








Intratumor heterogeneity of HPV integration in HPV-associated head and neck cancer

Received: 31 October 2023

Accepted: 10 January 2025

Published online: 26 January 2025

 Check for updates

Noah Sasa ^{1,2,3,4}, Toshihiro Kishikawa^{1,2,5}, Masashi Mori¹, Rie Ito^{1,6}, Yumie Mizoro ³, Masami Suzuki¹, Hirotaka Eguchi¹, Hidenori Tanaka ¹, Takahito Fukusumi¹, Motoyuki Suzuki¹, Yukinori Takenaka ¹, Keisuke Nimura ^{7,8}, Yukinori Okada ^{2,3,4,9,10} ✉ & Hidenori Inohara ¹ ✉

Integration of human papillomavirus (HPV) into the host genome drives HPV-positive head and neck squamous cell carcinoma (HPV⁺ HNSCC). Whole-genome sequencing of 51 tumors revealed intratumor heterogeneity of HPV integration, with 44% of breakpoints subclonal, and a biased distribution of integration breakpoints across the HPV genome. Four HPV physical states were identified, with at least 49% of tumors progressing without integration. HPV integration was associated with APOBEC-induced broad genomic instability and focal genomic instability, including structural variants at integration sites. HPV⁺ HNSCCs exhibited almost no smoking-induced mutational signatures. Heterozygous loss of ataxia-telangiectasia mutated (*ATM*) was observed in 67% of tumors, with its downregulation confirmed by single-cell RNA sequencing and immunohistochemistry, suggesting *ATM* haploinsufficiency contributes to carcinogenesis. PI3K activation was the major oncogenic mutation, with JAK-STAT activation in tumors with clonal integration and NF-kappa B activation in those without. These findings provide valuable insights into HPV integration in HPV⁺ HNSCC.

High-risk human papillomavirus (HPV) genome replicates inside the host cell as a circular episome. The viral oncoproteins E6 and E7 are essential for HPV-induced carcinogenesis. Integration of the viral genome into the host genome is frequently observed in most invasive cervical cancers, which are representative of HPV-associated cancers. The longer half-life of integrated viral transcripts compared to episomal transcripts is believed to promote cellular immortalization and transformation¹. Thus, HPV integration is considered a major driving factor in HPV-associated carcinogenesis. Frequent disruption of the HPV *E2* gene, which regulates *E6/E7*, due to integration is argued to be

responsible for strong *E6/E7* expression². Advances in next-generation sequencing have revealed that HPV integration occurs in most HPV18-positive (HPV18⁺) cervical cancers, but only in 75% of HPV16⁺ cervical cancers, based on data from The Cancer Genome Atlas (TCGA). Additionally, HPV integration breakpoints occur throughout the HPV genome without preferential disruption of *E2*^{1,3}. HPV integration events correlate with the number of somatic mutations, such as structural variants (SVs), suggesting that integration is associated with genome-wide genomic instability⁴. However, it remains unclear whether HPV integration precedes or follows genomic instability through *E6/E7*.

¹Department of Otorhinolaryngology-Head and Neck Surgery, Osaka University Graduate School of Medicine, Suita, Japan. ²Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan. ³Department of Genome Informatics, Graduate School of Medicine, the University of Tokyo, Tokyo, Japan. ⁴Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Tsurumi, Japan. ⁵Department of Head and Neck Surgery, Aichi Cancer Center Hospital, Nagoya, Japan. ⁶Department of Otorhinolaryngology-Head and Neck Surgery, Osaka Rosai Hospital, Sakai, Japan. ⁷Department of Genome Biology, Osaka University Graduate School of Medicine, Suita, Japan. ⁸Gunma University Initiative for Advanced Research, Gunma University, Maebashi, Japan. ⁹Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan. ¹⁰Premium Research Institute for Human Metaverse Medicine (WPI-PRIME), Osaka University, Suita, Japan. ✉ e-mail: yuki-okada@m.u-tokyo.ac.jp; hinothara@ent.med.osaka-u.ac.jp

Constitutive expression of E6/E7 initiates oncogenesis, but tumor progression to cancer requires somatically acquired mutations in the host genome. In both cervical cancer and HPV⁺ oropharyngeal squamous cell carcinoma (OPSCC), the most common cause of somatic mutations is the activity of apolipoprotein B mRNA editing enzyme catalytic polypeptide-like 3 (APOBEC3)⁵. The most important mutations caused by APOBEC in HPV⁺ OPSCC include oncogenic mutations in *PIK3CA*, which encodes the p110 α catalytic subunit of phosphoinositide 3-kinase (PI3K). Activation of PI3K signaling, driven by single nucleotide variants (SNVs) or copy number alterations (CNAs) in *PIK3CA*, is a frequent feature of HPV⁺ OPSCC and is thought to occur early in carcinogenesis^{5,6}. However, it remains unclear at what stage during HPV-associated oropharyngeal carcinogenesis APOBEC-induced mutations and PI3K activation occur. According to a previous report, somatic mutations in *PIK3CA* are a late event in cervical carcinogenesis⁷.

We performed whole-genome sequencing (WGS) of matched tumor and normal tissues from 14 Japanese patients with HPV⁺ OPSCC. We analyzed these alongside WGS datasets of 18 and 19 patients with HPV⁺ head and neck squamous cell carcinoma (HNSCC) in the Pan-Cancer Analysis of Whole Genomes (PCAWG)⁸ and Ohio cohorts⁵, respectively, to provide a comprehensive analysis of HPV integration events and somatic mutations. This study reveals intratumor heterogeneity of HPV integration and its association with focal genomic instability and APOBEC signatures; minimal impact of smoking on somatic mutations; the involvement of haploinsufficient ataxia-telangiectasia mutated (*ATM*); and early *PIK3CA* gain during carcinogenesis in HPV⁺ HNSCC.

Results

Demographics, clinical data, and sequencing data

We performed WGS of tumor and matched normal samples from 14 Japanese patients with HPV16⁺ OPSCC, primarily those with poor prognosis and metastasis or recurrence during the course of the disease. The cohort included 12 males and two females, with a median age of 67 years (interquartile range [IQR]: 61.3–71.0) at diagnosis (Supplementary Data 1). WGS data were analyzed alongside WGS datasets from 18 and 19 patients with HPV⁺ HNSCC in the PCAWG and Ohio cohorts, respectively (Supplementary Data 1; Supplementary Table 1). Some tumor samples had low tumor purity, but coverage was sufficient to detect somatic mutations. Tumor coverage corrected for ploidy and tumor purity did not significantly differ among the three cohorts (Supplementary Fig. 1; Supplementary Table 2). The HPV genome was detected in all tumors, with one HPV59⁺, three HPV33⁺, and the rest HPV16⁺ in the PCAWG and Ohio cohorts (Supplementary Table 3).

In addition, we performed bulk RNA-seq of primary tumors from 19 patients with HPV⁺ OPSCC in Japan (Supplementary Table 4).

HPV status and HPV16 sublineage distribution in Japan

Among the 14 tumors in Japan, the A4 HPV16 sublineage was the most common (57.1%, $n=8$), followed by A5 (14.3%, $n=2$; Fig. 1). The remaining sublineages were A2, A3, D2, and D3, were each found in 7.1% ($n=1$) of the tumors. A5 has been previously reported in cervical cancer cases in East and Southeast Asia^{9,10}. Although a previous analysis of HPV16 *L1* identified A5 in HPV16⁺ OPSCCs in Japan¹¹, our study provided supporting evidence using WGS of the HPV genome. The HPV16 sublineage distribution in Japan differs from that in the U.S., where A1 is predominant¹².

Intratumor heterogeneity of HPV integration

A total of 396 HPV16 integration breakpoints and seven HPV33 integration breakpoints were observed in 38/51 (74.5%) tumors (Supplementary Data 2). For HPV16 integration in 37 tumors, cancer cell fractions (CCFs) were estimated for 372 breakpoints, and 339

breakpoints with CCF ≥ 0.1 were included in downstream analyses (Fig. 2a). Of these, 56.0% of the breakpoints were classified as clonal (CCF ≥ 0.8), while 44.0% were subclonal (CCF < 0.8 ; Fig. 2b; Supplementary Fig. 2a), indicating that nearly one-half of the integration events occurred during subclonal evolution. As for HPV33 integration in a single tumor, CCFs were estimated for all seven breakpoints: four were classified as clonal, two as subclonal, and one was excluded (CCF < 0.1 ; Supplementary Fig. 2b). Only HPV16 integration breakpoints were used for downstream analyses.

Integration breakpoints in the HPV16 genome were detected in all HPV16 gene regions. Compared with the assumption that breakpoints were evenly distributed throughout the HPV16 genome, binomial tests showed no significant difference for any HPV16 gene when the clonality of the breakpoints was not considered (Supplementary Fig. 2c; Supplementary Table 5). However, when restricted to clonal breakpoints, they occurred significantly more frequently in the *E1* region compared to the random expected distribution (binomial two-tailed test $P=0.022$; Fig. 2c). In contrast, when restricted to subclonal breakpoints, they were significantly more frequent in the *L1* region ($P=0.013$) and significantly less frequent in the *E6* region ($P=0.037$). Additionally, when examining the direction of integration breakpoints in the HPV16 genome, we observed a tendency for them to occur in the opposite direction to the *E6/E7* regions (Supplementary Fig. 2d; Supplementary Table 6).

In terms of breakpoints in the human genome, HPV integration did not disrupt any coding sequences (CDSs; protein-coding regions; Fig. 2d). Untranslated regions were disrupted by 1.5% of all breakpoints; however, none of these breakpoints were located in cancer-related genes. The majority of the breakpoints (63.4%) occurred in intergenic regions. Clonal integration breakpoints occurred more frequently in intergenic regions, although this was not statistically significant when assuming an even distribution of breakpoints across the human genome (binomial two-tailed test $P=0.053$; Fig. 2e; Supplementary Table 7). In addition, 43.7% of the integration breakpoints were located in repetitive regions, suggesting sufficiently high sensitivity for detecting integration breakpoints in these regions.

To identify HPV integration hotspots in the human genome, we divided the human genome into 1-Mb bins and investigated whether regions where integration events occurred overlapped between tumors. We found overlaps in five regions (chr3:189M–190M, chr5:29M–30M, chr5:49M–50M, chr19:44M–45M, and chr21:9M–10M; Supplementary Fig. 2f). Cancer-related genes were present in some of these regions: chr3:189M–190M contains *TP63* and chr19:44M–45M contains *CBL3* and *BCL3*. A previous study also reported the region including *TP63* as one of the recurrent hotspots for HPV integration¹³.

