npj | Genomic Medicine

www.nature.com/npjgenmed

ARTICLE OPEN

Check for updates

Recurrent integration of human papillomavirus genomes at transcriptional regulatory hubs

Alix Warburton[1], Tovah E. Markowitz[2,3], Joshua P. Katz[4], James M. Pipas[4] and Alison A. McBride[1]✉

Oncogenic human papillomavirus (HPV) genomes are often integrated into host chromosomes in HPV-associated cancers. HPV genomes are integrated either as a single copy or as tandem repeats of viral DNA interspersed with, or without, host DNA. Integration occurs frequently in common fragile sites susceptible to tandem repeat formation and the flanking or interspersed host DNA often contains transcriptional enhancer elements. When co-amplified with the viral genome, these enhancers can form super-enhancer-like elements that drive high viral oncogene expression. Here we compiled highly curated datasets of HPV integration sites in cervical (CESC) and head and neck squamous cell carcinoma (HNSCC) cancers, and assessed the number of breakpoints, viral transcriptional activity, and host genome copy number at each insertion site. Tumors frequently contained multiple distinct HPV integration sites but often only one "driver" site that expressed viral RNA. As common fragile sites and active enhancer elements are cell-type-specific, we mapped these regions in cervical cell lines using FANCD2 and Brd4/H3K27ac ChIP-seq, respectively. Large enhancer clusters, or super-enhancers, were also defined using the Brd4/H3K27ac ChIP-seq dataset. HPV integration breakpoints were enriched at both FANCD2-associated fragile sites and enhancer-rich regions, and frequently showed adjacent focal DNA amplification in CESC samples. We identified recurrent integration "hotspots" that were enriched for super-enhancers, some of which function as regulatory hubs for cell-identity genes. We propose that during persistent infection, extrachromosomal HPV minichromosomes associate with these transcriptional epicenters and accidental integration could promote viral oncogene expression and carcinogenesis.

npj Genomic Medicine (2021)6:101 ; https://doi.org/10.1038/s41525-021-00264-y

INTRODUCTION

Persistent infection with high-oncogenic risk human papillomavirus (HPV) types is responsible for almost all cervical and ~70% oropharyngeal carcinomas[1]. One factor that can contribute to oncogenic progression of HPV-positive lesions is integration of the viral genome into host chromatin. Integration is associated with increased genetic instability in high-grade cervical intraepithelial neoplasia, and cervical and oropharyngeal carcinomas[2–9] due to dysregulated expression of the viral oncoproteins, E6 and E7. Many studies have compared the human genomic regions associated with HPV integration sites to elucidate the mechanisms that might promote integration and carcinogenesis. Here we have curated datasets from The Cancer Genome Atlas and other published sources to define a rigorous database of integration breakpoints and correlated these with "in-house" datasets of common fragile sites and enhancer elements defined in cervical carcinoma cells.

Integration of HPV DNA occurs in all human chromosomes; however, integration sites are often found within or in close proximity to common fragile sites[10–13]. Common fragile sites are regions of the genome that have difficulty completing replication and, as such, are susceptible to chromosome breakage in mitosis. They are prone to replication stress that can be due to a shortage of replication origins or clashes between replication and transcriptional processes. Therefore, they vary in fragility depending on cell type and disease state[14]. Most previous studies that documented an association of HPV integration sites with common fragile sites used the classical FRA (fragile) regions that were defined cytogenetically in lymphocytes[9–13,15]. As such, these fragile sites are not cell-type-specific and are often large, poorly defined regions that cover a large proportion of the human genome. FANCD2 is required for resolution of these genetically unstable sites and, as such, is a marker of common fragile sites[16–18]. Fragile sites are often still undergoing DNA synthesis during mitosis and novel datasets have recently been generated by analysis of nascent DNA synthesis in mitotic cells[19,20]. Here we use published datasets of mitotic DNA synthesis (MDS) in HeLa cells, as well as our own "in-house" datasets to define common fragile sites in cervical cancer cells.

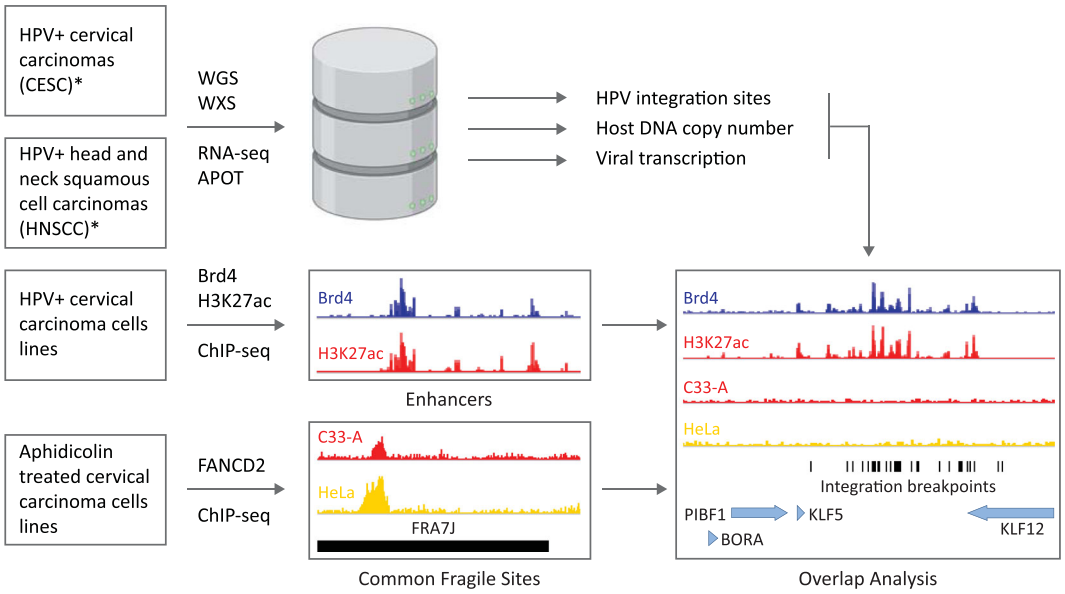
HPV integration sites occur frequently in amplified regions of the host genome and focal amplification of cellular flanking sequences at sites of viral integration are frequently observed in HPV-positive tumors[6,7,9,15]. Co-amplification of the viral genome and flanking cellular sequences can result from unlicensed initiation of replication at the viral origin resulting in endoreduplication[21–23]. Subsequent recombination can result in amplified tandem repeats. Genome amplification can also occur at common fragile sites by breakage-fusion-bridge cycles[24].

HPV integration is also enriched at transcriptionally active regions of the host genome[15,25]. We previously identified an HPV16 integration site in the W12 20861 cervical cell line that was adjacent to a cell-type-specific enhancer. Co-amplification of this regulatory element and the viral genome to ~25 copies resulted in the formation of a super-enhancer-like element to drive high viral oncogene expression[26,27]. This "enhancer-hijacking" is a novel mechanism by which HPV integration can promote oncogenesis;

[1]Laboratory of Viral Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 33 North Drive, MSC3209, Bethesda, MD 20892, USA. [2]NIAID Collaborative Bioinformatics Resource (NCBR), National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. [3]Advanced Biomedical Computational Science, Frederick National Laboratory for Cancer Research, Frederick, MD, USA. [4]Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. ✉email: amcbride@nih.gov

Published in partnership with CEGMR, King Abdulaziz University

npj nature partner journals



***Data Sources:** Bodelon, Holmes, Hu, Koneva, Lui, Ojesina, Olthof, Parfenov, The Cancer Genome Atlas, Xu

Fig. 1 Overview of datasets. Schematic representation of datasets used for overlap analysis of CESC and HNSCC integration sites with enhancers mapped in W12 cervical keratinocytes and FANCD2-associated common fragile sites mapped in C33-A and HeLa cervical carcinoma cell lines. APOT amplification of papillomavirus oncogene transcripts, WGS whole genome sequencing, WXS whole exome sequencing.

however, it is unclear how common this mechanism is in HPV-associated cancers.

The aim of this study was to examine the association among HPV integration loci, common fragile sites, and genome amplification, to determine whether insertion of HPV genomes adjacent to active cellular enhancers often resulted in viral "enhancer-hijacking," and whether genetic instability could result in co-amplification of viral-cellular regulatory repeats to drive oncogenic progression of HPV-associated cancers. We have extended our previously published work[28] to generate a common fragile-site dataset in cervical carcinoma cell lines using higher resolution mapping of FANCD2 binding and have mapped cellular enhancers and super-enhancers in an HPV16-positive cell line derived from a cervical lesion[29] using H3K27ac and Brd4 chromatin immunoprecipitation sequencing (ChIP-seq) (Fig. 1). Here we compare HPV integration sites with these cervical cell-type-specific enhancer and fragile-site datasets.

RESULTS

CESC and HNSCC tumors frequently contain multiple, clustered HPV integration breakpoints

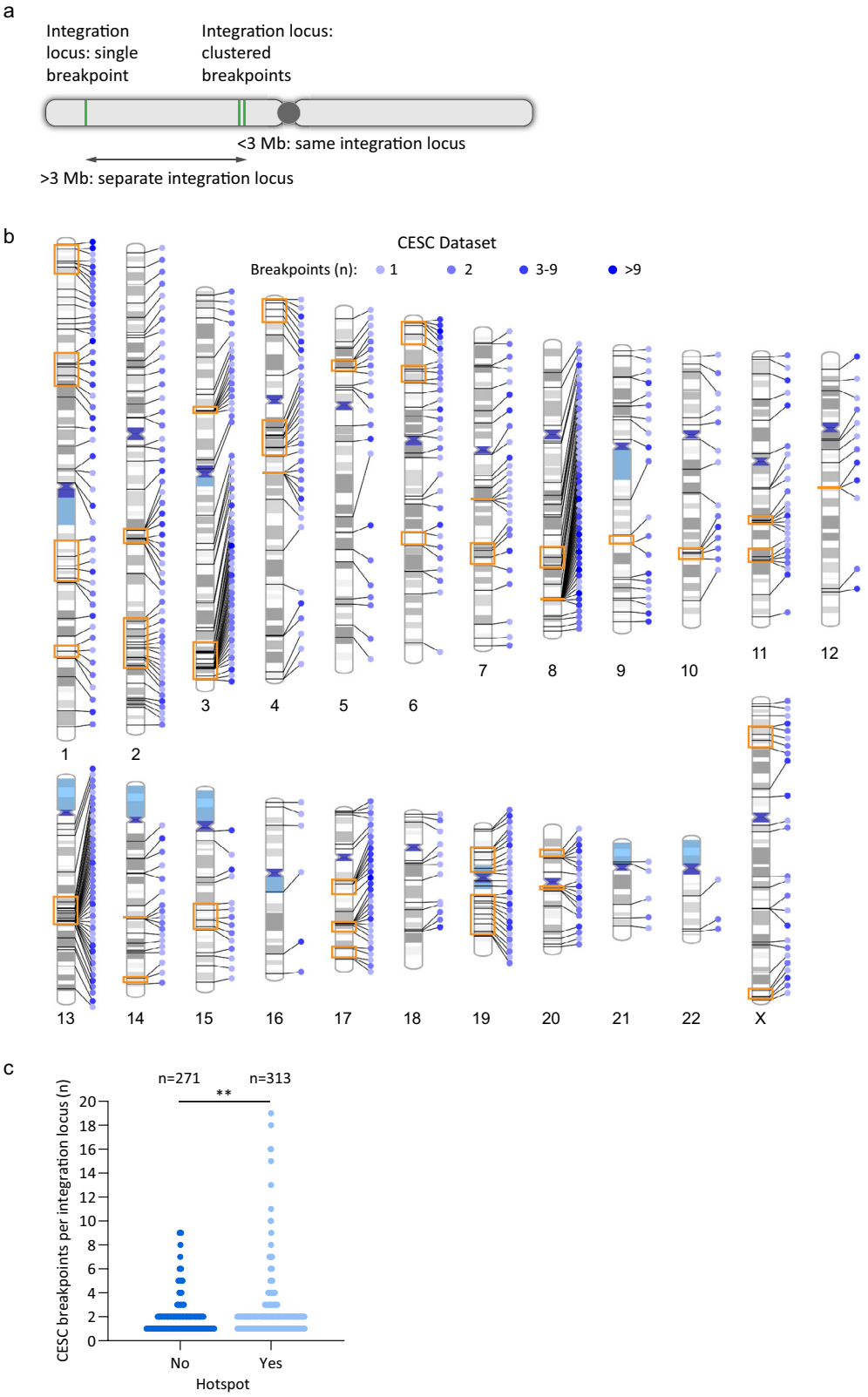
A dataset of HPV integration breakpoints was assembled from various sources[5,6,8,9,30–37], as outlined in Fig. 1 and Supplementary Data Table 1. Integration breakpoints were defined as the junctions between the viral and host chimeric reads within the human reference genome. A total of 1299 integration breakpoints from 333 cervical carcinomas (CESC) and 119 integration breakpoints from 41 head and neck squamous cell carcinomas (HNSCC) were included in this study (Supplementary Fig. 1 and Supplementary Data Tables 2 and 3). We found that many tumor samples contained multiple integration breakpoints that could have resulted from either independent integration events at different chromosomal loci or from amplification of a single integration site resulting in a cluster of multiple, closely spaced breakpoints. To classify this, we defined an integration locus as either a single HPV insertion breakpoint, or as multiple, closely spaced breakpoints (a cluster) (Fig. 2a). Clustered breakpoints within the same chromosome that had a maximum distance of 3 Mb between the most 5′- and 3′-breakpoints were classified as a single integration locus. This cutoff was determined initially by visual inspection of the datasets to assess which sites appeared to logically cluster together. Based on this classification, the total number of integration loci analyzed in our study was 584 for CESC samples and 58 for HNSCC samples. Tumors with multiple integration loci were observed in 28.8% of CESC and 22.0% of HNSCC tumors.

