Title

Analysis of Ugandan cervical carcinomas identifies human papillomavirus clade-specific epigenome and transcriptome landscapes.

Journal

Nature genetics, 52(8)

ISSN

1061-4036

Authors

Gagliardi, Alessia
Porter, Vanessa L
Zong, Zusheng
et al.

Publication Date

2020-08-01

DOI

10.1038/s41588-020-0673-7

Peer reviewed

ARTICLES
https://doi.org/10.1038/s41588-020-0673-7

nature genetics

Check for updates

Analysis of Ugandan cervical carcinomas identifies human papillomavirus clade–specific epigenome and transcriptome landscapes

Alessia Gagliardi[1,19], Vanessa L. Porter[1,2,19], Zusheng Zong[1,19], Reanne Bowlby[1,19], Emma Titmuss[1,19], Constance Namirembe[3], Nicholas B. Griner[4], Hilary Petrello[5], Jay Bowen[5], Simon K. Chan[1], Luka Culibrk[1], Teresa M. Darragh[6], Mark H. Stoler[7], Thomas C. Wright[8], Patee Gesuwan[4], Maureen A. Dyer[9], Yussanne Ma[1], Karen L. Mungall[1], Steven J. M. Jones[1,2], Carolyn Nakisige[3], Karen Novik[1], Jackson Orem[3], Martin Origa[3], Julie M. Gastier-Foster[5,10], Robert Yarchoan[11,12], Corey Casper[13], Gordon B. Mills[14], Janet S. Rader[15,20], Akinyemi I. Ojesina[16,17,18,20], Daniela S. Gerhard[4,20], Andrew J. Mungall[1,20] and Marco A. Marra[1,2,20] ✉

Cervical cancer is the most common cancer affecting sub-Saharan African women and is prevalent among HIV-positive (HIV$^+$) individuals. No comprehensive profiling of cancer genomes, transcriptomes or epigenomes has been performed in this population thus far. We characterized 118 tumors from Ugandan patients, of whom 72 were HIV$^+$, and performed extended mutation analysis on an additional 89 tumors. We detected human papillomavirus (HPV)-clade-specific differences in tumor DNA methylation, promoter- and enhancer-associated histone marks, gene expression and pathway dysregulation. Changes in histone modification at HPV integration events were correlated with upregulation of nearby genes and endogenous retroviruses.

Persistent HPV infection, in episomal or integrated form, is necessary but not sufficient for the development of cervical cancer[1]. HPV-16 and HPV-18 are detected in at least 70% of affected individuals[2]. HPV-16 (clade A9) is common in both squamous cell carcinomas and adenocarcinomas, while HPV-18 (clade A7) is associated with adenocarcinomas[2] and inferior survival[3–5].

Cervical cancer prevention strategies include vaccination and screening for HPV and treatment of high-grade precancer. Although effective[6], vaccine use remains low in low- and middle-income countries[7] where HIV is prevalent. Resource constraints similarly complicate screening, surgery[8] and radiotherapy[9], such that a 50% increase in cervical cancer mortality by 2040 is predicted[10].

Genomic cervical cancer studies, primarily conducted in non-African individuals[11,12], identified APOBEC mutational signatures, copy number amplifications of *CD274* (PD-L1) and *PDCD1LG2* (PD-L2), somatic alterations affecting the PI(3)K–MAPK and TGFβR2 pathways[11,12] and mutations in chromatin modifier genes[11–13]. Studies in HPV-infected individuals with head and neck squamous cell carcinomas linked HPV integration to changes in histone modification[14] and DNA methylation[15], suggesting the potential for similar findings in cervical cancer.

As part of the National Cancer Institute's (NCI's) HIV+ Tumor Molecular Characterization Project (HTMCP), we characterized the genomic, transcriptomic and epigenomic landscapes of cervical cancers from Ugandan patients. We identified previously uncharacterized differences in the epigenomes and transcriptomes of cervical tumors from individuals infected by different HPV clades and note that these clades appear relevant to prognosis.

Results

Patient samples and clinical data. Our cohort of 212 patients with cervical cancer received treatment at the Uganda Cancer Institute in Kampala. Of these, 118 made up our discovery cohort and 89 made up our extension cohort (Supplementary Tables 1 and 2, and Methods). HIV$^+$ patients (72/118, 61%) were 10 years younger, on average, than HIV-negative (HIV$^-$) patients (mean, 42.9 versus 52.4 years).

[1]Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, British Columbia, Canada. [2]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada. [3]Uganda Cancer Institute, Kampala, Uganda. [4]Office of Cancer Genomics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [5]Nationwide Children's Hospital, Columbus, OH, USA. [6]Department of Pathology, University of California at San Francisco, San Francisco, CA, USA. [7]Department of Pathology, University of Virginia, Charlottesville, VA, USA. [8]Department of Pathology and Cell Biology, Columbia University, New York, NY, USA. [9]Frederick National Laboratory for Cancer Research, Frederick, MD, USA. [10]Departments of Pathology and Pediatrics, The Ohio State University, Columbus, OH, USA. [11]Office of HIV and AIDS Malignancy, National Cancer Institute, National Institues of Health, Bethesda, MD, USA. [12]HIV and AIDS Malignancy Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [13]Infectious Disease Research Institute, Seattle, WA, USA. [14]Knight Cancer Institute, Oregon Health and Science University, Portland, OR, USA. [15]Department of Obstetrics and Gynecology, Medical College of Wisconsin, Milwaukee, WI, USA. [16]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, USA. [17]O'Neal Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL, USA. [18]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. [19]These authors contributed equally: Alessia Gagliardi, Vanessa L. Porter, Zusheng Zong, Reanne Bowlby, Emma Titmuss. [20]These authors jointly supervised this work: Janet S. Rader, Akinyemi I. Ojesina, Daniela S. Gerhard, Andrew J. Mungall, Marco A. Marra. ✉e-mail: mmarra@bcgsc.ca

Genomic alterations in HIV$^+$ and HIV$^-$ cervical cancers. Whole-genome sequencing of samples from our discovery cohort identified an average of 22,942 somatic mutations (range, 3,033–179,513) per sample, including 311 coding mutations (range, 30–2,683; Fig. 1a). We detected APOBEC mutation signatures 2 and 13 (refs. [16,17]), confirming previous reports[18] and in line with a mutational process driven by a cellular response to viral infections[19] (Fig. 1a). Tumors with a high proportion of mutations with APOBEC signatures (proportion ≥ 0.4) exhibited significantly more coding mutations (threefold increase per megabase, median; Wilcoxon, $P = 2.1 \times 10^{-7}$) than those with lower proportions (Extended Data Fig. 1a). Fifteen samples (13%) exhibited moderate to high homologous recombination deficiency (HRD) scores (>30), indicative of a dysfunctional HR repair pathway[20] (Fig. 1a and Methods). There were no differences in mutation burden, mutation signatures or HRD score between HIV$^+$ and HIV$^-$ samples.

Of the 12 significantly mutated genes (SMGs) in our cohort, *PIK3CA* was the most recurrent (Fig. 1a and Supplementary Table 3), as reported in other studies[11,12]. A higher proportion of HIV$^-$ tumors (45%, 20/45) than HIV$^+$ tumors (29%, 21/72) had *PIK3CA* mutations, and *PIK3CA* expression was 1.3 times higher in HIV$^-$ samples (Wilcoxon, $P = 1 \times 10^{-4}$; Extended Data Fig. 1b). Other SMGs included *FAT1*, *KMT2D*, *FBXW7*, *CASP8*, *MAPK1* and *ZNF750*, all previously reported in cervical cancer[11,12]. Notably, 87% of the cohort (101/118) had at least one mutation in an annotated chromatin modifier gene[20] (Supplementary Table 4 and Extended Data Fig. 1c).

We performed targeted sequencing of 2,735 selected genes in our extension cohort (HIV$^-$, $n = 73$; HIV$^+$, $n = 16$), confirming mutations in 11 of the 12 SMGs (Extended Data Fig. 1d) and observing similar mutation frequencies between the discovery and extension cohorts.

Analysis of copy number landscapes showed that broad copy number alterations were comparable between HIV$^+$ and HIV$^-$ samples, with shared amplifications of chromosomes 1, 8 and 20 and arms 3q, 5p and 19q and shared deletions of chromosome 11 and arms 3p, 4p, 19p and 21p (Fig. 1b, top two plots). We found six amplified and four deleted chromosome arms unique to HIV$^+$ samples and two amplified arms unique to HIV$^-$ samples (Fig. 1b and Supplementary Table 5). HIV$^+$ samples exhibited more unique focal amplifications and deletions than HIV$^-$ samples (Fig. 1c and Supplementary Table 5).

We compared the copy number landscapes of our HIV$^-$ samples to those of The Cancer Genome Atlas (TCGA) cervical cancers (HIV$^-$; Supplementary Table 6 and Methods). TCGA samples exhibited a larger number of significantly deleted regions, affecting 11 chromosomes, whereas only 21p was lost in our cohort. In comparison to TCGA, our HIV$^-$ cohort had three significantly amplified regions on chromosomes 3p, 8p and 15q (Fig. 1b). Five focal amplifications and nine deletions identified in the TCGA cohort were also detected in our HIV$^-$ cohort (Fig. 1c). Focal deletions unique to TCGA or our HIV$^-$ samples were comparable in number and more abundant than focal amplifications in either cohort (Supplementary Table 5). 11q22.1 and 11q22.2, containing *YAP1*, were amplified in HIV$^+$, HIV$^-$ and TCGA samples. Six of the 12 SMGs were impacted by copy number alterations; of these, *PIK3CA* was the most frequently altered gene, by mutation or copy number alteration (Fig. 1a).

Recurrent noncoding mutations. We leveraged whole-genome sequencing to identify seven high-confidence noncoding 'hotspots' (Methods and Fig. 2a), including two in the *TERT* promoter first described in melanomas[21,22], in 11% (13/118) of our discovery cohort samples. *TERT* transcript levels were not dysregulated in these samples. Two hotspots in a potential intronic enhancer (Methods) of *ADGRG6* were observed in 9% (11/118) of samples. These noncoding mutations, at chr6:142706206(G>A) and chr6:142706209(C>T), were located within palindromic sequences predicted to form hairpin loops, accessible to APOBEC enzymes[23] (Fig. 2b). These hotspots have been reported in approximately 3% of breast cancers[23] and 46% of bladder cancers[24,25] and have been associated with increased ADGRG6 protein expression and angiogenesis. We did not observe dysregulated *ADGRG6* mRNA expression in mutated samples. Three additional hotspots, on chromosomes 6, 8 and 11, were not associated with potential promoters or enhancers. All reported hotspots were present in HIV$^+$ and HIV$^-$ samples, and the samples with mutations (C>T or C>G) exhibited a moderate proportion of mutations with APOBEC signatures (Fig. 2a). Because *TERT* promoter mutations can create new binding sites for the c-ETS transcription factor[21,22], we investigated the potential for other noncoding hotspots to alter transcription factor binding sites[26] (Fig. 2c) and noted that POU and FOX family binding sites were either created or destroyed by mutations in five of the seven hotspots.

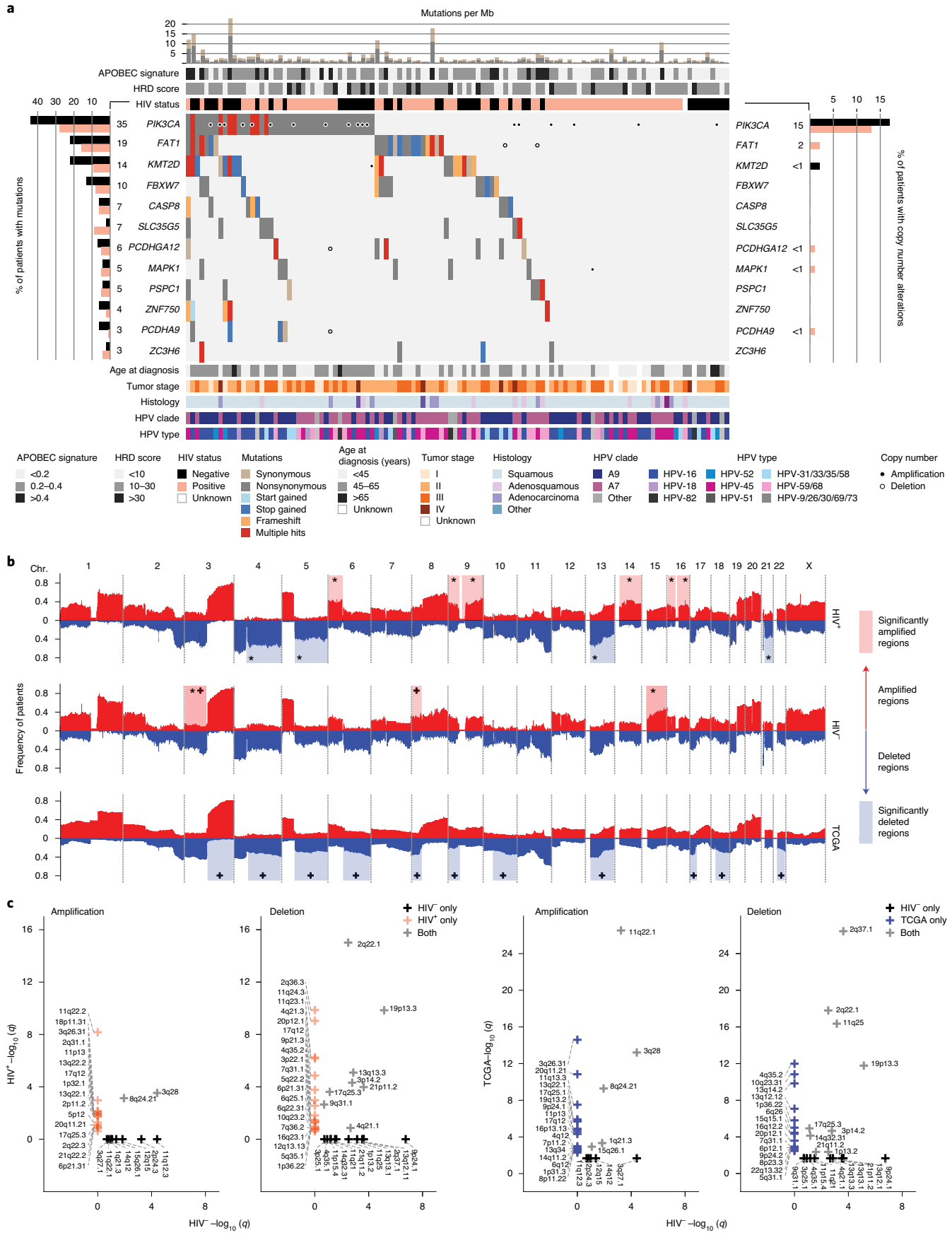
Distribution of HPV types. Whole-genome sequencing detected 17 HPV types and their associated clades in our cohort (Methods)[27,28]. High-risk HPV-16 (clade A9), HPV-18 and HPV-45 (clade A7) were the most abundant types (Fig. 3a), and clade A7 was more prevalent in our cohort than in the TCGA cohort, particularly among the squamous cell carcinomas (SCCs; Fig. 3b). Unlike previous reports[29,30], no difference in HPV types between HIV$^+$ and HIV$^-$ tumors was found (Extended Data Fig. 2a).