Four physical states of the HPV genome

The mean HPV genome copy number per cell within each tumor ranged from 0.8 to 1530 (median 76.2; Fig. 3a). The combination of HPV genome copy number and the presence of integration breakpoints can be used to infer the physical state of the HPV genome in the nucleus. The physical states were classified into three categories: episomal-only (no integration breakpoints), integrated-only (two integration breakpoints with no HPV coverage between them), and mixed. Mixed tumors were further subdivided into clonally-mixed (with clonal integration breakpoints) and subclonally-mixed (with only subclonal integration breakpoints and no clonal integration breakpoints). The distribution of the four HPV genome physical states in this study was as follows:

- (1) Clonally-mixed: episome + clonally integrated HPV genome (39.2%, $n=20$; Supplementary Fig. 3a).
- (2) Integrated-only: integrated HPV genome (11.8%, $n=6$; Fig. 3b).
- (3) Subclonally-mixed: episome + subclonally integrated HPV genome (23.5%, $n=12$; Supplementary Fig. 3b).

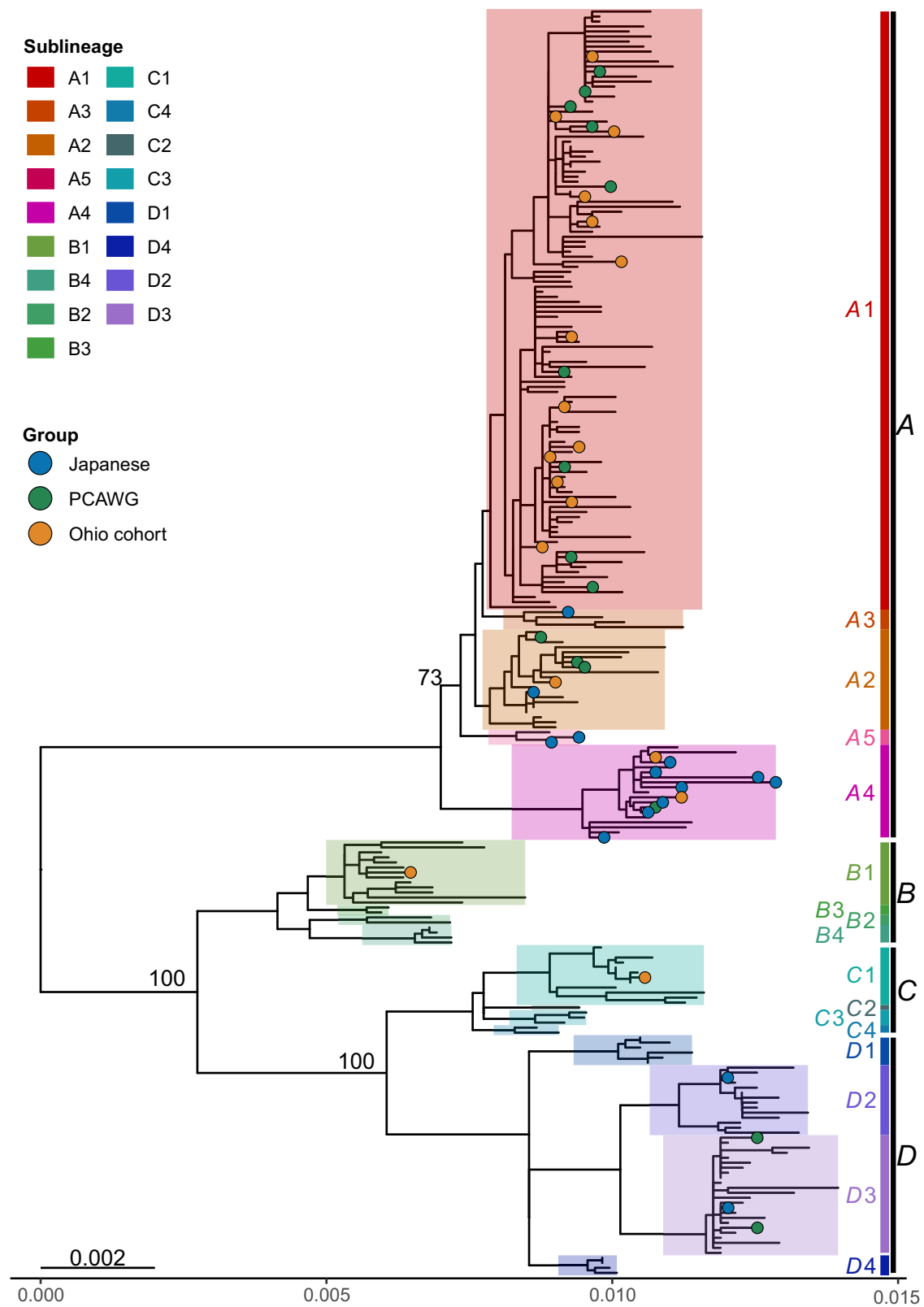


Fig. 1 | Maximum likelihood phylogeny of HPV16. The HPV16 genomes generated in this study were combined with publicly available genomes to construct a maximum likelihood phylogeny with 1000 bootstrap replicates. The phylogenetic tree is color-coded to distinguish sublineages, and bootstrap values > 70 are labeled on the corresponding nodes. Source data are provided as a Source Data file.

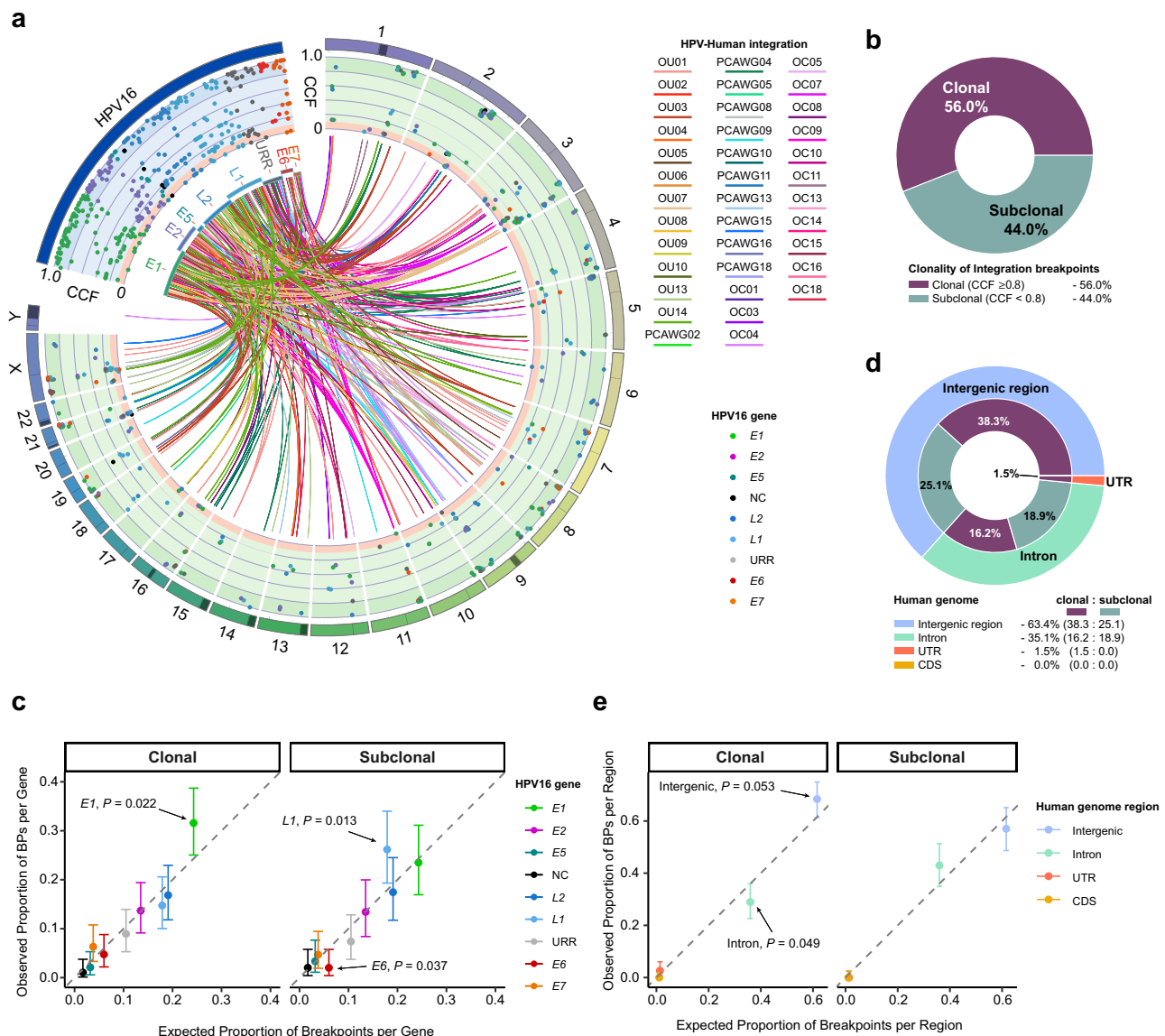


Fig. 2 | Clonality and distribution of HPV16 integration. **a** Circos plot of HPV16 integration breakpoints. The circle shows the positions of 396 integration breakpoints and their CCFs in the HPV16 and human genomes. Breakpoints for which CCFs could not be estimated are assigned a value of 0. The color of the dots indicates the HPV16 gene region where each breakpoint is located. Inside the circle, breakpoints in the HPV16 genome are connected with their corresponding locations in the human genome, with each tumor indicated by a different color. **b** Pie chart showing the clonality of integration breakpoints ($n_{CCF \geq 0.1} = 339$). Breakpoints for which CCF could not be estimated or with $CCF < 0.1$ ($n_{CCF < 0.1} = 57$) were excluded as false positives, and breakpoints with $CCF \geq 0.8$ were classified as clonal. BP, breakpoint. **c, e** Scatter plots of observed vs. expected proportions of integration breakpoints ($n_{CCF \geq 0.1} = 339$) for each HPV16 gene and each human genomic region,

