Y control or Y depleted, with X = (MYC, PVT1) and Y = (SCC1, CTCF), for a total of four tests. The processed data of transcriptional response (SLAM-seq) in SCC1 depleted versus control was downloaded from Thiecke et al. [35].

3. Results

3.1. HPV integrations are enriched in active chromatin regions and depleted in inactive chromatin regions

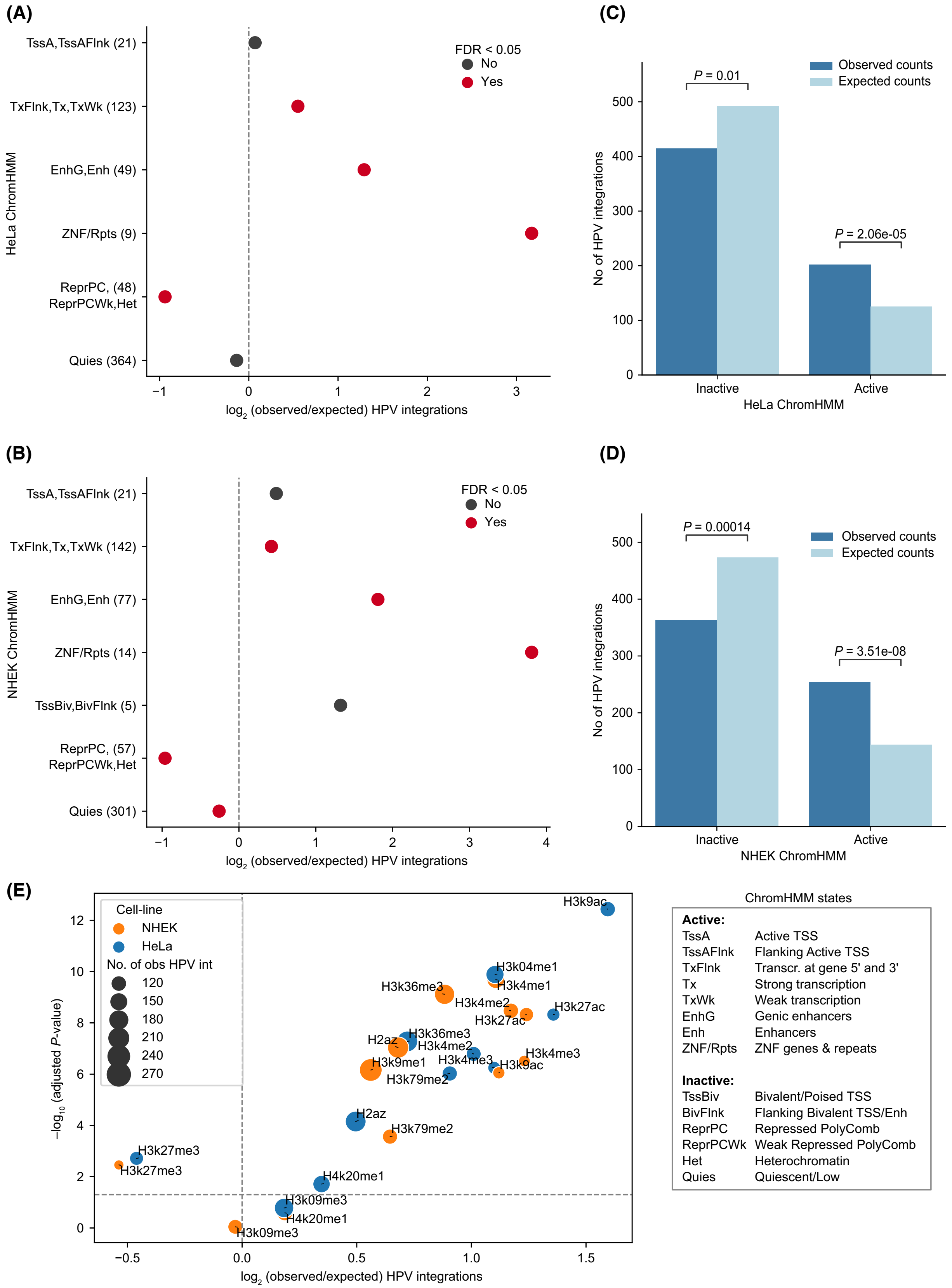
At first, we asked if the HPV integrations ($n = 617$ from 212 samples) detected from the genome-wide DNA-based approaches are distributed randomly or enriched in specific functional regions of the host genome. For this, we compared the HPV integration sites with ChromHMM annotations [22,36], which categorise the genome into broad functional annotations based on various histone modification profiles, from two closest cell lines available: HeLa (cervical adenocarcinoma) and NHEK (normal human epidermal keratinocytes); and with full-stack ChromHMM [23] representing a universal genome annotation unified from multiple cell-types. To check for the enrichment of integrations, we compared the number of observed HPV integrations overlapping with each of these annotations with the expected counts computed using random sites of equal size and similar GC content (see Section 2).

With both HeLa and NHEK ChromHMM annotations, we observed that compared to the expected, HPV integrations were significantly enriched (Chi-squared test, FDR < 0.05) in the transcriptionally active regions (TxFlnk, Tx, TxWk), enhancers (EnhG, Enh), and zinc finger protein gene/repeat regions (ZNF/Rpts); whereas a significant depletion (FDR < 0.05) was observed in polycomb repressed/heterochromatin regions (ReprPC, ReprPCWk, Het), and quiescent regions (Quies; in NHEK but not in HeLa) (Fig. 1A,B). We further merged all the ChromHMM annotations into two major categories – active and inactive – based on the gene-regulation/chromatin activity of the regions [22] (see Section 2). In both HeLa (Fig. 1C) and NHEK (Fig. 1D), HPV integrations were significantly enriched in active regions (one-sample Chi-squared test, $P < 0.0001$) and significantly depleted in inactive regions ($P < 0.05$) as compared to the expected counts. Similar results were obtained when we used universal genome annotation (full-stack ChromHMM, Fig. S1A,B), which provides a cell-type agnostic view of chromatin states. Together, this suggests that HPV integrations observed in the tumours are preferentially enriched in active chromatin regions.

Further, we checked specifically which histone modification marks associated with the above annotations are enriched with HPV integrations. We observed that the HPV integrations were significantly enriched (Chi-squared test, FDR < 0.05) in various active histone modification regions (such as H3K4me1, H3K4me2, H3K4me3, and H3K27ac) in both HeLa and NHEK (Fig. 1E). In contrast, HPV integrations were significantly depleted (FDR < 0.05) only in the repressive histone modification regions (H3K27me3) in both HeLa and NHEK (Fig. 1E). We further extended this analysis to other cervical cancer cell lines (SiHa, Ca Ski, S12, C-33 A), for which active histone modification marks (H3K27ac) were available [37] (Fig. S1C). Despite some of these cervical cancer cell lines being HPV positive (HeLa – HPV18; SiHa, Ca Ski, S12 – HPV16) and HPV negative (C-33 A), the results obtained were consistent with the above observation that HPV integrations are enriched in active chromatin regions. Together, this suggests that the host chromatin structure influences the HPV integration patterns regardless of malignancy.

Human papillomavirus integration has previously been shown to be enriched in regions of fragile sites [4], which contain DNA repeats that could form non-B DNA conformations and are associated with genomic instability. To check this further, we computed the enrichment of HPV integrations in the DNA regions predicted to form non-B DNA conformations (see Section 2). Among all the non-B forms of DNA, only direct repeats showed significant enrichment of HPV integration as compared to the expected (Chi-squared test, FDR < 0.05) (Fig. S1D). Further, to check this in the context of chromatin states, we calculated the odds ratio of observed versus expected HPV integrations in active versus inactive regions (for HeLa and NHEK, separately). This showed that even in the non-B DNA regions, integration tends to occur more frequently in the active regions as compared to inactive regions (odds ratio > 1 and Fisher exact test, FDR < 0.05), except for A phased repeats (with both HeLa and NHEK ChromHMM) and G quadruplex regions (only with HeLa ChromHMM) (Fig. S1E,F).

Taken together, these results suggest that HPV integrations are not randomly distributed across the genome. They are highly enriched in the active chromatin regions and depleted in repressed chromatin regions. This could be due to the combined or individual effect of DNA sequence context, DNA



Active:

TssA	Active TSS
TssAFlnk	Flanking Active TSS
TxFlnk	Transcr. at gene 5' and 3'
Tx	Strong transcription
TxWk	Weak transcription
EnhG	Genic enhancers
Enh	Enhancers
ZNF/Rpts	ZNF genes & repeats

Inactive:

TssBiv	Bivalent/Poised TSS
BivFlnk	Flanking Bivalent TSS/Enh
ReprPC	Repressed PolyComb
ReprPCWk	Weak Repressed PolyComb
Het	Heterochromatin
Quies	Quiescent/Low

Fig. 1. Enrichment of HPV integrations in functionally annotated regions. (A) Enrichment of HPV integrations in the ChromHMM annotated regions from the HeLa cell line. The *x*-axis represents the $\log_2$ of observed/expected number of HPV integrations. The *y*-axis represents the different annotations and the observed number of HPV integrations overlapping them (given in the bracket). The *P*-values were computed using Chi-squared goodness-of-fit test followed by FDR correction. The colour of the dots indicates whether the adjusted *P*-value is below the significance level of 5% or not. (B) Same as (A) but with ChromHMM annotations from the NHEK cell line. (C, D) Bar plot showing the frequency of observed and expected integrations in the active and inactive regions defined by combining ChromHMM annotations in HeLa (C) and NHEK (D) (see Section 2). The *P*-value was calculated using a one-sample Chi-squared test. (E) Enrichment of HPV integration in various histone modification regions from HeLa and NHEK. The *x*-axis represents the $\log_2$ of observed/expected number of HPV integrations and the *y*-axis represents the negative $\log_{10}$ of adjusted *P*-value (Chi-squared test followed by FDR correction). The horizontal dashed line represents FDR cut-off of 5%. The colour and size of the dots represent the cell line and the number of observed HPV integrations for each of the histone marks, respectively.

accessibility, DNA damage response (linked to the chromatin states) and positive selection.