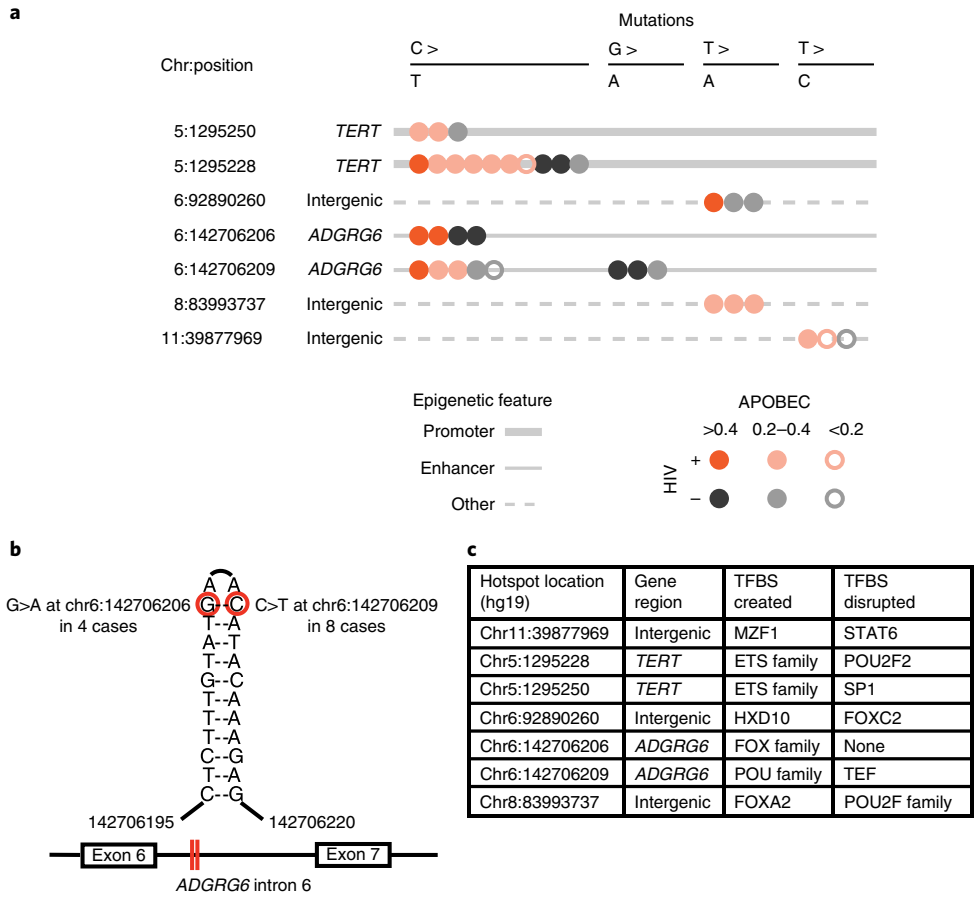
Expression and methylation profiles and HPV clades. We characterized expression and DNA methylation landscapes by performing unsupervised clustering on the most variably expressed genes ($n = 1{,}000$) and the most variably methylated probes ($n = 8{,}000$; 850k EPIC array), respectively, and correlated these with tumor features. We identified three gene expression clusters, enriched for adenocarcinomas ($q = 4.1 \times 10^{-8}$; cluster 1), non-keratinizing SCCs (cluster 3) or keratinizing SCCs ($q = 0.015$; cluster 2), similar to those reported previously[11] (Extended Data Fig. 2b). Additionally, cluster 1 was enriched for samples with clade A7 HPV ($q = 1.3 \times 10^{-9}$; Extended Data Fig. 2b) and cluster 3 was enriched for *PIK3CA*-mutated samples ($q = 3.1 \times 10^{-5}$). Two DNA methylation clusters (Fig. 3c) identified separation of SCCs with clade A9 HPV (cluster 1) from

Fig. 1 | Mutational landscape of cervical cancers from Ugandan patients. a, Mutations and copy number alterations for each tumor ($n = 118$) ordered by the frequency of alterations in SMGs. Synonymous and nonsynonymous mutation counts per megabase are shown above with the proportion of mutations with an APOBEC signature (COSMIC 2 and 13) and HRD score. HIV status, age at diagnosis, histology ('other' includes neuroendocrine and undifferentiated), tumor stage, and HPV clade and type are annotated below the oncoprint. Left bar chart, percentage of samples with mutations, by HIV status. Numbers to the right of the bar plot indicate the percentage of the entire cohort. Right bar chart, proportion of samples with copy number alterations in SMGs, by HIV status. Numbers to the left of the bar plot indicate the percentage of the entire cohort. **b**, Broad somatic copy number alterations in our HIV$^+$ and HIV$^-$ cohorts and the TCGA cohort. An asterisk indicates that the region is significantly amplified or deleted (as determined by GISTIC, FDR < 0.25; Methods) in only HIV$^+$ or HIV$^-$ samples, and a plus sign indicates that the region is significantly amplified or deleted in only TCGA or HIV$^-$ samples. **c**, Focal regions associated with significant copy number changes between HIV$^+$ and HIV$^-$ samples and between HIV$^-$ samples and TCGA. The numbers of tumor samples used to determine differences in **b** and **c** are as follows: HIV$^+$, $n = 72$; HIV$^-$, $n = 45$; TCGA, $n = 178$.

squamous and non-squamous carcinomas with clade A7 HPV (cluster 2; $q = 1.7 \times 10^{-13}$). Cluster 2 was also enriched in samples with higher tumor grade ($q = 0.020$).

We compared clade A7-infected samples to clade A9-infected samples (Supplementary Table 7) through differential methylation analysis[31–33] (Methods). We identified 107,685 differentially





Hotspot location (hg19)	Gene region	TFBS created	TFBS disrupted
Chr11:39877969	Intergenic	MZF1	STAT6
Chr5:1295228	*TERT*	ETS family	POU2F2
Chr5:1295250	*TERT*	ETS family	SP1
Chr6:92890260	Intergenic	HXD10	FOXC2
Chr6:142706206	*ADGRG6*	FOX family	None
Chr6:142706209	*ADGRG6*	POU family	TEF
Chr8:83993737	Intergenic	FOXA2	POU2F family

Fig. 2 | Recurrent noncoding mutations. a, Schematics of the seven noncoding hotspots found in our cohort. Each dot represents a sample carrying the base substitution reported on top of the plot. HIV status and the strength of the APOBEC signature for each sample with the mutation are indicated by the color and fill of the dot, respectively, and line type represents the epigenetic characteristics of the sequence surrounding each hotspot. **b**, Example of two hotspot mutations in *ADGRG6* intron 6, found in 11 samples (one sample has both mutations). Mutated nucleotides are circled in red, and red lines in the diagram of the *ADGRG6* gene highlight their location within intron 6. **c**, Predicted impact of the seven noncoding hotspot mutations on transcription factor binding sites (TFBS).

methylated probes (DMPs), including 46,639 DMPs in tumors with clade A9 HPV and 61,646 DMPs in tumors with clade A7 HPV (false discovery rate (FDR) < 0.05). The distribution of DMPs with respect to genomic features and proximity to CpG islands differed by clade. DMPs in 'open sea' and 'shelf' CpGs[34,35] (>2 kb from CpG islands) were more common for clade A7 (79% versus 45% for clade A9) and were often in intergenic regions. DMPs in CpG islands were more common for clade A9 (35% versus 3.7% for clade A7), with the majority of these residing in candidate transcriptional regulatory regions (Extended Data Fig. 2c; for example, <1,500 bp from a transcriptional start site (TSS), 5′ UTR or first exon of a gene).

Motivated by the differences in DMP distribution between clades, we detected 721 differentially expressed genes (Methods and Supplementary Table 8) between clades after accounting for histological differences, with approximately equal proportions of genes upregulated in each clade (A7, $n = 363$; A9, $n = 358$; Extended Data Fig. 2d). Functional enrichment analysis[36] (Fig. 3d and Supplementary Table 9) showed enrichment of ontologies linked to keratinocyte and epithelial differentiation in samples with clade A9 HPV. Genes with increased expression in these samples include keratin family genes as well as *AMTN*, *LCE3D* and *BCL2L10*. The tightly regulated keratinocyte differentiation pathways are known to be exploited during HPV infection for active production of the virus and later to direct uncontrolled cell growth in cervical epithelial cells[37]. Samples with clade A7 HPV had increased expression of *PROM1*, *TGFB2*, *PXDN* and *FN1* and differentially expressed genes were enriched for pathways linked to extracellular matrix organization and cell adhesion and migration, in line with the more aggressive tumor grades that correlated with clade A7 (Fig. 3c).

To relate the effect of DNA methylation at promoter regions to changes in gene expression, we identified differentially methylated regions (DMRs; Methods and Supplementary Table 10), defined as nearby probes exhibiting consistent methylation changes, and associated these regions with gene expression. In agreement with the higher number of DMPs at CpG islands in samples with clade A9 HPV, we identified 558 methylated DMRs in these samples (490 overlapping genes) in comparison to 53 methylated DMRs in samples with clade A7 HPV (48 overlapping genes). There were 26 upregulated genes in samples with A7 HPV that were associated with a methylated DMR in samples with A9 HPV, as compared to only 8 upregulated genes in samples with A9 HPV associated with a methylated DMR in samples with A7 HPV (Extended Data Fig. 2f). Thus, differential expression of genes between clades may result from differential methylation.

HPV viral gene expression influences host gene expression. HPV viral genes regulate epithelial cell differentiation and promote tumorigenesis[5,38]. To probe the impact of viral gene expression on tumor gene expression, we performed unsupervised clustering of

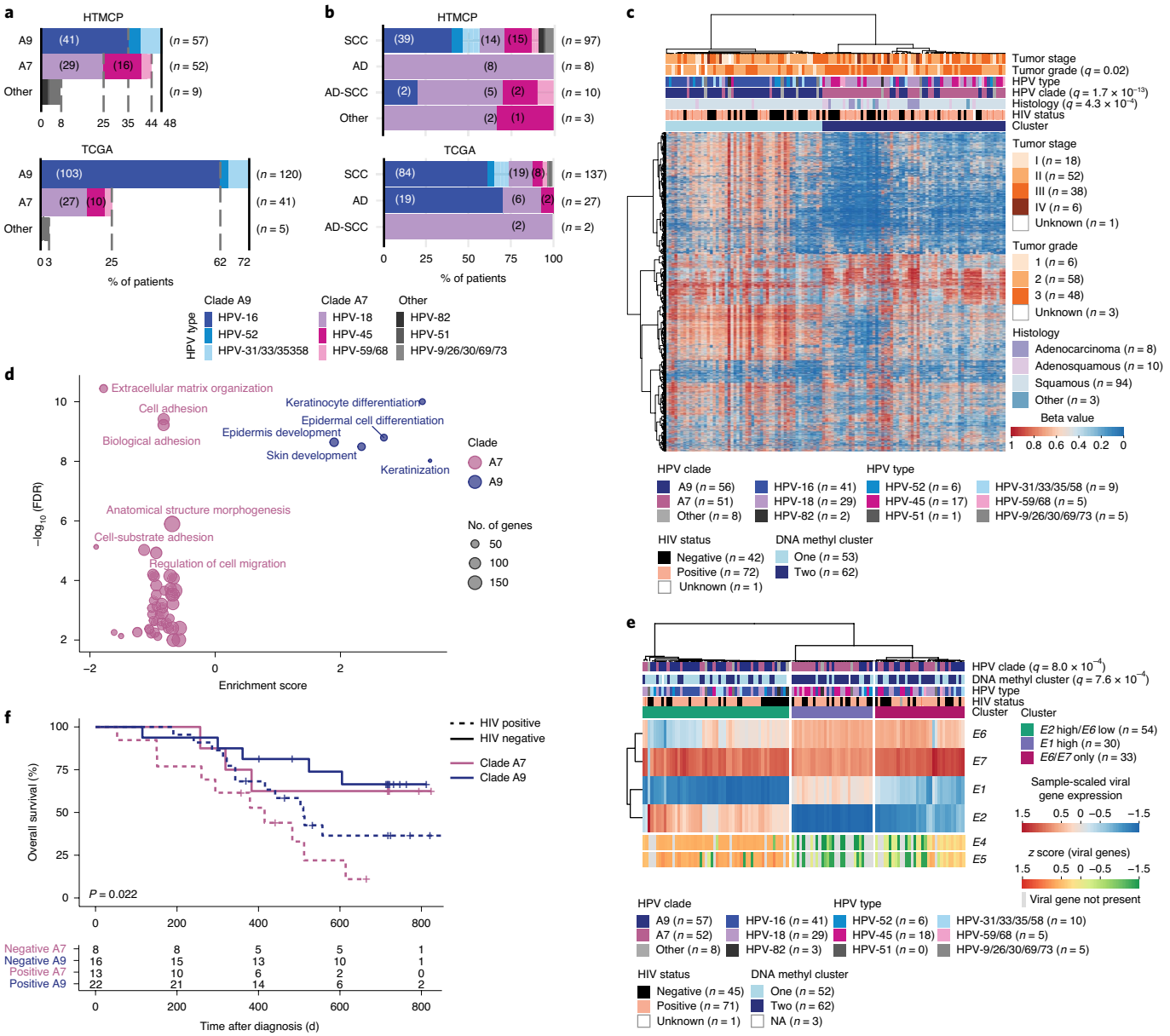


Fig. 3 | HPV-clade-specific molecular characteristics and prognosis. a,**b**, Distribution of HPV types in our cohort (top; $n=118$ tumors) and the TCGA cohort (bottom; $n=166$ HPV$^+$ tumors) (**a**) and proportion of types split by histology (**b**). SCC, squamous cell carcinoma; AD, adenocarcinoma; AD-SCC, adenosquamous carcinoma; other, neuroendocrine (2 samples) and undifferentiated (1 sample). For **a** and **b**, the *x* axis indicates the percentage of samples in that cohort infected by the indicated HPV type, and numbers in parentheses indicate the number of samples. **c**, Unsupervised clustering analysis of DNA methylation for the top 8,000 variable probes in 115 samples. HIV status, HPV type and clade, histology, and tumor grade and stage are annotated. **d**, Results from functional enrichment analysis of differentially expressed genes between samples with clade A9 versus clade A7 HPV (STRING; Methods). The size of the circles is proportional to the number of differentially expressed genes represented in each gene ontology. **e**, Unsupervised clustering analysis of sample-scaled HPV viral gene expression ($n=117$ samples). *z* scores for HPV genes not annotated in every HPV type are included below the clustering. HPV clade and type, DNA methylation clusters and HIV status are annotated. NA, not available. **f**, Overall survival of 59 patients stratified by the clade of HPV with which they were infected and HIV status. Kaplan–Meier overall survival statistics were determined by using a log-rank test, and *q* values for each variable on the heat maps were determined by Benjamini–Hochberg-corrected Fisher's exact test. All statistical tests were two sided.

viral *E1*, *E2*, *E6* and *E7* transcripts, which had annotations associated with them in GenBank (download date, December 2019) in all HPV types, in our cohort ($n=117$; Supplementary Table 11, Fig. 3e and Methods), identifying three clusters. Cluster 1, enriched for tumors with clade A9 HPV (Fisher's exact test, $q=0.0029$), exhibited high *E2* and low *E6* expression, indicating dominant HPV episome transcription. Cluster 2, enriched for tumors with clade A7 HPV (Fisher's exact test, $q=0.0029$), exhibited low *E2* and high *E1* expression. Cluster 3 contained tumors with HPV from both clades and exhibited low expression of both *E1* and *E2*. From the absence of *E2* expression, we inferred that clusters 2 and 3 reflected viral expression originating from the integrated form of HPV. The expression distributions of *E6* and *E7* were bimodal (Extended Data Fig. 2f,g) and were used to separate samples into high- and low-expressing tumors for each gene, on which we performed differential expression analysis. We identified 107 differentially expressed genes between tumors with high and low *E6* expression and 60 differentially expressed genes between tumors with high and

low *E7* expression. Genes from the group with high *E6* expression overlapped with clade A7-enriched pathways, while genes from the group with low *E7* expression overlapped with clade A9-enriched pathways (Extended Data Fig. 2f,g). Thus, clade-enriched tumor gene expression patterns may be influenced by the expression of HPV genes.