Sites of recurrent HPV DNA integration in different tumor samples are termed integration hotspots. We defined integration hotspots (five or more sites located <5 Mb apart) in our CESC dataset and compared them to previously defined hotspots from the literature[10,15,30,38–42]. This cutoff was determined initially by visual inspection of CESC integrations across each chromosome to assess which loci appeared to logically cluster together. We identified a total of 37 hotspots in CESC tumors (Supplementary Data Table 4), which represented 313/584 (53.6%) integration loci from our CESC dataset (Fig. 2b, c). Twenty-three hotspots overlapped previously defined sites of recurrent integration and 14 were novel hotspots (Supplementary Fig. 2 and Supplementary Data Tables 4 and 5). We were unable to define sites of recurrent integration in the HNSCC dataset because of the low number of integration loci in these tumors.

The distribution of clustered breakpoints at each integration locus and across each chromosome is shown in Fig. 2b for CESC samples and in Supplementary Fig. 3 for HNSCC samples. Most integration sites had one to two breakpoints in both the CESC (81.1%) and HNSCC (75.9%) datasets. Sites of recurrent integration are indicated by orange boxes for the CESC samples. The integration loci at these hotspots were more likely to contain clustered breakpoints compared to integration sites elsewhere in the genome for CESC tumors (Fig. 2c; $p = 0.004$). Higher numbers of clustered breakpoints at sites of recurrent integration suggests that these regions are susceptible to genomic instability.

Most tumors with integrated HPV DNA have a single driver integration

Constitutive expression of the viral oncogenes from the integration locus is required for clonal selection and oncogenic



progression. Transcriptionally silent HPV integration loci can be considered to be passenger sites[37]. To identify driver versus passenger integrations, the transcriptional activity of each integration locus was determined for the subset of samples that had matched RNA sequencing data (CESC, $n = 144$; HNSCC, $n = 35$). The transcription status of each integration locus is indicated in Supplementary Data Tables 2 and 3. Integration loci in which no viral–host chimeric junctions were detected by RNA sequencing (RNA-seq) analysis were classified as inactive or passenger loci. In CESC, all samples with a single integration locus ($n = 86$) were

Fig. 2 HPV integration loci frequently contain clustered insertional breakpoints. a Schematic representation of HPV integration breakpoints and loci. Green lines represent integration breakpoints. Integration loci are defined as either a single breakpoint, or multiple, closely spaced breakpoints (a cluster). Samples with clustered breakpoints within the same chromosome are classified as a single integration locus if the 5′ and 3′ most breakpoints are within 3 Mb of each other. **b** Schematic representation of clustered breakpoints at CESC integration loci across the human genome. Lines connecting to each chromosome represent different integration loci. Blue circles represent the indicated number of breakpoints per integration locus; orange boxed regions represent integration hotspots. See Supplementary Fig. 3 for the distribution of clustered breakpoints at integration loci in HNSCC tumors. **c** Scatter plot showing the frequency of single and clustered breakpoints per integration locus for CESC tumors grouped according to whether they overlap integration hotspots. The *p*-value is based on a non-parametric, unpaired *t*-test (two-tailed; $^{**}P < 0.01$).

transcriptionally active (Fig. 3a). For samples with multiple integration loci (CESC, $n = 58$; HNSCC, $n = 13$), more than one transcriptionally active integration locus was observed in 35 (60.3%) CESC and 4 (30.8%) HNSCC tumors. Three HNSCC samples had no driver integrations; one sample (4.5%) had a single integration locus and two samples (15.4%) had multiple integration loci (Fig. 3b). Overall, the majority of CESC and HNSCC tumors with integrated HPV genomes had only a single transcriptionally active integration locus. This implies that most tumors with integrated HPV DNA have a single driver integration. Viral oncogene transcription was analyzed at integration hotspots in CESC, which showed that sites of recurrent integration can have both driver and passenger integrations but are more likely to be transcriptionally active (Fig. 3c).

Clustered integration breakpoints are associated with amplified regions of the host genome in CESC and HNSCC

HPV integration loci often have amplification and/or rearrangements of the flanking cellular sequences at the insertion sites[6,7,9,15]. We determined the frequency with which integration loci in our datasets were associated with somatic copy number alterations for the subset of samples that had matched host genome copy number data (235 CESC and 22 HNSCC tumors; Supplementary Data Tables 6 and 7). In this subset, most integration breakpoints occurred within amplified regions of cellular DNA in both CESC and HNSCC relative to regions with a normal genomic copy number or adjacent to deletions (Fig. 4a, b). In addition, the number of breakpoints per integration locus was significantly different at amplified regions of the host genome relative to loci with a normal genomic profile for both CESC and HNSCC samples; higher numbers of breakpoints per cluster occurred at amplified regions (Fig. 4a, b). No significant difference was found in the number of clustered breakpoints per integration loci in regions with a normal genomic copy number relative to those with genomic copy number losses in CESC samples. We conclude that integration loci associated with flanking host DNA amplification were more likely to contain clustered breakpoints and to have higher numbers of breakpoints per integration locus.

Integration loci with associated focal amplifications were found on all chromosomes in CESC and on 15 chromosomes in HNSCC. The distribution of clustered integration breakpoints relative to host genome amplification and sites of recurrent integration in CESC is shown (Fig. 4c, d). There was an enrichment of integration loci with associated host DNA amplifications at integration hotspots in CESC (Fig. 4e) and higher numbers of clustered breakpoints were common at these regions (Fig. 4c). Host DNA copy number at sites of recurrent integration in CESC are indicated in Supplementary Data Table 6. Genomic instability in these regions most likely results in co-amplification of both viral and host DNA.

Identification of common fragile sites in cervical cancer cell lines

Genomic instability occurs frequently at common fragile sites. We previously used FANCD2 ChIP-chip to map fragile sites in an aphidicolin-treated cervical carcinoma cell line (C33-A)[28]. Here we have extended the FANCD2 dataset using ChIP-seq analysis of both C33-A and HeLa cervical carcinoma cells (Supplementary Data Tables 8 and 9) and combined these results with those previously mapped in C33-A cells by ChIP-chip[28]. Despite the reported chromothrypsis in HeLa cells[43], there was good overlap between the FANCD2 peaks mapped in the HeLa and C33-A datasets ($p < 0.0001$). In total, we defined 513 FANCD2-enriched regions between the two cervical carcinoma cell lines and they are listed in Supplementary Data Table 10.

We compared our cervical carcinoma cell line derived FANCD2 mapped fragile sites (genomic coverage 7.9%) to the 77 aphidicolin-induced common fragile sites (FRA regions) defined cytogenetically in lymphocytes and reported in the HGNC (HUGO Gene Nomenclature Committee) database (genomic coverage 48.3%) (Fig. 5a). A total of 115 (22.4%) FANCD2-enriched regions derived from C33-A and HeLa cells overlapped with 55.8% ($n = 43/77$) FRA regions (Fig. 5b). FRA regions that overlapped with FANCD2-associated fragile sites are listed in Supplementary Data Table 11. Permutation testing was used to determine the significance in overlap between our FANCD2 dataset and traditional FRA regions. The association between these genomic features did not reach significance ($p = 0.0634$), which reflects differences in replication stress at these regions in different cell types.

Recent studies used high-resolution MiDA-seq (next-generation sequencing of EdU incorporation at sites of mitotic DNA synthesis) to map fragile sites in HeLa cells[19,20] and show that they colocalize with FRA regions and FANCD2 foci in cells treated with aphidicolin. We compared the overlap between our dataset of FANCD2-enriched regions and the mitotic DNA synthesis regions profiled in HeLa cells (total genomic coverage of 4.4%, Fig. 5a). A total of 120 (23.4%) FANCD2-enriched regions overlapped with mitotic DNA synthesis regions, which represented 48.3% ($n = 112/232$) of mitotic DNA synthesis regions (Fig. 5b). Permutation testing was used to determine the significance in overlap between FANCD2-enriched regions and mitotic DNA synthesis regions ($p < 0.0001$). Mitotic DNA synthesis regions that overlapped with FANCD2-associated fragile sites are indicated in Supplementary Data Table 12. Collectively, these data show good correlation between our FANCD2-associated fragile sites and regions of the genome susceptible to genetic instability.