3.2. HPV integration affects host transcriptional activity regardless of the host chromatin state

Previous studies have shown that HPV integration can affect host transcriptional activity in its immediate vicinity [11,13]; however, how this is influenced by the host cell-type specific chromatin states has not been studied systematically yet. To check this, we used TCGA cervical cancer (CESC) samples (only WGS with matched gene expression data available, $n = 50$) with 151 HPV integration regions and plotted the host endogenous normalised total RNA expression from all transcript types that overlap with the 10 kb flanking regions around the integration [11] (see Section 2). We observed that the samples with HPV integration showed increased expression in their neighbouring 10 kb regions as compared to mean normalised expression from the samples without HPV integration in the same region, regardless of whether the HPV integration was in an active (Mann–Whitney *U* test, $P = 0.0048$ for HeLa and $P = 7.28\text{e-}05$ for NHEK) or inactive ($P = 0.0065$ for HeLa and $P = 0.057$ for NHEK) region (Fig. 2A; Fig. S2A). Following that, we asked whether the activity of host regulatory elements (like enhancers) are also enhanced around the HPV integration regions. To test this, we looked at the level of endogenous eRNA (indicative of the functional activity of enhancers [38]) transcribed from the annotated super enhancer (SE) regions, if any, within 10 kb flanking regions around the integration in TCGA-CESC samples (see Section 2). We observed an increased SE eRNA expression in the HPV integrated samples compared to the samples without integrations (Fig. 2B; Fig. S2B). Again as shown above, it was not influenced by the chromatin state of the integrated region.

We additionally noticed that the control samples (without HPV integration) showed significantly higher expression in active as compared to inactive regions ($P = 0.00019$ for HeLa, Fig. 2A and $P = 0.00031$ for NHEK, Fig. S2A), as expected. This underscores that the ChromHMM annotations from cell lines (HeLa/NHEK) match well with the tumour tissues in terms of transcriptional activity observed in the active and inactive regions. More importantly, in samples with HPV integration, we observed a higher expression in the active as compared to the inactive regions ($P = 0.055$ for HeLa, Fig. 2A and $P = 0.0061$ for NHEK, Fig. S2A). This indicates that the HPV integration in active chromatin regions further enhances the host transcription activity in its vicinity. Further, we asked whether all HPV integrations in a sample could lead to overexpression in its immediate vicinity or not. For this, we computed the expression fold change for each HPV integration region (ratio of expression in the 10 kb flanks around integration regions with the mean expression from other samples without integration in the same region) [11]. This showed that the expression association with HPV integration was highly variable and not all HPV integration regions associated with higher transcriptional activity in its vicinity (Fig. 2C; Fig. S2C). This could be due to the epigenetic suppression of certain integrated HPV regions or impaired regulatory activity [39].

Taken together, these results suggest that the HPV integration leads to transcriptional upregulation and enhanced enhancer activity in its immediate vicinity, regardless of the host chromatin state. Nevertheless, the transcriptional activity associated with HPV integration was relatively higher in active chromatin regions, as compared to inactive regions.

3.3. Transcriptional activity associated with HPV integration is mostly confined to the same TAD

We next asked whether the HPV integration can mediate or alter the host long-range chromatin interactions (such as enhancer-promoter or promoter-promoter), and thereby dysregulate gene expression in tumours.

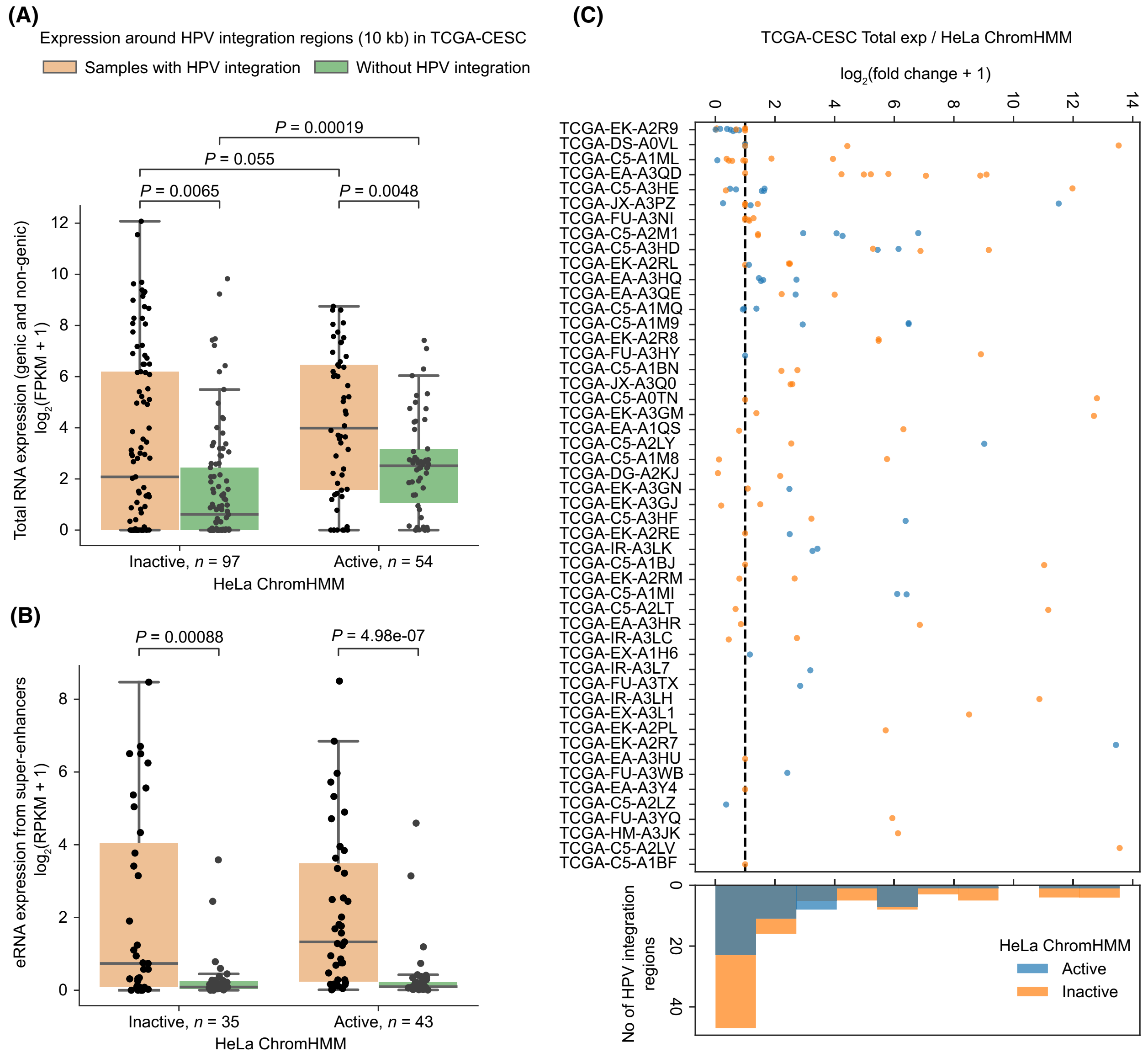


Fig. 2. Enhanced transcriptional activity near HPV integration in the context of chromatin states from HeLa. (A) Boxplot showing the total expression in the 10 kb flanking region around the HPV integration regions as compared to mean expression from TCGA-CESC samples without HPV integration in the same region. In the boxplot, the horizontal middle line indicates the median, the height of the shaded box indicates the interquartile range (IQR) and the whiskers indicate 1.5 × IQR. The *x*-axis represents whether the HPV integration is located in an inactive ($n = 97$) or active ($n = 54$) chromatin regions with respect to HeLa ChromHMM. The *P*-values were computed using Mann–Whitney *U* test (two-sided). (B) Same as (A) but for the eRNA expression from SEs (if any) within 10 kb on either side of the HPV integration regions located in inactive ($n = 35$) or active ($n = 43$) chromatin regions. (C) Expression fold change associated with each of the HPV integration regions. The *x*-axis represents the $\log_2$ fold change, which was calculated as the total expression in the 10 kb flanking region around HPV integration regions divided by the mean expression from other samples without HPV integration in the same genomic region. The *y*-axis represents the individual sample-id of TCGA-CESC samples. The colour of the dots indicates if the integration overlaps an active or inactive ChromHMM region of HeLa. The black vertical line represents the value of $\log_2(\text{fc} + 1) = 1$. The histogram at the bottom shows the frequency of integration regions at different fold-change bins. Each dot in (A–C) represents a HPV integration region from a sample.