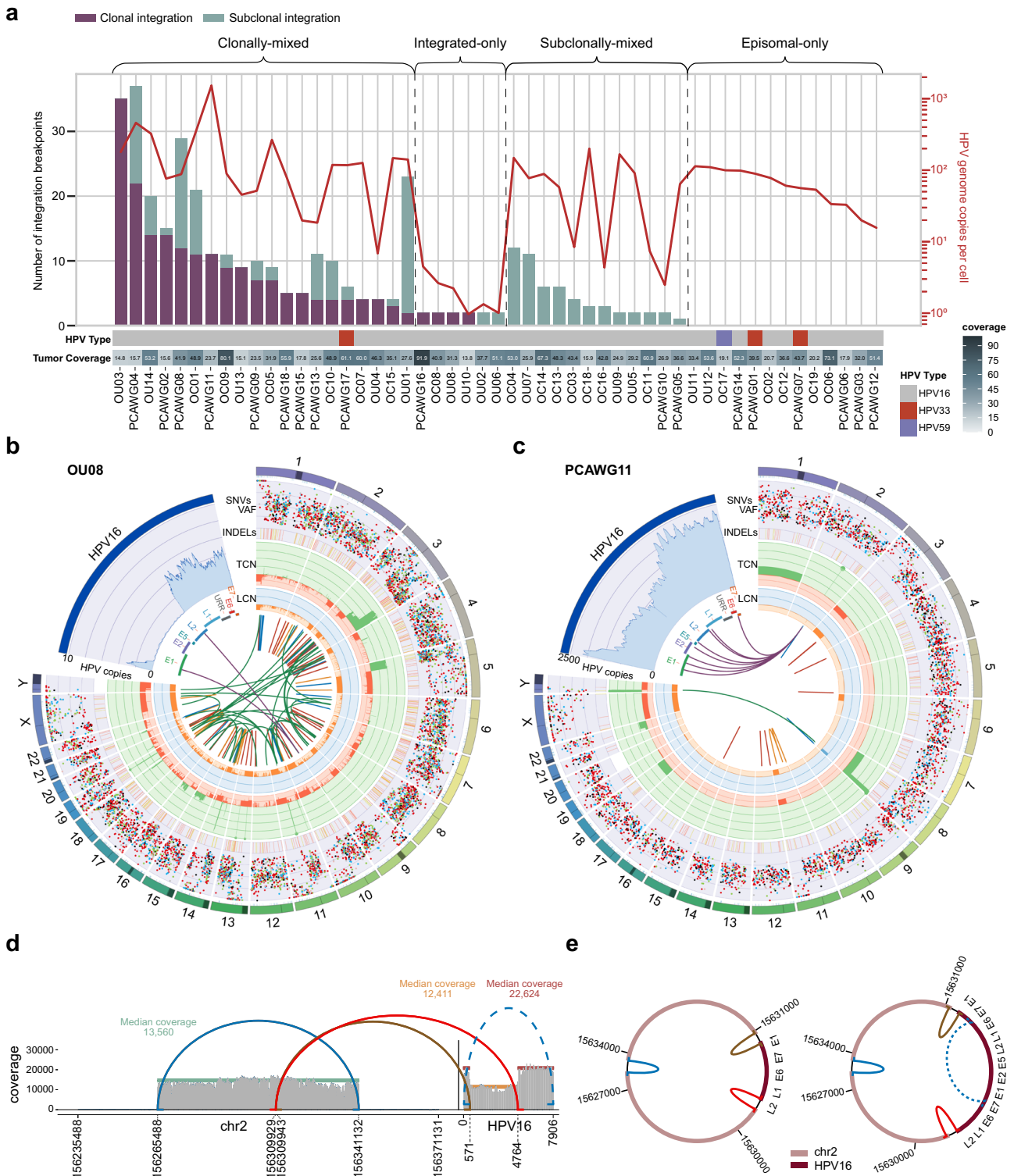
with 95% confidence interval (CI) represented by error bars. The expected proportion of breakpoints per HPV16 gene or human genomic region is plotted on the x-axis, assuming that breakpoints are equally distributed across each HPV16 gene or each human genomic region. The observed proportion is plotted on the y-axis. Left and right panels show results restricted to clonal and subclonal breakpoints, respectively. Two-sided binomial tests were performed for each HPV16 gene and each human genomic region. The slope of the dashed line is 1. Observed and expected values, as well as two-sided P -values, are provided in Supplementary Tables 5 and 7. *URR* upstream regulatory region, *NC* noncoding region. **d** Pie chart showing distribution of integration breakpoints ($n_{CCF \geq 0.1} = 339$) in the human genome. *UTR* untranslated region, *CDS* coding sequence. Source data are provided as a Source Data file.

(4) Episomal-only: circular episome (25.5%, $n = 13$; Supplementary Fig. 3c).

Some clonally-mixed tumors had multiple clonal integration breakpoints even with the tumor coverage of ~ 15 , suggesting that the tumor coverage (tumor purity) in episomal-only tumors was sufficient to detect integration breakpoints.

In at least 25/51 tumors (49.0%; episomal-only and subclonally-mixed), carcinogenesis occurred through episomal maintenance of HPV without clonal HPV integration. Thus, HPV integration is not essential for carcinogenesis.

In the HPV genome, the copy number of episomal-only tumors was 20–100, similar to that of subclonally-mixed tumors. Clonally-mixed tumors tended to have a slightly higher copy numbers than other tumor types. The presence of tumors with HPV copy numbers exceeding 100 aligns with the findings from the entire Ohio cohort's report, which also reported similarly high HPV copy numbers, ranging from 0 to 830 per cell, and confirmed these values through qPCR¹⁴. Integrated-only tumors, where all HPV genomes were considered integrants, had low HPV genome copy numbers. Interestingly, in all six integrated-only tumors, *E6/E7* were conserved while *E1* and/or *E2* were



disrupted (both *E1* and *E2* were disrupted in three tumors, *E2* in two tumors, and *E1* in one tumor; Fig. 3b).

In PCAWG11, the average number of HPV copies was exceptionally high at 1530 (Fig. 3c). Coverage around the integration breakpoints on chr2 was also very high due to a duplication (DUP), while the HPV genome showed an increase in coverage similar to that of chr2 at the integration breakpoint (Fig. 3d). This DUP could be a tandem DUP, but the high number of repeats suggested that it was in the form of extrachromosomal DNA (ecDNA) containing a truncated HPV16 integrant (Fig. 3e left). In this ecDNA, the entire HPV16 genome may be

disrupted in the truncated integrant, leading to a significant increase in the overall HPV copy number (Fig. 3e right). These would be challenging to verify with short-read WGS, and will require future validations by other methods such as long-read sequencing.

Integration breakpoints in two integrated-only tumors, OU02 and OU06, were considered subclonal (CCF < 0.8). The mechanism by which episomes are eliminated during cancer evolution in integrated-only tumors is elusive. One possibility is that episomes are eliminated from subclonally-mixed tumors. However, it is generally accepted that HPV-associated cancers contain HPV genomes in all cancer cells; thus,

Fig. 3 | Physical states of the HPV genome. **a** Combined bar and line plot showing number of integration breakpoints and mean HPV genome copy number. Four HPV genome physical states were identified: 1) clonally-mixed: episome + clonal integrant; 2) integrated-only: integrant; 3) subclonally-mixed: episome + subclonal integrant; and 4) episomal-only: episome. Heatmaps below show HPV types and tumor median coverage corrected for purity and ploidy. **b** Circos plot showing HPV copy number, HPV integration events, and somatic mutations in the integrated-only tumor OU08. The upper left shows HPV genome coverage per cell, indicating low HPV16 copy number and *E2* disruption. Of the rest, the first outer circle shows somatic variants: an outer ring for SNVs and their VAFs (0–1, corrected for purity) and an inner ring for insertions/deletions (INDELS). SNVs are colored by base change type (e.g., C > T/G > A in red)¹⁰¹. Insertions are colored in yellow and deletions in red. The second circle shows purity/ploidy-adjusted total copy number (TCN): $TCN < 2$ in red, > 2 in green, and > 6 as 6 with a green dot. The third circle

shows minor allele copy number (MCN): $MCN < 1$ in orange as a loss and > 1 in blue. The innermost circle displays SVs and HPV integration events: translocations (TRAs) in green, deletions (DELS) in red, insertions (INSS) in black, duplications (DUPS) in blue, inversions (INVs) in yellow, clonal integration events in purple, and subclonal integration events in pale green. **c**, Circos plot for PCAWG11 with a notably high HPV copy number. CNA gain is observed in chr2 around the integration breakpoints, and similarly in the HPV genome. **d** Coverage plot of chr2 and the HPV genome around HPV integration in PCAWG11. chr2 shows an increase in coverage due to a DUP, while the HPV genome shows a similar increase in coverage due to HPV integration. **e** Schematic of extrachromosomal DNA (ecDNA) with HPV16 integration. The increase in HPV coverage at the integration breakpoints can be explained by the integration of a truncated HPV16 genome into ecDNA (left). The high *E2* and *E5* copy number can also be explained by the duplication of the entire HPV16 genome within the integrant. Source data are provided as a Source Data file.

the integration breakpoints in integrated-only tumors would be clonal in nature. In other words, we suspect an underestimation of their CCFs, which may be caused by insufficient VAFs due to lack of power to detect split reads or discordant read-pairs across the human-HPV junctions that suggest HPV integration. In addition, these two tumors had integration breakpoints in regions of chr3q CNA gain (Supplementary Fig. 3d), and underestimation of the tumor total copy number (TCN) was also a possible cause. In this study, we treated the breakpoints in these two tumors as clonal.

HPV integration and focal genomic instability

As shown in Fig. 2a, HPV integration breakpoints with various CCF values in the same tumors may have been concentrated in localized regions of the human genome. To confirm this, we investigated the distances between integration breakpoints and SV breakpoints in each tumor and found that integration breakpoints in each tumor tended to be localized on the human genome, regardless of clonality. SV breakpoints also appeared adjacent to each integration breakpoint (Fig. 4a; Supplementary Fig. 4). We confirmed in advance that no high-confidence chromothripsis events had occurred in any tumor (Supplementary Data 3).

The observed distances between integration and SV breakpoints were significantly shorter than those obtained when each breakpoint was randomly placed on the human genome (median *P*-value from Wilcoxon signed-rank tests across 100 rounds of randomizations = 4.7×10^{-33} ; Supplementary Fig. 5a). Based on the distribution of distances between integration and SV breakpoints, we used ± 100 kb as the size of the region adjacent to an integration breakpoint (ITG region) for downstream analysis. The ± 100 -kb region around each integration breakpoint was defined as the ITG region and divided into clonal and subclonal-only subregions (Fig. 4b). The percentage of SV breakpoints in the ITG regions was significantly higher than that of randomly allocated SV breakpoints in the ITG regions, regardless of SV type (Supplementary Fig. 5b). This indicates that SVs are more likely to occur in regions where integration occurs.