The instability of fragile sites is often due to transcription-replication conflicts that frequently occur at long genes[44]. We and others have previously shown that FANCD2-enriched regions overlap with transcriptionally active long genes in C33-A[28] and U2OS cells[45]. Here we extended that association to include genes that are >0.3 Mb in length[46]. A total of 184/513 (35.9%) FANCD2-enriched regions overlapped with protein-coding genes that were ≥0.3 Mb, which corresponded to 185/782 (23.7%) long genes. A Fisher's exact test was used to determine the significance in overlap between FANCD2-enriched regions and long genes (two-tailed, $p = 4.22E-13$). Of the long genes that overlapped with FANCD2-enriched regions, 121/185 (65.4%) were expressed in C33-A and/or HeLa cells (Fig. 5c), further validating our FANCD2 peaks as sites of genetic instability in cervical carcinoma cells. Supplementary Data Table 13 lists the long genes used in this analysis and their association with common fragile sites and

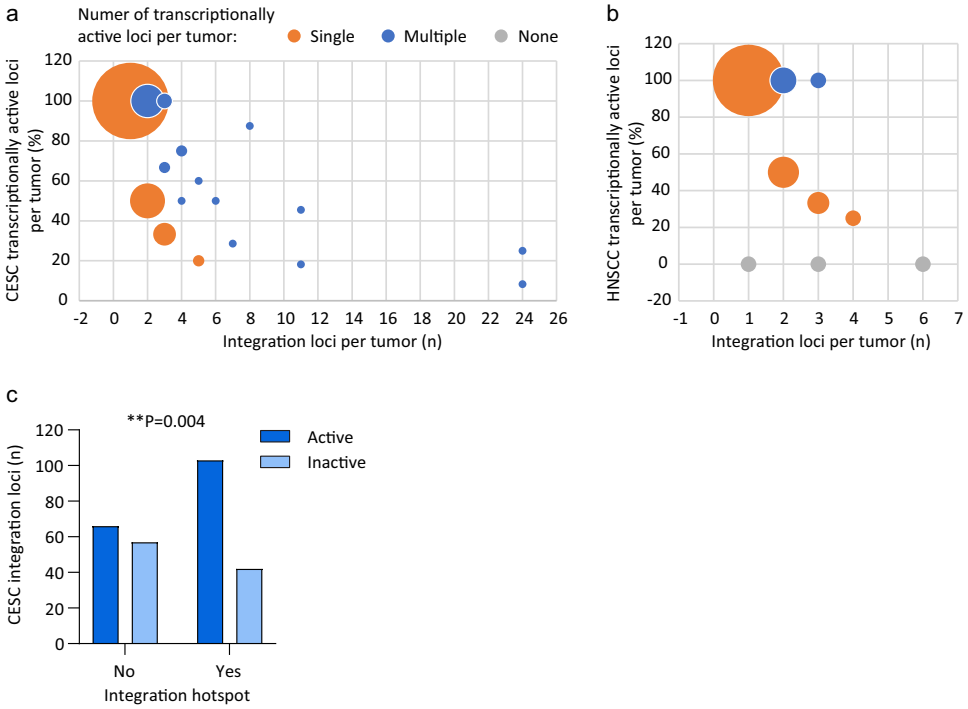


Fig. 3 Transcription status of integrated viral genomes. a, b Bubble graph showing the percentage of transcriptionally active integration loci per tumor in CESC (**a**) and HNSCC (**b**) samples relative to the number of integration loci per tumor; 100% indicates that all integration loci are active in that tumor. Orange, blue and gray circles represent tumors with a single, multiple or no transcriptionally active loci, respectively. Circle size indicates the number of samples per grouping (for CESC, largest, $n = 86$ and smallest, $n = 1$; for HNSCC, largest, $n = 21$ and smallest, $n = 1$). Three HNSCC samples (*TCGA-CR-6482*, *TCGA-CN-5374*, and *TCGA-CR-7404*) were reported as integration negative from RNA-seq[32] but had a single or multiple integration loci detected through WGS[5] and were therefore classified as transcriptionally inactive. **c** Bar chart showing the number of CESC integration loci that are transcriptionally active or inactive for viral oncogene expression at integration hotspots. Association between viral oncogene transcription and integration hotspots was based on a Fisher's exact test (two-tailed; $^{**}P < 0.01$).

expression in C33-A and HeLa cells from RNA-seq[28] (Expression Atlas, https://www.ebi.ac.uk/gxa/home). Example alignments of our FANCD2-associated fragile sites relative to common FRA regions, mitotic DNA synthesis regions and long genes are shown in Fig. 5d.

Integration breakpoints are enriched at FANCD2-associated fragile sites

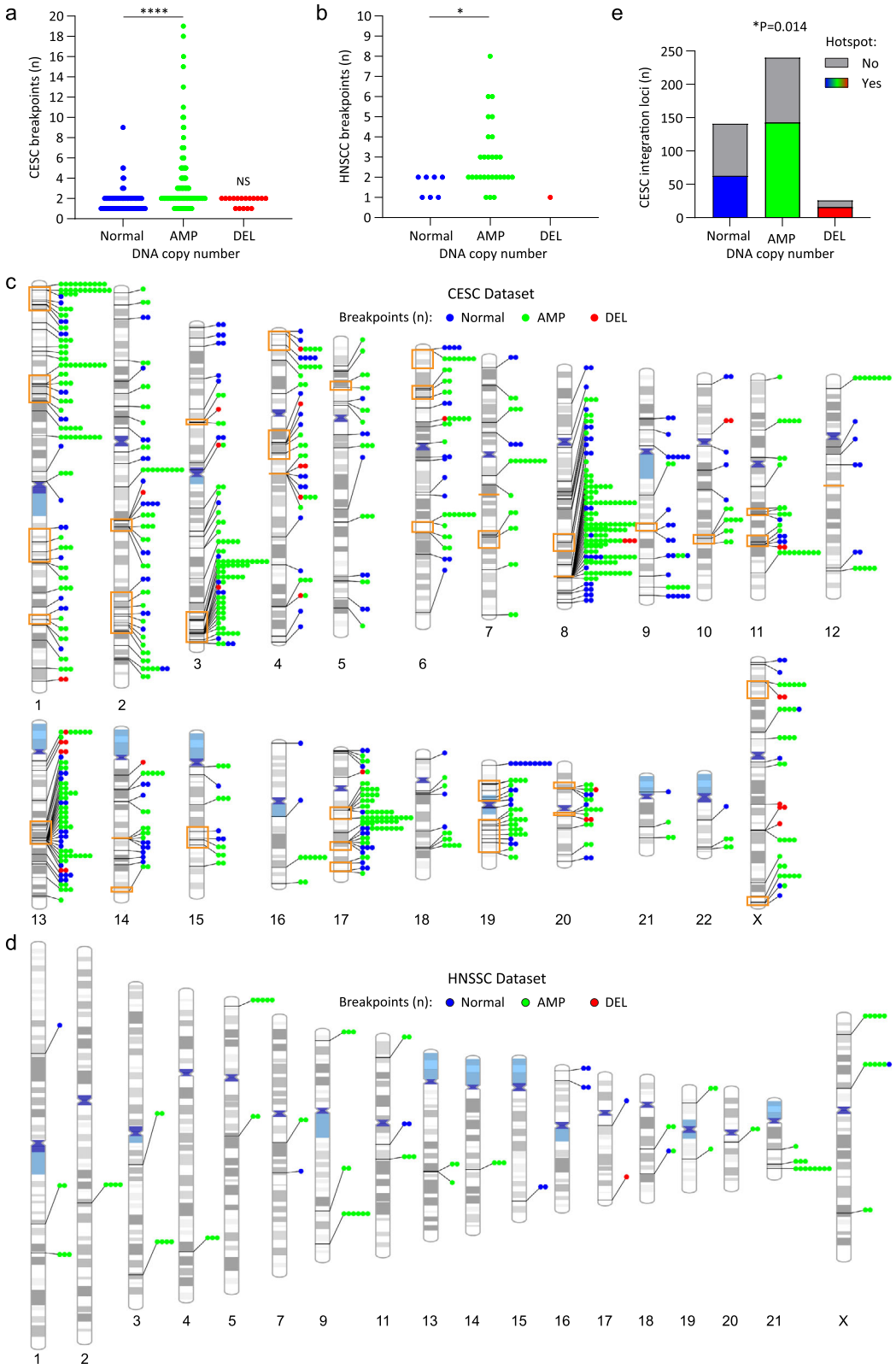
To assess the association of our cervical carcinoma-specific fragile-site dataset with HPV integration sites, we calculated the frequency with which an integration breakpoint occurred within 50 Kb[9] of the C33-A and HeLa FANCD2-enriched regions (Supplementary Data Table 10). Approximately 18% integration breakpoints were associated with FANCD2-enriched regions in both the CESC and HNSCC datasets. The data were permutated 10,000 times to create an expected distribution of the overlap between integration breakpoints and FANCD2-enriched regions. This showed that the FANCD2-associated fragile sites were significantly enriched for both CESC and HNSCC HPV integration sites when each breakpoint was analyzed independently from each other (Table 1).

To remove bias resulting from overrepresentation of integration loci containing clusters of breakpoints, the integration loci were simplified into defined subsets for significance testing. Integration loci contain either a single breakpoint or a cluster of breakpoints (Fig. 2a). In the latter group, the clustered breakpoints were condensed into a single site represented by the most 5′- and 3′-breakpoints. The final category combined the single and condensed categories, in which each integration locus was represented just once. The integration loci in each category were tested independently for their association with FANCD2-enriched regions. For the CESC dataset, sites containing single or clustered breakpoints were significantly associated with FANCD2-enriched regions (Table 1). In contrast, only sites with clustered breakpoints reached significance for the HNSCC dataset (Table 1).

Generation of a cervical keratinocyte enhancer dataset using Brd4 and H3K27ac ChIP-seq

It has been noted previously that HPV integration occurs frequently at transcriptionally active regions[15,25], and we have demonstrated that HPV integration can capture and amplify cellular enhancers to drive viral oncogene expression[27]. Brd4 is a marker of cell lineage-specific enhancers[47–49] and HPV E2 tethering sites[28,50]. Moreover, we, and others, have shown previously that Brd4 and the HPV E2 replication protein bind to transcriptionally active chromatin within the host genome[50,51] that overlap many FANCD2-associated fragile sites[28]. Viral replication factories form adjacent to these sites[28] and we have proposed that tethering of the viral genome to these unstable sites would increase the chances of integration at these regions.

To further examine the association of cellular enhancers with HPV integration in CESC and HNSCC, we generated an "in-house" enhancer dataset using Brd4 and H3K27ac ChIP-seq in four different subclones of W12 cervical keratinocytes. We defined 6935 enhancer consensus peaks in the four W12 subclones (Supplementary Data Table 14). The resulting H3K27ac and Brd4 ChIP-seq signals were compared to enhancers in the NHEK (normal human epidermal keratinocytes) ENCODE dataset[52], which showed that 83.5% ($p < 0.0001$) W12 Brd4/H3K27ac



enriched peaks overlapped ENCODE defined enhancers (Supplementary Fig. 4). Approximately 40% integration breakpoints and loci from the CESC and HNSC datasets were significantly associated with our cervical keratinocyte enhancer dataset (Table 2).

Integration hotspots are associated with gene loci related to cell development and identity

Enhancers that were associated with integration loci were analyzed using the Genomic Regions Enrichment of Annotations Tool (GREAT)[53] to identify common functional significance based

Fig. 4 Clustered integration breakpoints are associated with amplified regions of the host genome in CESC and HNSCC. For the subset of CESC and HNSCC samples that had matched somatic copy number alteration data, HPV integration breakpoints were grouped according to the associated host DNA copy number status. Normal, AMP (amplification) and DEL (deletion) refers to the genomic profile of the host DNA at the integration locus. **a**, **b** Scatter plots showing the number of breakpoints per locus grouped according to the somatic copy number alteration status of the integration locus for CESC (**a**) and HNSCC (**b**) tumors. For CESC, the number of integration loci per grouping was Normal, $n = 140$; AMP, $n = 240$ and DEL, $n = 17$. For HNSCC, the number of integration loci per grouping was Normal, $n = 7$; AMP, $n = 28$ and DEL, $n = 1$. *P*-values are based on non-parametric, unpaired *t*-tests (two-tailed; $^{*}p < 0.05$, $^{****}p < 0.0001$, NS = nonsignificant). All statistical tests were performed relative to integration loci with a normal genomic profile. DNA amplification associated with integration loci ranged from 693 bp to 54.2 Mb (average, 1.6 Mb; median 47.4 Kb) in CESC and 6.5 Kb to 102.7 Mb (average, 3.3 Mb; median 43.1 Kb) in HNSCC. **c**, **d** Schematic representation of clustered breakpoints at integration loci that have associated host somatic copy number alterations. Lines connecting to each chromosome represent different integration loci for the CESC (**c**) and HNSCC (**d**) datasets. The number of circles represents the number of breakpoints per integration locus. Blue, green and red colored circles respectively represent integration sites that have a normal genomic profile or associated amplifications or deletions. Orange boxed regions represent integration hotspots. **e** Stacked bar chart showing the number of CESC integration loci that overlap integration hotspots grouped according to whether they have associated somatic copy number alterations. Association between somatic copy number alterations and integration hotspots was based on a chi-square test ($^{*}p < 0.05$).

on proximity testing. This identified 52 putative enhancer target genes for the CESC dataset, representing 28 gene loci of which 60.7% overlapped integration hotspots (Fig. 6a). Twelve of the gene loci are previously defined sites of recurrent integration and include the *KLF5*, *KLF12*, *MYC*, *TP63*, *RAD51B*, and *HMGA2* genes, and five of the gene loci are novel integration hotspots and include the *CAMK1G*, *FOXQ1*, *EXOC2*, *GRHL2*, *ID1*, *COX4I2*, *HM13*, and *NFIA* genes (Fig. 6a). For HNSCC, 19 target genes were identified through GREAT Gene Ontology analysis, representing 8 gene loci of which 50% overlapped integration hotspots profiled in CESC (Supplementary Fig. 4). Six target genes (*KLF5*, *KLF12*, *FAM84B*, *POU5F1B*, *TUBD1*, and *VMP1*) were common across CESC and HNSCC. Gene Ontology analysis of our enhancer regions identified epithelium development, epithelial cell differentiation, and negative regulation of keratinocyte differentiation to be significantly enriched biological processes associated with CESC integration loci (Supplementary Data Table 15). Ectoderm development and differentiation, epidermal and epithelial differentiation, tongue morphogenesis, and negative regulation of spreading epidermal cells during wound-healing were biological processes significantly associated with enhancer enrichment at HNSCC integration loci (Supplementary Data Table 15). This indicates that sites of recurrent HPV integration are often associated with cellular pathways relevant to host cell development and differentiation.