To test this, we compared the host gene expression at the level of TADs, obtained from HeLa and NHEK cell lines [25]. TADs act as functional units of genome organisation by restricting the interactions between regulatory elements and thereby controlling gene regulation [25,40]. TAD boundaries are commonly bound

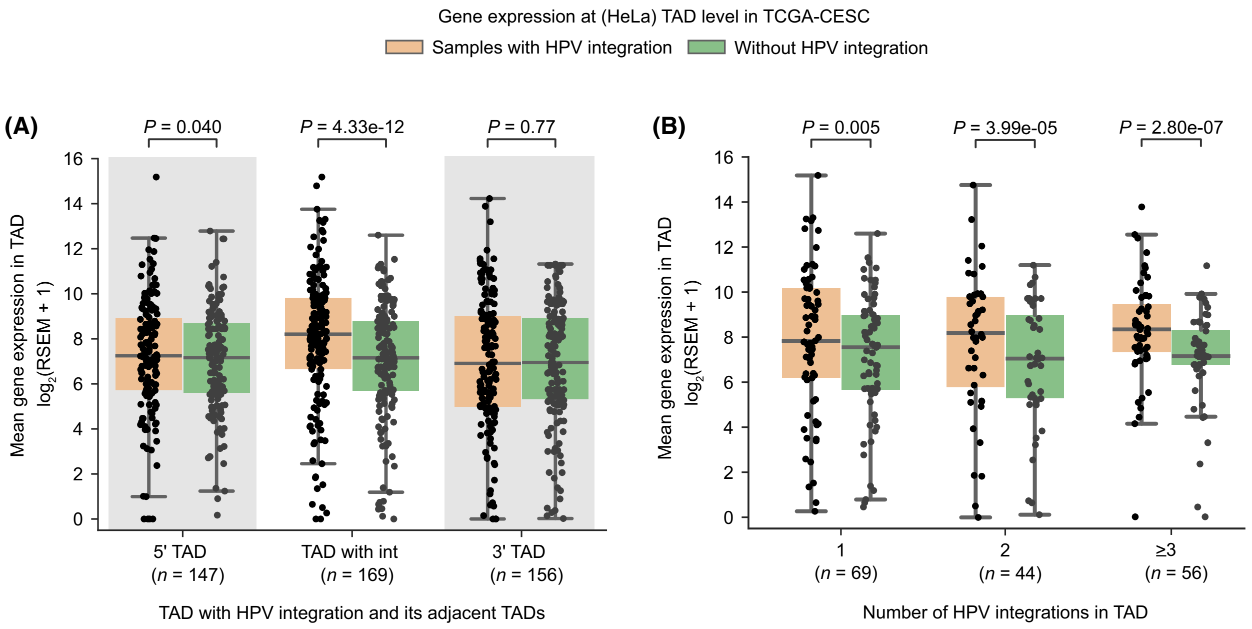


Fig. 3. HPV integration associated host gene overexpression with respect to HeLa TAD domains. (A) TAD level gene expression in the TCGA-CESC samples with HPV integration compared to the mean expression from the samples without HPV integration in the same TADs (*n* = 169), also for the neighbouring upstream (5′, *n* = 147) and downstream (3′, *n* = 156) TADs. (B) TAD level gene expression in the TCGA-CESC samples with HPV integration, separated by whether the TAD had one (*n* = 69), two (*n* = 44) or more than two (*n* = 56) integrations compared to the mean expression from the samples without HPV integration in the same TADs. The TAD information was obtained from the HeLa cell line for (A, B). Each dot in (A) and (B) represents a unique HeLa TAD-tumour sample combination. In each boxplot, the horizontal middle line indicates the median, the height of the shaded box indicates the interquartile range (IQR) and the whiskers indicate 1.5 × IQR. The *P*-values shown at the top were computed using the Wilcoxon signed-rank test (two-sided).

by insulator proteins like CTCF that prevent interactions across the TADs. We hypothesised that the transcriptional overexpression associated with the HPV integration would be majorly restricted to the TADs where the HPV integration is localised. To test this, we plotted the average gene expression at the TAD level in the TCGA-CESC samples with HPV integration and compared it with the mean expression from samples without integration in the same TAD (see Section 2). We further extended this analysis to the immediate upstream (5′) and downstream (3′) TADs for comparison. Only the TADs with HPV integration showed overall increased expression compared to the samples without integration (Wilcoxon signed-rank test, $P < 0.0001$) (Fig. 3A; Fig. S3A). However, neither the upstream nor the downstream neighbouring TADs showed any effect at TAD level expression of genes ($P > 0.01$) (Fig. 3A; Fig. S3A).

Further, we asked whether increase in the number of integrations in a TAD would result in more perturbations in gene regulation and overexpression. For this, we separated TADs based on whether they had 1, 2, or more than 2 integrations and compared the TAD level gene expression among them, along with the samples without integration. This showed that indeed an increase in the number of integrations in a TAD was associated with an increase in gene overexpression as compared to the samples without integration in the same TADs (Fig. 3B; Fig. S3B).

Direct HPV integration in the host genes (or fusions) can lead to overexpression of the target genes. To remove the influence of these, we repeated the above analysis after removing the genes within the 10 kb regions flanking the integration sites (Fig. S3C,D). Nevertheless, we found that the gene expression was higher in the TADs with integration as compared to samples without integration, albeit lower than the above (Fig. 3A; Fig. S3A), suggesting the long-range cis-regulatory potential of the HPV integration.

Taken together, these results indicate the potential of HPV integration to enhance expression of nearby host genes but mostly within the same TAD, likely due to the constraint imposed by the TAD boundaries or genome organisation. Further, within the TAD, the overexpression observed could come from both the direct and long-range effect of HPV integration on target genes through chromatin contacts in 3D nuclear space.

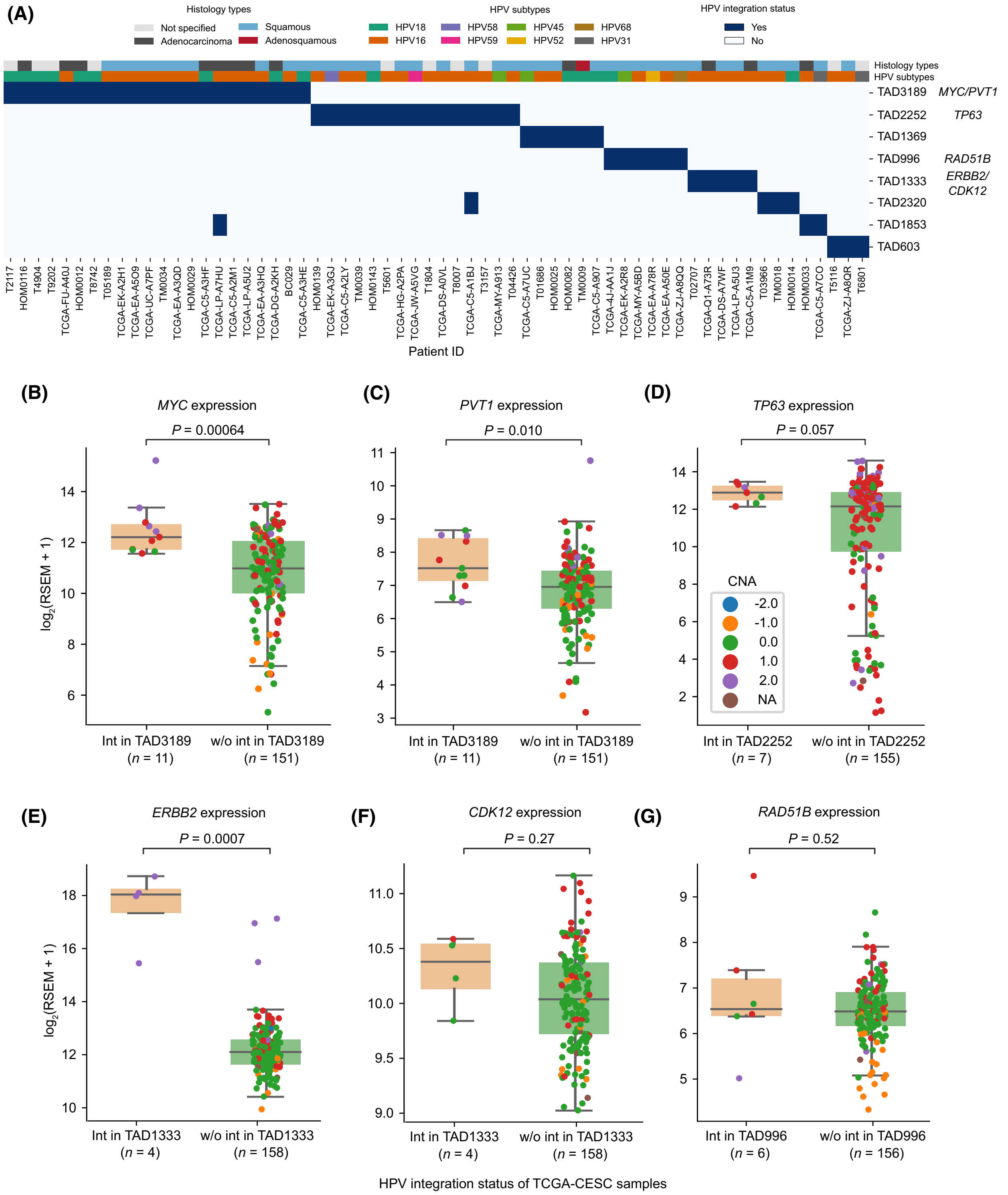

(A)
Histology types
Not specified
Adenocarcinoma
Squamous
Adenosquamous
HPV subtypes
HPV18
HPV16
HPV58
HPV59
HPV45
HPV52
HPV68
HPV31
HPV integration status
Yes
No
Histology types
HPV subtypes
TAD3189 MYC/PVT1
TAD2252 TP63
TAD1369
TAD996 RAD51B
TAD1333 ERBB2/ CDK12
TAD2320
TAD1853
TAD603
T2117
HOM0116
T4904
T9202
TCGA-FU-A40J
HOM0012
T8742
T05189
TCGA-EK-A2H1
TCGA-EA-A5O9
TCGA-UC-A7PF
TM0034
TCGA-EA-A3QD
HOM0029
TCGA-C5-A3HF
TCGA-LP-A7HU
TCGA-C5-A2M1
TCGA-LP-A5U2
TCGA-EA-A3HQ
TCGA-DG-A2KH
BC029
TCGA-C5-A3HE
HOM0139
TCGA-EK-A3GJ
TCGA-C5-A2LY
TM0039
HOM0143
T5601
TCGA-HG-A2PA
TCGA-JW-A5VG
T1804
TCGA-DS-A0VL
T8007
TCGA-C5-A1BJ
T3157
TCGA-MY-A913
T04426
TCGA-C5-A7UC
T01686
HOM0025
HOM0082
TM0009
TCGA-C5-A907
TCGA-4J-AA1J
TCGA-EK-A2R8
TCGA-MY-A5BD
TCGA-EA-A78R
TCGA-EA-A50E
TCGA-ZJ-A8QQ
T02707
TCGA-Q1-A73R
TCGA-DS-A7WF
TCGA-LP-A5U3
TCGA-C5-A1M9
T03966
TM0018
HOM0014
HOM0033
TCGA-C5-A7CO
T5116
TCGA-ZJ-A8QR
T6801
Patient ID
(B)
MYC expression
P = 0.00064
log2(RSEM + 1)
6
8
10
12
14
Int in TAD3189 (n = 11)
w/o int in TAD3189 (n = 151)
(C)
PVT1 expression
P = 0.010
3
4
5
6
7
8
9
10
11
Int in TAD3189 (n = 11)
w/o int in TAD3189 (n = 151)
(D)
TP63 expression
P = 0.057
2
4
6
8
10
12
14
CNA
-2.0
-1.0
0.0
1.0
2.0
NA
Int in TAD2252 (n = 7)
w/o int in TAD2252 (n = 155)
(E)
ERBB2 expression
P = 0.0007
log2(RSEM + 1)
10
12
14
16
18
Int in TAD1333 (n = 4)
w/o int in TAD1333 (n = 158)
(F)
CDK12 expression
P = 0.27
9.0
9.5
10.0
10.5
11.0
Int in TAD1333 (n = 4)
w/o int in TAD1333 (n = 158)
(G)
RAD51B expression
P = 0.52
5
6
7
8
9
Int in TAD996 (n = 6)
w/o int in TAD996 (n = 156)
HPV integration status of TCGA-CESC samples