HPV clades are linked to prognosis. In line with our observation that tumors with clade A7 HPV displayed expression profiles indicative of a more aggressive phenotype (Fig. 3c,d), these tumors also appeared to be more aggressive clinically, with A7-infected patients exhibiting inferior prognosis in comparison to A9-infected patients (hazard ratio (HR) = 1.83, confidence interval (CI) = 1.02–3.30, $P = 0.04$, log-rank test). This observation held true even after accounting for other covariates in our cohort, including HIV status (Fig. 3f), stage and histology (Methods and Extended Data Fig. 2h). Despite the relatively small number of patients available for analysis, the impact of HPV clade (HR = 1.75, $P = 0.14$) on overall survival was similar to that of disease stage (HR = 2.10, $P = 0.19$), the latter of which is an established prognostic factor[39]. Similar observations have been reported previously[40].

Histone modification profiles associated with HPV clades. Motivated by our DNA methylation results (Fig. 3c) and the preponderance of somatic mutations in chromatin modifier genes, we investigated whether histone modifications also exhibited clade-specific differences. By using chromatin immunoprecipitation and sequencing (ChIP–seq), we profiled four histone modification marks associated with transcriptional activation (H3K4me1, H3K4me3, H3K27ac and H3K36me3) and two marks associated with repression (H3K9me3 and H3K27me3) in 52 samples. Cluster of clusters analysis[41] of these marks identified a clustering solution that resembled the individual solutions for the promoter- and enhancer-associated marks H3K4me1, H3K4me3 and H3K27ac, but not H3K36me3, H3K9me3 and H3K27me3 (Extended Data Fig. 3a,b). We thus performed the analysis again while using only these three active marks, which identified four clusters (Methods and Fig. 4a). Cluster 1 was enriched in tumors with clade A9 HPV ($q = 4.9 \times 10^{-5}$), while cluster 2 included an equal number of tumors with clade A7 (mostly HPV-45) and clade A9 HPV. The remaining tumors with clade A7 HPV were found in clusters 3 and 4, which was enriched for non-SCC tumors (Fisher's exact test, $P = 4.4 \times 10^{-6}$). None of the clusters were enriched for somatic mutations in genes encoding members of chromatin-modifying complexes, although cluster 4 conspicuously lacked alterations in SEC, NuRD, HDAC and ISWI complex members (Fig. 4a and Extended Data Fig. 3a).

With HPV-clade-specific differences observed, we assessed whether clades were associated with differential abundance[42] of histone marks at regulatory regions, including active promoters (H3K27ac and H3K4me3) and enhancers (H3K4me1). We identified differential abundance of 15,245 H3K4me1 peaks, 9,902 H3K27ac peaks and 7,736 H3K4me3 peaks (adjusted $P < 0.01$, fold change > 2) between clades, after normalizing for histology differences (Methods, Fig. 4b and Supplementary Table 12).

Clade A7-infected samples had three times more H3K4me1-enriched regions (11,530 for clade A7 versus 3,715 for clade A9) and approximately double the number of H3K27ac-enriched regions (6,405 for clade A7 versus 3,497 for clade A9) in comparison to samples infected with clade A9, suggesting an enrichment for enhancers. In contrast, tumors with clade A9 HPV had almost twice the number of differential H3K4me3 peaks (4,997 for clade A9 versus 2,739 for clade A7), and one-quarter of these (1,271 peaks) overlapped with H3K27ac-enriched regions (chi-squared test, $P < 2.2 \times 10^{-16}$; Fig. 4b), indicating enrichment for promoter marks in tumors with clade A9 HPV.

The two most frequently mutated chromatin modifier genes in our cohort, *KMT2C* and *KMT2D*, deposit H3K4me1 at enhancer regions[43–46]. We therefore sought to investigate the impact of loss-of-function mutations in *KMT2C* and *KMT2D* on the number of H3K4me1-marked regions. We observed that, for the 15 samples with loss-of-function mutations in *KMT2C* or *KMT2D*, the number of primed enhancer regions (marked by only H3K4me1) was lower than in tumors with no mutations in chromatin modifier genes or with mutations in other chromatin modifier genes, whereas there was no difference in the number of active enhancer regions (marked by H3K4me1 and H3K27ac) (Methods and Fig. 4c). Despite different peak enrichments between HPV clades (Fig. 4b), the effect of *KMT2C* and *KMT2D* mutations on primed enhancers was not clade specific.

To relate chromatin mark profiles to differential expression, we mapped the regions with differential H3K27ac and H3K4me3 peaks to nearby genes (Methods and Supplementary Table 13). This identified 769 regions with differential H3K27ac and H3K4me3 peaks enriched in samples with clade A7 HPV, of which 18 were near the TSS of genes upregulated in samples with clade A7 HPV, including the invasion and extracellular matrix genes *SRCRB4D*, *PXDN* and *CXCL2* (Fig. 4d and Extended Data Fig. 3c). We also identified 1,271 regions with differential H3K27ac and H3K4me3 peaks in samples with clade A9 HPV, 25 of which were near the TSS of genes upregulated in samples with clade A9 HPV (Fig. 4d and Extended Data Fig. 3c), including *TMPRSS11A*, *WNT2B* and *MEI1*. We similarly observed clade-specific correlations between differential H3K4me1-marked regions and expression of the nearest gene (Fig. 4e). Relationships between histone modification and gene expression differences between clades are displayed in Fig. 4f and Extended Data Fig. 3d, using the *PXDN* and *TMPRSS11A* genes as examples.

Our results thus indicate that DNA methylation (Fig. 3c) and epigenetic modification patterns attributed to H3K27ac, H3K4me3 and H3K4me1 (Fig. 4b) are altered in an HPV-clade-specific manner in our cohort.

Altered RNA and histone profiles at HPV integration sites. We studied the genomic impacts of 1,010 unique HPV integration sites in 109 of the 118 tumors (Supplementary Table 2). Grouping of integration sites near one another (<500 kb apart) within samples resulted in the identification of 257 'integration events' (median length, 2.6 kb; range, 1 bp–409 kb; Methods and Supplementary Table 14). Clade A7 integration events contained more integration sites per event than clade A9 events (Wilcoxon, $P = 0.043$; Extended Data Fig. 4a).

Of these events, 155 (60%) were within 10 kb of one or more genes (Methods and Extended Data Fig. 4b). *KLF12*, *TP63*, *RAD51B* and *MYC* were among 16 genes that were the closest in proximity to an integration event in multiple samples, as previously reported[12,47] (Fig. 5a and Supplementary Table 15). Of the genes near integration events, 61 (from 45 events) displayed significantly higher expression in samples with integration (fold change ≥ 2, adjusted $P \leq 0.05$; Extended Data Fig. 4c, Supplementary Table 15 and Methods). Furthermore, events containing a higher number of integration sites were associated with increased fold change in expression (ANOVA, $P = 9.8 \times 10^{-4}$; Fig. 5b). Clade A7 integration events appeared to have a more pronounced effect on expression than clade A9 events (Wilcoxon, $P = 0.025$; Fig. 5c), which may result from the higher number of integration sites per event in this clade. Of the 16 genes identified in multiple samples near integration events (Fig. 5a), 8 were significantly upregulated in integrated samples (Extended Data Fig. 4d), including the oncogenes *ERBB2* (69-fold increase, adjusted $P = 0.033$) and *TP63* (8.3-fold increase, adjusted $P = 0.033$).

To explore possible epigenomic mechanisms of altered gene expression at HPV integration events, we examined the fold change

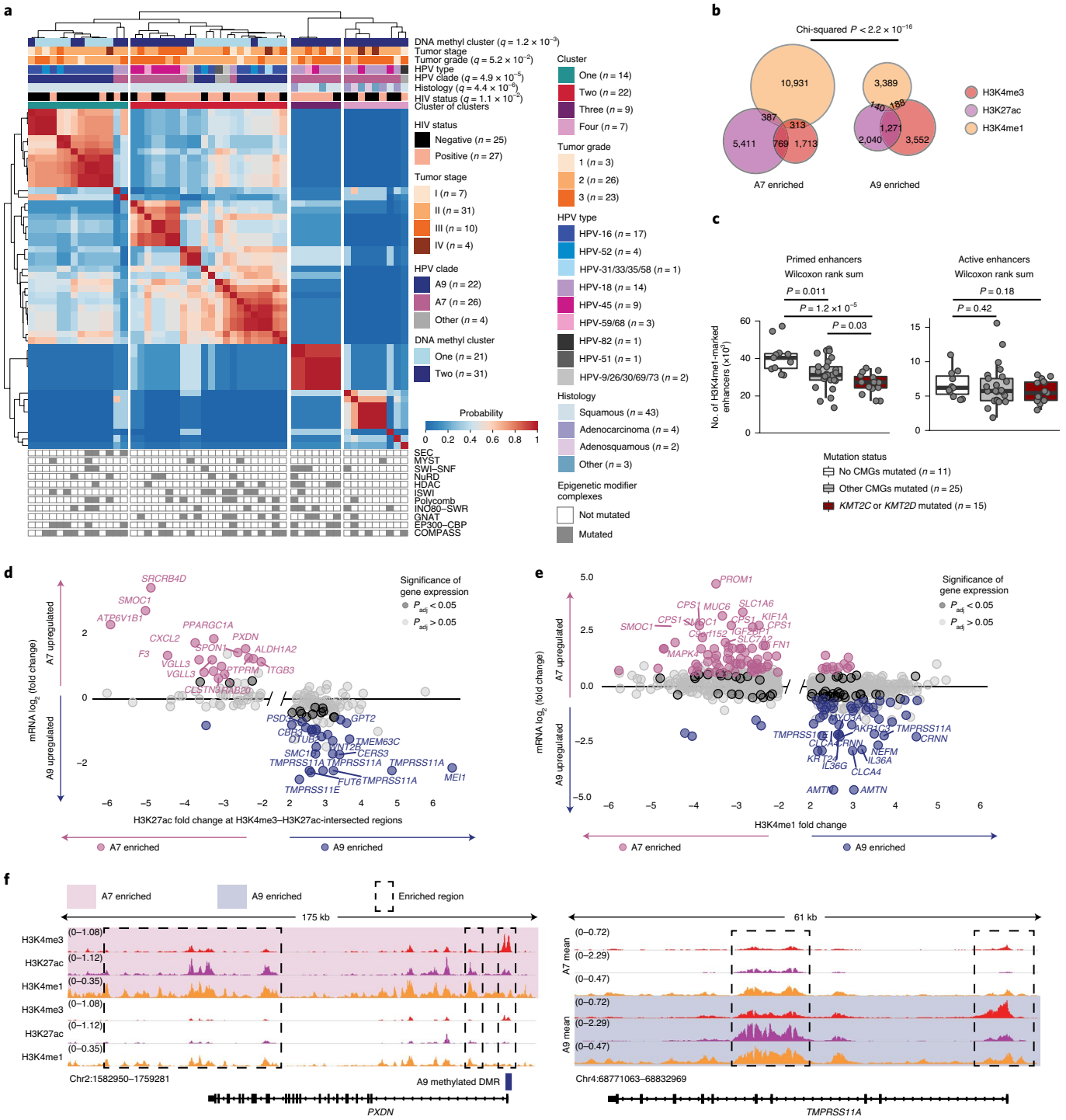


Fig. 4 | HPV-clade-specific histone mark landscapes. a, Cluster of clusters analysis showing 27 consensus clustering solutions for 3 active histone marks in 52 samples ($k = 2$–10 for each mark). HIV status, histology, HPV type and clade, tumor grade and stage, DNA methylation cluster and the mutation status of genes encoding members of epigenetic modifier complexes are annotated. q values for each variable were determined by Benjamini–Hochberg-corrected Fisher's exact test. **b**, Overlap of H3K27ac, H3K4me3 and H3K4me1 peaks significantly enriched for each clade. **c**, Number of H3K4me1-marked enhancers at primed (H3K4me1-only) regions (left) or active (H3K4me1/H4K27ac) regions (right). The samples are divided by the mutation status of chromatin modifier genes (CMGs). Box plots represent the median and upper and lower quartiles of the distribution; whiskers represent the limits of the distribution (1.5 times the interquartile range). P values were calculated by Wilcoxon rank-sum test. **d**,**e**, Fold change in histone mark abundance and gene expression between clades associated with the TSS of genes found between −5 to +20 kb with respect to intersecting H3K4me3 and H3K27ac peaks (**d**) and between −20 and +20 kb with respect to differential H3K4me1 peaks (**e**). Sample numbers used for differential analyses (and derivation of adjusted P values) were as follows: expression: A7, $n = 52$; A9, $n = 57$; H3K4me3, H3K27ac and H3K4me1: A7, $n = 25$; A9, $n = 22$. Genes with Bonferroni–Hochberg-adjusted $P < 0.05$ (DESeq2; Methods) are highlighted. **f**, Examples of differential active histone marks (H3K4me3, H3K27ac and H3K4me1) near *PXDN* (left) and *TMPRSS11A* (right), which are differentially expressed between clades. All statistical tests were two sided.