Keratinocyte-specific super-enhancers are enriched at integration loci

Large enhancer clusters were characteristic of the integration targets identified through GREAT analysis. We therefore defined super-enhancers in our Brd4-defined W12 enhancer dataset based on relative peak height of the H3K27ac and Brd4 ChIP-seq datasets using the Rank Ordering of Super-Enhancers (ROSE) tool[54,55]. We defined 338 super-enhancers in W12 cervical keratinocytes (Supplementary Data Table 16). Intersect analysis showed that 89/584 (15.2%) CESC integration loci overlapped with super-enhancers profiled in W12 cells, and of these loci 72 (80.9%) were classified as sites of recurrent integration (Fig. 6b). A total of 25/37 (67.6%) integration hotspots contained super-enhancers and permutation testing showed that both CESC integration loci and sites of recurrent integration were significantly associated with these regulatory domains ($p < 0.0001$). For HNSCC, 8/57 (14%) integration loci were associated with W12 super-enhancers and 62.5% of these loci overlapped integration hotspots profiled in CESC, including the *KLF5*/*KLF12*, *MYC*, *ERBB2*, and *VMP1* gene loci. Permutation testing showed that HNSCC integration loci were significantly associated with super-enhancers profiled in W12 cells ($p < 0.0001$). Thus, keratinocyte-specific super-enhancers are enriched at integration loci in HPV-associated tumors and are frequently found at integration hotspots.

The association of CESC integration loci with super-enhancers and FANCD2-enriched fragile sites at integration hotspots was also addressed (Fig. 6c). This showed that the frequency of FANCD2 enrichment at integration loci was comparable for sites of recurrent (50/313 loci; 16.0%) and non-recurrent integration (48/271 loci; 17.7%) in CESC, whereas the association of super-enhancers was augmented at integration loci that occurred within hotspots (72/313 loci; 23.0%) relative to non-recurrent sites of integration (17/271 loci; 6.3%). Most integration loci that overlapped super-enhancers were active for viral oncogene expression (driver integrations) and were more frequently observed at integration hotspots (Fig. 6d). Furthermore, several cancer driver genes, including *ASXL1*, *CACNA1A*, *IRF6*, *KANSL1*, *KLF5*, *KRT222*, *MYC*, *PPM1D*, *PTCH1*, and *PTPDC1*[56] were located within 1 Mb of super-enhancers that overlapped with integration hotspots (Supplementary Fig. 5 and Supplementary Data Table 17). Alignment of FANCD2-associated fragile sites, super-enhancers, and associated target genes at integration hotspots are shown in Fig. 6e and Supplementary Figs. 4 and 5. Collectively, these data show that transcriptionally active chromatin and/or regions of genetic instability are common features of HPV integration sites. Moreover, integration hotspots are commonly associated with super-enhancers, several of which regulate cancer driver and/or cell-identity genes.

Super-enhancers are frequently amplified at integration hotspots in CESC

The association of super-enhancers at integration hotspots was compared with the host somatic copy number alteration in CESC samples. Super-enhancers were more frequently observed at those CESC integration loci with associated host DNA amplifications (53/240; 22.1%) relative to those that had either a normal genomic profile (16/140; 11.4%) or deletions within the host DNA flanking sequences (1/17; 5.9%) (Fig. 6f). Of the amplified CESC integrations that had associated super-enhancers, 43 (81.1%) were sites of recurrent integration and represented 43/143 (30.1%) hotspot and 10/97 (10.3%) non-hotspot loci with associated host genome amplifications (Fig. 6f). Super-enhancer overlap was also more frequently observed at integration hotspots (11/62; 17.7%) than non-hotspots (5/78; 6.4%) for loci with a normal genomic profile, representing 68.8% of loci that overlapped super-enhancers for this subgroup. However, for integration loci that had associated host deletions, no super-enhancers were observed at sites of recurrent integration (Fig. 6f). This data shows that amplification of super-enhancers is frequently observed at integration loci in CESC, particularly at sites of recurrent integration.

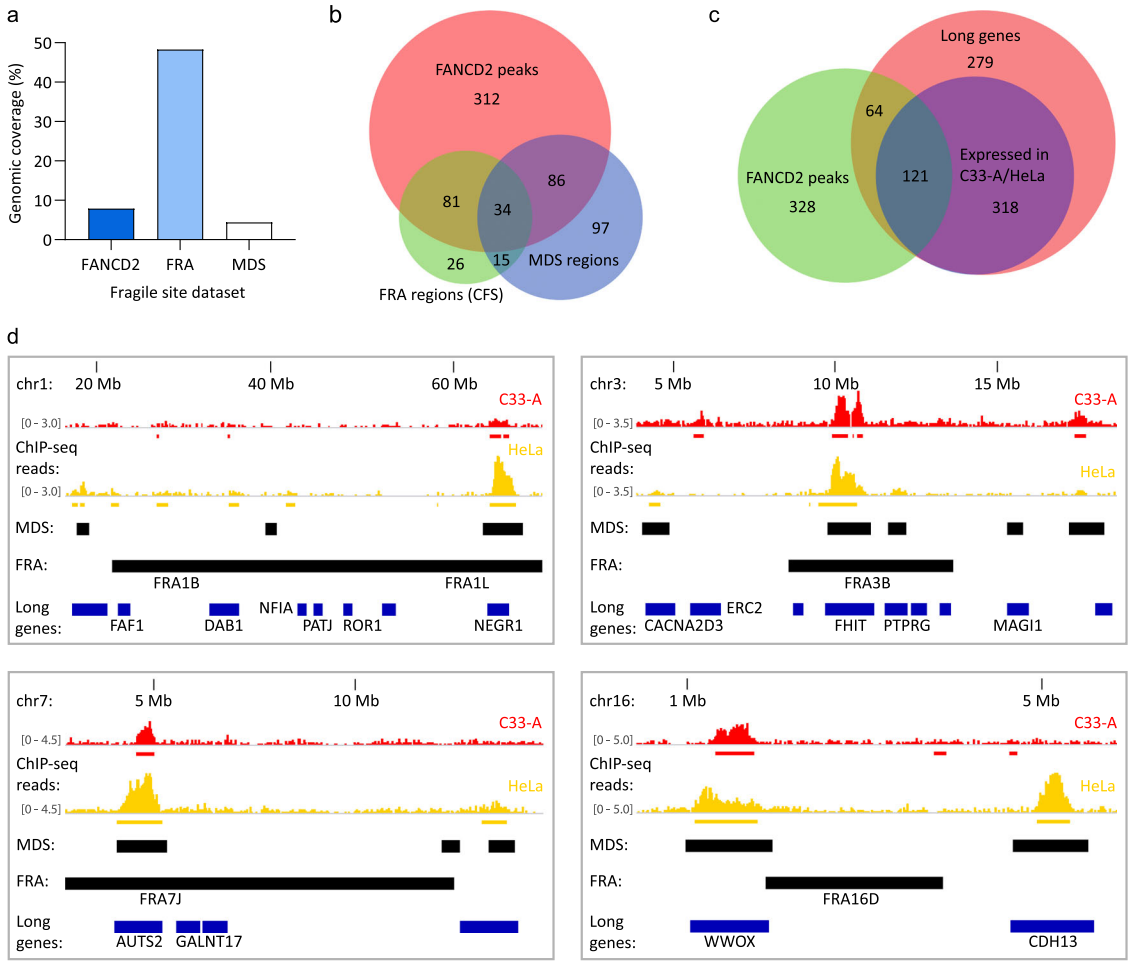


Fig. 5 Cervical keratinocyte-specific fragile sites mapped by FANCD2 ChIP-seq. HPV-negative (C33-A) and HPV18-positive (HeLa) cervical carcinoma cells were treated for 24 h with 0.2 μM aphidicolin and FANCD2-enriched regions were identified by ChIP-seq. All analyses were performed using the combined C33-A and HeLa FANCD2 dataset. **a** Bar graph showing the genomic coverage of FANCD2-enriched regions relative to common fragile sites (FRA regions) and mitotic DNA synthesis (MDS) regions reported in the literature[19,20]. **b** Venn diagram showing the regions of overlap between fragile sites identified in the FANCD2, FRA, and MDS datasets. **c** Venn diagram showing the overlap of FANCD2-associated fragile sites with protein-coding genes longer or equal to 0.3 Mb. Red circle represents all long genes; blue circle represents long genes that are expressed in C33-A and/or HeLa cells. **d** Alignment of FANCD2-enriched regions in C33-A (red) and HeLa (yellow) cells with associated genes (blue bars represent genes >0.3 Mb; genes expressed in C33-A and/or HeLa cells are indicated by gene name), and FRA and MDS regions (black bars). Red and yellow bars below the FANCD2 ChIP-seq signal tracks represent peaks mapped by SICER analysis in the corresponding cell lines. Relative ChIP-seq peak heights are indicated in square parentheses.

DISCUSSION

Many studies have documented the "landscape" of HPV integration sites with respect to traditional common fragile sites, host genome amplification, and transcription and related regulatory elements[9–13,15,25]. Here we combined and curated DNA sequencing and RNA-seq datasets from HPV-positive CESC and HNSCC tumors and compared them with novel "in-house" datasets of common fragile sites defined by FANCD2 ChIP-seq, and enhancers and super-enhancers defined by Brd4 and H3K27ac ChIP-seq in cervical carcinoma-derived cells. We show that viral integration sites in CESC are enriched at FANCD2-associated fragile sites in cervical cells. We also show that cervical cell enhancers are overrepresented at HPV integration sites and that HPV integration is often associated with super-enhancers, particularly at integration hotspots enriched for cell-identity genes. Furthermore, we show that the flanking host DNA that is enriched for enhancers and super-enhancers is frequently amplified in CESC tumors.

HPV genomes replicate as extrachromosomal nuclear minichromosomes at every stage of the infectious cycle. The virus relies on the host replication and transcriptional machinery and it is thought that the HPV genome localizes to regions of the nucleus that facilitate these processes[57]. At different stages of infection, the viral DNA associates with nuclear ND10 bodies and interphase and mitotic host chromatin, and highjacks the DNA damage repair processes to amplify viral DNA[58,59]. Concomitantly, the viral E6 and E7 proteins induce cell proliferation and replication stress, abrogate cell cycle checkpoints, and inhibit the innate immune response[60]. E6 and E7 proteins also modify the epigenetic landscape of the host genome by changing the levels of different histone-modifying enzymes[61]. Together, these activities could promote the accidental integration of viral DNA that is closely associated with host chromatin.