Fig. 4. Recurrently integrated TADs and expression alteration of associated cancer genes. (A) Heatmap shows the HeLa TADs with recurrent HPV integrations. The *x*-axis represents the sample-id and *y*-axis represents the TADs (denoted with distinct numbers to differentiate each TAD domain). Blue box indicates HPV integration(s) in a particular TAD, and in a particular sample. The top two rows represent the tumour histology and HPV subtype in each of the samples, respectively. (B, C) Gene expression of MYC (B) and PVT1 (C) in TCGA-CESC samples with (n = 11) and without (n = 151) HPV integration in TAD3189. (D) Gene expression of TP63 in TCGA-CESC samples with (n = 7) and without (n = 155) HPV integration in TAD2252. (E, F) Gene expression of ERBB2 (E) and CDK12 (F) in TCGA-CESC samples with (n = 4) and without (n = 158) HPV integration in TAD1333. (G) Gene expression of RAD51B in TCGA-CESC samples with (n = 6) and without (n = 156) HPV integration in TAD996. The *P*-values shown in panels (B–G) were calculated using the Wilcoxon rank-sum test (two-sided). The colour of each dot in the box plot (B–G) represents the relative copy number status of the gene in the respective TCGA-CESC samples (−2 deep deletion, −1 deletion, 0 copy neutral, 1 amplification, 2 high amplification). In each boxplot, the horizontal middle line indicates the median, the height of the shaded box indicates the interquartile range (IQR) and the whiskers indicate 1.5 × IQR.

3.4. Recurrent HPV integrations in TADs are associated with oncogene overexpression

Recurrent HPV integrations near cancer-related genes (such as MYC and TP63) have been previously reported in cervical cancers [4,5]. In those studies, the recurrence was defined mostly based on the HPV integration directly overlapping with or in close proximity (at a defined distance cut-off) of the target genes. This can potentially miss out the integrations that are far away and can have a long-range effect on the target genes. To overcome this, we performed recurrent integration analysis at the TAD level. For this analysis, we considered HPV integrations (n = 1324 from 326 samples) collated from both DNA-based studies (WGS, WXS and Hybrid capture) and RNA-seq (see Section 2). The TADs which had HPV integration(s) in at least three samples were considered recurrent (Fig. 4A). This resulted in eight recurrent TADs having integrations from 62 samples. TAD3189 (with 22 samples) and TAD2252 (with 15 samples) were the top two most recurrently integrated TADs. Interestingly, these TADs also exhibited mutually exclusive patterns of HPV integration, indicating that these recurrent TADs were not affected in the same patients. Moreover, we found that certain HPV subtypes were frequently integrated in these TADs (Fig. 4A). HPV16 was predominantly observed in TAD3189 (13/22), TAD2252 (11/15), TAD1333 (5/5) and TAD2320 (3/4), whereas HPV18 was predominant in TAD1369 (3/6). In the TAD3189, we observed both HPV16 and HPV18; however, the proportion of HPV18 was much higher (41%) in TAD3189 as compared to the overall HPV18 positive samples (16%) in TCGA-CESC cohort (Chi-squared test, P = 0.011). This is likely due to the preferential integration sites for different HPV strains [41]. Analysis of an independent dataset of small cell cervical carcinoma [21] also revealed TAD3189 to have the most recurrent HPV integration (with HPV18) among other TADs (Fig. S4A).

To further understand if the recurrent HPV integrations are associated with tumorigenesis, we looked at the presence of cancer-related genes (from Cancer Gene Census [42] and Cancer LncRNA Census [43]) in these TADs. Interestingly, TAD3189 harbours multiple important coding (MYC) and non-coding (PVT1 and CCAT1) oncogenes. Overexpression of these oncogenes have been previously reported in cervical cancers and were also associated with poor prognosis [44–46]. Thus, to test if HPV integration in TAD3189 is affecting the expression of these oncogenes, we divided the TCGA-CESC samples into two groups, one which had integration in TAD3189 and other which did not. Expression of both MYC (Wilcoxon rank-sum test, P = 0.0006) (Fig. 4B) and PVT1 (P = 0.010) (Fig. 4C) was significantly higher in the samples with integration in TAD3189 as compared to samples without integration in TAD3189. Similarly in TAD2252, TP63 (oncogene) expression was higher (P = 0.057) (Fig. 4D) in samples with integration in TAD2252. In TAD1333, ERBB2 (oncogene) expression was significantly higher (P = 0.0007) in samples with integration, whereas CDK12 (tumour suppressor gene) in the same TAD did not show any change in expression (P = 0.27) (Fig. 4E,F). Also, RAD51B (tumour suppressor gene) in TAD996 did not show any change in expression level (P = 0.52) (Fig. 4G). This suggests that the HPV integration preferentially enhances the expression of oncogenes in these recurrent TADs. Interestingly, in these TADs (TAD3189 and TAD2252) with oncogenes, increased gene expression (z-score > 0; above the average expression level across samples) was evident regardless of whether the HPV integration occurred directly at the gene or further away in the respective TADs (Fig. S5), suggesting that the HPV integration can affect gene expression both locally and at longer distances. Further, we checked the effect of these recurrent TADs with integration on the host gene expression at the pathway level (using MSigDB

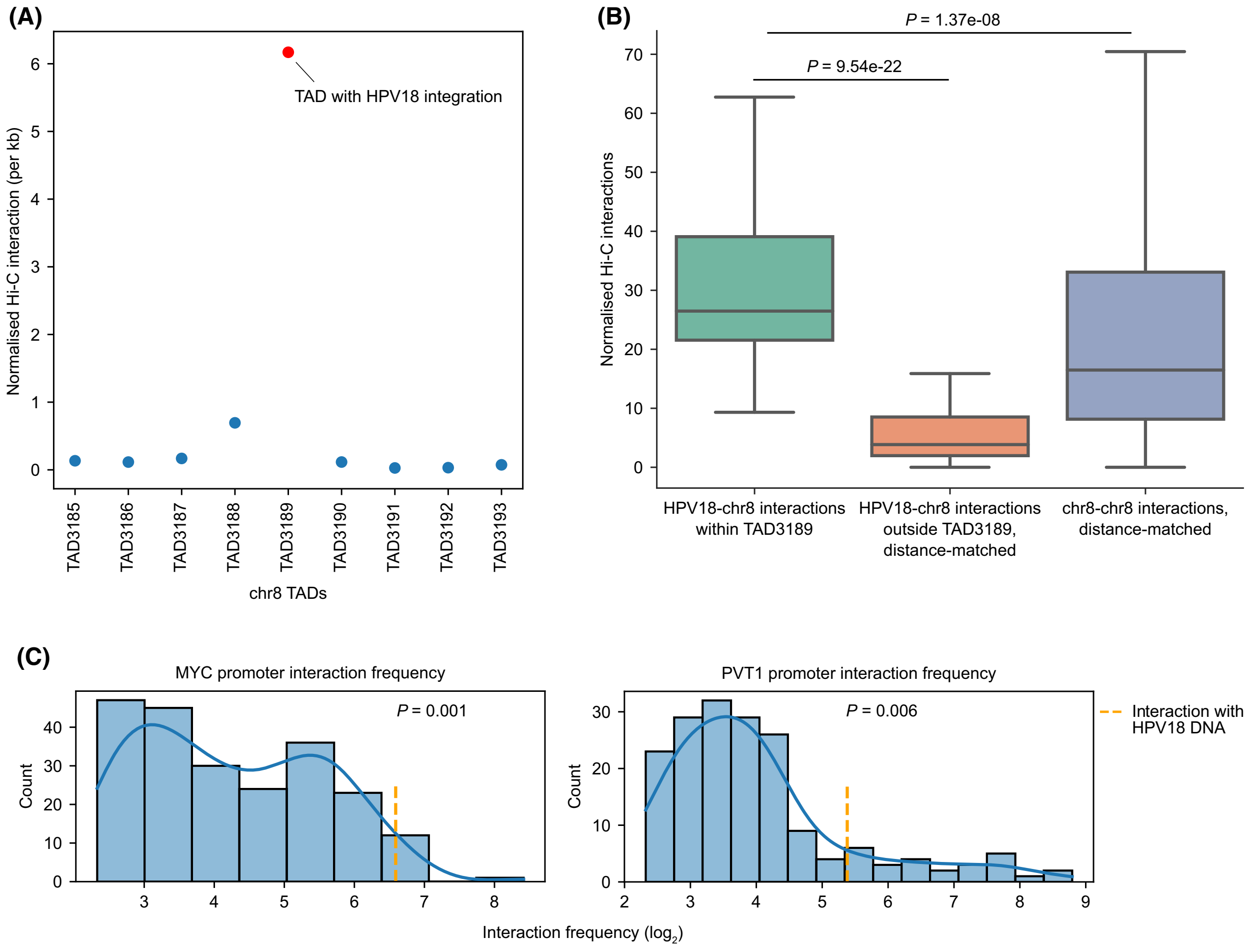


Fig. 5. Hi-C analysis reveals highly localised chromatin interactions between integrated HPV18 DNA and host genome in HeLa. (A) Scatterplot shows the interaction frequency per kb between TADs on chromosome 8 and the HPV18 genome. All the TADs are arranged in a linear manner (5′ to 3′ direction of the genome). TAD3189 contains the HPV integration (red dot) which showed highest interaction as compared to the adjacent TADs (and also at the chromosome level, Fig. S6B). (B) Boxplot shows the interaction frequency between HPV18 and TAD3189 regions (n = 74), distance-matched interactions between HPV18 and regions outside of TAD3189 (n = 74), and between genomic regions within chromosome 8 (n = 1000, random sampled). The P-values were computed using Mann–Whitney U test (two-sided). In the boxplot, the horizontal middle line indicates the median, the height of the shaded box indicates the interquartile range (IQR) and the whiskers indicate 1.5 × IQR. (C) Histogram distribution plot shows the interaction frequency of the bins which overlap the promoters of MYC and PVT1 with all the interacting bins on chromosome 8 and the integrated HPV18 (orange line). Only those interacting bins which were supported by more than 5 reads were shown in the plot. Empirical P-value indicates the likelihood of finding interaction frequency greater than the observed between HPV18-gene promoters in the distance-matched random sampling (n = 1000) of chromatin interactions between genomic regions within chromosome 8.