In order to confirm that clonal and subclonal integration breakpoints tend to occur close together, we further divided the clonal subregions into clonal-only and overlapped subregions (Supplementary Fig. 5c and 5d). In the overlapped subregions, clonal breakpoints were located within 100 kb of subclonal breakpoints. We also defined overlapped subregions for SV breakpoints as well as integration breakpoints, and compared the percentages of breakpoints assigned to these overlapped subregions for both integration and SV. The percentage of integration and SV breakpoints in the overlapped subregions was significantly higher than that of randomly allocated breakpoints (Supplementary Fig. 5e). This indicates that the clonal and subclonal breakpoints for both integration and SV tend to occur within 100 kb of each other. Additionally, the percentage of clonal integration breakpoints in the overlapped subregions was higher than that of clonal SV breakpoints (Supplementary Fig. 5f). This suggests that while

subclonal breakpoints for both integration and SV tended to occur within 100 kb of clonal breakpoints, subclonal breakpoints are more frequently found within 100 kb of clonal integration breakpoints compared to clonal SV breakpoints.

We then evaluated the number of SV breakpoints within the ITG region to assess the correlations between integration and SV using Spearman's rank correlation coefficient. The number of SV breakpoints in the ITG region positively correlated with the number of integration breakpoints (Spearman's rank correlation coefficient $\rho = 0.78$; Benjamini–Hochberg (BH)-adjusted permutation test $P = 8.9 \times 10^{-7}$; Fig. 4c; Supplementary Fig. 6a); however, the number of SV breakpoints in the background showed no correlation with the number of integration breakpoints ($\rho = -0.067$; BH-adjusted permutation test $P = 0.82$). SVs in the background were probably caused by mechanisms unrelated to HPV integration. Consistently, this trend was observed in DELs, DUPS, and INVs.

Since a larger number of integration breakpoints is associated with a larger ITG region, this may in turn lead to a greater number of SV breakpoints in the ITG region. Therefore, we normalized the number of integration and SV breakpoints in the ITG region by the size of the ITG region before making comparisons. Normalized counts of SV breakpoints per 200 kb of the ITG region positively correlated with normalized counts of integration breakpoints ($\rho = 0.71$; BH-adjusted permutation test $P = 2.7 \times 10^{-6}$; Fig. 4d; Supplementary Fig. 6b). Consistently, this trend was observed in DUPS. Thus, SVs occur more densely in regions where integration breakpoints are dense. Among SVs, DUPS are particularly associated with HPV integration. Moreover, DUPS were relatively more common in clonal subregions than in the background (Fig. 4e). This suggests a mechanism of CNA gain around integration breakpoints and the presence of ecDNA harboring HPV integrants.

We then focused on the clonality of SV breakpoints within the ITG region. Comparison of clonal breakpoint percentage between ITG region and background by SV type (excluding tumors without SVs detected in each region), showed that clonal breakpoint percentage was significantly lower in DELs in the ITG region than those in the background (Wilcoxon rank sum test $P = 3.7 \times 10^{-4}$; Supplementary Fig. 7a; Supplementary Table 8). For ITG subregions, this trend was observed in DELs in the clonal subregion ($P = 1.3 \times 10^{-4}$; Fig. 5a; Supplementary Table 9). Thus, HPV integration events may be followed by DELs in the ITG region.

Because both breakpoints of a pair were counted, the smaller the size of SV, the greater the likelihood of double counting. Nevertheless, the sizes of DELs and DUPS in the ITG region were larger than those in the background (Fig. 5b). Comparing percentage SVs > 10 kb for each tumor (excluding tumors without SVs detected in each region), DUPS in the ITG region were larger than those in the background (Wilcoxon rank sum test $P = 1.2 \times 10^{-4}$; Supplementary Fig. 7b; Supplementary Table 10). Comparison of SVs separately by clonality showed that clonal DUPS were significantly larger in the ITG region than background ($P = 1.2 \times 10^{-5}$; Supplementary Fig. 7c; Supplementary Table 11).

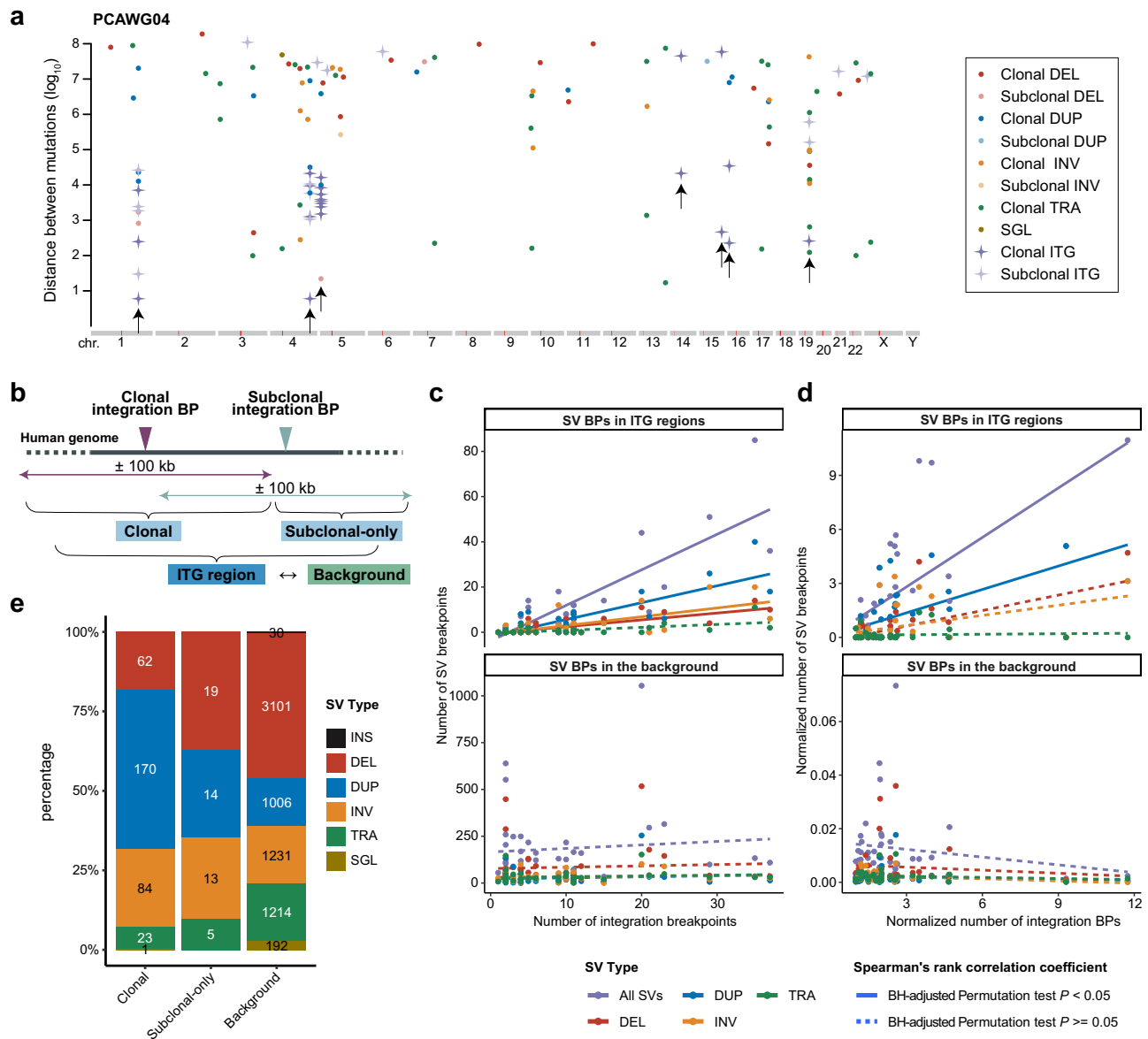


Fig. 4 | HPV integration and SV breakpoints around the integration site.

a Rainfall plot of 37 integration breakpoints in the human genome and 57 SV breakpoints for PCAWG04, which is representative of tumors because PCAWG04 had the largest number of integration breakpoints. Rainfall plots summarizing the integration and SV breakpoints for each cohort are shown in Supplementary Fig. 4. Both breakpoints of TRA are plotted, while only 5' breakpoints of other SVs are plotted. Integration breakpoints are clustered regardless of clonality, and SV breakpoints are observed within clusters of integration breakpoints (arrows showing clusters that contained integration breakpoints and had a minimum distance $< 10^5$ bp). SGL, single breakend; ITG, integration. **b** Schematic to determine regions adjacent to integration breakpoints (ITG region). The ITG region was divided into clonal subregion within 100 kb from the clonal integration breakpoint and subclonal-only subregion more than 100 kb from the clonal integration

breakpoint and within 100 kb from subclonal integration breakpoint. **c** Scatter plots of the two-sided Spearman's rank correlation coefficient between the number of integration breakpoints and the number of breakpoints for each SV type in ITG regions or the background for 37 tumors with HPV16 integration breakpoints. **d** Scatter plot of the two-sided Spearman's rank correlation coefficient between the normalized counts of integration breakpoints per 200 kb of ITG region and the normalized counts of breakpoints for each SV type per 200 kb of ITG region or the background for 37 tumors with HPV16 integration breakpoints. The coefficients with Benjamini–Hochberg (BH)-adjusted $P < 0.05$ are shown as solid lines, and those with BH-adjusted $P \geq 0.05$ are shown as dashed lines in **c** and **d**. The exact numbers of breakpoints, ρ values, and BH-adjusted P -values are provided in Supplementary Fig. 6. **e** Stacked bar plots showing percentage of SV types with their breakpoints in each region. Source data are provided as a Source Data file.

Subclonal SVs of each type were relatively large in the ITG region. When compared by ITG subregions, clonal DUPs were significantly larger in the clonal subregion than background ($P = 2.3 \times 10^{-5}$; Fig. 5c; Supplementary Table 12). Subclonal DELs were larger in the subclonal-only subregion than background ($P = 0.010$). Thus, large DUPs occur in regions where HPV integrations occur, followed by various types of large SVs.

These results indicate that focal genomic instability in regions with HPV integration may cause SVs, such as large DUPs, during tumor progression to cancer, followed by subclonal integration events and SVs, such as DELs, in the same region during cancer progression. This is consistent with the results of long-read sequencing of cervical cancer, where enrichment of SVs, especially large DUPs, occurred near clonal HPV integration breakpoints¹⁵.