HPV genomes replicate using two different modes: in maintenance replication the genomes replicate bidirectionally at low copy number, but this switches to a unidirectional recombination-directed mechanism in the amplification stage[62,63]. The formation of tandem repeats at integration sites could be related to these processes and over-replication of viral and host sequences could result from repeated initiation of replication at the viral replication origin, especially if the HPV E1 and E2 proteins are expressed[23,64]. In fact, unscheduled firing of replication origins and increased replication fork stalling has been shown to occur in both viral and

Table 1. Overlap of integration breakpoints with FANCD2-associated fragile sites.

Dataset	Integration subgroups	Breakpoints or loci per subgroup (*n*)	Overlap with FANCD2 sites (*n*)	%	*p*-Value
CESC	All breakpoints	1299	229	17.6	**0.0001**
	Single breakpoints	258	39	15.1	**0.0044**
	Clustered breakpoints	1041	190	18.3	**0.0001**
	Condensed clustered breakpoints	326	59	18.1	**0.0001**
	Combined single and condensed breakpoints	584	98	16.8	**0.0001**
HNSCC[a]	All breakpoints	118	21	17.8	**0.0039**
	Single breakpoints	27	2	7.4	0.5011
	Clustered breakpoints	91	19	20.9	**0.0009**
	Condensed clustered breakpoints	30	5	16.7	0.2293
	Combined single and condensed breakpoints	57	7	12.3	0.3700

CESC and HNSCC integration breakpoints (±50 Kb flank regions) were grouped by the number of breakpoints per integration locus; single indicates one breakpoint and clustered indicates two or more breakpoints. Integration breakpoints from each subgroup were intersected with FANCD2-enriched regions and the frequency of overlap calculated. For the "All breakpoints," "Single breakpoints (not clustered)," and "Clustered breakpoints" each breakpoint was tested independently for its overlap with FANCD2-enriched regions, regardless of whether it was part of a cluster or not. For the "Condensed clustered breakpoints" subgroup, the region spanning the most 5′- and 3′-breakpoints of an integration locus was used to test for the overlap with FANCD2-enriched regions. For the "Combined single and condensed breakpoints" subgroup, the "Single breakpoints (not clustered)," and "Condensed clustered breakpoints" subgroups were combined for overlap analysis. The data was permutated 10,000 times to create an expected distribution of overlap. Bold font indicates significant *p*-values.
[a]A single integration breakpoint on chromosome Y was excluded from this analysis.

host sequences at HPV integration sites in the MYC locus[22], which is frequently amplified in HPV-associated cancers. Tandem repeating units of co-amplified viral and cellular DNA could result from this endoreduplication, replication fork arrest, and homologous recombination. Highly rearranged integrations are also consistent with the breakage-fusion-bridge–type model of genome amplification[21]. At fragile sites, perturbed replication dynamics could also generate focal amplifications and/or rearrangements of viral–host sequences.

In this study, we generated a dataset of aphidicolin-induced common fragile sites in two cervical carcinoma cell lines, C33-A and HeLa, and found a significant association between these sites and integration breakpoints in CESC, particularly at those loci with clustered breakpoints. Common fragile sites are susceptible to somatic copy number alterations[65,66] likely due to replication stress that arises from perturbed replication dynamics in conflict with transcription of long genes[67]. Accordingly, our C33-A and HeLa common fragile sites were overrepresented at long genes expressed in these cells. We did not observe an enrichment of HPV integration sites in HNSCC samples with our FANCD2-associated common fragile sites, although an association was previously noted between traditional FRA regions and integration sites in oropharyngeal squamous cell carcinomas[10]. This difference could reflect the larger genomic coverage of FRA regions used in the previous study, or the limited number of samples in our HNSCC dataset. Moreover, common fragile sites are likely distinct in cervical and oropharyngeal derived keratinocytes, or alternatively these findings could reflect differences in the biology of HPV infection and mechanisms of oncogenic progression in the different tissue types.

HPV integration often occurs in transcriptionally active chromatin within the host genome[15,25,68] and we previously described an example of enhancer-hijacking and co-amplification of cellular and viral regulatory sequences at an HPV integration site in cervical lesion derived cells[27]. The association of viral integration breakpoints with putative enhancer regions in HPV-associated cancers has been reported[15], but the enhancer regions used were based on ENCODE histone modifications and therefore did not reflect the specific enhancer profiles of HPV-positive cervical cells. Here, we defined keratinocyte enhancers in HPV16-positive W12 cervical keratinocytes by H3K27ac and Brd4 enrichment and show that these specific enhancers are significantly overrepresented at HPV integration loci. In some cases, these loci were associated with focal amplification of host DNA, providing evidence for potential enhancer-capture. A recent study showed enrichment of active histone marks at HPV integration loci in cervical tumors, which correlated with upregulation of local gene expression, and increased gene expression levels at loci with increased breakpoints[69]. Kamal et al.[38] also found increased local host gene expression at loci with multiple junction copies (analogous to clustered breakpoints). Furthermore, integration loci with associated somatic copy number alterations have also been shown to have increased gene expression[31]. These observations could represent enhancer-hijacking. The three-dimensional interactions between host and viral DNA at integration loci can also perturb local and long-range cellular gene expression[70], and can also result in enhancer-hijacking in HPV-associated cancer[71]. Integration hotspots often contain genes that drive cancer[56] and accidental integration at these regions could result in disruption of proximal and distal regulatory regions, clonal expansion and selection due to perturbation of these oncogenic pathways. Thus, enhancer-hijacking can drive expression of both viral and cellular genes.

Super-enhancers are large clusters of enhancers, rich in Brd4 binding and H3K27ac modification, which often control cell-identity genes and are coopted in tumorigenesis[54,55]. We defined super-enhancers in our W12 cervical cell line datasets and showed that they were strongly associated with integration hotspots, including the *MYC*, *KLF5*/*KLF12*, and *ERBB2* gene loci, which are important regulators of cell cycle, proliferation, and apoptosis[72–74]; *TP63*, which is a master regulator of epidermal keratinocyte proliferation and differentiation[75]; and *RAD51B*, which is a key regulator of homologous recombination repair[76]. We propose that, during persistent infection, extrachromosomal HPV genomes specifically localize at key transcriptional regulatory hubs within the host genome, several of which are important for keratinocyte biology.

The cellular Brd4 protein is involved in many of the cellular and viral processes described in this study. Brd4 is a chromatin scaffold protein that modulates transcriptional initiation and elongation and is a major component of super-enhancers[77]. Brd4 is also important at multiple stages of the HPV infectious cycle[78], binds to

Table 2. Overlap of integration breakpoints with keratinocyte-specific enhancers.

Dataset	Integration subgroups	Breakpoints or loci per subgroup (*n*)	Overlap with enhancers (*n*)	%	*p*-Value
CESC	All breakpoints	1,299	495	38.1	**0.0001**
	Single breakpoints	28	67	26.0	**0.0001**
	Clustered breakpoints	1041	428	41.1	**0.0001**
	Condensed clustered breakpoints	326	162	49.7	**0.0001**
	Combined single and condensed breakpoints	584	229	39.2	**0.0001**
HNSCC[a]	All breakpoints	118	43	36.4	**0.0001**
	Single breakpoints	27	9	33.3	**0.0052**
	Clustered breakpoints	91	34	37.4	**0.0001**
	Condensed clustered breakpoints	30	13	43.3	**0.0011**
	Combined single and condensed breakpoints	57	22	38.6	**0.0003**

CESC and HNSCC integration breakpoints (±50 Kb flank regions) were grouped by the number of breakpoints per integration locus; single indicates one breakpoint and clustered indicates two or more breakpoints. Integration breakpoints from each subgroup were intersected with cellular enhancers and the frequency of overlap calculated. For the "All breakpoints," "Single breakpoints (not clustered)," and "Clustered breakpoints" subgroups, each breakpoint was tested independently for its overlap with enhancer regions, regardless of whether it was part of a cluster or not. For the "Condensed clustered breakpoints" subgroup, the region spanning the most 5′- and 3′-breakpoints of an integration locus was used to test for the overlap with enhancer regions. For the "Combined single and condensed breakpoints" subgroup, the "Single breakpoints (not clustered)" and "Condensed clustered breakpoints" subgroups were combined for overlap analysis. The data were permutated 10,000 times to create an expected distribution of overlap. Bold font indicates significant *p*-values.
[a]A single integration breakpoint on chromosome Y was excluded from this analysis.

common fragile sites in C33-A cells[28], and is important for tethering HPV genomes to mitotic chromatin[79,80]. Brd4 is enriched at the HPV16 integration site/super-enhancer in W12 cells and inhibition of Brd4 binding reduces E6/E7 transcription and cell growth[26]. As such, inhibitors against Brd4 have great therapeutic potential in HPV-associated cancers. Therefore, Brd4 is an example of a factor that is crucial for key cell and viral chromatin-related processes and the juxtaposition of these processes could promote integration of viral DNA and oncogenic progression.

In conclusion, many factors contribute to the integration of a viral genome that eventually drives oncogenesis. Cancer genomes often contain multiple HPV integration sites but usually only one is transcriptionally active (this study and refs. [31,37]). In CESC-derived cells, expression of the viral E6 and E7 oncoproteins is necessary for cell proliferation, survival, and maintenance of the tumor phenotype[81]. Therefore, integration is common, but requires the right genomic location for constitutive viral oncogene expression. For example, the viral oncogenes are usually expressed from a viral–host fusion transcript that requires a splice acceptor and polyadenylation signal in the flanking host DNA[82,83]. Most likely, most integration events do not lead to dysregulated viral oncogene expression and many that do are silenced by DNA methylation[57,84]. Therefore, HPV integrants require a combination of events and processes that are dependent on the genetic and/or epigenetic landscape of the flanking host chromatin to drive oncogenesis.