hallmark gene sets), and it showed that samples with integration in TAD3189 (with MYC) have upregulation of MYC target genes, whereas TAD2252 (with TP63) have upregulation of interferon (alpha and gamma) signalling, as compared to samples without integration (Fig. S4B,C). This indicates that the upregulation of oncogenes observed within these TADs indeed causes downstream transcriptional changes at the pathway level and also exhibits different pathological responses.

3.5. HPV18-induced chromatin interactions in HeLa are confined locally and target oncogenes

Next, we asked whether the long-distance effect of HPV integration on the oncogene expression in the above recurrent TADs could be mediated by chromatin looping. To study this, we chose the HeLa cell line (as a model) which has HPV18 integration in TAD3189 (chromosome 8), the most frequently HPV integrated TAD in cervical cancers (Fig. 4A). To obtain an unbiased and global overview of all the genomic interactions between the integrated HPV DNA and the host genome, we leveraged the available Hi-C data from HeLa [47], which captures overall genomic interaction at a particular resolution (or regions that are in close proximity in 3D space). For the Hi-C data analysis, we constructed a hybrid human-HPV18 genome, and mapped reads to this to compute the contact frequency (see Section 2).

First, we looked at the chromosome level to identify which of the chromosomes have contacts with the integrated HPV18 DNA. This showed that the majority of the contacts involving HPV18 were associated with chromosome 8 (chromosome with HPV18 integration), followed by intra-HPV18 contacts (Fig. S6A). Second, to gain further insights into these interactions within chromosome 8, we plotted the normalised contact frequency (see Section 2) between the HPV18 DNA and host genome for all the TADs on chromosome 8. This showed that the highest interaction frequency with the integrated HPV18 DNA were observed with genomic regions in TAD3189, where the HPV integration localised (Fig. 5A; Fig. S6B,C). Further, to check if this is expected due to the proximity of HPV integration in the linear genome, we compared the normalised interaction frequency between HPV18 and TAD3189 with distance-matched randomly sampled interactions from genomic regions within chromosome 8. This showed that the HPV18-TAD3189 interactions were significantly higher (Mann–Whitney U test, P = 1.37e-08) as compared to random interactions of similar distance (Fig. 5B). Further, to check if the TAD boundary constrains the interaction of HPV18 to be mostly within the TAD3189, we compared the normalised interaction frequency between HPV18 and TAD3189 with the interactions between HPV18 and regions outside of TAD3189 but within chromosome 8 at similar distances. Again, this showed that the HPV18-TAD3189 interactions were significantly higher (Mann–Whitney U test, P = 9.54e-22), suggesting that the TAD boundaries limit the interaction of HPV18 DNA within TAD3189 (Fig. 5B). Together, these results demonstrate that the chromatin interactions involving the integrated HPV18 are mostly confined to the same TAD containing the HPV18 integration, thus possibly supporting the TAD level upregulation of genes observed in tumours (see Fig. 3A; Fig. S3A).

We wanted to further understand the specific host regions within TAD3189 which looped with the integrated HPV18 with a high frequency. TAD3189 has two oncogenes MYC and PVT1 which are ~ 500 kb far away from the integration. We focused on the promoter region of these two oncogenes and observed that their interaction with HPV18 (among other regions on chromosome 8) was in the top 1 percentile for both of them (Fig. 5C). Also, the normalised interaction frequency was significantly higher between HPV18 and promoters of MYC (P = 0.001) and PVT1 (P = 0.006) as compared to the distance-matched randomly sampled interactions within chromosome 8. This indicates a potentially strong chromatin interaction between these oncogenes promoter with the integrated HPV18 DNA in HeLa.

3.6. 4C-seq reveals haplotype-specific chromatin looping between integrated HPV18 and host genome

To further characterise the chromatin looping interaction mediated by the HPV18 integration at the local scale with higher resolution, we performed a 4C-seq experiment with the integrated HPV18 DNA as an anchor point in HeLa. At first, we wanted to check the extent of HPV integration induced chromatin interactions at the TAD level. For this, we plotted per bp coverage of 4C-seq reads in TAD3189 (with HPV integration) and in the neighbouring TADs (4 upstream and 4 downstream). Most of the 4C-seq signal was observed from the TAD with HPV integration similar to Hi-C analysis (Fig. S7A,B). Even though the HPV integration was found near the left boundary of TAD3189, the coverage of reads in the immediate upstream TAD was quite low. This suggests the role of chromatin structure (TAD boundaries) in confining the regulatory effects of the integrated HPV DNA within the same TAD predominantly. Further within the TAD3189, higher level of interaction was observed between the integrated HPV18 and all the three oncogenes (CCAT1, MYC and PVT1) in that TAD (Fig. 6A, 4C track). This further supports the previous studies which showed an interaction between HPV18 and MYC locus using ChIA-PET [16] and 3C analysis [15].

In HeLa, integration of HPV18 is observed in only one of the two haplotypes of chromosome 8. Based on this, we expected that the chromatin interactions observed between the integrated HPV18 DNA and the host genome could be haplotype-specific. To check this, we performed haplotype-specific 4C-seq coverage analysis, using heterozygous SNPs (as marker positions) from the HeLa genome (see Section 2). This revealed that almost all reads from the 4C-seq (~ 99–100%) mapped to the allele from the Haplotype A (which has the HPV integration) (Fig. S7C,D). This suggests that the HPV integration-mediated chromatin interactions are not only localised mostly within the same TAD but are also haplotype specific.