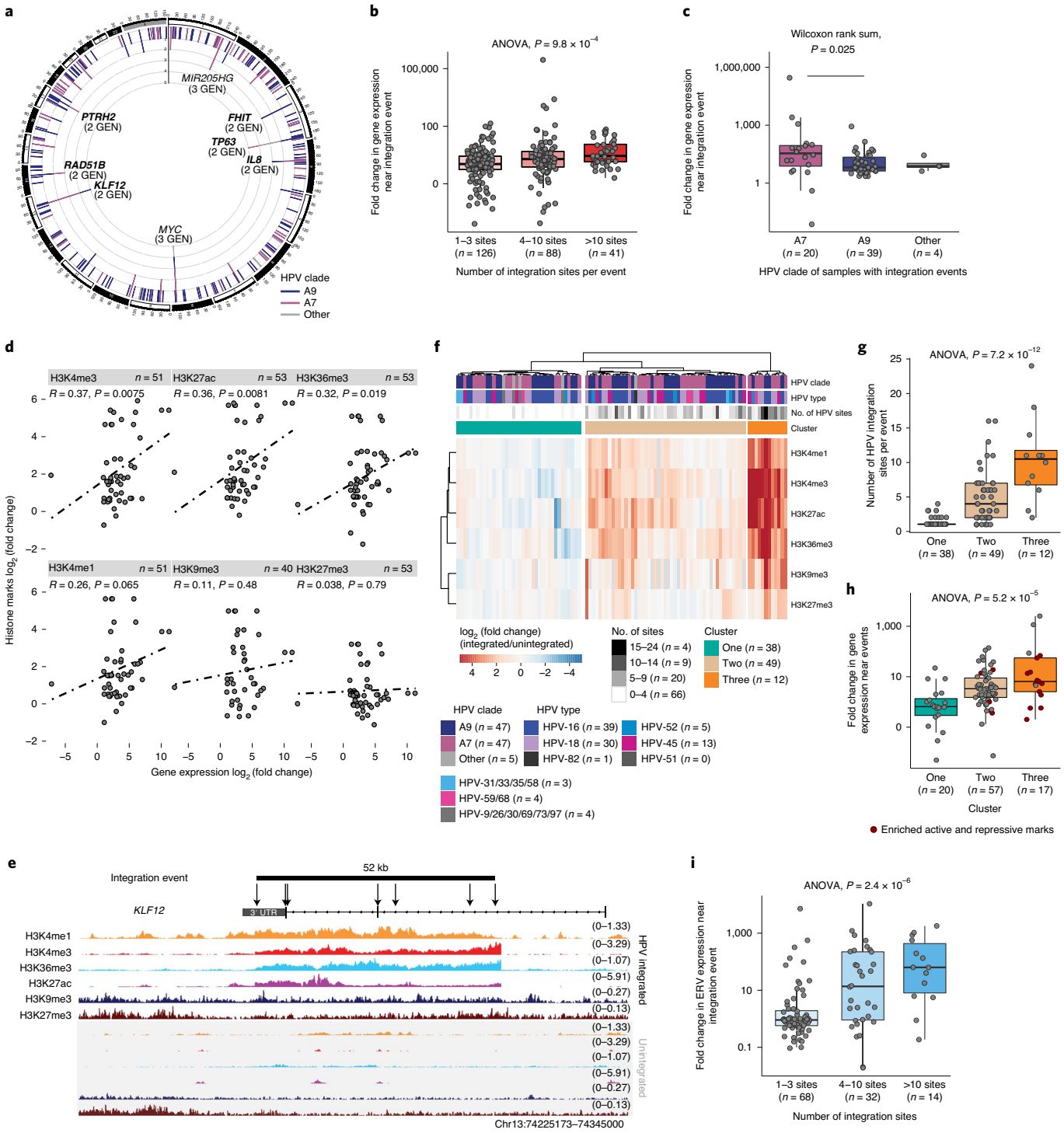


Fig. 5 | HPV integration alters local histone modifications and expression. a, HPV integration events in 109 samples collapsed into frequent regions within 500 kb of one another. The number of integrations, colored by clade, is presented radially. The number of unique genes closest to integration events is labeled (GEN), and upregulated genes are highlighted (bold). **b**,**c**, Fold change in expression of genes near integration events, by the number of integration sites per event (**b**) and clade (**c**). **d**, Fold change in local gene expression and histone mark coverage at events. Statistics were determined by Spearman test. **e**, Histone mark coverage of the 3′ region of *KLF12*. Arrows indicate individual integration sites within the event (line). Top tracks show a sample with an event in this region, and bottom tracks show a sample without an event. **f**, Unsupervised clustering of the fold change in histone mark coverage at integration events ($n = 99$). **g**, Number of integration sites per event in each cluster in **f**. **h**, Fold change in expression of genes near events by the clusters in **f**. **i**, Fold change in ERV expression near events by the number of sites within the event. All fold change values refer to the integrated sample ($n = 1$) versus the cohort ($n = 117$ for expression, $n = 46$ for H3K9me3, $n = 50$ for H3K4me1 and H3K4me3, $n = 51$ for H3K27ac, H3K27me3 and H3K36me3). Box plots in **b**, **c** and **g**–**i** represent the median and upper and lower quartiles of the distribution; whiskers represent the limits of the distribution (1.5 times the interquartile range). Where relevant, all statistical tests were two sided.

in histone mark enrichment within integration events (Methods). Fold changes in histone mark enrichment of H3K27ac, H3K4me3, H3K4me1 and H3K36me3 were positively correlated with gene expression changes (Fig. 5d). In our unsupervised clustering analyses, we noted that increased H3K36me3, typically associated with transcription[48,49], was associated with local transcriptional dysregulation at integration events but not with global dysregulation (Extended Data Fig. 3b). The 3′ region of *KLF12* provides a visual example, highlighting the relationship between an HPV integration event, altered histone modifications (mean $\log_2$(fold change) = 2.8 for H3K27ac, H3K4me3, H3K4me1 and H3K36me3; Fig. 5e) and increased gene expression (fold change = 112, adjusted $P = 0.033$; Methods and Extended Data Fig. 4d).

Unsupervised clustering of the fold changes in histone modification near integration events identified three clusters with varying levels of histone mark enrichment (Fig. 5f and Supplementary Tables 16 and 17). Cluster 3 contained 12 events with increased coverage of H3K27ac, H3K36me3, H3K4me3 and H3K4me1. Seven of these events also had significant enrichment of the repressive modifications H3K9me3 and H3K27me3. Cluster 2 included 49 events with a lower degree of enrichment of all active histone modifications in comparison to cluster 3 (14/49 versus 12/12 events), while cluster 1 included events with no enrichment of all active histone modifications (0/38 events). The average number of HPV integration sites per event was significantly different between the three clusters, with cluster 3 having the highest number (ANOVA, $P = 7.2 \times 10^{-12}$; Fig. 5g). In line with our observation that a higher number of HPV integration sites was associated with increased expression of nearby genes (Fig. 5b), events in cluster 3 were associated with genes exhibiting the highest increases in expression (ANOVA, $P = 5.2 \times 10^{-5}$; Fig. 5h). Events enriched for active and repressive marks appeared to have dampened upregulation of local genes as compared to events without enrichment of repressive marks (Wilcoxon, $P = 0.04$; Fig. 5h; $n = 18$ with active and repressive marks, $n = 76$ with only active marks).

Thirty-two percent (32/99) of integration events in samples with available ChIP data were not within 10 kb of a protein-coding gene but were associated with locally altered histone modification patterns (clusters 2 and 3). We thus sought other genomic features potentially influenced by these events, identifying endogenous retroviral sequences (ERVs) near 114 of 257 events (44%). ERVs are epigenetically silenced in the human genome, and their reactivation is associated with induction of antiviral pathways, such as double-stranded RNA (dsRNA) response signaling[50]. We analyzed the expression of ERVs near integration events and observed that expression was significantly higher in integrated samples and was also positively correlated with the number of HPV insertions within the event (Fig. 5i and Supplementary Table 18). As with genes, upregulation of ERVs near integration events was associated with histone modification changes (ANOVA, $P = 0.081$; Extended Data Fig. 4g,h).

To explore the tumor microenvironment and to determine whether increased ERV expression was associated with increased immune cell presence, as described previously[50], we inferred immune expression scores with RNA-seq[51] (Methods). Samples with upregulated ERVs at integration events had higher total T cell scores (Extended Data Fig. 4i, left), but, owing to the lower estimated tumor content for these samples (Methods and Extended Data Fig. 4i, right) and the inverse correlation between tumor content and total immune scores in our samples, we could not confidently assess the relationship between ERV upregulation and immune cell abundance. We also examined the expression of genes in pathways involved in ERV recognition, including type I interferon signaling (GO:0060337) and dsRNA-sensing pathways (GO:0043330), but we did not observe increased expression of such genes in samples with an ERV integration event, nor did we observe evidence of immune escape through point mutation, deletion or methylation of these genes.

As HIV infection targets CD4+ T cells, we compared CD4+ T cell scores between samples from HIV+ and HIV− patients. Total CD4+ T cell scores were lower in HIV+ than HIV− samples (Wilcoxon, $P = 0.0030$; Extended Data Fig. 4j), particularly follicular helper T cell scores (Wilcoxon, Benjamini–Hochberg-corrected $P = 0.0094$; Extended Data Fig. 4k, left), which is consistent with HIV infection primarily affecting these cells[52]. The only immune score found to be higher in HIV+ samples was of neutrophils (Wilcoxon, Benjamini–Hochberg-corrected $P = 0.024$; Extended Data Fig. 4k, right), the role of which is unclear in the mucosal environment of the genital tract[53].

These observations indicate that HPV integration sites in tumor genomes are associated with local histone modification changes that correlate with altered expression of genes and ERVs, including known oncogenes such as *ERBB2*, which may contribute to tumor progression.

Discussion

We characterized the genomic, transcriptomic and epigenomic landscapes of 118 cervical cancers from an understudied population of HIV+ and HIV− Ugandan patients and identified HPV-clade-associated dysregulation. Large-scale genomics studies like this are important, particularly in under-represented ancestry groups, to understand molecular phenotypes of these cancers, which can lead to improved treatment options.

The composition of this cohort, including comparable representation of clade A9- and clade A7-infected samples, allowed us to describe molecular characteristics associated with clades. Clade A7-infected samples exhibited distinct gene expression patterns converging on pathways linked to the extracellular matrix and to cell adhesion and migration, indicating a more aggressive phenotype. While inferior prognosis associated with clade A7 has previously been reported in invasive cervical cancer[40], our study provides insight into the cellular pathways that may promote the aggressive phenotype in these tumors. Genes upregulated in samples with clade A7 HPV, such as *PXDN*, are upregulated in cancers that have more potential to progress through the epithelial–mesenchymal transition[54,55]. DNA methylation, for which we also observed clade-specific patterns, is tightly regulated through cell differentiation[56]. It is therefore reasonable to hypothesize that these two HPV clades may push epithelial cells to replicate at the two ends of the epithelial differentiation spectrum, with clade A7 driving a less differentiated phenotype.

We related distinct patterns of viral gene expression to HPV clades and linked these to dysregulated genes in the tumor. The absence of *E2* expression in the A7-enriched cluster supports the current understanding that tumors with HPV-18 (clade A7) are always associated with integration, which leads to loss of *E2* expression[57]. Conversely, only about 76% of cervical tumors with HPV-16 (clade A9) show evidence of HPV integration[11], supporting the presumed presence of episomal HPV DNA due to the persistent expression of the *E2* gene observed in our samples. The presence of episomal HPV is associated with epithelial differentiation and active HPV infection[58]. This higher expression of episomal HPV genes in clade A9-infected samples further supports the hypothesis that molecular characteristics associated with clade A9 indicate more epithelial differentiation than with clade A7-infected samples, and such samples may have more active HPV infection.

We found that HPV clades exhibit distinct histone modification profiles. HPV viral proteins have been reported to interact with different epigenetic modifiers, including CREBBP, CHD4, KAT2B, EP300 and SNW1 (ref. [59]); however, clade-specific interactions that may explain our epigenomic changes at distinct genomic regions remain unexplored. The high frequency of samples with mutations

in chromatin modifiers in our cohort (87%) may suggest a mechanism beyond simple transcriptome dysregulation, perhaps encompassing variation in chromatin accessibility or three-dimensional chromatin structure that could promote HPV infection or cancer progression[14]. Such ideas await future studies.

Enrichment of active histone marks in close proximity to HPV integration events was associated with increased expression of nearby genes and ERVs. In our study, we cannot distinguish between the possibility that we are observing several distinct HPV integrations in similar regions in different cells within the same tumor (type 1 integrations[57,60]) and the possibility of multiple HPV integrations in tandem within single cells (type 2 integration[57,60]). However, the increased upregulation of genes and ERVs in tumors with a higher number of integration sites per event suggests that it may be the latter. Local alterations associated with HPV integration events may also result from focal amplification of the integrated viral genome and neighboring regions in the human genome[61]. These amplifications may support the enhanced recruitment of chromatin modifiers to viral regulatory regions[61], contributing to increased histone modifications in the regions surrounding HPV integration.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0673-7.

Received: 23 October 2019; Accepted: 26 June 2020;
Published online: 3 August 2020

References

1. Bodily, J. & Laimins, L. A. Persistence of human papillomavirus infection: keys to malignant progression. *Trends Microbiol.* **19**, 33–39 (2011).
2. de Sanjose, S. et al. Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. *Lancet Oncol.* **11**, 1048–1056 (2010).
3. Wright, J. D. et al. Human papillomavirus type and tobacco use as predictors of survival in early stage cervical carcinoma. *Gynecol. Oncol.* **98**, 84–91 (2005).
4. Yang, S.-H., Kong, S.-K., Lee, S.-H., Lim, S.-Y. & Park, C.-Y. Human papillomavirus 18 as a poor prognostic factor in stage I–IIA cervical cancer following primary surgical treatment. *Obstet. Gynecol. Sci.* **57**, 492–500 (2014).
5. Lai, C.-H. et al. Role of human papillomavirus genotype in prognosis of early-stage cervical cancer undergoing primary surgery. *J. Clin. Oncol.* **25**, 3628–3634 (2007).
6. Garland, S. M. et al. Impact and effectiveness of the quadrivalent human papillomavirus vaccine: a systematic review of 10 years of real-world experience. *Clin. Infect. Dis.* **63**, 519–527 (2016).
7. Bruni, L. et al. Global estimates of human papillomavirus vaccination coverage by region and income level: a pooled analysis. *Lancet Glob. Health* **4**, e453–e463 (2016).
8. Nakisige, C., Schwartz, M. & Ndira, A. O. Cervical cancer screening and treatment in Uganda. *Gynecol. Oncol. Rep.* **20**, 37–40 (2017).
9. Zubizarreta, E. H., Fidarova, E., Healy, B. & Rosenblatt, E. Need for radiotherapy in low and middle income countries—the silent crisis continues. *Clin. Oncol. (R. Coll. Radiol.)* **27**, 107–114 (2015).
10. Ferlay, J. et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int. J. Cancer* **144**, 1941–1953 (2019).
11. Cancer Genome Atlas Research Network. Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**, 378–384 (2017).
12. Ojesina, A. I. et al. Landscape of genomic alterations in cervical carcinomas. *Nature* **506**, 371–375 (2014).
13. Li, X. Emerging role of mutations in epigenetic regulators including *MLL2* derived from The Cancer Genome Atlas for cervical cancer. *BMC Cancer* **17**, 252 (2017).
14. 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).
15. Lleras, R. A. et al. Unique DNA methylation loci distinguish anatomic site and HPV status in head and neck squamous cell carcinoma. *Clin. Cancer Res.* **19**, 5444–5455 (2013).
16. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).
17. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
18. Henderson, S., Chakravarthy, A., Su, X., Boshoff, C. & Fenton, T. R. APOBEC-mediated cytosine deamination links *PIK3CA* helical domain mutations to human papillomavirus-driven tumor development. *Cell Rep.* **7**, 1833–1841 (2014).
19. Wallace, N. A. & Münger, K. The curious case of APOBEC3 activation by cancer-associated human papillomaviruses. *PLoS Pathog.* **14**, e1006717 (2018).
20. Zhang, H.-M. et al. AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res.* **43**, D76–D81 (2015).
21. Huang, F. W. et al. Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
22. Horn, S. et al. *TERT* promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
23. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
24. Garinet, S. et al. High prevalence of a hotspot of noncoding somatic mutations in intron 6 of *GPR126* in bladder cancer. *Mol. Cancer Res.* **17**, 469–475 (2019).
25. Wu, S. et al. Whole-genome sequencing identifies *ADGRG6* enhancer mutations and *FRS2* duplications as angiogenesis-related drivers in bladder cancer. *Nat. Commun.* **10**, 720 (2019).
26. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–3849 (2015).
27. Chu, J. et al. BioBloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters. *Bioinformatics* **30**, 3402–3404 (2014).
28. Schiffman, M., Clifford, G. & Buonaguro, F. M. Classification of weakly carcinogenic human papillomavirus types: addressing the limits of epidemiology at the borderline. *Infect. Agent Cancer* **4**, 8 (2009).
29. Maranga, I. O. et al. HIV infection alters the spectrum of HPV subtypes found in cervical smears and carcinomas from Kenyan women. *Open Virol. J.* **7**, 19–27 (2013).
30. Clifford, G. M. et al. Effect of HIV infection on human papillomavirus types causing invasive cervical cancer in Africa. *J. Acquir. Immune Defic. Syndr.* **73**, 332–339 (2016).
31. Morris, T. J. et al. ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics* **30**, 428–430 (2014).
32. Tian, Y. et al. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* **33**, 3982–3984 (2017).
33. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
34. Sandoval, J. et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**, 692–702 (2011).
35. Shen, J. et al. Exploring genome-wide DNA methylation profiles altered in hepatocellular carcinoma using Infinium HumanMethylation 450 BeadChips. *Epigenetics* **8**, 34–43 (2013).
36. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
37. Doolittle-Hall, J. M., Cunningham Glasspoole, D. L., Seaman, W. T. & Webster-Cyriaque, J. Meta-analysis of DNA tumor–viral integration site selection indicates a role for repeats, gene expression and epigenetics. *Cancers* **7**, 2217–2235 (2015).
38. Moody, C. A. & Laimins, L. A. Human papillomavirus oncoproteins: pathways to transformation. *Nat. Rev. Cancer* **10**, 550–560 (2010).
39. Monk, B. J., Tian, C., Rose, P. G. & Lanciano, R. Which clinical/pathologic factors matter in the era of chemoradiation as treatment for locally advanced cervical carcinoma? Analysis of two Gynecologic Oncology Group (GOG) trials. *Gynecol. Oncol.* **105**, 427–433 (2007).
40. Rader, J. S. et al. Genetic variations in human papillomavirus and cervical cancer outcomes. *Int. J. Cancer* **144**, 2206–2214 (2019).
41. Hoadley, K. A. et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).
42. Ross-Innes, C. S. et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
43. Lin-Shiao, E. et al. KMT2D regulates p63 target enhancers to coordinate epithelial homeostasis. *Genes Dev.* **32**, 181–193 (2018).