METHODS

HPV integration datasets

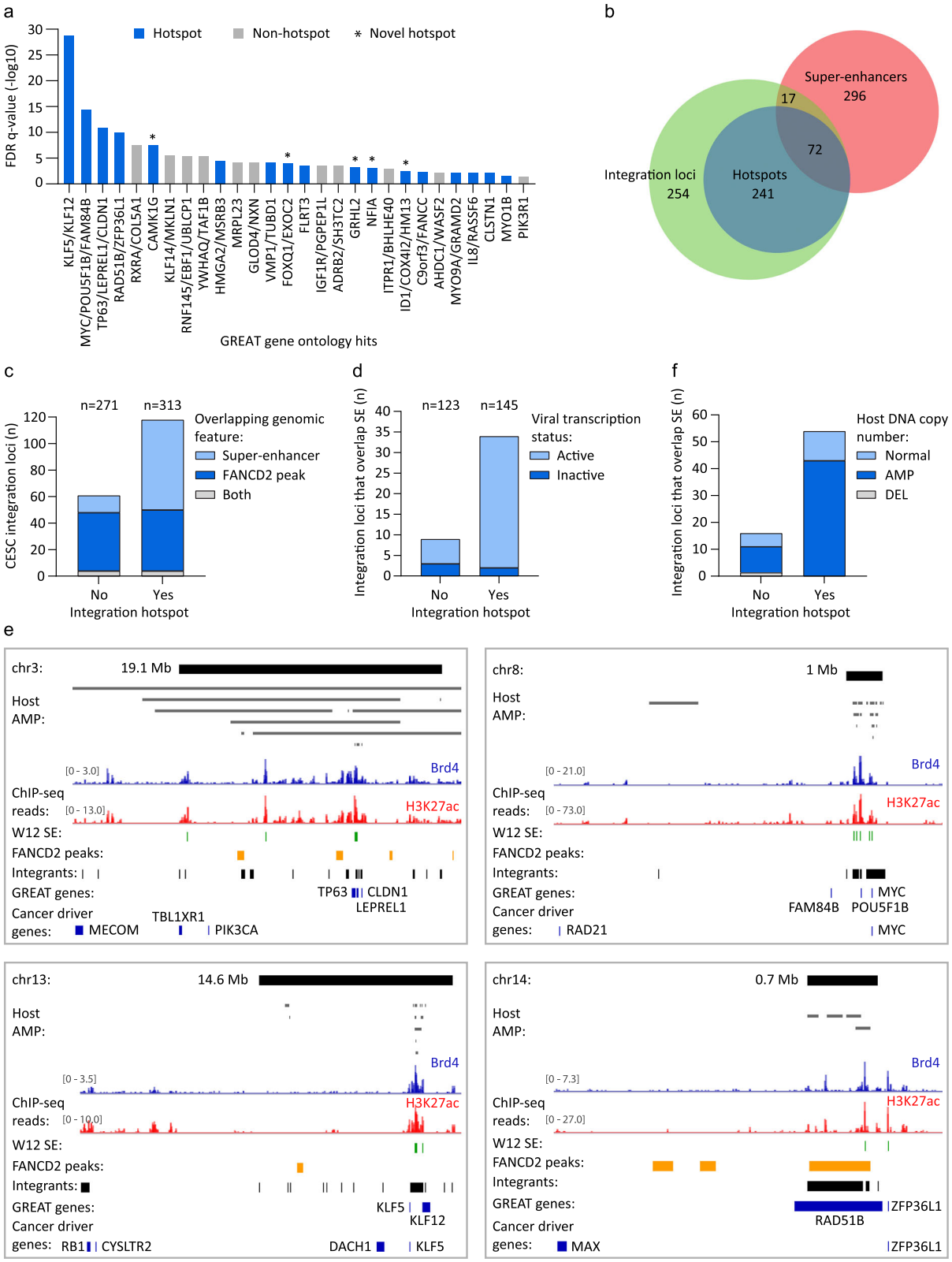
A systematic literature review identified genomic datasets from HPV-positive CESC and HNSCC that contained information on HPV type and integration breakpoints within the host and viral genomes, which were identified by sequencing (Supplementary Data Table 1)[5,6,8,9,30–37]. The integration breakpoints used in this study were originally identified from both RNA-seq and DNA sequencing methods, including APOT (amplification of papillomavirus oncogene transcripts), DIPS (detection of integrated papillomavirus sequences), and next-generation sequencing technologies. The use of hybrid-capture technologies for detection of viral integration sites has been reported to give high rates of false positives[35,36] and so insertion breakpoints identified by this method were only included if they were validated by other means, such as Sanger sequencing. The methodology used to identify each integration breakpoint is referenced in Supplementary Data Tables 2 and 3. RNA-based sequencing methods give an approximation of the insertion site based on the closest splice acceptor site within the host genome; therefore, integration breakpoints identified by RNA-seq and/or APOT were only used to determine the viral transcription status of an integration site for samples with matched DNA sequencing data. For The Cancer Genome Atlas (TCGA) CESC and HNSCC samples, unmapped reads were extracted from RNA-seq, whole genome sequencing (WGS), and whole exome sequencing BAM files (https://portal.gdc.cancer.gov/legacy-archive; accessed 01/01/2013) and pre-processed with prinseq-lite.pl version 0.20.2 to remove low-quality reads[85]. Pre-processed reads were mapped with Bowtie 2 using the very-sensitive preset option against the Viral Refseq database[86,87]. All unmapped reads were subjected to BLASTN with default parameters against the Viral Refseq database[88]. All aligned reads were then subjected to BLASTN against the human hg19 reference assembly. Bowtie 2 and BLASTN reports were passed into SummonChimera using a 1000 bp deletion size for integration detection[89]. The SummonChimera reports were manually parsed to remove chimeric junctions with lower than 20 read coverage, chimeric junctions with no cross-analysis verification, and ambiguously reported integration predictions. Finally, a unique ID was provided to all uniquely detected chimeric junctions. For analysis of the association of integration breakpoints with different genomic features of interest, only integration breakpoints identified by DNA sequencing methods were used. The characteristics of samples included in this study, categorized by histology type, HPV type, tumor location, and sequencing methods are summarized in Supplementary Fig. 1. CESC and HNSCC integration breakpoints included in this study are listed in Supplementary Data Tables 2 and 3.

Integration hotspot dataset

Integration loci from CESC tumors that were within 5 Mb of each other were collapsed into a single genomic interval to define integration hotspot boundaries. Exceptions to this size cutoff for collapsing adjacent integration loci were permitted to reflect previously defined hotspots from the literature. Five or more integration loci per hotspot (or three or more integration loci for sites that overlapped previously defined hotspots from the literature) were used to define sites of recurrent integration. Integration hotspots defined from our CESC dataset and previously in the literature are listed in Supplementary Data Tables 4 and 5, respectively.

Somatic copy number alteration datasets

The amplification status of the cellular sequences flanking integration breakpoints had been assessed in a subset of samples by comparative genomic hybridization (CGH) or SNP array datasets. For CGH array data, we defined twofold or more focal amplification or deletion of the host genome at an integration locus as having an associated somatic copy



number alteration[8,30]. Matched SNP6 copy number segment data for TCGA CESC and HNSCC tumor samples were downloaded from the Broad Institute on 04/07/2020, http://firebrowse.org/ (genome_wide_snp_6-segmented_scna_minus_germline_cnv_ hg19)[90,91]. Positive and negative mean segment values above 0.3 and below −0.3 represented copy number gains and losses, respectively, and mean segment values between 0.3 and −0.3 were considered noise[90,91]. Deletions that occurred within 50 Kb of an integration breakpoint were included in this analysis. Deletions that directly overlapped an integration breakpoint were excluded; these likely reflected the alternative chromosome as sequencing of the chimeric viral–host junction was available for the associated HPV insertion sites. Integration loci defined as having associated somatic copy number

Fig. 6 Integration hotspots are associated with cellular super-enhancers. H3K27ac- and Brd4-enriched regions were profiled in HPV16-positive cervical derived W12 keratinocyte subclones by ChIP-seq. Enhancer regions were defined as peaks that overlapped in both H3K27ac and Brd4 datasets, and that were identified across the four W12 subclones. **a** GREAT (Genomic Regions Enrichment of Annotations Tool) Gene Ontology analysis was performed using W12 enhancers that overlapped CESC integration breakpoints (±50 Kb flanks) as input and compared against all W12 enhancers, to identify putative target genes associated with these *cis*-regulatory regions based on enhancer frequency. Bars represent putative target genes plotted against their FDR (false discovery rate) adjusted *p*-values (*q*-value). Blue and gray bars represent genes that overlap integration hotspots and sites of non-recurrent integration, respectively. Enriched target genes within the same genomic locus were grouped (e.g., KLF5 and KLF12) and plotted using the most significant *q*-value. **b** Venn diagram showing the regions of overlap between integration loci, integration hotspots, and super-enhancers mapped in W12 subclones. **c** Bar chart showing the number of CESC integration loci that were grouped according to whether or not they are integration hotspots and plotted based on their overlap with super-enhancers, FANCD2-associated fragile sites, or both genomic features. Numbers above the graph indicate the total number of integration loci within each grouping. **d** Bar chart showing the number of CESC integration loci that are associated with super-enhancers (SE) that were grouped according to whether they are integration hotspots and plotted based on their viral transcription status. Numbers above the graph indicate the total number of integration loci that have associated viral transcription data within each grouping. **e** Alignment of Brd4 (blue) and H3K27ac (red) ChIP-seq signals mapped in W12 cervical keratinocytes at integration hotspots (top black bars; size indicated in Mb) in cervical carcinomas. Relative ChIP-seq peak heights are indicated in square parentheses. Gray bars represent amplified (AMP) host DNA in different CESC tumors from The Cancer Genome Atlas. Green, yellow, and black bars below the ChIP-seq signal tracks represent super-enhancers (SE) mapped in W12 subclones, FANCD2-associated fragile sites mapped in C33-A and HeLa cells, and CESC integration loci, respectively. Genes identified from GREAT Gene Ontology analysis[53] and cancer driver genes[56] are indicated by blue bars. Each integration hotspot is characterized in Supplementary Fig. 5 and Supplementary Data Table 4. **f** Bar chart showing the number of CESC integration loci that are associated with super-enhancers (SE) that were grouped according to whether they are integration hotspots and plotted based on their host somatic copy number alternation status. The number of integration loci per grouping was normal, $n = 140$, amplification (AMP), $n = 240$, and deletion (DEL), $n = 17$.

alterations in CESC and HNSCC are detailed in Supplementary Data Tables 6 and 7, respectively.

Cell culture

Subclones (20831, 20861, 20862, and 20863) derived from HPV16-positive W12 cervical keratinocytes[82,92] (a gift from Dr. Paul Lambert, McArdle Laboratory for Cancer Research, WI, USA) were maintained in F-medium (3 : 1 [vol/vol] F-12–Dulbecco's modified Eagle's medium, 5% fetal bovine serum, 0.4 µg/ml hydrocortisone, 5 µg/ml insulin, 8.4 ng/ml cholera toxin, 10 ng/ml epidermal growth factor, 24 µg/ml adenine, 100 U/ml penicillin, and 100 µg/ml streptomycin). All cells were grown in the presence of irradiated 3T3-J2 feeder cells. C33-A and HeLa cervical carcinoma-derived cell lines were purchased from ATCC and maintained in Dulbecco's modified Eagle's medium, supplemented with 10% fetal bovine serum, 100 U/ml penicillin, and 100 µg/ml streptomycin. To induce replication stress, C33-A and HeLa cells were treated for 24 h with 0.2 µM aphidicolin (Sigma A0781) prior to collecting for FANCD2 ChIP-seq experiments, described below.

ChIP-seq: FANCD2

Aphidicolin-treated C33-A and HeLa cells were processed for ChIP as previously described[26]. Briefly, cells were crossed-linked with 1% formaldehyde and chromatin was isolated and sheared to 100–500 bp DNA fragments using a Bioruptor sonicator (Diagonode) on high power settings. Chromatin samples (25 µg per ChIP) were incubated overnight at 4 °C with an antibody against FANCD2 (Bethyl, A302–174A, 2.5 µg). Rabbit IgG (Jackson ImmunoRes, 011-000-003) was used to determine nonspecific binding to control regions (although not sequenced). Chromatin immunocomplexes were precipitated for 1 h at 4 °C with blocked Dynabeads Protein G (Invitrogen), subjected to multiple wash steps and the chromatin eluted in elution buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA pH 8.0, 1% SDS). Chromatin was reverse cross-linked overnight at 65 °C in 0.2 M NaCl, followed by RNase A and proteinase K treatment, and the DNA purified using the ChIP DNA Clean & Concentrator kit (Zymo Research). ChIP DNA from two biological replicates were pooled and subjected to 2 × 150 bp paired-end read sequencing on the Illumina HiSeq-4000 platform (Genomics Resource Center, Institute for Genome Sciences, University of Maryland) to a sequencing depth of >14 million reads per sample. FANCD2 ChIP-seq datasets are accessible through GEO Series accession number GSE183048.

ChIP-seq: Brd4 and H3K27ac

W12 chromatin samples were isolated as described above, and have been described previously[27]. However, only the sequence data flanking the HPV16 integration sites was previously analyzed and published. Here we analyze the same dataset but for the entire human genome. Antibodies used were Brd4 (Bethyl Laboratories A301-985A, 3 µg) or H3K27ac (Millipore 07–360, 3 µl). No antibody controls were included to monitor nonspecific binding. Brd4 and H3K27ac ChIP-seq datasets are accessible through GEO Series accession number GSE183048.