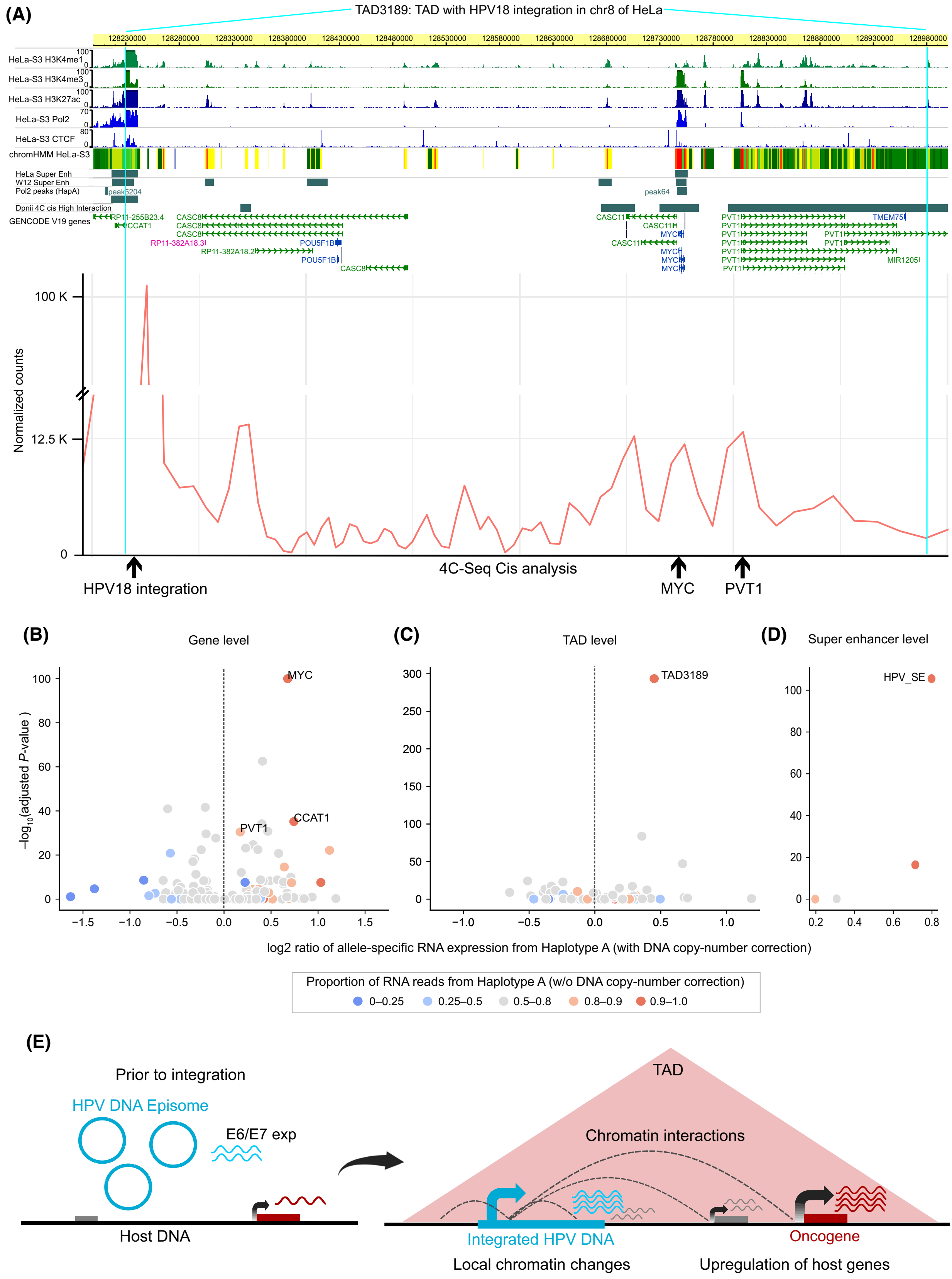

(A)
TAD3189: TAD with HPV18 integration in chr8 of HeLa
128230000
128280000
128330000
128380000
128430000
128480000
128530000
128580000
128630000
128680000
128730000
128780000
128830000
128880000
128930000
128980000
HeLa-S3 H3K4me1
HeLa-S3 H3K4me3
HeLa-S3 H3K27ac
HeLa-S3 Pol2
HeLa-S3 CTCF
chromHMM HeLa-S3
HeLa Super Enh
W12 Super Enh
Pol2 peaks (HapA)
Dpnii 4C cis High Interaction
GENCODE V19 genes
peak5204
peak64
RP11-255B23.4
CCAT1
CASC8
RP11-382A18.3
RP11-382A18.2
POU5F1B
CASC11
MYC
PVT1
TMEM75
MIR1205
100 K
12.5 K
0
Normalized counts
HPV18 integration
4C-Seq Cis analysis
MYC
PVT1
(B)
Gene level
−log10(adjusted P-value)
MYC
CCAT1
PVT1
(C)
TAD level
TAD3189
(D)
Super enhancer level
HPV_SE
log2 ratio of allele-specific RNA expression from Haplotype A (with DNA copy-number correction)
Proportion of RNA reads from Haplotype A (w/o DNA copy-number correction)
0–0.25
0.25–0.5
0.5–0.8
0.8–0.9
0.9–1.0
(E)
Prior to integration
HPV DNA Episome
E6/E7 exp
Host DNA
TAD
Chromatin interactions
Integrated HPV DNA
Oncogene
Local chromatin changes
Upregulation of host genes

Fig. 6. HPV18 integration mediated chromatin interactions from 4C-seq and allele-specific expression of oncogenes and SE in HeLa. (A) Regions in the TAD3189 which show high interactions with the integrated HPV18 DNA by 4C-seq are shown in the bottom line plot (Cis analysis). The HPV integration region, MYC and PVT1 is marked by an arrow. Various histone modifications, Pol2 and CTCF binding tracks from HeLa are also overlaid. Pol2 peaks (HapA) track shows the Pol2 peaks which were significantly haplotype specifically binding to Haplotype A. (B–D) Allele-specific expression analysis for (B) individual genes, (C) TADs, and (D) super-enhancers on chromosome 8. Values on the *x*-axis represent the $\log_2$ ratio calculated as the fraction of RNA reads coming from haplotype A out of the total RNA reads divided by the fraction of DNA reads coming from haplotype A out of the total DNA reads (sum over all the heterozygous SNPs in a particular feature). The *y*-axis represents the negative $\log_{10}$ of Bonferroni corrected *P* values, after a binomial test of RNA read counts from Haplotype A against the proportion of DNA read counts from Haplotype A, for each of the features. (E) Model summarising HPV integration mediated changes in host chromatin structure and gene expression dysregulation. HPV integration can lead to local chromatin changes resulting in transcriptional upregulation of host genes (also in fusion with viral genes) in its vicinity. In addition, the integrated HPV DNA can mediate long-range chromatin interactions resulting in the upregulation of genes at a distance.

3.7. Haplotype-specific cis-regulating activity of HPV integration is associated with allele-specific oncogenes overexpression

Further, we asked if the above haplotype-specific chromatin looping interaction could lead to haplotype-specific regulatory changes and gene expression alterations. To check this, we performed haplotype-specific analysis for RNA polymerase II (Pol2) binding on chromosome 8 in HeLa. This resulted in three peaks of Pol2 in and around the TAD3189 which showed significant haplotype-specific binding to Haplotype A (> 98% reads from Haplotype A in all 3), as compared to the expected proportion based on DNA copy number. Interestingly, these peaks overlapped with the CCAT1 and MYC genes and also the super-enhancer overlapping the HPV integration (Fig. 6A, Pol2 peaks (HapA) track), suggesting that the integrated HPV DNA drives gene regulation in haplotype specific manner. Further, we asked if this results in preferential expression of genes from the haplotype with HPV integration. To test this, we performed allele-specific gene expression analysis (taking into account the DNA copy number) on chromosome 8 in HeLa. This revealed MYC having significantly higher expression from Haplotype A (Fig. 6B) (see Section 2). The other oncogenic lncRNAs in the TAD3189, CCAT1 and PVT1, also showed a similar pattern (Fig. 6B). Further, we performed allele-specific gene expression analysis at the TAD level, combining all the genes in a TAD together. It also revealed TAD3189 to have the significantly higher expression from haplotype A among all the TADs on chromosome 8 (Fig. 6C). These results are in line with the TAD specific chromatin interactions (from Hi-C and 4C-seq) reported above (Fig. 5B; Fig. S7A).

Furthermore, to test if the chromatin interactions are essential for the overexpression of MYC/PVT1, we utilised the Promoter Capture Hi-C data generated from HeLa with depletion of SCC1 (RAD21, a subunit of Cohesin) and CTCF, separately, by using auxin-inducible degron system [35]. Given that both Cohesin and CTCF are important for chromatin looping, we asked whether the chromatin interactions between HPV18 and promoters (TSS ±10 kb) of MYC/PVT1 change upon Cohesin or CTCF depletion. This showed that SCC1 (Cohesin) depletion resulted in a significant decrease in chromatin interactions between HPV18 DNA and the promoter regions of MYC (Fisher's exact test, *P*-value = 0.004, OR = 3.8) and PVT1 (*P*-value = 0.004, OR = 6.4) (Fig. S8A), whereas with CTCF depletion no significant changes were observed (*P*-value = 0.9, OR = 0.93/*P*-value = 0.9, OR = 1.1, respectively) (Fig. S8B). Furthermore, the transcriptional response (captured using SLAM-seq) in the SCC1 depletion condition (versus control) revealed a significant down-regulation of MYC expression ($\log_2$ fold change = −1.78 and FDR adjusted *P*-value = 3.07×10^{-21}) (Fig. S8C), suggesting that the chromatin interaction mediated by cohesin is directly influencing the MYC overexpression. However, PVT1 did not show any change in the expression, likely due to the low expression of lncRNA (as compared to mRNAs) and the sensitivity of SLAM-seq to capture it in a short time interval.

A recent study has reported that HPV integrations are enriched in super-enhancers (SE) which ultimately control lineage determining genes [4]. Hence, we asked whether the HPV18 integration in HeLa affects the SE activity on chromosome 8. TAD3189 also harbours multiple SEs, one of which overlaps HPV integration. Interestingly, this SE is the most strongest in terms of signal (enhancer marks) on chromosome 8, and the third most strongest across the genome in HeLa (SEdb [24]). Allele-specific expression analysis using GRO-seq data revealed this SE in TAD3189 to be the most significantly and highly expressed from Haplotype A among all the SEs on chromosome 8 (Fig. 6D). Taken together, these results indicate that the HPV integration in chromosome 8 lead to changes in regulatory activity and chromatin interactions, resulting in allele-specific expression of SE and oncogenes within the same TAD (TAD3189).

4. Discussion

Human papillomavirus DNA integration into the host nuclear DNA is often found in the tumours of advanced cervical cancers, as compared to their early stages where the HPV is mostly present in the episomal form [3,39,48]. Thus, the HPV DNA integration is considered as an important oncogenic event in the transformation and progression of cervical cancer, however, the molecular mechanism underlying this is not yet fully understood. On one hand, it can be attributed to the overexpression of viral oncogenes (E6 and E7) from the integrated HPV DNA which could affect the cellular functions (such as cell proliferation and immune response) [7,49]. On the other hand, the integrated HPV DNA itself can affect expression of nearby host genes in *cis* and thereby contribute to tumour development [13,50]. However, the extent of the latter at the genome-wide level, particularly in the context of host chromatin structure, is not well understood. Thus, in this study, we attempted to investigate the impact of HPV integration on host chromatin structure and gene regulation genome-wide, by using the HPV integration from cervical tumour samples combined with the chromatin structure information from closest matching cancer and normal cell lines.

First, we showed that the distribution of HPV integrations across the genome is non-random. They are significantly enriched in the active chromatin regions and depleted in the inactive chromatin regions. Though these results are consistent with the previous meta-analysis studies [17,18,51], our analysis tried to remove any bias emerging from the integration detection methods (such as RNA-seq and APOT) by considering only whole-genome DNA-based assays. Additionally, for the enrichment analysis, random integration sites were generated by matching the GC content around the observed HPV integration sites to account for the influence of local sequence content (since the HPV integrations are associated with microhomology-mediated processes [5]). This approach is robust as compared to the previous studies which used uniform random distributions. Still the significant enrichment of HPV integration observed at the active chromatin regions in the cervical tumours could be explained by multiple factors acting prior to and/or during the clonal selection. Prior events may include the closer proximity of the episomal HPV DNA to the active chromatin regions due to the interaction between viral E2 protein and host BRD4 chromatin proteins (likely for the utilisation of host transcription/replication factors for viral transcription and replication), DNA breaks at the transcriptionally active regions due to replication-transcription conflicts or torsional stress, DNA repair and microhomology between the HPV and host DNA [5,39,52]. All these factors can favour the accidental integration of HPV DNA into the host genome. After that, if the host locus of the HPV integration is favourable for the expression of viral oncogenes (E6/E7) and also for upregulation of nearby host genes (especially oncogenes discussed below) that provides the selective growth advantage and favours the clonal expansion, then these integrations are likely to undergo positive selection.