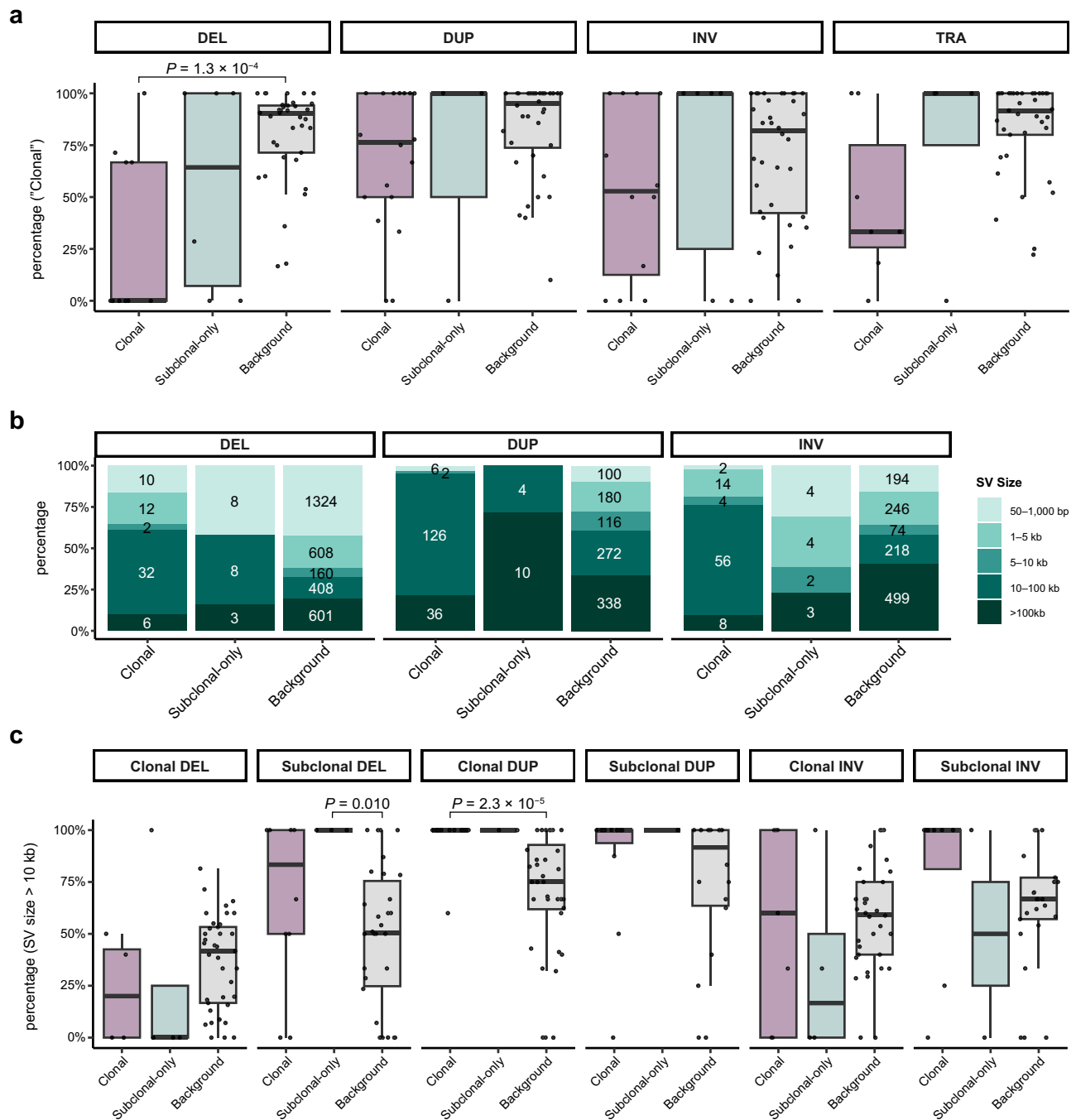


Fig. 5 | Clonality and size of SVs around HPV integration events. a Scatter and box plots comparing clonal breakpoint percentages in ITG subregions and background by SV type using the two-sided Wilcoxon rank sum test. Boxes denote the interquartile range (IQR) and the median is shown as horizontal bars; whiskers extend to 1.5 times the IQR. The number of tumors used for the calculation and clonal breakpoint percentages in each SV type and region are provided in Supplementary Table 9. **b** Stacked bar plots showing percentage of SV size in each

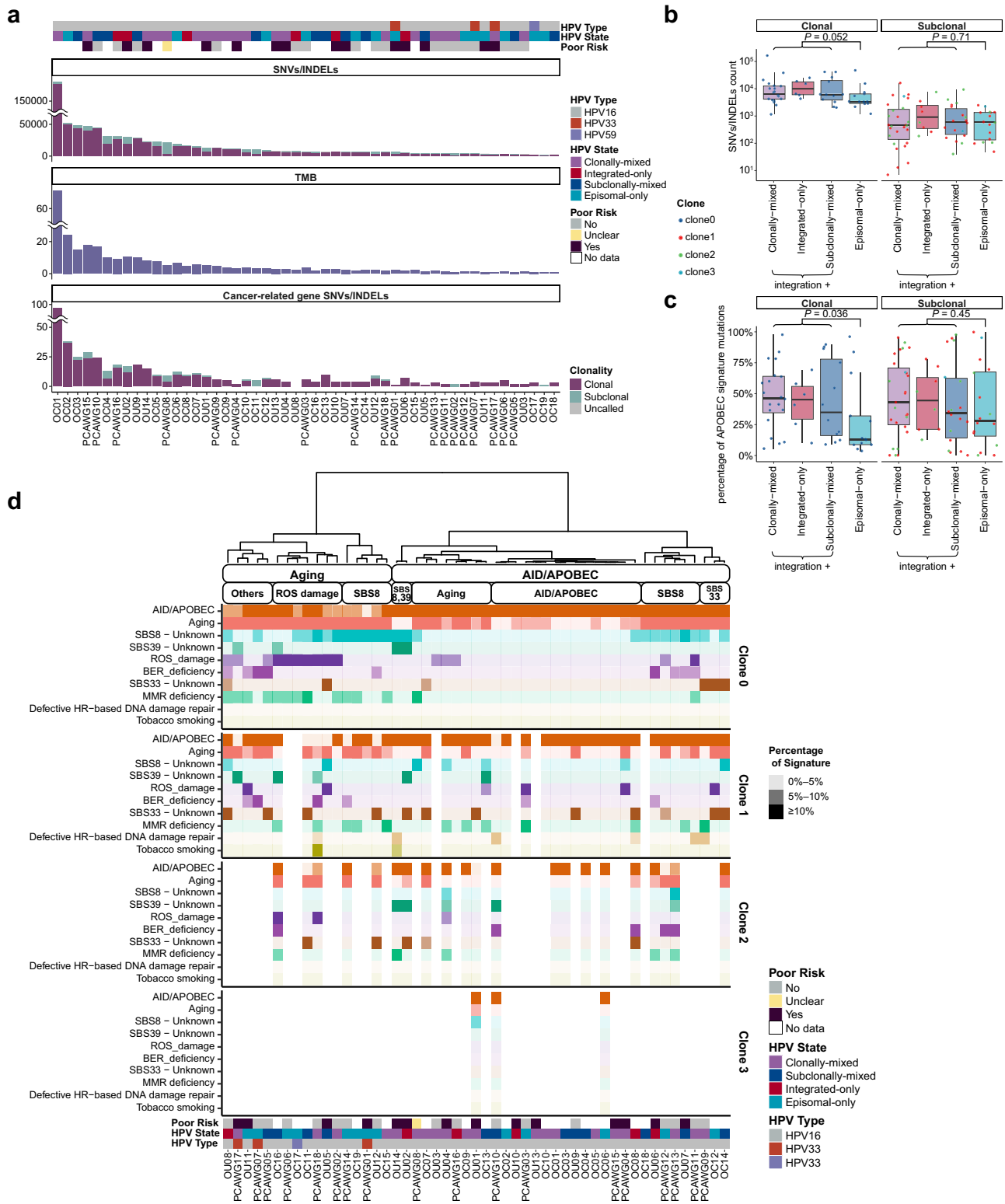
region by SV type. **c** Scatter and box plots comparing percentage of SVs >10 kb in ITG subregions and background by SV type using the two-sided Wilcoxon rank sum test. Boxes denote the IQR and the median is shown as horizontal bars; whiskers extend to 1.5 times the IQR. The number of tumors used for the calculation and percentages of SVs >10 kb in each SV type and region are provided in Supplementary Table 12. Source data are provided as a Source Data file.

Somatic short variants and HPV physical states

We focused on evaluating somatic SNVs and INDELS in the human genome. The number of SNVs/INDELS in each tumor varied from 2000 to >180,000 (median count 7429 [4348–17,868]). Tumor mutational burden (TMB) correlated with total number of SNVs/INDELS (Spearman's rank correlation coefficient $\rho = 0.98$; permutation test $P < 1.0 \times 10^{-9}$; Fig. 6a). The number of SNVs/INDELS in cancer-related genes correlated with their total number and TMB ($\rho = 0.83$ and 0.85 ;

permutation test $P < 1.0 \times 10^{-9}$ and $< 1.0 \times 10^{-9}$). We speculate that a large number of mutations tend to correlate with worse prognosis.

Wilcoxon rank sum test analysis revealed that the total number of SNVs/INDELS was significantly smaller in episomal-only tumors than other tumor types ($P = 0.034$; Supplementary Fig. 8a; Supplementary Table 13). Although there was no significant difference when considering clonality, a similar trend was observed in clonal SNVs/INDELS ($P = 0.052$; Fig. 6b). Interestingly, the number of SNVs/INDELS in



cancer-related genes was not particularly small in episomal-only tumors, implying that a certain number is required for carcinogenesis, regardless of HPV physical state.

Mutational signatures and intratumor heterogeneity

Considering the VAFs of each SNV or INDEL, mutational signature analysis and intratumor heterogeneity estimation were performed simultaneously (Supplementary Fig. 8b). Up to three subclones were observed in each tumor. Tumor samples with no subclones had lower

tumor coverage adjusted for ploidy and tumor purity than those with subclones (median coverage 15.5 [14.83–17.37] vs. 39.5 [27.26–51.84]; Wilcoxon rank sum test $P = 3.4 \times 10^{-5}$), implying that tumor coverage was too low to detect subclones. In general, many subclones and high intratumor heterogeneity are associated with treatment resistance and poor prognosis¹⁶; however, in this study, the accuracy of subclone detection was highly dependent on the purity of tumor samples, and no correlation was found between the number of detected subclones and prognosis.