44. Herz, H.-M. et al. Enhancer-associated H3K4 monomethylation by Trithorax-related, the *Drosophila* homolog of mammalian Mll3/Mll4. *Genes Dev.* **26**, 2604–2620 (2012).
45. Hu, D. et al. The MLL3/MLL4 branches of the COMPASS family function as major histone H3K4 monomethylases at enhancers. *Mol. Cell. Biol.* **33**, 4745–4754 (2013).
46. Lee, J.-E. et al. H3K4 mono- and di-methyltransferase MLL4 is required for enhancer activation during cell differentiation. *eLife* **2**, e01503 (2013).
47. 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).
48. Pokholok, D. K. et al. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**, 517–527 (2005).
49. Gates, L. A., Foulds, C. E. & O'Malley, B. W. Histone marks in the 'driver's seat': functional roles in steering the transcription cycle. *Trends Biochem. Sci.* **42**, 977–989 (2017).
50. Hurst, T. P. & Magiorkinis, G. Activation of the innate immune response by endogenous retroviruses. *J. Gen. Virol.* **96**, 1207–1218 (2015).
51. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
52. Okoye, A. A. & Picker, L. J. $CD4^+$ T-cell depletion in HIV infection: mechanisms of immunological failure. *Immunol. Rev.* **254**, 54–64 (2013).
53. Hensley-McBain, T. & Klatt, N. R. The dual role of neutrophils in HIV infection. *Curr. HIV/AIDS Rep.* **15**, 1–10 (2018).
54. Sitole, B. N. & Mavri-Damelin, D. Peroxidasin is regulated by the epithelial–mesenchymal transition master transcription factor Snai1. *Gene* **646**, 195–202 (2018).
55. Zheng, Y.-Z. & Liang, L. High expression of PXDN is associated with poor prognosis and promotes proliferation, invasion as well as migration in ovarian cancer. *Ann. Diagn. Pathol.* **34**, 161–165 (2018).
56. Gifford, C. A. et al. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* **153**, 1149–1163 (2013).
57. McBride, A. A. & Warburton, A. The role of integration in oncogenic progression of HPV-associated cancers. *PLoS Pathog.* **13**, e1006211 (2017).
58. Kajitani, N., Satsuka, A., Kawate, A. & Sakai, H. Productive lifecycle of human papillomaviruses that depends upon squamous epithelial differentiation. *Front. Microbiol.* **3**, 152 (2012).
59. Ou, H. D., May, A. P. & O'Shea, C. C. The critical protein interactions and structures that elicit growth deregulation in cancer and viral replication. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **3**, 48–73 (2011).
60. 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).
61. Groves, I. J., Knight, E. L. A., Ang, Q. Y., Scarpini, C. G. & Coleman, N. HPV16 oncogene expression levels during early cervical carcinogenesis are determined by the balance of epigenetic chromatin modifications at the integrated virus genome. *Oncogene* **35**, 4773–4786 (2016).

Methods

Ethical compliance, consent and cohort enrolment. The study was approved by the Fred Hutchinson Cancer Research Center Institutional Review Board (7662) and complies with ethical regulation. Accrual received institutional and governmental approval, and informed consent was obtained from all patients. Approval was obtained from the BC Cancer Research Ethics Board (UBC BC Cancer REB H16-02279) for molecular characterization. Initially, 212 patients were enrolled and split into discovery ($n=123$) and extension ($n=89$) groups before further exclusions.

Pathology and molecular review. Formalin-fixed, paraffin-embedded (FFPE) tumor blocks or unstained sections were submitted for histopathological review by the Uganda Cancer Institute to the University of California, San Francisco. Hematoxylin and eosin (H&E) and p16 immunohistochemistry (IHC) slides sent to Nationwide Children's Hospital were imaged at ×40 with an Aperio scanner and assigned to three pathologists (M.H.S., T.C.W. and T.M.D.) for consensus review. Tumors were evaluated for histological type, subtype, grade and p16 immunoreactivity.

Gene expression analysis flagged 17 tumors for re-review that appeared discordant with initial H&E diagnosis of p16-positive poorly differentiated SCC. Twelve tumors were queried again with IHC stains for p63 and p40 (SCC) as well as BER-EP4, MOC 31 and B72.3 (adenocarcinoma). For two of these, analysis of neuroendocrine markers (synaptophysin and chromogranin) was also performed, leading to revised histological classification of two samples as adenocarcinoma, seven samples as adenosquamous carcinoma, two samples as neuroendocrine carcinoma and one sample as an undifferentiated carcinoma. Of the re-reviewed cases, five, negative for high-risk HPV types by BBT[27], were excluded from the final discovery cohort ($n=118$). Of these, four were uterine primary tumors and one was an equivocal cervical/uterine primary tumor.

A manual of standard operating procedures for NCI's Office of Cancer Genomics Cancer Genome Characterization Initiative is available at https://ocg.cancer.gov/sites/default/files/HTMCP_SOP_manual.pdf.

Clinical data and survival analyses. Clinical data, including overall survival, were obtained from https://cgci-data.nci.nih.gov/PreRelease/HTMCP-CC in January 2020. To mitigate clinical bias, we excluded patients who had not received a therapeutic intervention ($n=83$), as these individuals had more advanced disease ($P=0.01$, chi-squared test), had poorer prognosis ($P=0.00082$, log-rank test) and would not have been followed with curative intent. To account for prognostic differences between histologies, the analyses described only assessed survival differences in a subset of the patients with SSCs ($n=66$). HR and P values were determined by log-rank test.

Whole-genome sequencing library construction. PCR-free whole-genome sequencing libraries were constructed with the TruSeq DNA PCR-free kit (E6875-6877B-GSC, New England Biolabs), automated on a Microlab NIMBUS liquid handling robot (Hamilton), as previously described[62]. Libraries were purified with paramagnetic beads (Aline Biosciences), and concentrations were quantified before sequencing with a qPCR Library Quantification kit (KAPA, KK4824).

Genome library construction for custom capture. DNA (500 ng) was sonicated (Covaris) to fragments of 250–350 bp in size, purified with PCRClean DX magnetic beads (Aline Biosciences), end-repaired, phosphorylated and bead purified before A-tailing with a custom NEB Paired-End Sample Prep Premix kit. Samples were ligated to Illumina sequencing adaptors overnight at 16 °C, bead purified, enriched by six cycles of PCR with indexed primers enabling library pooling and sequenced with paired-end 125-base reads in a single flow cell lane.

Poly(A) RNA library construction. Poly(A) mRNA was purified from total RNA and cDNA was synthesized as previously described[62].

Native ChIP–seq. Fifty-two tumor samples were lysed in buffer with 0.1% Triton X-100 and 0.1% deoxycholate plus protease inhibitors (PI). Extracted chromatin was digested with micrococcal nuclease (MNase) enzyme (New England Biolabs), and the reaction was quenched with 250 µM EDTA. 1% Triton X-100 and 1% deoxycholate were mixed and added to the samples on ice. Four percent of the digested chromatin was used as an input control, and the remainder was precleared with Protein A/G Dynabeads (Invitrogen) in IP buffer (20 mM Tris-HCl (pH 7.5), 2 mM EDTA, 150 mM NaCl, 0.1% Triton X-100, 0.1% deoxycholate, PI) at 4 °C for 1.5 h. Supernatants were transferred to a 96-well plate containing antibody–bead complex, and plates were incubated overnight at 4 °C with agitation. Immunoprecipitated samples were washed twice with low-salt buffer (20 mM Tris-HCl (pH 8.0), 0.1% SDS, 1% Triton X-100, 2 mM EDTA, 150 mM NaCl) and twice with high-salt buffer (same, but with 500 mM NaCl). DNA–antibody complexes were eluted in elution buffer (100 mM $NaHCO_3$, 1% SDS) at 65 °C for 1.5 h with mixing (1,350 r.p.m.). Qiagen Protease was used to digest protein in the eluted DNA at 50 °C for 30 min with mixing (600 r.p.m.). ChIP DNA was purified on Sera-Mag beads (Fisher Scientific) with 30% PEG before library construction as described for custom capture.

Amplified libraries were purified as described above (Aline Biosciences), and DNA quality and quantity were determined with the Caliper LabChip GX DNA High-Sensitivity assay (PerkinElmer) and the Quant-iT dsDNA High-Sensitivity assay (Thermo Fisher Scientific).

Whole-genome, transcriptome and ChIP sequencing. Tumor genomes were sequenced to a target depth of 80× coverage and normal blood samples were sequenced to 40× coverage with 125-bp reads. Transcriptomes were sequenced with 75-bp paired-end reads. ChIP libraries were normalized and pooled before sequencing. All sequencing was performed on an Illumina HiSeq 2500.

Estimation of tumor content. Tumor purity and ploidy were estimated with Ploidetect (https://github.com/lculibrk/ploidetect). Tumor reads and heterozygous SNP allele frequencies in non-overlapping bins (~100 kb with equal coverage in the matched normal samples) were computed for each case. Read counts were modeled with Gaussian mixture models (GMMs) and modified to restrict component means as a fixed depth apart and component variances as equal to one another. Allele frequencies were modeled with a separate GMM, incorporating priors from the first. Models were generated for each possible value of tumor purity and scored on the basis of the mean likelihood of both the depth and allele frequency GMMs. All results were verified by review of GMM parameters and their fit to the data. Estimates were congruent with observed copy number data in 104 of 118 samples. In the remaining samples, purity and ploidy were determined by review of alternate models.

Somatic alteration detection. Tumor and normal sequencing reads were aligned to the human reference genome (hg19) with BWA-MEM v0.7.6a[63]. Read duplicates were marked with sambamba[64] (v0.5.5). Somatic single-nucleotide variants (SNVs) were identified with Strelka (v1.0.6)[65]. A panel of 2,735 genes including mutated oncogenes, tumor suppressors, epigenetic modifiers, splicing factors and other genes recurrently mutated (≥3 samples) in this cohort were selected for targeted sequencing in the extended cohort. The coding mutation rate was reported for each tumor as the number of coding SNVs (low, moderate or high SNPeff annotation[66]) per megabase.

Custom capture validation of SNVs. DNA samples from the 89 extension libraries were pooled before hybridization capture of 2,735 target genes with SureSelect XT custom probes (Agilent) and RNA probes at 65 °C for 24 h. Streptavidin-coated magnetic beads (Dynal, MyOne) were used for custom capture, followed by purification on MinElute columns (Qiagen) and enrichment with ten PCR cycles using primers that maintained library-specific indices. Pooled libraries were sequenced, generating 125-bp paired-end reads. To capture the *KMT2D* gene and noncoding hotspots, 544 120-bp xGen Lockdown probes were designed and synthesized (Integrated DNA Technologies) for targeted capture sequencing as above.

Significantly mutated genes. SMGs were identified with MutSig2CV (https://software.broadinstitute.org/cancer/cga/mutsig) as previously described[62].

Expression profiling. RNA-seq reads were aligned to the human reference genome (hg19) and converted to RPKM (reads per kilobase per million mapped reads) as described previously[67].

ChIP–seq alignment and peak calling. ChIP–seq reads (75 nucleotide) were aligned to the human reference genome (hg19) with BWA-MEM[63] (v0.7.6a; parameter -M). Read duplicates were marked with sambamba[64] (v0.5.5). Forty-seven samples had all six histone marks (four broad: H3K4me1, H3K9me3, H3K27me3 and H3K36me3; two narrow: H3K4me3 and H3K27ac), and five had a subset of these.

Peaks were called with MACS2 (v2.1.1)[68] using default parameters, comparing each mark to its control. Bedgraph output files were converted to library-size-normalized bigWig format for manual inspection with the UCSC and IGV genome browsers[69,70].

ChIP–seq data quality was assessed with ENCODE guidelines[71]. Samples had a minimum of 50 million sequenced reads for narrow marks and 100 million sequenced reads for broad marks. The percentage of uniquely mapped reads was above 70%, and the percentage of duplicated reads varied between 1% and 10%. The nonredundant fraction, fraction of reads in peaks (FRIP) and sequencing saturation determined with preseq v2.0.2 (https://github.com/smithlabcode/preseq) were also assessed.

HPV typing and expression. Microbial detection, HPV typing and HPV integration detection were performed with BBT (v2.0.11b)[27]. Where two or more HPV types were integrated ($n=3$), the dominant type was determined by *E6*/*E7* expression. Where no integration was found ($n=9$), the dominant HPV type was determined as the type with the most read evidence.

To determine expression of HPV genes, fasta genome references and gff annotation files were downloaded from NCBI for 16 HPV strains. HPV-51 did not have a gff file, and the one sample infected with this was therefore excluded (samples with HPV expression, $n=117$). Samples were aligned to their HPV strain with BWA-mem v0.7.6a sambamba. The fraction of reads with sequencing quality

greater than Q10 within gene boundaries was counted and normalized to reads per kilobase of exons per million reads mapped to HPV (RPKM).

Mutation signatures and HRD score. SNVs were categorized into 96 mutation classes on the basis of 6 variant types and 16 trinucleotide contexts. For each sample, values for the 96 classes were used to compute a non-negative least-squares deconvolution on the basis of 30 previously described mutational signatures (COSMIC)[16,17]. The APOBEC signature reported for each sample is the maximum exposure value of signature 2 or 13.

HRD scores were computed with HRDtools (v0.0.0.9, R), as previously described[72].

Copy ratio landscape comparisons. Copy number alterations between cohorts were called and analyzed with GATK4 (ref. [73]; v4.0.9; https://gatkforums.broadinstitute.org/gatk/discussion/9143/how-to-call-somatic-copy-number-variants-using-gatk4-cnv) and GISTIC2.0 (ref. [74]; v2.0.17). Genomic intervals were prepared by dividing the reference genome into equally sized bins (1,000 bp). A panel of normal samples was generated to median sample reference counts. Allele counts were collected independently for the tumor and matching normal sample. Continuous segments were modeled with both the allelic and copy ratios.

Germline copy number alterations previously identified in the TCGA cervical cancer (CESC) study[11] were filtered out to remove any potential germline copy number variations in this cohort. Segments were excluded if 75% or more of the segment overlapped with these.

Somatic copy number alterations in TCGA CESC tumors were determined previously with SNP 6.0 arrays[11], and these were downloaded from the Broad GDAC website. The 178 samples in the TCGA CESC core set were used for comparison to our HIV$^-$ samples.