ChIP-seq processing and peak calling

Reads were trimmed with Cutadapt version 1.18[93]. All reads aligning to the ENCODE hg19 v1 blacklist regions[94] were identified by alignment with BWA version 0.7.17[95] and removed with Picard SamToFastq, https://broadinstitute.github.io/picard/. Remaining reads were aligned to an hg19 reference genome using BWA. Reads with a mapQ score less than 6 were removed with SAMtools version 1.6[96] and PCR duplicates were removed with Picard MarkDuplicates. Peaks were called by comparing each ChIP sample to its matching input sample. For FANCD2, the mean fragment size was estimated by Phantompeakqualtools version 2.0[97]. Peaks were called using SICER version 1.1[98] with the following parameters: redundancy threshold of 100, effective genome fraction of 0.75, window size of 25,000 bp, and gap size of 50,000 bp. H3K27ac and Brd4 peaks were called using macsBroad (macs version 2.1.1 from 09/03/2016)[99] with the following parameters: -broad-cutoff 0.01 -f "BAMPE". Data was converted into bigwigs for viewing and normalized by reads per genomic content (RPGC) using deepTools version 3.0.1[100] using the following parameters: -binSize 25 -smoothLength 75 -effectiveGenomeSize 2700000000 -centerReads -normalizeUsing RPGC. RPGC-normalized input values were subtracted from RPGC-normalized ChIP values of matching cell-type genome-wide using Deeptools with -binSize 25.

FANCD2-associated fragile-site dataset

FANCD2 ChIP-seq peaks were filtered by a −log10 *q*-value of 10 or above, to remove low-confidence calls. Filtered C33-A ChIP-seq peaks were combined with previously mapped aphidicolin-induced FANCD2 peaks identified by ChIP-chip in these cells[28]. C33-A (Supplementary Data Table 8) and HeLa FANCD2-enriched regions (Supplementary Data Table 9) were combined and overlapping peaks merged using bedtools MergeBED[101]. Association of ChIP peaks between the three FANCD2 datasets were determined by permutation testing using regioneR[102]. Combined FANCD2 peaks from C33-A and HeLa are listed in Supplementary Data Table 10. FANCD2-enriched regions were compared to aphidicolin-induced common fragile sites characterized in lymphoblast cells (FRA regions) and mitotic DNA synthesis regions characterized in HeLa cells[19,20], which are listed in Supplementary Data Tables 11 and 12, respectively, using regioneR[102]. FRA regions were downloaded from the HGNC (HUGO Gene Nomenclature Committee) database (https://www.genenames.org/download/custom/) on 27/08/2020 using advanced filtering: *gd_locus_type = 'fragile site'*.

Overlap analysis of FANCD2-enriched regions with long genes

The Gencode Release 19 human reference genome (GRCh37) was filtered for protein-coding genes greater than or equal to 0.3 Mb in length, including untranslated regions (Supplementary Data Table 13), and used to determine the overlap with FANCD2-enriched regions.

Enhancer dataset

Consensus peak sets for H3K27ac were defined as overlapping regions found in at least four out of eight W12 samples using DiffBind[103,104]. Enhancers were defined as genomic intervals that overlapped between H3K27ac peaks and Brd4 peaks, and are listed in Supplementary Data Table 14. Proximal and distal enhancer-like *cis*-Regulatory Elements by ENCODE for NHEK and HeLa cells were downloaded from https://screen.encodeproject.org/# (accessed September 2020)[52]. ENCODE GRCh38 enhancer files were converted to hg19 using the UCSC Genome Browser LiftOver tool, http://genome.ucsc.edu/cgi-bin/hgLiftOver. W12 enhancers were compared to ENCODE NHEK enhancers using regioneR[102].

Overlap analysis of integration breakpoints with fragile sites and enhancers

The intersect between the genomic coordinates of HPV integration breakpoints (±50 Kb flank regions[7,9]) with enhancers (Supplementary Data Table 14) or FANCD2-enriched regions (Supplementary Data Table 10) was analyzed using the *Overlapping Pieces of Intervals* function in the Galaxy genomics platform (https://usegalaxy.org/). Samples with multiple reported breakpoints within the same chromosome were classified as a single integration locus if the 5′ and 3′ most breakpoints were within 3 Mb of each other. This 3 Mb cutoff was based on manual analysis of the distance between clustered breakpoints identified by WGS and/or hybrid-capture technologies for the CESC and HNSCC datasets. Each integration locus was assigned a unique integration ID so that the number of breakpoints per integration could be categorized as a cluster. Each integration breakpoint was analyzed independently for their association with the genomic feature of interest, as well as with adjacent HPV breakpoints, which were classified as belonging to the same integration locus/cluster. For significance testing, the data were permutated 10,000 times to create an expected distribution of the overlap between integration breakpoints and loci with FANCD2-enriched regions or enhancers using regioneR[102].

Super-enhancer dataset

Super-enhancers were defined in the Brd4 and H3K27ac W12 ChIP-seq datasets using the ROSE tool, using default parameters[54,55]. Enhancers defined in W12 cells (Supplementary Data Table 14) were used as the input list of enhancers. Super-enhancers were defined by Brd4 and/or H3K27ac consensus peaks that mapped in at least six out of 12 W12 samples for the 20831, 20862, and 20863 subclones, and are listed in Supplementary Data Table 16. Super-enhancers mapped in the 20861 subclone were excluded as they were masked by the amplified viral–host derived super-enhancer-like element at the HPV integration site in these cells[27].

Gene Ontology analysis of W12 enhancers

W12 enhancers that overlapped with CESC integration breakpoints (±50 Kb flanks) were analyzed using the GREAT using default parameters, accessed July 2021, http://great.stanford.edu/public/html/. All enhancers profiled in W12 cells (Supplementary Data Table 14) were used as the input list of background regions.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The FANCD2 and Brd4 ChIP-seq data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus[105] and are accessible through GEO Series accession number GSE183048 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=%20GSE183048). TCGA datasets analyzed in this study are accessible from the database of Genotypes and Phenotypes (dbGaP), accession number phs000178[31,34]. SNP6 copy number segment data for TCGA datasets are accessible from http://firebrowse.org/[90,91]. The aggregate data analyzed in this study are available from the corresponding author on reasonable request.

CODE AVAILABILITY

All software used in this study is detailed in the "Methods" section. Collapsing of nearby HPV integration sites into a single genomic interval was performed in R. Permutation testing of HPV integration sites with different genomic features was implemented using regioneR[102]. Source codes are available at https://github.com/NCBR-FNLCR/McBride_Integration_Analysis.

Received: 30 August 2021; Accepted: 28 October 2021;
Published online: 30 November 2021

REFERENCES

1. Viens, L. J. et al. Human papillomavirus-associated cancers - United States, 2008-2012. *MMWR Morb. Mortal. Wkly Rep.* **65**, 661–666 (2016).
2. Alazawi, W. et al. Genomic imbalances in 70 snap-frozen cervical squamous intraepithelial lesions: associations with lesion grade, state of the HPV16 E2 gene and clinical outcome. *Br. J. Cancer* **91**, 2063–2070 (2004).
3. Peter, M. et al. Frequent genomic structural alterations at HPV insertion sites in cervical carcinoma. *J. Pathol.* **221**, 320–330 (2010).
4. Thomas, L. K. et al. Chromosomal gains and losses in human papillomavirus-associated neoplasia of the lower genital tract - a systematic review and meta-analysis. *Eur. J. Cancer* **50**, 85–98 (2014).
5. Parfenov, M. et al. Characterization of HPV and host genome interactions in primary head and neck cancers. *Proc. Natl Acad. Sci. USA* **111**, 15544–15549 (2014).
6. Ojesina, A. I. et al. Landscape of genomic alterations in cervical carcinomas. *Nature* **506**, 371–375 (2014).
7. Akagi, K. et al. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res.* **24**, 185–199 (2014).
8. Holmes, A. et al. Mechanistic signatures of HPV insertions in cervical carcinomas. *NPJ Genom. Med.* **1**, 16004 (2016).
9. Bodelon, C. et al. Chromosomal copy number alterations and HPV integration in cervical precancer and invasive cancer. *Carcinogenesis* **37**, 188–196 (2016).
10. Gao, G. et al. Common fragile sites (CFS) and extremely large CFS genes are targets for human papillomavirus integrations and chromosome rearrangements in oropharyngeal squamous cell carcinoma. *Genes Chromosomes Cancer* **56**, 59–74 (2017).
11. Thorland, E. C., Myers, S. L., Gostout, B. S. & Smith, D. I. Common fragile sites are preferential targets for HPV16 integrations in cervical tumors. *Oncogene* **22**, 1225–1237 (2003).
12. Thorland, E. C. et al. Human papillomavirus type 16 integrations in cervical tumors frequently occur in common fragile sites. *Cancer Res.* **60**, 5916–5921 (2000).
13. Smith, P. P., Friedman, C. L., Bryant, E. M. & McDougall, J. K. Viral integration and fragile sites in human papillomavirus-immortalized human keratinocyte cell lines. *Genes Chromosomes Cancer* **5**, 150–157 (1992).
14. Le Tallec, B. et al. Updating the mechanisms of common fragile site instability: how to reconcile the different views? *Cell. Mol. Life Sci.* **71**, 4489–4494 (2014).
15. Bodelon, C., Untereiner, M. E., Machiela, M. J., Vinokurova, S. & Wentzensen, N. Genomic characterization of viral integration sites in HPV-related cancers. *Int J. Cancer* **139**, 2001–2011 (2016).
16. Madireddy, A. et al. FANCD2 facilitates replication through common fragile sites. *Mol. Cell* **64**, 388–404 (2016).
17. Pentzold, C. et al. FANCD2 binding identifies conserved fragile sites at large transcribed genes in avian cells. *Nucleic Acids Res.* **46**, 1280–1294 (2018).
18. Chan, K. L., Palmai-Pallag, T., Ying, S. & Hickson, I. D. Replication stress induces sister-chromatid bridging at fragile site loci in mitosis. *Nat. Cell Biol.* **11**, 753–760 (2009).
19. Ji, F. et al. Genome-wide high-resolution mapping of mitotic DNA synthesis sites and common fragile sites by direct sequencing. *Cell Res* https://doi.org/10.1038/s41422-020-0357-y (2020).
20. Macheret, M. et al. High-resolution mapping of mitotic DNA synthesis regions and common fragile sites in the human genome through direct sequencing. *Cell Res.* https://doi.org/10.1038/s41422-020-0358-x (2020).
21. Herrick, J. et al. Genomic organization of amplified MYC genes suggests distinct mechanisms of amplification in tumorigenesis. *Cancer Res.* **65**, 1174–1179 (2005).
22. Conti, C., Herrick, J. & Bensimon, A. Unscheduled DNA replication origin activation at inserted HPV 18 sequences in a HPV-18/MYC amplicon. *Genes Chromosomes Cancer* **46**, 724–734 (2007).