Fig. 6 | Somatic short variants and mutational signatures of each clone.

a Stacked bar plots showing the number of SNVs/INDELS, tumor mutational burden (TMB), and number of SNVs/INDELS in cancer-related genes for each tumor. Correlations were investigated using Spearman's rank correlation coefficient. The upper heat maps show HPV types, HPV physical states, and prognosis. **b,c** Scatter and box plots comparing the number of mutations and the percentage of APOBEC signature mutations (SBS2 + SBS13) for clonal SNVs/INDELS and subclonal SNVs/INDELS between episomal-only tumors and other tumor types using the two-sided Wilcoxon rank sum test. Boxes denote the IQR and the median is shown as horizontal bars; whiskers extend to 1.5 times the IQR. The number of clones used for the calculation and counts of SNVs/INDELS or percentages of APOBEC signature mutations in each physical state and clonality are provided in

Supplementary Table 13. **d** Heat maps indicating percentage of 10 signature groups, consisting of the 9 most frequent signature groups and the tobacco smoking signature group per clone (numbered in order of increasing cluster CCF in each sample), dividing each group into three ranges: 0–5%, 5–10%, and $\geq 10\%$. Hierarchical clustering was based on the relative contribution of clonal mutational signatures (clone 0). Heat maps at the bottom show HPV types, HPV physical states, and prognosis. AID/APOBEC, SBS2 + SBS13; aging, SBS1 + SBS5; ROS damage, SBS18; BER deficiency, SBS30 + SBS36; MMR deficiency, SBS6 + SBS14 + SBS15 + SBS20 + SBS21 + SBS26 + SBS44; defective homologous recombination (HR)-based DNA damage repair, SBS3 + ID6; tobacco smoking, SBS4 + ID3. Source data are provided as a Source Data file.

Tumors with more mutations had more activation-induced deaminase (AID)/APOBEC mutational signatures (SBS2 and SBS13), and the higher the proportion of mutations attributed to APOBEC3 activity, the larger the overall number of mutations (Spearman's rank correlation coefficient $\rho = 0.76$; $P < 1.0 \times 10^{-9}$; Supplementary Table 14). The proportion of APOBEC signature mutations was significantly smaller in episomal-only tumors than other tumor types (Wilcoxon rank sum test $P = 0.014$). When considering clonality, a significant difference was observed only for clonal mutations, suggesting that APOBEC in carcinogenesis may be involved in HPV integration, including subclonal integration (Wilcoxon rank sum test $P = 0.036$ for clonal and 0.45 for subclonal; Fig. 6c). The significant difference in the number of SNVs/INDELS (mentioned above) may also be due to somatic hypermutation by APOBEC signature. Most of the mutations in the tumor with the extremely high number of mutations (OC01) were APOBEC signature mutations (Supplementary Fig. 8b).

Hierarchical clustering was performed based on the relative contribution of clonal mutational signatures (clone 0) in each tumor. The signatures were further grouped by etiology, and heat maps were created for 10 signature groups, comprising 9 most frequent signature groups and the tobacco smoking signature group (Fig. 6d). The tumors were divided into two major groups based on whether clonal mutational signatures were primarily attributed to aging or APOBEC3. The subclonal mutational signatures were not always consistent with the clonal mutational signatures; however, most were predominantly caused by aging or APOBEC3.

Many of the 32 patients with available metadata on smoking had a history of smoking, and some had smoked for >50 pack-years. However, both clonal and subclonal mutations showed almost no smoking-induced mutational signatures. It is also interesting to note the only clone with a smoking signature rate $>10\%$ was a subclone of laryngeal cancer, and the patient was a current smoker. SBS8, SBS33, and SBS39, which are attributed to unknown mutational processes, were highly prevalent among both clonal and subclonal mutations. SBS33 is a mutational signature with a high incidence in HNSCC¹⁷. Although cluster classification did not identify the poor prognosis group that developed distant metastasis or recurrence, two patients with SBS8 and SBS39 (i.e., OU14 and OU02) had distant metastasis at diagnosis. The mutational processes of SBS8 and SBS39 although unknown, may be poor prognostic factors in HPV⁺ OPSCC.

CNA and haploinsufficiency of tumor suppressor genes

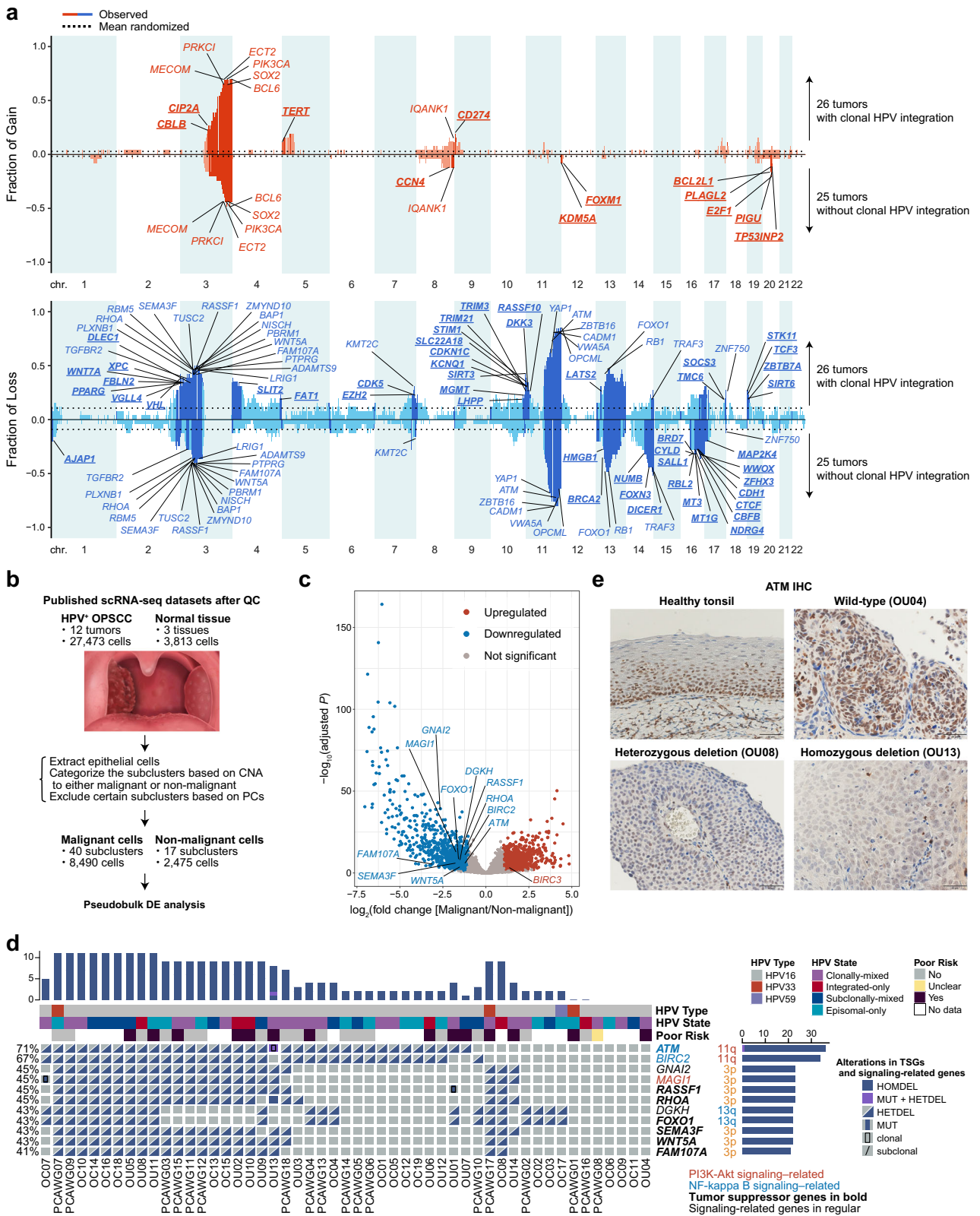
Regions with significant gains were detected in chr3q (*PIK3CA* [29/51; G-test $q = 1.4 \times 10^{-35}$], *SOX2* [28/51; G-test $q = 1.6 \times 10^{-34}$], etc.) and chr8q (*IQANK1* [6/51; G-test $q = 3.6 \times 10^{-6}$]), whereas significant losses in chr11q (*ATM* [37/51; two hits in 1/51; G-test $q = 2.9 \times 10^{-24}$], *BIRC3* [34/51; two hits in 0/51; G-test $q = 5.8 \times 10^{-19}$], etc.), chr3p (*NISCH* [22/51; two hits in 3/51; G-test $q = 7.7 \times 10^{-6}$], *GNAI2* [21/51; two hits in 0/51; G-test $q = 2.2 \times 10^{-5}$], etc.), chr13q (*RBI* [25/51; two hits in 2/51; G-test $q = 1.7 \times 10^{-9}$]), chr14q (*TRAF3* [17/51; two hits in 2/51; G-test $q = 3.7 \times 10^{-13}$]), chr16q (*CYLD* [13/51; two hits in 3/51; G-test $q = 0.0034$]), etc. (Supplementary Fig. 9; Supplementary

Data 4 and 5). These results are consistent with those of the entire Ohio cohort study with WGS of HPV⁺ oral cavity cancer and oropharyngeal cancer⁵.

When comparing tumors with and without clonal integration, high proportions of chr11q loss and chr3q gain were commonly observed (Fig. 7a). However, notable differences in the recurrent CNAs were seen between the two groups. Tumors with clonal integration frequently harbored chr5p gain (*TERT* [3/26 in tumors with clonal integration vs. 0/25 in tumors without clonal integration; G-test $q = 0.014$ vs. 0.76; Fisher's exact test mid- $P = 0.12$]), chr9p gain (*CD274* [5/26 vs. 0/25; G-test $q = 6.1 \times 10^{-4}$ vs. 0.58; Fisher's exact test mid- $P = 0.028$]), and chr19p loss (*STK11* [7/26 vs. 0/25; two hits in 2/26 vs. 0/25; G-test $q = 5.3 \times 10^{-6}$ vs. 0.65; Fisher's exact test mid- $P = 0.0057$]). In contrast, tumors without clonal integration often harbored chr20q gain (*BCL2L1* [2/26 vs. 4/25; G-test $q = 0.38$ vs. 0.0016; Fisher's exact test mid- $P = 0.40$]) and chr16q loss (*CYLD* [5/26 vs. 8/25; two hits in 0/26 vs. 3/25; G-test $q = 0.28$ vs. 0.034; Fisher's exact test mid- $P = 0.32$]). Despite prominent losses of chr4p, chr4q, and chr11p in some tumors with clonal integration, and losses of chr14q and chr16q (excluding the regions containing *TRAF3* and *CYLD*) in tumors without clonal integration, no TSGs with two hits were identified in these regions across multiple tumors. Notable TSGs included *FAT1* (5/26 vs. 2/25; two hits in 0/26 vs. 0/25; G-test $q = 0.0086$ vs. 0.26; Fisher's exact test mid- $P = 0.28$), *CDHI* (5/26 vs. 7/25; two hits in 0/26 vs. 0/25; G-test $q = 0.19$ vs. 0.039; Fisher's exact test mid- $P = 0.48$), and *WWOX* (5/26 vs. 8/25; two hits in 0/26 vs. 0/25; G-test $q = 0.073$ vs. 8.8×10^{-4} ; Fisher's exact test mid- $P = 0.32$). While *FAT1* mutation is associated with prognosis in HPV⁻ HNSCC¹⁸, the role of heterozygous deletion in *FAT1* in HPV⁺ HNSCC is unknown. *CDHI* is targeted for inactivation by the HPV16 E7 oncoprotein¹⁹, and its heterozygous deletion may promote carcinogenesis driven by HPV16 E7. *WWOX* expression is reduced in cervical carcinoma²⁰, and a similar pattern may occur in HPV⁺ HNSCC. The heterozygous deletion of these TSGs may favor HPV episomal carcinogenesis.