To determine regions of variance in copy number alterations between cohorts (HIV$^-$ versus HIV$^+$, HIV$^-$ versus TCGA), analyses were performed on each cohort separately according to the GISTIC2.0 (v2.0.22) documentation (http://software.broadinstitute.org/cancer/software/genepattern/modules/docs/GISTIC_2.0) with parameters -qvt 0.25, -genegistic 1, -broad 1, -brlen 0.5, -conf 0.99, -armpeel 1, -savegene 1, -gcm extreme and -maxseg 3000. The genome was binned into 1-kb segments, and the fraction of patients with a copy gain (>0.1) or loss (<−0.1) was calculated on the basis of the mean segment values for each cohort. Significantly amplified and deleted chromosome arms were identified with a threshold of FDR < 0.25. Unique arms and cytobands were identified as those significant in one cohort but not the other.

Specific copy number alterations in samples. Regions of copy number alteration in individual samples were identified as previously described[62].

Noncoding mutation hotspots. Noncoding variants annotated by SNPeff[66] as 5′ flank, 3′ flank, IGR, 3′ UTR, intron, 5′ UTR, RNA, splice site and translation start site were used as input to Rainstorm[75], with all parameters set at default values ($k = 4$).

In the 3,094 hotspot regions identified, we focused on 3,539 potential point mutation hotspots, present in three or more samples. These were filtered for those called by both Strelka and MutationSeq[76] and did not reside in centromeric regions. Further filtering removed any variant called in a normal sample, reducing the potential noncoding hotspots to 404, of which 7 (high confidence) were confirmed by manual review.

Hotspots were annotated as 'potential promoter', 'potential enhancer' or 'intergenic' with ChIP–seq data (enhancer, intersect of H3K4me1 and H3K27ac peaks; promoter, H3K4me3 peaks).

We assessed the potential for other noncoding hotspots to alter transcription factor binding with motifBreakR[26].

DNA methylation analysis. Human DNA methylation analysis using the EPIC array (Illumina) was performed by the Centre for Applied Genomics at the Hospital for Sick Children (Toronto, Canada). The DNA methylation beta values for 115 samples were binarized as unmethylated (beta ≤ 0.25) or methylated (beta > 0.25). The 8,000 most variable probes were clustered with ConsensusClusterPlus[77] (v1.38.0, R) using a 'binary' distance and 'ward.D2' clustering method with 1,000 iterations.

DMPs and DMRs between clade A7- and clade A9-infected samples were determined with CHAMP[31,32] (v2.10.2, R; $q < 0.05$); DMR determination used the 'bumphunter' method. For DMPs, associated genes, genomic features and CpG island features came from CHAMP[31,32]. DMRs were intersected with protein-coding genes (hg19 Ensembl (v75), $n = 20{,}232$) by using bedtools (v2.27.1)[78].

Human and viral gene expression and gene ontology enrichment analyses. Clustering analysis was performed with ConsenusClusterPlus[77] (v1.38.0, R) by using $\log_{10}$ (RPKM) values with the 'Pearson' method and 'ward.D2' linkage with 1,000 iterations. Human genes used included the top 1,000 genes with the most variable expression (RPKM > 5 in at least one sample). All 118 samples were included in human gene clustering, and 117 samples were included in viral gene clustering (no gff file was available for HPV-51).

Differential gene expression between groups (A7 versus A9; *E6* and *E7* high versus low) was performed with DESeq2 (v1.14.1, R)[33]. Genes were filtered for those with adjusted $P < 0.05$, fold change > 1.5 in mean expression and baseMean expression > 1,000. For the A7 versus A9 comparison, the differential analysis was normalized for histology with a multifactorial approach. Results from the normalized analysis were compared to those obtained when using only squamous samples with A7 and A9 to ensure the histology correction was only removing expression differences attributed to histology (89% concordance).

Functional enrichment analysis of the significantly differentially expressed genes in the A7 versus A9 comparison was performed with STRING (v11.0)[36]. For visualization, enrichment scores for A7-enriched ontologies were set to negative values. Functional enrichment analysis of the significantly differentially expressed upregulated and downregulated gene lists for the *E6* and *E7* comparison was performed with HOMER (v4.10.3)[79].

ChIP clustering analyses. The union of peaks for each histone mark was found by concatenating peak files and merging overlapping regions with bedtools v2.27.1 (ref. [78]). The normalized coverage of each sample in the peak union was determined with deeptools (v3.0.2)[80]. For each mark, the top 1% most variable peaks were clustered with ConsensusClusterPlus (v1.38.0, R) using the 'pearson' distance and 'complete' clustering method with 1,000 iterations for $k = 2$–10 clusters. The 54 consensus clustering solutions (6 marks × 9 solutions) were then analyzed with cluster of clusters analysis (COCA)[41]. For active marks, pairwise probabilities were generated for 27 solutions (3 marks × 9 solutions). For marks for which some samples had missing data, pairwise comparisons were normalized to exclude samples for those marks. Matrices of probabilities (54 × 52, 27 × 52) were clustered with pheatmap (v1.0.10, R) using the 'pearson' distance and 'complete' clustering method.

H3K4me3, H3K27ac and H3K4me1 peaks differentially present between HPV clades (A7 versus A9) were determined by using DiffBind (http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf) with the DESeq2 method (v2.2.12, R; FDR < 0.01, fold change > 2)[42]. Coverage at peaks was counted for the 500 bp around the center of the peak, and a multifactorial experimental design was performed to normalize histology differences (referred to as blocking factor). Significantly differential peaks were intersected with bedtools (v2.27.1)[78]. Associated genes were defined as the genes with the nearest TSS to the differential H3K4me1 and intersected H3K4me3 and H3K27ac regions, identified with bedtools[78] (v2.27.1) with respect to RefSeq's hg19 annotation.

H3K4me1-marked enhancer regions. H3K4me1-marked enhancer regions were selected from the union of H3K4me1 peaks[81]. In each sample, H3K4me1 and H3K27ac were overlapped in this union to identify primed (H3K4me1) and active (H3K4me1 and H3K27ac) regions, excluding those overlapping with H3K27me3. To eliminate regions marking promoters, the regions were filtered for median H3K4me1:H3K4me3 ratio coverage > 1 and a distance of >2,000 bp from a TSS ($n = 324{,}447$ regions).

HPV integration events and ChIP. HPV integration sites were determined with chimeric reads mapping to both human and HPV genomes. Within each sample, integration sites were merged into a single integration event ($n = 257$) if they were <500 kb apart. HPV integration hotspots were determined by counting the number of events that fell within a 500-kb bin across the genome.

ChIP–seq alterations at HPV integration events were clustered by using the $\log_2$ (fold change) in normalized coverage (RPM) of the integrated sample versus the mean RPM of the unintegrated samples with pheatmap (v1.0.10, R) using a 'ward.D2' clustering method. Events of <20 kb in length were extended to 20 kb to obtain adequate coverage of the region.

For each mark in an event (6 modifications in 99 events), a control peak set was made by randomly selecting 1,000 peak regions of the same mark on the same chromosome as the event and extending the peaks from their center to be the same size as the event. Normalized ChIP–seq coverage of the histone modification at these 1,000 random peaks was determined in 52 samples, and the $\log_2$ (fold change) in coverage was calculated for the integrated sample. A P value was calculated for the fold change in the integration event on the basis of the distribution of fold change values in the control peak set. Benjamini–Hochberg-adjusted $P < 0.05$ was regarded as significant.

HPV integration events and expression. For each integration event ($n = 257$), we identified all protein-coding genes (hg19 Ensembl (v75), $n = 20{,}232$) that fell within the event ± 10 kb (ref. [37]), which identified 255 genes near integration events. Fold changes in integrated samples were calculated on the basis of the mean expression for all samples lacking events, and P values were derived from the distribution of expression of the gene across all samples. Oncogenes were identified with OncoKB[82]. The same method was applied to identify ERVs upregulated at HPV integration events ($n = 34$ events). Samples were labeled as having a statistically significant integration event if they had fold change ≥ 2 and Benjamini–Hochberg-adjusted $P \leq 0.05$ for genes and fold change ≥ 10 and Benjamini–Hochberg-adjusted $P \leq 0.05$ for ERVs, on the basis of the distribution of fold change values for each (Extended Data Fig. 4). Samples with significant events

were correlated with T cell infiltration scores from CIBERSORT[51] and with genes from the gene ontologies for dsRNA-sensing pathways (GO:0043330) and type I interferon signaling (GO:0060337).

ERV quantification. A total of 5,467,457 repeat elements and hg19 coordinates (chromosomes 1–22 and X) were downloaded from RepeatMasker Open v4.0.5 (http://www.repeatmasker.org/faq.html). To minimize read count bias from nearby expressed protein-coding genes, we filtered for ERVs >10 kb away from their nearest gene. Raw expression values were calculated by counting the number of reads that mapped unambiguously (mates mapped within 10 kb) to each region and were normalized for sequencing depth and length by conversion to RPKM.

Estimation of immune cell content. CIBERSORT (v1.0.4)[51] was used to quantify leukocyte expression signatures on the expression RPKM data as previously described[62]. Total $CD4^+$ T cell content was taken as the sum of the content for the following cells; naive, memory resting, memory activated, follicular helper and regulatory T cells.

Visualization. All heat maps were visualized with pheatmap (v1.0.10, R).

Statistical analyses. No sample sizes were predetermined. Unless otherwise stated, all statistical tests correspond to two-sided tests. *P*-value methods and multiple-test correction are reported in the text. Wilcoxon in the text refers to the Wilcoxon rank-sum test.

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

Data availability

All molecular and clinical data used in this publication can be found on the National Cancer Institute's Genome Data Commons Publication Page at https://gdc.cancer.gov/about-data/publications/CGCI-HTMCP-CC-2020. Data from this publication are publicly available for download through dbGaP (phs000528), as part of the NCI Cancer Genome Characterization Initiative (CGCI; phs000235). Sample metadata are reported in Supplementary Table 2. TCGA cervical cancer data (file name: CESC.snp__genome_wide_snp_6__broad_mit_edu__Level_3__segmented_scna_minus_germline_cnv_hg19__seg.seg.txt) were obtained from http://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/CESC/20160128/. Source data are provided with this paper.

Code availability

Bioinformatics analyses in this study were conducted with open-source software, with the exception of tumor purity and ploidy estimation, which was performed with Ploidetect (https://github.com/lculibrk/ploidetect).

References

62. Pleasance, E. et al. Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nat. Cancer* **1**, 452–468 (2020).
63. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
64. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
65. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
66. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
67. Chun, H.-J. E. et al. Genome-wide profiles of extra-cranial malignant rhabdoid tumors reveal heterogeneity and dysregulated developmental pathways. *Cancer Cell* **29**, 394–406 (2016).
68. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
69. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
70. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).
71. Landt, S. G. et al. ChIP–seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
72. Zhao, E. Y. et al. Homologous recombination deficiency and platinum-based therapy outcomes in advanced breast cancer. *Clin. Cancer Res.* **23**, 7521–7530 (2017).
73. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
74. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
75. Arthur, S. E. et al. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat. Commun.* **9**, 4001 (2018).
76. Ding, J. et al. Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. *Bioinformatics* **28**, 167–175 (2012).
77. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).
78. Quinlan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.1–11.12.34 (2014).
79. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
80. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
81. Pellacani, D. et al. Analysis of normal human mammary epigenomes reveals cell-specific active enhancer states and associated transcription factor networks. *Cell Rep.* **17**, 2060–2074 (2016).
82. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* https://doi.org/10.1200/PO.17.00011 (2017).

Acknowledgements

This project has been funded in whole or in part with US federal funds from the National Cancer Institute, National Institutes of Health, under contract no. HHSN261200800001E and HHSN261201500003I. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US government. We gratefully acknowledge the Fred Hutchinson Cancer Research Center and the Uganda Cancer Institute for overseeing sample and data collection in Uganda. We are grateful for contributions from the other members of the HTMCP Cervical Cancer Working Group at the Department of Epidemiology, University of Alabama at Birmingham, the Pancreas Centre BC and various groups at Canada's Michael Smith Genome Sciences Centre, including those from the Biospecimen, Library Construction, Sequencing, Bioinformatics, Technology Development, Quality Assurance, LIMS, Purchasing and Project Management teams. We thank the AIDS and Cancer Specimen Resource for logistical coordination and support of this project through NIH grants U01CA066535, U01CA096230 and UM1CA181255. L.C. and V.L.P. are the recipients of CIHR Frederick Banting and Charles Best Canada Graduate Scholarships GSD-164207 and GSD-152374, respectively. S.J.M.J. is the recipient of the Canada Research Chair in Computational Genomics. This research was supported by the Intramural Research Program of the NIH, National Cancer Institute (R.Y.). C.C. is supported by NIH grant P30AI027757. G.B.M. is supported by NCI grants U01CA217842 and P50CA098258. M.A.M. is the recipient of the Canada Research Chair in Genome Science. This work was supported in part by funding provided by the Canadian Institutes for Health Research (CIHR award FDN-143288) to M.A.M. A.I.O. was supported in part by the Endlichhofer Trust (OCCC 3120957) and a V Foundation grant (DVP2018-007).

Author contributions

A.G., V.L.P., Z.Z., R.B. and E.T. contributed equally to this work. J.S.R., A.I.O., D.S.G., A.J.M. and M.A.M. equally supervised this work. The HTMCP Cervical Cancer Working Group contributed collectively to this work. Project management and data coordination: K.N., M.A.D. and P.G. Cohort and clinical data collection: C.C., C. Nakisige, C. Namirembe, J.O., M.O., N.B.G., H.P., J.B. and J.M.G.-F. Pathology and molecular review: T.M.D., M.H.S., T.C.W. and R.B. Data were generated by Canada's Michael Smith Genome Sciences Centre at BC Cancer and analyses were performed by V.L.P., Z.Z., R.B. and E.T. Contribution to analyses: G.B.M., R.Y., S.J.M.J., Y.M., K.L.M., A.G., S.K.C. and L.C. A.G., A.J.M., V.L.P., E.T. and M.A.M. wrote the manuscript. All authors reviewed and edited the manuscript.

Competing interests

G.B.M. reports the following potentially competing interests: SAB/consultant: AstraZeneca, Chrysallis Biotechnology, ImmunoMET, Ionis, Lilly, PDX Pharmaceuticals, Signalchem Lifesciences, Symphogen, Tarveda, Zentalis; stock/options/financial: Catena Pharmaceuticals, ImmunoMet, SignalChem, Tarveda; licensed technology: HRD assay to Myriad Genetics, DSP patents with Nanostring; sponsored research: Nanostring Center of Excellence, Ionis (provision of tool compounds). R.Y. reports the following potentially competing interests: research support from a CRADA with Celgene/BMS. T.C.W. reports the following potentially competing interests: consultant to Roche, BD and Inovio with respect to HPV diagnostic tests and therapeutic vaccines.