Second, we observed that the transcriptional upregulation in the flanking region (up to 10 kb) of HPV integration was evident in both active and inactive chromatin regions. This could be due to the local genomic alterations (amplifications/translocations) associated with the HPV integration, and also changes in the local chromatin environment. For example, the host transcription factors can bind to the integrated HPV DNA and thereby increase the local chromatin accessibility [53–55] and subsequently upregulate the expression of nearby host genes. Alternatively, this can be due to the fusion of nearby host genes with the viral genes [11]. Also, it is possible that the HPV integration at host regulatory regions (such as enhancers) could lead to the formation of super-enhancers [56,57] and thereby enhance the expression of host genes. Our findings expand this observation in cervical cancers as we observed a higher expression of SE eRNAs in the immediate vicinity of the HPV integration regions, regardless of the host chromatin state.

Third, we explored the long-range effect of HPV integration on expression of the host genes. For this, instead of defining arbitrary length cut-offs around the integration sites [4,11], we used TAD boundaries as demarcation points. Overall, we found that the TADs with HPV integration showed higher gene expression as compared to samples without integration in the same TADs. And this upregulation is positively correlated with the number of HPV integrations within the TAD. However, genes in the neighbouring TADs were mostly unaffected. This may be due to the fact that the host chromatin structure is influencing the HPV-associated genomic alterations (amplification/translocation) during the integration processes or limiting the HPV integration mediated chromatin interaction changes to intra-TAD because of the insulation property of TAD boundaries. Recent studies have shown that HPV integration can cause changes in the local TAD structures in advanced cervical cancers [58] and also in human cell lines: HPV16 integration in W12 (prior to clonal selection) [14] and HPV16 integration in SiHA cell line [59]. Our results further extend this observation genome-wide in cervical tumours and show that the majority of the HPV integration induced chromatin changes and associated gene expression changes are mostly confined to the same TAD.

We found few loci with recurrent HPV integration at the TAD level and were associated with overexpression of oncogenes within them. Further, looking at the distribution of HPV integration within these TADs revealed that the expression of oncogenes (such as MYC and PVT1) were not only affected by HPV integration directly at or closer to these genes, but also by integration farther away (~ 500 kb) but within the same TAD. To further understand the mode of regulation of these oncogenes by HPV integration, we chose HeLa as the model system as it had the integration in the most recurrently integrated TAD3189 in cervical tumours. Previous studies have shown chromatin interaction between the integrated HPV DNA and the MYC gene (~ 500 kb away from the integration) by using ChIA-PET and 3C assays in HeLa [15,16]. However, whether this interaction is localised and specific to MYC regions or the integrated HPV DNA can interact with other genomic regions is not known. Our unbiased analysis, by using available Hi-C data from HeLa, revealed that the majority of the chromatin interactions were localised within the same chromosome, specifically within the same TAD as integration (Fig. 5A; Fig. S6A). Further, 4C-seq analysis by taking integrated HPV DNA as a viewpoint, showed haplotype-specific chromatin interaction between integrated HPV DNA and host genomic regions in HeLa. This is further supported by the allele-specific RNA pol II binding enrichment, SE activity and overexpression of MYC and PVT1 genes within the TAD3189. This allele-specific activity can be extrapolated to the cervical tumour samples as well, because TAD3189 is the most recurrently integrated TAD among cervical tumours (Fig. 4A), and we also observed the overexpression of oncogenes MYC and PVT1 within them (Fig. 4B,C). We propose that this may be one of the many ways by which HPV integration influences the process of tumorigenesis, as the role of both of these oncogenes along with the CCAT1 in cervical carcinogenesis have already been established [44–46]. In line with this, a recent study has shown that the MYC overexpression in other tumour types could be driven by the somatic structural variant mediated changes in long-range chromatin interactions [60]. Similar to cervical cancers, the HPV integrations in head and neck squamous cell carcinoma also showed oncogenic transcriptional upregulation near the integrated sites and epigenetic changes in the host genome regions interacting with integrated HPV DNA [57].

However, the limitations of this study include: (a) The HPV integrations we analysed were mostly detected from the advanced stage cervical tumours, thus their genome-wide distribution with respect to the chromatin states and dysregulation of cancer genes expression observed here could be influenced by the factors acting prior and during the clonal selection. Future studies which test for *de novo* HPV integration (for example, by using cell line either infected with HPV or genome-wide profiling of early-stage cervical tumours) might shed light on the interplay between viral and host factors which drives the integration processes and associated chromatin changes; (b) the overexpression of host regions observed near the HPV integration in active and inactive regions could have contributions from the extrachromosomal DNA (ecDNA), which have a hybrid of viral-host genome [11,37]. Perhaps in future, the application of long-read DNA/RNA sequencing could help to disentangle the DNA structural conformations of HPV integrations and better quantify the contributions from the ecDNA; (c) the TADs from cell lines HeLa and NHEK covers only 47.5% and 56% of the genome respectively, and so we were limited to analyse those HPV integration that fall within that TAD region. Chromatin interaction maps from the matched tumour samples could help to better understand the long-range effect of HPV integrations genome-wide and also to study the changes of the TAD structure (whether the integration lead to split of existing TAD or merging of adjacent TADs).

5. Conclusions

To conclude, this study reveals the cis-regulatory potential of HPV integration on the host gene dysregulation through changes in the host chromatin structure and gene-regulatory interactions (Fig. 6E). On the basis of our results and previous findings we propose that HPV integration is a strong driver which mediates multiple modes of dysregulation (including overexpression of E6/E7 viral oncogene, cis-regulatory effect of HPV integrations, local copy number changes, and amplification of regulatory elements) that can affect the host cellular functioning and thereby providing selective growth advantages to the cancer cells. This study also demonstrates that TADs can be used to identify host genes at a distance that are likely to be affected by the HPV integrations through looping or chromatin contacts, instead of arbitrary distance cut-offs. This will help to identify more recurrent integrations at the TAD level and also to associate orphan HPV integrations with new target genes. Moreover, our findings reveal the significance of insulated neighbourhoods in the form of TADs and their key role in safeguarding the genome from spurious transcriptional changes driven by viral integrations.

Acknowledgements

We thank the members of RS and DN labs for feedback and suggestions. The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The genome sequence described/used in this research was derived from a HeLa cell line. Henrietta Lacks, and the HeLa cell line that was established from her tumour cells without her knowledge or consent in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research. This study was reviewed by the NIH HeLa Genome Data Access Working Group. The genomic datasets used for analysis described in this manuscript were obtained from the database of Genotypes and Phenotypes (dbGaP) through dbGaP accession number phs000640.v1.p1. This work was supported by the DBT/Wellcome Trust India Alliance Fellowship [grant number IA/I/20/1/504928] awarded to RS. We also acknowledge support of the Department of Atomic Energy, Government of India, under Project Identification No. RTI 4006 and intramural funds from NCBS-TIFR.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

AKS and RS conceived and designed the study. AKS performed most of the data analysis, interpreted the results, and prepared figures. DT and GC contributed for the Hi-C data analysis. KW and DN performed 4C-seq experiment. AKS and RS wrote the manuscript with input from other authors. All authors read and approved the final manuscript.

Data accessibility

The HPV integrations data used in study are given in Table S1. The data resources for the individual analysis were mentioned in Section 2. 4C-seq data generated from this study is deposited to GEO (GSE247605). The codes used for the analyses and generation of figures are given here https://github.com/onkoslab/cisHPV.

References

1 Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;**71**:209–49.

2 Bhatla N, Lal N, Bao Y-P, Ng T, Qiao Y-L. A meta-analysis of human papillomavirus type-distribution in women from South Asia: implications for vaccination. *Vaccine*. 2008;**26**:2811–7.

3 Moody CA, Laimins LA. Human papillomavirus oncoproteins: pathways to transformation. *Nat Rev Cancer*. 2010;**10**:550–60.

4 Warburton A, Markowitz TE, Katz JP, Pipas JM, McBride AA. Recurrent integration of human papillomavirus genomes at transcriptional regulatory hubs. *NPJ Genom Med*. 2021;**6**:101.

5 Hu Z, Zhu D, Wang W, Li W, Jia W, Zeng X, et al. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat Genet*. 2015;**47**:158–63.

6 Cancer Genome Atlas Research Network, Albert Einstein College of Medicine, Analytical Biological Services, Barretos Cancer Hospital, Baylor College of Medicine, Beckman Research Institute of City of Hope, et al. Integrated genomic and molecular characterization of cervical cancer. *Nature*. 2017;**543**:378–84.

7 Yim E-K, Park J-S. The role of HPV E6 and E7 oncoproteins in HPV-associated cervical carcinogenesis. *Cancer Res Treat*. 2005;**37**:319–24.

8 Hoppe-Seyler K, Bossler F, Braun JA, Herrmann AL, Hoppe-Seyler F. The HPV E6/E7 oncogenes: key factors for viral carcinogenesis and therapeutic targets. *Trends Microbiol*. 2018;**26**:158–68.

9 Akagi K, Li J, Broutian TR, Padilla-Nash H, Xiao W, Jiang B, et al. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res*. 2014;**24**:185–99.

10 McBride AA, Warburton A. The role of integration in oncogenic progression of HPV-associated cancers. *PLoS Pathog*. 2017;**13**:e1006211.

11 Nguyen N-PD, Deshpande V, Luebeck J, Mischel PS, Bafna V. ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer. *Nucleic Acids Res*. 2018;**46**:3309–25.

12 Tang K-W, Alaei-Mahabadi B, Samuelsson T, Lindh M, Larsson E. The landscape of viral expression and

host gene fusion and adaptation in human cancer. *Nat Commun*. 2013;**4**:2513.

13 Ojesina AI, Lichtenstein L, Freeman SS, Pedamallu CS, Imaz-Rosshandler I, Pugh TJ, et al. Landscape of genomic alterations in cervical carcinomas. *Nature*. 2014;**506**:371–5.

14 Groves IJ, Drane ELA, Michalski M, Monahan JM, Scarpini CG, Smith SP, et al. Short- and long-range cis interactions between integrated HPV genomes and cellular chromatin dysregulate host gene expression in early cervical carcinogenesis. *PLoS Pathog*. 2021;**17**: e1009875.