These differences in recurrent CNAs between tumors with and without clonal integration were observed, but no significant differences were observed for individual cancer-related or signaling-related genes after correcting for multiple testing using the Fisher's exact test. We then investigated whether the expression of these genes, with or without integration, differed using bulk RNA-seq data from primary tumor tissues of 19 Japanese patients with HPV⁺ OPSCC. No differentially expressed cancer-related or signaling-related genes were identified (Supplementary Fig. 10a). It is important to note that the RNA-seq data were derived from different patients than those analyzed by WGS in this study, and subclonal integration may have been treated as integration in these RNA-seq samples.

CNA loss was more common than CNA gain in HPV⁺ HNSCCs. We hypothesize that the loss of function of certain tumor suppressor genes (TSGs) or signaling-related genes due to CNA loss, especially through heterozygous deletion where the wild-type allele is retained, may play a significant role in the carcinogenesis of HPV⁺ HNSCC. Many tumors harbored heterozygous losses of chr11q, chr13q, chr3p, etc.,



and heterozygous deletions were observed in a large number of genes (Supplementary Fig. 11). To identify genes that were actually downregulated in HPV⁺ HNSCC, we used publicly available single-cell RNA-seq (scRNA-seq) datasets²¹. We collected published scRNA-seq data from 12 HPV⁺ OPSCCs and three normal tissues adjacent to each tumor, integrated these datasets, and performed pseudobulk differential expression (DE) analysis between malignant cells from HPV⁺ OPSCCs

and non-malignant epithelial cells from both HPV⁺ OPSCCs and normal tissues (Fig. 7b; Supplementary Fig. 10b). Among TSGs and signaling-related genes that were mutated in ≥ 21 tumors, along with *TRAF3* and *CYLD* (as listed in Supplementary Fig. 11), we identified seven TSGs and four signaling-related genes, including *ATM* and *BIRC2*, that were significantly downregulated in HPV⁺ OPSCC malignant cells (Fig. 7c; Supplementary Data 6). Conversely, only *BIRC3* was significantly

Fig. 7 | Recurrent CNAs and heterozygous mutations in downregulated TSGs and signaling-related genes. **a** Histograms showing cumulative fractions of tumors with CNA gain (red) and loss (blue) in 100-kb bins for 26 tumors with clonal integration (positive) and 25 without (negative). Fractions (y-axis) were calculated as tumors with $TCN > \text{ploidy} \times 2^{0.8}$ (gain) or $TCN < \text{ploidy} \times 2^{-0.8}$ (loss), divided by total. Dark-colored bins indicate significantly higher fractions than expected using the G-test of goodness of fit (BH-adjusted $P < 0.05$ and odds ratio > 1). Proto-oncogenes in significant gain bins are in red. Among the genes in significant loss bins, TSGs with two hits in multiple tumors, mutated in ≥ 1 tumors along with *TRAF3* and *CYLD* (in Supplementary Fig. 11), and mutated in ≥ 5 tumors exclusively in either group are in blue. Genes exclusive to one group are in bold and underlined. **b** Schematic of scRNA-seq analysis using publicly available datasets: 12 tumors from 11 patients with HPV⁺ OPSCC versus three normal tissues. QC, quality control; PC, principal component; DE, differential expression. **c** Volcano plot from pseudobulk differential expression analysis using a two-sided Wald test. Significantly upregulated (shrunken \log_2 fold change > 1) and downregulated (shrunken \log_2 fold

change < -1) genes (BH-adjusted $P < 0.05$) in HPV⁺ OPSCC malignant cells are in red and blue, respectively. Among TSGs and signaling-related genes listed in Supplementary Fig. 11, the upregulated and downregulated genes are labeled. **d** Heat map showing mutations, including heterozygous deletions, in the seven downregulated TSGs and the four downregulated signaling-related genes. Upper bar plot and heat maps show the mutation counts, HPV types, HPV physical states, and prognosis. HOMDEL, homozygous deletion ($TCN < 0.5$, no alleles); MUT + HETDEL, non-synonymous mutation and heterozygous deletion ($TCN < 1.5$; no wild-type alleles); HETDEL, heterozygous deletion (only one wild-type allele); MUT, nonsynonymous mutation (only one wild-type allele). **e** ATM immunohistochemistry showing lower expression in a tumor with a heterozygous deletion of *ATM* (OU08) compared to the epithelial cells in a healthy tonsil or a tumor with wild-type *ATM* (OU04), similar to homozygous deletion (OU13). Representative results were obtained from four groups, with samples from each. Scale bars, 50 μm . IHC, immunohistochemistry. Source data are provided as a Source Data file.

upregulated. Among the malignant cells, 68.9% were *HPV_{on}* cells with HPV gene expression and 31.1% were *HPV_{off}* cells without HPV gene expression²¹. There was no difference in the expression of these DE genes between *HPV_{on}* and *HPV_{off}* cells (Supplementary Fig. 10c; Supplementary Data 6). We then investigated whether the expression of these DE genes differed between HPV⁺ OPSCC and HPV⁻ OPSCC using bulk RNA-seq data from primary tumor tissues of 19 Japanese patients with HPV⁺ OPSCC and 17 with HPV⁻ OPSCC. *BIRC3* was upregulated in HPV⁺ OPSCC compared to HPV⁻ OPSCC, while no significant differences were observed in the genes that were downregulated compared to non-malignant epithelial cells (Supplementary Data 6).

The mutational status of seven downregulated TSGs and four downregulated signaling-related genes showed that the majority of mutations were heterozygous losses, particularly in *ATM* (67%) and *BIRC2* (67%) (Fig. 7d). To confirm decreased expression in tumors with heterozygous loss of *ATM*, immunohistochemistry (IHC) was used to examine ATM expression in a healthy tonsil, as well as in one case each of tumors with wild-type *ATM* (OU04), a heterozygous deletion of *ATM* (OU08), and a homozygous deletion of *ATM* (OU13; Fig. 7e). The tumor with a heterozygous deletion of *ATM* showed markedly decreased ATM expression compared to the epithelial cells in the stratum basale of the healthy tonsil or the tumor with wild-type *ATM*, with expression levels approaching those observed in the tumor with a homozygous deletion of *ATM*. These results suggest that these genes, especially *ATM* and *BIRC2*, may be haploinsufficient in HPV⁺ HNSCCs due to heterozygous loss. Consistently, the HiPred scores predicting haploinsufficiency for these eleven downregulated genes are greater than 0.5, with exception of *RASSF1* (Supplementary Data 6)²². Similarly, pHaplo scores (predicted probability of haploinsufficiency, or deletion intolerance) from DECHIPHER for these genes are also high, except for *GNAI2*, *RASSF1*, *RHOA*, and *FAM107A*²³.

Driver genes in HPV⁺ HNSCC

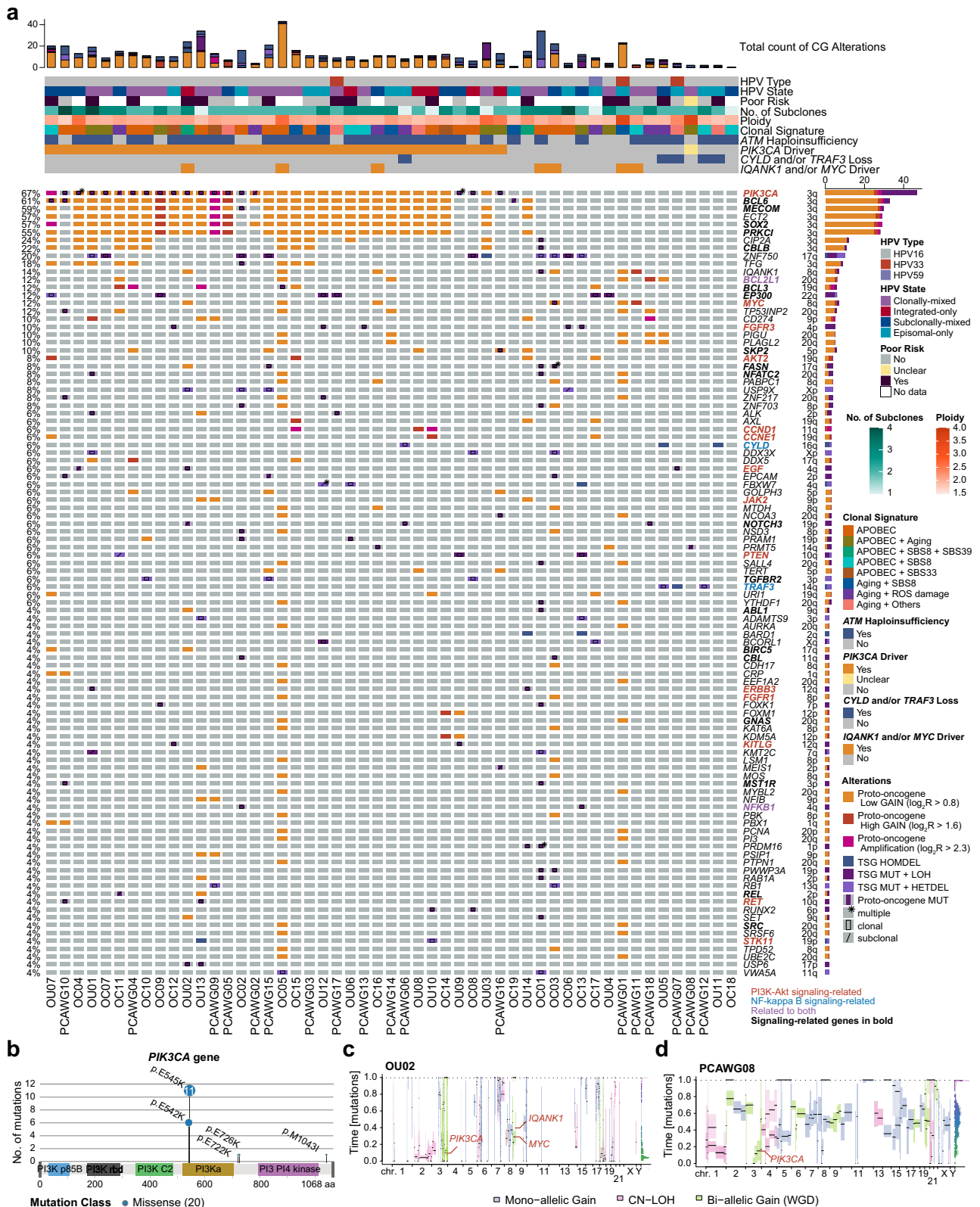
We investigated the recurrence of mutations in cancer-related genes (Fig. 8a; Supplementary Data 7). Oncogenic mutations in *PIK3CA* were detected in 34 tumors (66.7%), including 16 (31.4%) with clonal missense mutations and 29 (56.9%) with CNA gains. Eleven tumors (21.6%) had both. These gains included 2 amplifications ($TCN > \text{tumor ploidy} \times 2^{2.3}$), 2 high copy number gains ($TCN > \text{tumor ploidy} \times 2^{1.6}$), and 25 low copy number gains ($TCN > \text{tumor ploidy} \times 2^{0.8}$). This finding is consistent with previous studies on cervical cancer and the entire Ohio cohort^{5,24}. Most of the *PIK3CA* gains were low copy number gains, as were gains in other proto-oncogenes. A nonsynonymous-to-synonymous substitution (dN/dS) ratio test for each variant subtype in each gene showed that *PIK3CA* and *ZNF750* were significantly mutated genes in HPV⁺ HNSCCs (BH-adjusted P integrating all mutation types $< 1.0 \times 10^{-16}$ and 7.8×10^{-8} , respectively). Although the dN/dS ratio for mutations not undergoing