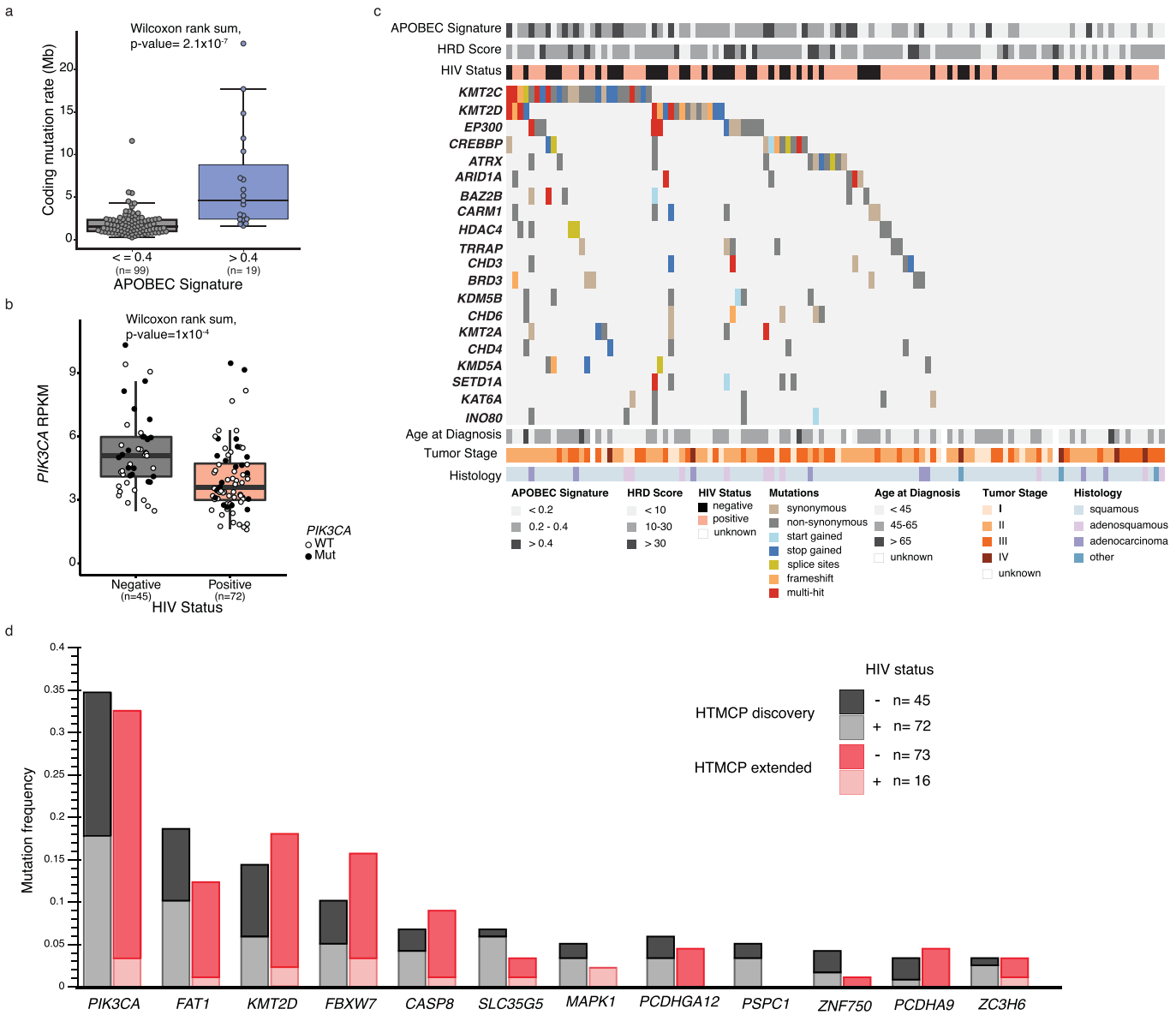
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-020-0673-7.

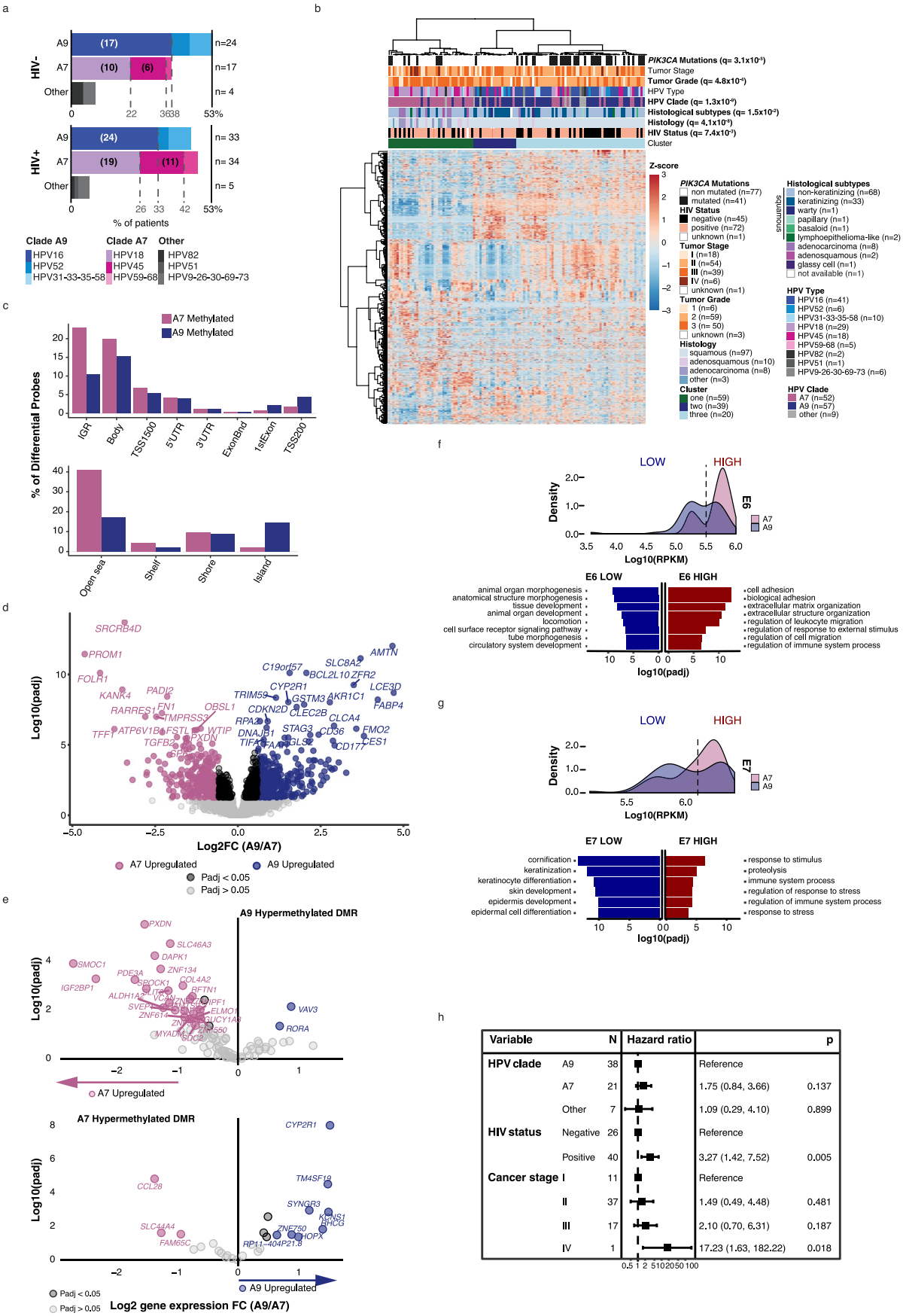
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0673-7.

Correspondence and requests for materials should be addressed to M.A.M.

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

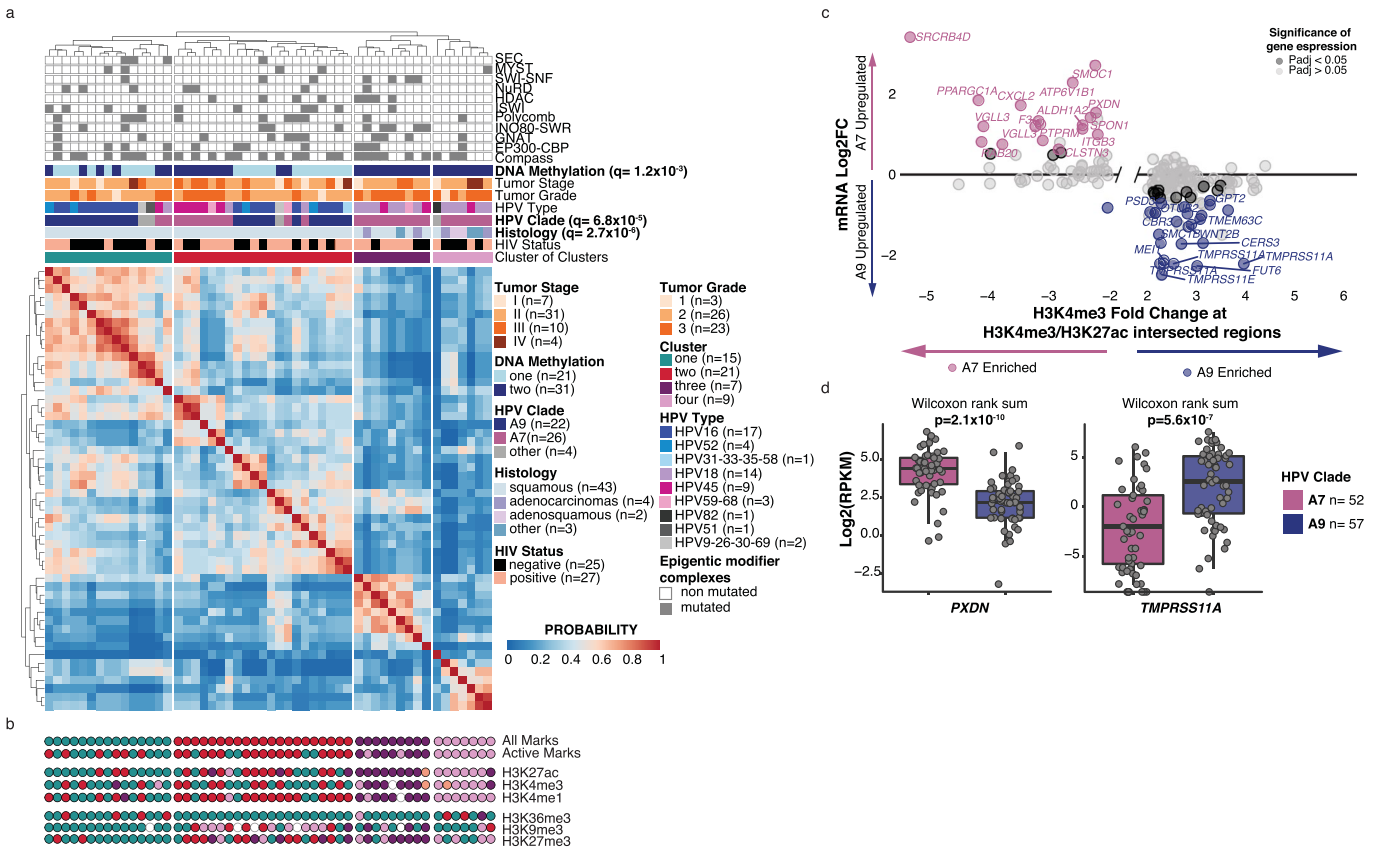


Extended Data Fig. 1 | Additional characteristics of the HTMCP discovery and extension cohorts. a. Coding mutations per Mb in samples exhibiting low (≤ 0.4) and high (> 0.4) APOBEC signatures. **b**. Difference in *PIK3CA* expression by HIV status. **c**. Mutations in the top 20 most mutated epigenetic modifiers, ordered by frequency of alterations for the cohort (n = 118). APOBEC signature proportion and homologous recombination deficiency (HRD) scores are reported above. HIV status, age at diagnosis, tumor histology ("other" includes neuroendocrine and undifferentiated) and stage are also annotated. **d**. Comparison of mutation frequencies of the 12 SMGs in the discovery vs. extension cohorts. Boxplots in **a** and **b** represent the median, upper and lower quartiles of the distribution and whiskers represent the limits of the distribution (1.5-times interquartile range), and statistics were determined using two-sided Wilcoxon rank sum tests.

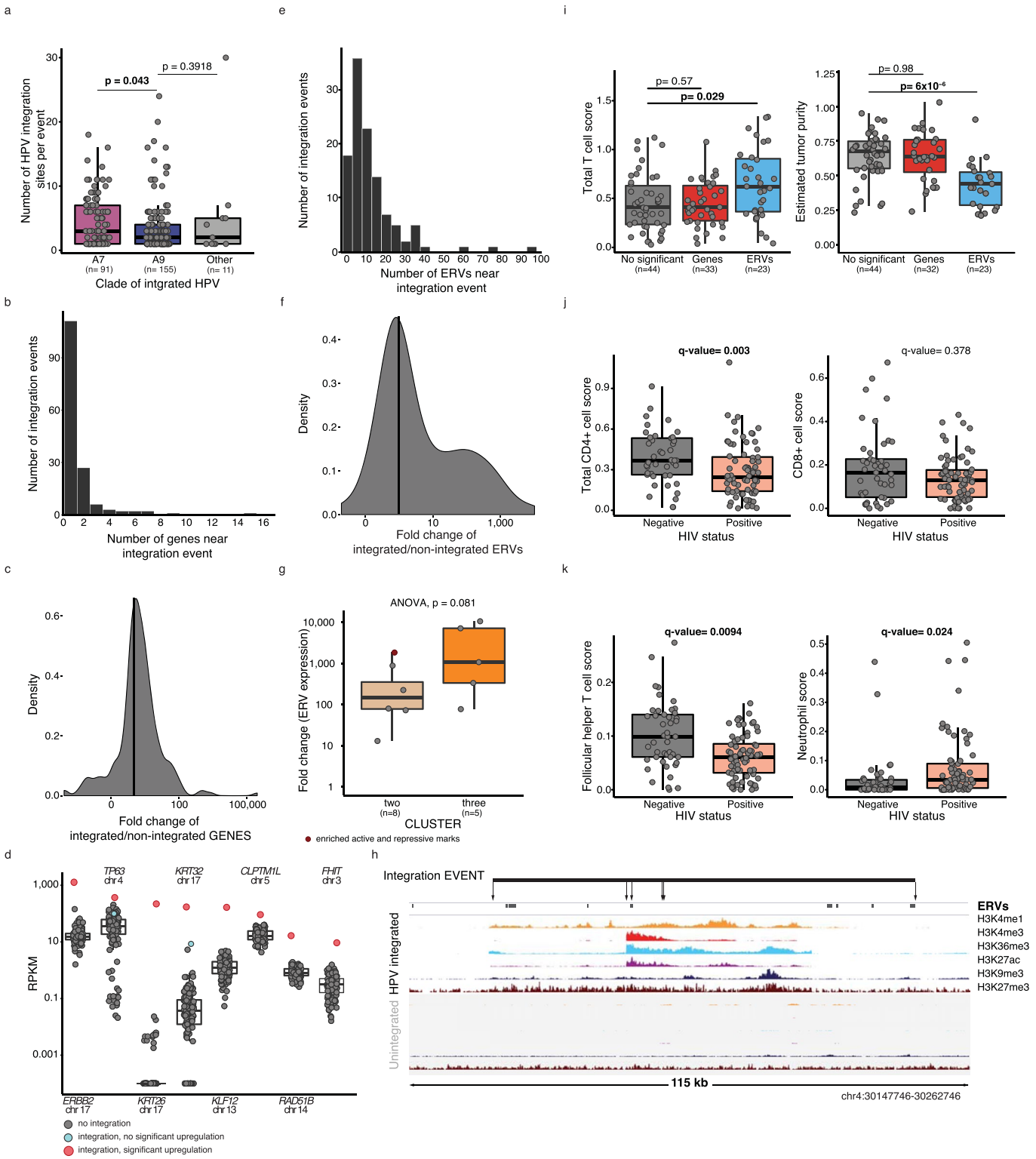


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Association of HPV clades with HIV status, gene expression, DNA methylation and survival. a. HPV types in our cohort separated by HIV status (n = 72 positive samples, n = 45 negative), and clade. The x axis indicates the percentage of samples in that cohort infected by the indicated HPV type, and in brackets is the number of samples. **b**. Unsupervised clustering of the top 1,000 most variable genes across our cohort (n = 118 samples). q-values were determined using Benjamini-Hochberg (BH) corrected Fisher exact tests. **c**. Percentage of differentially methylated probes between clades (A7 = 51 samples, A9 = 56 samples) at different genomic features, by HPV clade. **d**. Log2 fold change and adjusted (BH) p-value of differentially expressed genes between clade A7- (n = 52) and A9-infected (n = 57) samples. **e**. Volcano plots showing the log2 fold change and adjusted p-value (BH) of differentially expressed genes between clade A7- (n = 52) and A9-infected (n = 57) samples associated with A9 hypermethylated (top), and A7 hypermethylated (bottom) differentially methylated regions (DMRs). **f, g**. *top:* Kernel density of *E6* (**f**) and *E7* (***g***) expression in the HTMCP cohort separates samples into high- and low expressing cases. *bottom*: gene ontologies enriched in differentially expressed genes in samples with low / high E6 (n = 68 / n = 48) (**f**) and E7 (n = 58 / n = 59) (**g**). **h**. Multivariate cox proportional hazards model for HPV clade, HIV status and disease stage for 66 patients. Hazard ratios and p-values reported for each variable were determined using log-rank tests. Where relevant, all statistical tests were two-sided.