15 Shen C, Liu Y, Shi S, Zhang R, Zhang T, Xu Q, et al. Long-distance interaction of the integrated HPV fragment with MYC gene and 8q24.22 region upregulating the allele-specific MYC expression in HeLa cells. *Int J Cancer*. 2017;**141**:540–8.

16 Adey A, Burton JN, Kitzman JO, Hiatt JB, Lewis AP, Martin BK, et al. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*. 2013;**500**:207–11.

17 Doolittle-Hall JM, Cunningham Glasspoole DL, Seaman WT, Webster-Cyriaque J. Meta-analysis of DNA tumor-viral integration site selection indicates a role for repeats, gene expression and epigenetics. *Cancers*. 2015;**7**:2217–35.

18 Bodelon C, Untereiner ME, Machiela MJ, Vinokurova S, Wentzensen N. Genomic characterization of viral integration sites in HPV-related cancers. *Int J Cancer*. 2016;**139**:2001–11.

19 Christiansen IK, Sandve GK, Schmitz M, Dürst M, Hovig E. Transcriptionally active regions are the preferred targets for chromosomal HPV integration in cervical carcinogenesis. *PLoS One*. 2015;**10**:e0119566.

20 Zapatka M, Borozan I, Brewer DS, Iskar M, Grundhoff A, Alawi M, et al. The landscape of viral associations in human cancers. *Nat Genet*. 2020;**52**: 320–30.

21 Wang X, Jia W, Wang M, Liu J, Zhou X, Liang Z, et al. Human papillomavirus integration perspective in small cell cervical carcinoma. *Nat Commun*. 2022;**13**:5968.

22 Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;**518**:317–30.

23 Vu H, Ernst J. Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *Genome Biol*. 2022;**23**:9.

24 Jiang Y, Qian F, Bai X, Liu Y, Wang Q, Ai B, et al. SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res*. 2019;**47**:D235–43.

25 Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 2014;**159**:1665–80.

26 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;**26**:841–2.

27 McCole RB, Erceg J, Saylor W, Wu C-T. Ultraconserved elements occupy specific arenas of three-dimensional mammalian genome organization. *Cell Rep*. 2018;**24**:479–88.

28 Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*. 2013;**14**:7.

29 Van der Auwera GA, O'Connor BD. Genomics in the cloud: using docker, GATK, and WDL in Terra. Sebastopol: O'Reilly Media; 2020.

30 Duttke SHC, Lacadie SA, Ibrahim MM, Glass CK, Corcoran DL, Benner C, et al. Human promoters are intrinsically directional. *Mol Cell*. 2015;**57**:674–84.

31 Farooq U, Saravanan B, Islam Z, Walavalkar K, Singh AK, Jayani RS, et al. An interdependent network of functional enhancers regulates transcription and EZH2 loading at the INK4a/ARF locus. *Cell Rep*. 2021;**34**:108898.

32 Raviram R, Rocha PP, Müller CL, Miraldi ER, Badri S, Fu Y, et al. 4C-ker: a method to reproducibly identify genome-wide interactions captured by 4C-seq experiments. *PLoS Comput Biol*. 2016;**12**:e1004780.

33 Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*. 2012;**9**:999–1003.

34 Servant N, Lajoie BR, Nora EP, Giorgetti L, Chen C-J, Heard E, et al. HiTC: exploration of high-throughput "C" experiments. *Bioinformatics*. 2012;**28**:2843–4.

35 Thiecke MJ, Wutz G, Muhar M, Tang W, Bevan S, Malysheva V, et al. Cohesin-dependent and -independent mechanisms mediate chromosomal contacts between promoters and enhancers. *Cell Rep*. 2020;**32**:107929.

36 Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*. 2012;**9**:215–6.

37 Tian R, Huang Z, Li L, Yuan J, Zhang Q, Meng L, et al. HPV integration generates a cellular super-enhancer which functions as ecDNA to regulate genome-wide transcription. *Nucleic Acids Res*. 2023;**51**:4237–51.

38 Li W, Notani D, Ma Q, Tanasa B, Nunez E, Chen AY, et al. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*. 2013;**498**:516–20.

39 Warburton A, Della Fera AN, McBride AA. Dangerous liaisons: long-term replication with an extrachromosomal HPV genome. *Viruses*. 2021;**13**:1846. https://doi.org/10.3390/v13091846

40 Li Y, Hu M, Shen Y. Gene regulation in the 3D genome. *Hum Mol Genet*. 2018;**27**:R228–33.

41 Ferber MJ, Thorland EC, Brink AATP, Rapp AK, Phillips LA, McGovern R, et al. Preferential integration

of human papillomavirus type 18 near the c-myc locus in cervical carcinoma. *Oncogene*. 2003;**22**:7233–42.

42 Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer*. 2018;**18**:696–705.

43 Carlevaro-Fita J, Lanzós A, Feuerbach L, Hong C, Mas-Ponte D, Pedersen JS, et al. Cancer LncRNA census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. *Commun Biol*. 2020;**3**:56.

44 Narisawa-Saito M, Inagawa Y, Yoshimatsu Y, Haga K, Tanaka K, Egawa N, et al. A critical role of MYC for transformation of human cells by HPV16 E6E7 and oncogenic HRAS. *Carcinogenesis*. 2012;**33**:910–7.

45 Iden M, Fye S, Li K, Chowdhury T, Ramchandran R, Rader JS. The lncRNA PVT1 contributes to the cervical cancer phenotype and associates with poor patient prognosis. *PLoS One*. 2016;**11**:e0156274.

46 Shen H, Wang L, Xiong J, Ren C, Gao C, Ding W, et al. Long non-coding RNA CCAT1 promotes cervical cancer cell proliferation and invasion by regulating the miR-181a-5p/MMP14 axis. *Cell Cycle*. 2019;**18**:1110–21.

47 Yardımcı GG, Ozadam H, Sauria MEG, Ursu O, Yan K-K, Yang T, et al. Measuring the reproducibility and quality of hi-C data. *Genome Biol*. 2019;**20**:57.

48 Daniel B, Rangarajan A, Mukherjee G, Vallikad E, Krishna S. The link between integration and expression of human papillomavirus type 16 genomes and cellular changes in the evolution of cervical intraepithelial neoplastic lesions. *J Gen Virol*. 1997;**78**(Pt 5):1095–101.

49 Rodrigues C, Joy LR, Sachithanandan SP, Krishna S. Notch signalling in cervical cancer. *Exp Cell Res*. 2019;**385**:111682.

50 Groves IJ, Coleman N. Human papillomavirus genome integration in squamous carcinogenesis: what have next-generation sequencing studies taught us? *J Pathol*. 2018;**245**:9–18.

51 Schmitz M, Driesch C, Jansen L, Runnebaum IB, Dürst M. Non-random integration of the HPV genome in cervical cancer. *PLoS One*. 2012;**7**:e39632.

52 Canela A, Maman Y, Jung S, Wong N, Callen E, Day A, et al. Genome organization drives chromosome fragility. *Cell*. 2017;**170**:507–521.e18.

53 Johannsen E, Lambert PF. Epigenetics of human papillomaviruses. *Virology*. 2013;**445**:205–12.

54 Karimzadeh M, Arlidge C, Rostami A, Lupien M, Bratman SV, Hoffman MM. Human papillomavirus integration transforms chromatin to drive oncogenesis. *bioRxiv*. https://doi.org/10.1101/2020.02.12.942755

55 Gagliardi A, Porter VL, Zong Z, Bowlby R, Titmuss E, Namirembe C, et al. Analysis of Ugandan cervical carcinomas identifies human papillomavirus clade-specific epigenome and transcriptome landscapes. *Nat Genet*. 2020;**52**:800–10.

56 Warburton A, Redmond CJ, Dooley KE, Fu H, Gillison ML, Akagi K, et al. HPV integration hijacks and multimerizes a cellular enhancer to generate a viral-cellular super-enhancer that drives high viral oncogene expression. *PLoS Genet*. 2018;**14**:e1007179.

57 Mima M, Okabe A, Hoshii T, Nakagawa T, Kurokawa T, Kondo S, et al. Tumorigenic activation around HPV integrated sites in head and neck squamous cell carcinoma. *Int J Cancer*. 2023;**152**:1847–62.

58 Cao C, Hong P, Huang X, Lin D, Cao G, Wang L, et al. HPV-CCDC106 integration alters local chromosome architecture and hijacks an enhancer by three-dimensional genome structure remodeling in cervical cancer. *J Genet Genomics*. 2020;**47**:437–50.

59 Xu X, Han Z, Ruan Y, Liu M, Cao G, Li C, et al. HPV16-LINC00393 integration alters local 3D genome architecture in cervical cancer cells. *Front Cell Infect Microbiol*. 2021;**11**:785169.

60 Xu Z, Lee D-S, Chandran S, Le VT, Bump R, Yasis J, et al. Structural variants drive context-dependent oncogene activation in cancer. *Nature*. 2022;**612**:564–72.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Enrichment of HPV integrations in full-stack ChromHMM and non-B form DNA regions.

Fig. S2. Enhanced transcriptional activity near HPV integration in the context of chromatin states from NHEK.

Fig. S3. HPV integration associated host gene overexpression with respect to HeLa and NHEK TAD domains.

Fig. S4. Recurrently integrated TADs in small cell cervical cancer and the pathway level dysregulation in TADs with integration.

Fig. S5. Relationship between distance of HPV integration and the expression of oncogenes.

Fig. S6. TAD level interaction frequency between HPV18 integrated DNA and host genome in HeLa.

Fig. S7. 4C-seq analysis reveals haplotype-specific chromatin interactions mediated by integrated HPV18 DNA in HeLa.

Fig. S8. Promoter Capture Hi-C analysis in HeLa depleted for Cohesin and CTCF.

Table S1. List of HPV integrations in the human genome (hg19) from cervical tumour samples.

Table S2. HeLa and NHEK TAD co-ordinates (hg19).

Table S3. Primers used in 4C-seq experiment.