23. Kadaja, M., Isok-Paas, H., Laos, T., Ustav, E. & Ustav, M. Mechanism of genomic instability in cells infected with the high-risk human papillomaviruses. *PLoS Pathog.* **5**, e1000397 (2009).
24. Coquelle, A., Pipiras, E., Toledo, F., Buttin, G. & Debatisse, M. Expression of fragile sites triggers intrachromosomal mammalian gene amplification and sets boundaries to early amplicons. *Cell* **89**, 215–225 (1997).
25. Christiansen, I. K., Sandve, G. K., Schmitz, M., Durst, M. & Hovig, E. Transcriptionally active regions are the preferred targets for chromosomal HPV integration in cervical carcinogenesis. *PLoS ONE* **10**, e0119566 (2015).
26. Dooley, K. E., Warburton, A. & McBride, A. A. Tandemly integrated HPV16 can form a Brd4-dependent super-enhancer-like element that drives transcription of viral oncogenes. *MBio* **7** https://doi.org/10.1128/mBio.01446-16 (2016).
27. Warburton, A. 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.* **14**, e1007179 (2018).
28. Jang, M. K., Shen, K. & McBride, A. A. Papillomavirus genomes associate with BRD4 to replicate at fragile sites in the host genome. *PLoS Pathog.* **10**, e1004117 (2014).
29. Stanley, M. A., Browne, H. M., Appleby, M. & Minson, A. C. Properties of a non-tumorigenic human cervical keratinocyte cell line. *Int. J. Cancer* **43**, 672–676 (1989).
30. Hu, Z. 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.* **47**, 158–163 (2015).
31. Cancer Genome Atlas Research Network et al. Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**, 378–384 (2017).
32. Koneva, L. A. et al. HPV integration in HNSCC correlates with survival outcomes, immune response signatures, and candidate drivers. *Mol. Cancer Res.* **16**, 90–102 (2018).
33. Olthof, N. C. 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* **9**, e88718 (2014).
34. Cancer Genome Atlas Network Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).
35. Liu, Y., Lu, Z., Xu, R. & Ke, Y. Comprehensive mapping of the human papillomavirus (HPV) DNA integration sites in cervical carcinomas by HPV capture technology. *Oncotarget* **7**, 5852–5864 (2016).
36. Liu, Y. et al. Genome-wide profiling of the human papillomavirus DNA integration in cervical intraepithelial neoplasia and normal cervical epithelium by HPV capture technology. *Sci. Rep.* **6**, 35427 (2016).
37. Xu, B. et al. Multiplex identification of human papillomavirus 16 DNA integration sites in cervical carcinomas. *PLoS ONE* **8**, e66693 (2013).
38. Kamal, M. et al. Human papilloma virus (HPV) integration signature in cervical cancer: identification of MACROD2 gene as HPV hot spot integration site. *Br. J Cancer* https://doi.org/10.1038/s41416-020-01153-4 (2020).
39. Li, W. et al. Characteristic of HPV integration in the genome and transcriptome of cervical cancer tissues. *Biomed. Res. Int.* **2018**, 6242173–6242173 (2018).
40. Kraus, I. et al. The majority of viral-cellular fusion transcripts in cervical carcinomas cotranscribe cellular sequences of known or predicted genes. *Cancer Res.* **68**, 2514–2522 (2008).
41. Schmitz, M., Driesch, C., Jansen, L., Runnebaum, I. B. & Dürst, M. Non-random integration of the HPV genome in cervical cancer. *PLoS ONE* **7**, e39632 (2012).
42. Zhang, R. et al. Dysregulation of host cellular genes targeted by human papillomavirus (HPV) integration contributes to HPV-related cervical carcinogenesis. *Int. J. Cancer* **138**, 1163–1174 (2016).
43. Landry, J. J. et al. The genomic and transcriptomic landscape of a HeLa cell line. *G3 (Bethesda)* **3**, 1213–1224 (2013).
44. Helmrich, A., Ballarino, M. & Tora, L. Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol. Cell* **44**, 966–977 (2011).
45. Okamoto, Y. et al. Replication stress induces accumulation of FANCD2 at central region of large fragile genes. *Nucleic Acids Res.* **46**, 2932–2944 (2018).
46. Brison, O. et al. Transcription-mediated organization of the replication initiation program across large genes sets common fragile sites genome-wide. *Nat. Commun.* **10**, 5693 (2019).
47. Lee, J.-E. et al. Brd4 binds to active enhancers to control cell identity gene induction in adipogenesis and myogenesis. *Nat. Commun.* **8** https://doi.org/10.1038/s41467-017-02403-5 (2017).
48. Brown, J. D. et al. BET bromodomain proteins regulate enhancer function during adipogenesis. *Proc. Natl Acad. Sci. USA* **115**, 2144–2149 (2018).
49. Najafova, Z. et al. BRD4 localization to lineage-specific enhancers is associated with a distinct transcription factor repertoire. *Nucleic Acids Res.* **45**, 127–141 (2017).
50. Jang, M. K., Kwon, D. & McBride, A. A. Papillomavirus E2 proteins and the host BRD4 protein associate with transcriptionally active cellular chromatin. *J. Virol.* **83**, 2592–2600 (2009).
51. Helfer, C. M., Yan, J. & You, J. The cellular bromodomain protein Brd4 has multiple functions in E2-mediated papillomavirus transcription activation. *Viruses* **6**, 3228–3249 (2014).
52. Consortium, E. P. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
53. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
54. Whyte, WarrenA. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
55. Lovén, J. et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013).
56. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e318 (2018).
57. Warburton, A., Della Fera, A. N. & McBride, A. A. Dangerous liaisons: long-term replication with an extrachromosomal HPV genome. *Viruses* **13**, 1864 (2021).
58. Moody, C. A. Impact of replication stress in human papillomavirus pathogenesis. *J. Virol.* **93** https://doi.org/10.1128/JVI.01012-17 (2019).
59. Anacker, D. C. & Moody, C. A. Modulation of the DNA damage response during the life cycle of human papillomaviruses. *Virus Res.* **231**, 41–49 (2017).
60. Schiffman, M. et al. Carcinogenic human papillomavirus infection. *Nat. Rev. Dis. Prim.* **2**, 16086 (2016).
61. Mac, M. & Moody, C. A. Epigenetic regulation of the human papillomavirus life cycle. *Pathogens* **9** https://doi.org/10.3390/pathogens9060483 (2020).
62. Flores, E. R. & Lambert, P. F. Evidence for a switch in the mode of human papillomavirus type 16 DNA replication during the viral life cycle. *J. Virol.* **71**, 7167–7179 (1997).
63. Sakakibara, N., Chen, D. & McBride, A. A. Papillomaviruses use recombination-dependent replication to vegetatively amplify their genomes in differentiated cells. *PLoS Pathog.* **9**, e1003321 (2013).
64. Kadaja, M. et al. Genomic instability of the host cell induced by the human papillomavirus replication machinery. *EMBO J.* **26**, 2180–2191 (2007).
65. Myllykangas, S. et al. DNA copy number amplification profiling of human neoplasms. *Oncogene* **25**, 7324–7332 (2006).
66. Hellman, A. et al. A role for common fragile site induction in amplification of human oncogenes. *Cancer Cell* **1**, 89–97 (2002).
67. Wilson, T. E. et al. Large transcription units unify copy number variants and common fragile sites arising under replication stress. *Genome Res.* **25**, 189–200 (2015).
68. Kelley, D. Z. et al. Integrated analysis of whole-genome ChIP-Seq and RNA-Seq data of primary head and neck tumor samples associates HPV integration sites with open chromatin marks. *Cancer Res.* **77**, 6538–6550 (2017).
69. Gagliardi, A. et al. Analysis of Ugandan cervical carcinomas identifies human papillomavirus clade-specific epigenome and transcriptome landscapes. *Nat. Genet.* **52**, 800–810 (2020).
70. Groves, I. J. 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.* **17**, e1009875 (2021).
71. Cao, C. 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* **47**, 437–450 (2020).
72. Schmidt, E. V. The role of c-myc in cellular growth control. *Oncogene* **18**, 2988–2996 (1999).
73. Holbro, T. The ErbB receptors and their role in cancer progression. *Exp. Cell Res.* **284**, 99–110 (2003).
74. Dong, J. T. & Chen, C. Essential role of KLF5 transcription factor in cell proliferation and differentiation and its implications for human diseases. *Cell Mol. Life Sci.* **66**, 2691–2706 (2009).
75. Soares, E. & Zhou, H. Master regulatory role of p63 in epidermal development and disease. *Cell Mol. Life Sci.* **75**, 1179–1190 (2018).
76. Takata, M. et al. The Rad51 paralog Rad51B promotes homologous recombinational repair. *Mol. Cell Biol.* **20**, 6476–6482 (2000).
77. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
78. McBride, A. A., Warburton, A. & Khurana, S. Multiple roles of Brd4 in the infectious cycle of human papillomaviruses. *Front. Mol. Biosci.* **8** https://doi.org/10.3389/fmolb.2021.725794 (2021).
79. You, J., Croyle, J. L., Nishimura, A., Ozato, K. & Howley, P. M. Interaction of the bovine papillomavirus E2 protein with Brd4 tethers the viral DNA to host mitotic chromosomes. *Cell* **117**, 349–360 (2004).
80. Baxter, M. K., McPhillips, M. G., Ozato, K. & McBride, A. A. The mitotic chromosome binding activity of the papillomavirus E2 protein correlates with interaction with the cellular chromosomal protein, Brd4. *J. Virol.* **79**, 4806–4818 (2005).
81. Goodwin, E. C. & DiMaio, D. Repression of human papillomavirus oncogenes in HeLa cervical carcinoma cells causes the orderly reactivation of dormant tumor suppressor pathways. *Proc. Natl Acad. Sci. USA* **97**, 12513–12518 (2000).

82. Jeon, S. & Lambert, P. F. Integration of human papillomavirus type 16 DNA into the human genome leads to increased stability of E6 and E7 mRNAs: implications for cervical carcinogenesis. *Proc. Natl Acad. Sci. USA* **92**, 1654–1658 (1995).
83. Wentzensen, N. et al. Characterization of viral-cellular fusion transcripts in a large series of HPV16 and 18 positive anogenital lesions. *Oncogene* **21**, 419–426 (2002).
84. Chaiwongkot, A. et al. Differential methylation of E2 binding sites in episomal and integrated HPV 16 genomes in preinvasive and invasive cervical lesions. *Int. J. Cancer* **132**, 2087–2094 (2013).
85. Schmieder, R. E. R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
86. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
87. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**, 421–432 (2019).
88. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, W20–W25 (2004).
89. Katz, J. P. & Pipas, J. M. SummonChimera infers integrated viral genomes with nucleotide precision from NGS data. *BMC Bioinformatics* **15**, 348 (2014).
90. Harvard, B. I. o. M. a. Broad Institute TCGA Genome Data Analysis Center (2016): SNP6 copy number analysis (GISTIC2); cervical squamous cell carcinoma and endocervical adenocarcinoma. https://doi.org/10.7908/C16D5SCD (2016).
91. Harvard, B. I. o. M. a. Broad Institute TCGA Genome Data Analysis Center (2016): SNP6 copy number analysis (GISTIC2); head and neck squamous cell carcinoma. https://doi.org/10.7908/C1V987FP (2016).
92. Jeon, S., Allen-Hoffmann, B. L. & Lambert, P. F. Integration of human papillomavirus type 16 into the human genome correlates with a selective growth advantage of cells. *J. Virol.* **69**, 2989–2997 (1995).
93. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17** https://doi.org/10.14806/ej.17.1.200 (2011).
94. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
95. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
96. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
97. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351–1359 (2008).
98. Xu, S., Grullon, S., Ge, K. & Peng, W. Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. *Methods Mol. Biol.* **1150**, 97–111 (2014).
99. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
100. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
101. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
102. Gel, B. et al. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32**, 289–291 (2016).
103. Stark, R. & Brown, G. D. DiffBind: differential binding analysis of ChIP-Seq peak data (2011).
104. Ross-Innes, C. S. et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
105. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).

ACKNOWLEDGEMENTS

We thank Justin Lack (NIH/NCBR/NCI/NIAID/FNL) for advice on statistical analyses and Susan Huse (NIH/NCBR/NCI/NIAID/FNL) for preliminary analysis of enhancer-capture at integration sites. The results published here are in part based upon data generated by the TCGA Research Network, https://www.cancer.gov/tcga. This work was funded by the Intramural Research Program of NIAID, NIH grant number ZIA AI001223 LVD, and NIH grant number AI153156 (J.M.P.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

A.A.M. supervised the project. A.A.M. and A.W. conceived and designed the study. A.W. compiled datasets and performed ChIP-seq experiments. T.E.M. processed ChIP-seq data. T.E.M. designed and advised on statistical analyses. A.W. and T.E.M. performed statistical analysis. A.A.M., A.W., and T.E.M. analyzed the data. J.P. and J.K. identified integration breakpoints from CESC and HNSCC TCGA datasets. A.A.M. and A.W. wrote the manuscript. All authors discussed, critically revised, and approved the final version of the article for publication.

FUNDING

Open Access funding provided by the National Institutes of Health (NIH)

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41525-021-00264-y.

Correspondence and requests for materials should be addressed to Alison A. McBride.

Reprints and permission information is available at http://www.nature.com/reprints

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