Extended Data Fig. 3 | Correlations between histone modifications and gene expression. a. Cluster of clusters analysis for 54 consensus clustering solutions for all histone marks on 52 samples (solutions with k = 2 to 10 for each mark). The heatmap color indicates the sample probabilities in the consensus matrix. q-values for each variable were determined using Benjamini-Hochberg corrected Fisher exact tests. **b**. Schematic showing the cluster of clusters solution (k = 5 for H3K27ac and H3K4me3 and k = 4 for the other marks) for all histone marks and for the 3 active marks. Each dot represents a sample and dot color represents the cluster membership of the sample. Hollow circles indicate no available ChIP data for that sample. **c**. Fold change of H3K4me3 abundance and gene expression between clades associated with TSS of genes (−5/+20 kb) found at intersecting H3K4me3 and H3K27ac peaks. Sample Ns used for differential analyses (and derivation of adjusted p-values) were: expression A7 = 52, A9 = 57; H3K4me3 and H3K27ac A7 = 25, A9 = 22. Genes with BH-adjusted p-values <0.05 (DESeq, Methods) are highlighted. **d**. Expression of the genes reported in Fig. 4f separated by HPV clade. Boxplots represent the median, upper and lower quartiles of the distribution and whiskers represent the limits of the distribution (1.5-times interquartile range), and p-values were calculated by Wilcoxon rank sum tests. Where relevant, all statistical tests were two-sided.



Extended Data Fig. 4 | HPV integration events and tumor microenvironments. a. Number of HPV integration sites per event separated by HPV clade. **b**, **c**, **e**, **f**. Distribution of the number (**b**, **e**) and fold change in integrated samples (**c**, **f**) of genes (**b**, **c**) and ERVs (**e**, **f**) near integration events. **d**. Expression (RPKM) of selected genes near HPV integration events in each sample (n = 118). **g**. Fold change of ERVs nearby integration events separated based on the clusters identified in Fig. 5f. **h**. Histone mark coverage of a 115 kb genomic region containing ERVs. The line represents an integration event, and arrows indicate individual integration sites. Top tracks refer to a case with integration, and the bottom to a control case without integration. **i**. Total T-cell scores and estimated tumor content of samples with HPV integration events that are associated with significant changes in expression of ERVs or genes, and those that are not. **j, k**. CIBERSORT scores for all CD4 + T-cells (sum) and CD8 + T-cells **(j)**, Follicular helper T-cells and neutrophils **(k)** separated by HIV status (HIV + n = 72, HIV- n = 45). Boxplots in **a**, **d**, **g**, **i**–**k** represent the median, upper and lower quartiles of the distribution and whiskers represent the limits of the distribution (1.5-times interquartile range). All p-values were determined by Wilcoxon tests unless otherwise stated, and q-values were corrected using the Benjamini-Hochberg method. Where relevant, all statistical tests were two-sided.

nature research

Corresponding author(s): Marco Marra

Last updated by author(s): May 26, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Sequencing was performed on the HiSeq2500 and bases were called using Illumina bcl2fastq. No custom code was used for data collection.
Data analysis	All analyses used previously published tools, with the exception of tumor content and ploidy (https://github.com/lculibrk/ploidetect). Tools used for each analysis are as follows: DNA-Seq alignment, somatic variant calling: Human reference genome (hg19) BWA-MEM (v0.7.6a) Sambamba (version 0.5.5) Strelka (v1.0.6) in house consensus caller Genome Validator (v2.3) Identification of significantly mutated genes (SMGs): MutSig2CV (v3.11) Genes meeting significant criteria (q<0.1) were selected as SMGs Mutation signatures and HRD score: SNVs signatures were deciphered using a published framework (Based on the Catalog of Somatic Mutations in Cancer (COSMIC) available from https://cancer.sanger.ac.uk/cosmic/signatures_v2) R package HRDtools (v0.0.0.9) was used to compute HRD scores as the arithmetic sum of loss of heterozygosity (LOH), TAI, and LST scores. Copy number landscape and analysis of HTMCP HIV+, HIV-, and TCGA core set cohorts: GATK4 (v4.0.9)

GISTIC_2.0 (v2.0.22)

Copy number alterations in samples:
CNAseq (v0.0.6)

Non-coding mutation hotspot analysis and transcription factor binding site prediction:
snpEFF (v4.1)
Rainstorm (september 28, 2017)
MutationSeq (v4.3.8)
Strelka (v1.0.6)
motifBreakR (v1.13.3)

RNA-Seq alignment, expression profiling:
BWAMEM (v0.7.6a)
software developed in house JAGuaR v1.7.5
Ensembl v69 was used for gene annotation

ChIP-Seq peak calling and QC
callpeak command of MACS2 (2.1.1) (https://github.com/jsh58/MACS) "broad" option was used with the broad mark libraries (H3K4me1, H3K9me3, H3K27me3 and H3K36me3)
preseq 2.0.2(https://github.com/smithlabcode/preseq)

HPV typing
BioBloom tool (v2.0.11b)

Hierarchical clustering and heat-map plotting:
ConsensusClusterPlus2 R package (v1.38.0)
pheatmap R package (v1.0.10)

DNA methylation analysis:
CHAMP (v2.10.2) R package
DMRs were intersected with protein-coding genes (hg19 Ensembl (v75), n=20,232) using bedtools (v2.27.1)

Human and viral gene expression and gene ontology analyses:
DESeq2 R package (v1.14.1)
Proteins with Values/Ranks in STRING (v11.0)
HOMER (v4.10.3)

ChIP-sequencing analyses:
deeptools (multiBigWigSummary BED-file v3.0.2)
DiffBind R package (v2.2.12)

Analysis of HPV integration events:
bedtools v2.27.1
genes (Ensembl (v75)) within +/-10 kb of an integration event were associated with the event.
ERVs (from RepeatMasker Open v.4.0.5 (http://www.repeatmasker.org/faq.html)) within +/-10 kb were associated with the event.
Immune cell content based on RNA-seq:
CIBERSORT (v1.0.4)

Estimation of tumor content:
Tumor purity and ploidy were estimated using in house tool Ploidetect (https://github.com/lculibrk/ploidetect).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All molecular and clinical data used in this publication can be found on the National Cancer Institute's Genome Data Commons Publication Page https://gdc.cancer.gov/about-data/publications/CGCI-HTMCP-CC-2020. Data from this publication is publicly available for download through dbGaP (phs000528), as part of the NCI Cancer Genome Characterization initiative (CGCI, phs000235). Sample metadata is reported in Supplementary Table 2. Source data for all Figures and Extended Data Figures are presented with the paper.
TCGA cervical cancer data (file name:
CESC.snp__genome_wide_snp_6__broad_mit_edu__Level_3__segmented_scna_minus_germline_cnv_hg19__seg.seg.txt) was

obtained from http://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/CESC/20160128/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine cohort size, however power analysis indicated that we had 92% power to detect an effect size of 0.5 in our HIV- samples, and 99% power to detect an effect of this size in the HIV+ samples, at a significance of 0.05. These numbers demonstrate that we had a sufficiently large cohort to determine moderate differences, but less frequent events and smaller effects may require a larger cohort to obtain statistically significant results. This cohort size was sufficiently large, however, to robustly detect HPV clade associated patterns in multiple data types (RNA, ChIP, methylation) with sufficiently small p-values after multiple test correction. Additionally, we were adequately equipped to identify significantly mutated genes at a frequency of 3% in the cohort.
Data exclusions	Each participant who consented to the study had a sample undergo local pathology review, as well as at a central laboratory. The participants continued enrollment in the study was contingent upon a confirmatory result from the central pathology laboratory. Five samples, with discordant pathology and molecular review were re-reviewed and excluded from the final discovery cohort (n=118) because four were uterine primaries and one an equivocal cervical/uterine primary.
Replication	The extension cohort included 89 cases from the same geographical region. These samples were used to validate genes mutated in the discovery cohort by targeted sequencing. Of the 12 significantly mutated genes identified in the discovery cohort, 11 were confirmed in the extension cohort. PSPC1 is the only significantly mutated gene that did not validate. Mutations in this gene were present in only 5% of the discovery cohort, and mutations at this frequency may require a higher number of samples to have the power to detect them.
Randomization	We performed a prospective observational cohort study which did not require randomization of the samples. The covariates studied in this cohort represented inherent clinical features (such as HIV status, HPV clade, tumor grade and stage) and so samples were not allocated into experimental groups.
Blinding	Investigators were not blinded in this study as samples were studied based on inherent features rather than different treatments. These inherent characteristics including HIV status and HPV clade are not subject to bias by investigators. When appropriate, unsupervised clustering analyses where performed on different data types to identify patterns independently of known characteristics.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	These description refers to the 118 cases in the discovery cohort: HIV+ patients 72/118; HIV- patients 45/118, 1 patient unknown, Median age of HIV+ patients was 43 years old, and HIV- patients was 54 years old. Cervical cancer histology: 97 samples were categorized as squamous cell carinoma, 10 samples as adenosquamous cell carinoma, 8 samples as adenocarcinoma, 2 samples as neuroendocrine, 1 sample as undifferentiated.

	Cancer grade: 50 samples were cancer grade G3, 59 samples were cancer grade G2, 6 samples were G1 and 3 unknown. Cancer stage: 5 samples were stage IB1, 13 samples were stage IB2, 2 samples were stage IIA, 3 samples were stage IIA1, 6 samples stage IIA2, 43 samples stage IIB, 2 samples stage IIIA, 37 samples stage IIIB, 4 samples stage IVA, 2 samples stage IVB, 1 sample was not available.
Recruitment	Any patient undergoing cervical cancer diagnostic assessment or who has been diagnosed with cervical cancer at the Uganda Cancer Institute (UCI - Kampala, Uganda) and provided informed consent were enrolled for this study. We are not aware of any self-selection or other biases in patient selection. Cervical cancer is one of the HIV-associated cancers and HIV+ is prevalent in Uganda. Therefore this cohort has enabled insight into a previously under-explored cohort of cervical cancers.
Ethics oversight	Fred Hutchinson Cancer Research Center (Institutional review board - IR file #7662)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	ChIP-seq data access can be requested via dbGAP at this link https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000528.v11.p2
Files in database submission	305 BAM and 305 bigwig files were submitted at the NIH data coordinating centre (DCC). 52 bam and 52 bigwig files were submitted for each of the following marks: H3K27ac, H3K36me3 and H3K27me3; 51 bam and 51 bigwig files were submitted for each of the following marks: H3K4me3 and H3K4me1 47 bam and 47 bigwig files were submitted for H3K9me3.
Genome browser session (e.g. UCSC)	https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr20%3A100000%2D150000&hgsid=843747109_eKEYdQ2aKzVfAvGPlHVJhLa15BrD

Methodology

Replicates	52 cervical cancer samples were sequenced once per ChIP-seq experiment. 52 samples have ChIP-seq data for H3K27ac, H3K36me3 and H3K27me3; 51 samples have ChIP-seq data for H3K4me3 and H3K4me1 marks; 47 samples have ChIP-seq data for H3K9me3 mark.
Sequencing depth	The narrow peaks (H3K27ac and H3K4me3) were sequenced to a target depth of ~50 million reads (mean = 71,708,388), and the broad peaks (H3K4me1, H3K36me3, H3K27me3, H3K9me3) were sequenced to a target depth of 100 million reads (mean = 134,313,487). There was an average rate of uniquely mapped reads with a q >= 10 of 94%. We sequenced with a read length of 75nt using paired-end sequencing.
Antibodies	H3K27ac from Hiroshi Kimura Lab H3K27me3 from Diagenode #C15410195 (lot #A1811-001P) H3K36me3 from Diagenode #C15410192 (lot# A1857P) H3K4me1 from Diagenode #C15410037 (lot# A1657D) H3K4me3 from Cell SignalingTechnology #9751S (lot# 10) H3K9me3 from Diagenode #C15410056 (lot# A1675-001P)
Peak calling parameters	Peak calling was done with the callpeak command of MACS2 (2.1.1) (https://github.com/jsh58/MACS). Default parameters for callpeak were used except that bedgraph files were chosen as the output format and the --broad option was used with the broad mark libraries (H3K4me1, H3K9me3, H3K27me3 and H3K36me3).
Data quality	The following quality metrics were calculated from the ChIP-Seq BAMs: * percentage of mapped reads: above 70% * Percent of total reads that are uniquely aligned with mapping quality >= 10: above 70% * Percent of mapped reads that were duplicates: acceptable value below 10%, higher values for H3K9me3 due to this mark is associated with repetitive regions of the genome. * Fraction of total mapped reads in peak regions (FRiPs). FRiPs were calculated by first excluding all reads that were unmapped, not of primary alignment, failed platform/vendor quality checks, flagged as PCR/optical duplicates, or were supplementary alignment. Note that these reads can be filtered by the following command: samtools view -F 3844 /path/to/BAM. Let this number be represented by 'n.' FRiPs would be calculated by dividing the number of 'n' reads in peaks by total number of reads 'n.' FRiP values for each mark: H3K4me3: median = 48%; range = 19% - 74% H3K4me1: median = 55%; range = 39% - 64% H3K27ac: median = 51%; range = 27% - 69% H3K36me3: median = 41%; range = 3% - 79% H3K27me3: median = 23%; range = 2% - 47% H3K9me3: median = 24%; range = 3% - 51%

*Saturation curves per each mark were generated with preseq (v2.0.2) and plot inspected.
*NFR was calculated by dividing the number of uniquely mapped reads by the number of mapped reads. All the samples were above 0.5 (no concerns about library complexity).

Software

Sequence Alignment and QC
BWA-MEM (v0.7.6a), sambamba (version 0.5.5)

Peak Calling:
MACS2 (2.1.1)

bigWig Conversion and Sample Depth:
deeptools (bamCoverage v3.0.2)
deeptools (multiBigWigSummary BED-file v3.0.2)

Differential Peak Analysis:
'DiffBind' R package (v2.2.12)

BED File Comparisons:
bedtools (v2.27.1)

Clustering:
'ConsensusClusterPlus' R package (v1.38.0)
Cluster of Clusters Analysis (custom code based on description in Hoadley et al., 2014)

Data Visualization:
Integrated Genome Browser (IGV; v2.4.14)