natural selection is approximately 1, the dN/dS ratio for missense mutations in *PIK3CA* was 42.9. Missense mutations included p.E545K and p.E542K, found in 11 and six tumors, respectively, which were frequent in other cancer types in the PCAWG and other HNSCC cohorts (Fig. 8b)²⁵. These were C > T substitutions that appeared in the tCw motif, consistent with the characteristics of APOBEC3-induced mutations. Consistently, all mutations were inferred to be attributable to AID/APOBEC signatures. The relative timing of *PIK3CA* missense mutations and gains during cancer evolution was estimated using the MutationTimeR package in R²⁶ (Fig. 8c; Supplementary Data 8). The results indicated that most of the *PIK3CA* missense mutations and gains occurred in early carcinogenesis, supporting the hypothesis that *PIK3CA* mutations and gains are crucial drivers of carcinogenesis in HPV⁺ HNSCC.

In addition to *PIK3CA* alterations, we identified oncogenic mutations in several cancer-related genes associated with PI3K-Akt signaling (hsa04151 in KEGG database²⁷) in multiple tumors. These mutations occurred in *BCL2L1* (6/51), *MYC* (6/51), *FGFR3* (5/51), *AKT2* (4/51), *CCND1* (3/51), *CCNE1* (3/51), *EGF* (3/51), *JAK2* (3/51), *PTEN* (3/51), *ERBB3* (2/51), *FGFR1* (2/51), *KITLG* (2/51), *NFKB1* (2/51), *RET* (2/51), and *STK11* (2/51). Additionally, oncogenic mutations were found in genes potentially involved in PI3K signaling, including *SOX2* (29/51), a transcriptional regulator important for stemness that is thought to cooperate with PI3K signaling in promoting carcinogenesis²⁸; *SKP2* (5/51), a ubiquitin ligase that induces resistance to PI3K inhibitors by reactivating *AKT*²⁹; *ZNF217* (4/51), a transcriptional regulator that activates PI3K by modulating *ERBB3*³⁰; and *ALK* (3/51), an upstream activator of PI3K signaling. These findings are consistent with previous studies, which suggest that HPV E6/E7 proteins contribute to alterations in the PI3K-Akt signaling pathway³¹.

In addition to oncogenic mutations in PI3K signaling-related genes, mutations in NF-kappa B (NF- κ B) signaling-related genes (hsa04064) were also observed in multiple tumors. These included *BCL2L1* (6/51; 2/26 in tumors with clonal integration vs. 4/25 in tumors without clonal integration; Fisher's exact test mid- $P = 0.40$; overlapping with PI3K-Akt signaling), *CYLD* (3/51; 0/26 vs. 3/25; Fisher's exact test mid- $P = 0.11$), *TRAF3* (3/51; 0/26 vs. 3/25; Fisher's exact test mid- $P = 0.11$), and *NFKB1* (2/51; 0/26 vs. 2/25; Fisher's exact test mid- $P = 0.24$; overlapping with PI3K-Akt signaling). No clonal integration events were observed in *CYLD* or *TRAF3* loss-of-function tumors, nor in *NFKB1* gain-of-function tumors, despite the Fisher's exact test showing no significant differences. In line with our findings, previous research suggests that NF- κ B activation supports the maintenance of HPV episomes and HPV-driven carcinogenesis, independent of PI3K activation³².

While the CNA gains of the proto-oncogenes *IQANK1* and *MYC* (gained in 6 tumors) were not estimated to occur as early as *PIK3CA* gain (Supplementary Data 9), they may still act as oncogenic drivers in



HPV⁺ HNSCC. This aligns with a previous study where *IQANK1* (lncRNA FAM83H-AS1) was overexpressed in an E6-dependent manner in HPV16⁺ cervical cancer cell lines, and its high expression correlated with worse overall survival³³. In three tumors with both the *PIK3CA* gain and the *IQANK1* and *MYC* gains (OU02, OC05, and OC16), the *PIK3CA* gain occurred earlier than the *IQANK1* and *MYC* gains (Fig. 8c; Supplementary Data 8 and 9).

Oncogenic mutations in cancer-related genes involved in epidermal differentiation were observed in two or more tumors, including *ZNF750* (10/51), *EP300* (6/51), *DDX3X* (3/51), *EGF* (3/51), *TRAF3* (3/51; overlapping with NF- κ B signaling), and *RET* (2/51; overlapping with PI3K signaling). Other cancer-related genes mutated in multiple tumors included *BCL3* (6/51; with STAT3-mediated oncogenic effects in cervical cancer³⁴), *CD274* (5/51; PD-L1 (programmed cell death-1 ligand

Fig. 8 | Recurrent mutations in cancer-related genes and mutations in *PIK3CA*. **a** Driver gene mutation profiles of HPV⁺ HNSCC. Bar plot shows total mutations in cancer-related genes, excluding heterozygous mutations in TSGs. Heat maps above show HPV types, HPV physical states, prognosis, number of subclones, tumor ploidy, clonal signatures, one- or two-hit loss in *ATM*, and mutations in *PIK3CA*, *CYLD* and/or *TRAF3*, as well as *IQANK1* and *MYC*. Heat map below shows mutations in cancer-related genes in multiple tumors. CG, cancer-related gene; \log_2R , $\log_2(TCN / \text{tumor ploidy})$; Low GAIN, CNA gain ($TCN > \text{tumor ploidy} \times 2^{0.8}$); High GAIN, CNA gain ($TCN > \text{tumor ploidy} \times 2^{1.6}$); Amplification, CNA gain ($TCN > \text{tumor ploidy} \times 2^{2.3}$); MUT + LOH, nonsynonymous mutation combined with copy-neutral loss of heterozygosity ($TCN > 1.5$ and $MCN < 0.5$; no wild-type alleles). **b** Lollipop

plot illustrating 20 *PIK3CA* nonsynonymous mutations. **c, d** MutationTimeR results for OU02 with both *PIK3CA* and *IQANK1* & *MYC* gains and PCAWG08 with *PIK3CA* gain and whole genome duplication (WGD). Horizontal lines in the plot show estimated timing of CNA gains (0 for earliest gain, 1 for latest gain). Boxes denote 95% CI. Histogram on the right shows the distribution of timing events. Monoallelic gains are in blue, copy-neutral loss of heterozygosity in pink, and biallelic gains (WGD) in green. For OU02, the estimated timing of CNA gain region containing *PIK3CA* (in chr3q) was 0.12 and that of CNA gain region containing *MYC* (chr8q) was 0.29. Missense mutation in *PIK3CA* (p.E542K) was considered an early clonal mutation, occurring earlier than *PIK3CA* gain (Supplementary Data 8). Source data are provided as a Source Data file.

1) highly expressed in HNSCC via JAK2-STAT1 signaling³⁵), *DDX5* (3/51; p68, a transcription factors co-activator upregulated in HNSCC³⁶) and *UR1* (3/51; a poor prognostic factor for progression and metastasis in cervical cancer³⁷). Since all five tumors with *CD274* gain had clonal integration events (Fig. 7a), HPV integration may be associated with JAK-STAT signaling. Both tumors with *UR1* gain and available prognostic data had poor outcomes.

Several tumors harbored mutations in genes associated with HPV oncoprotein-induced carcinogenesis. Although pRB is well-known to be functionally inactivated by HPV E7, *RBI* mutations were still found in 45.1% of tumors (23/51), with 21 tumors carrying one hit and two tumors carrying two hits, consistent with findings from the entire Ohio cohort⁵. Gains in *CCND1* (3/51) and *CCNE1* (3/51), both related to PI3K signaling and known to suppress *RBI*, were also observed. These gains may play a role in tumor progression by further impairing the RB1 pathway. Additionally, five tumors showed *PIGU* gain (5/51), which is overexpressed in HPV E7-transformed uroepithelial cells³⁸, and two tumors showed *BIRC5* gain (2/51), which encodes Survivin, a protein previously reported to be transcriptionally activated by HPV16 E6³⁹.

We investigated the mutation profiles of genes related to PI3K-Akt (hsa04151), NF- κ B (hsa04064) and JAK-STAT (hsa04630) signaling pathways, in addition to cancer-related genes (Supplementary Figs. 12 and 13). CNA gains were observed in many of these signaling-related genes, often occurring alongside nearby cancer-related genes (e.g., CNA gains on chr3q with *PIK3CA* and chr8q with *IQANK1*). Notably, more tumors harbored CNA gains in the PI3K signaling-related genes on chr5p (*FGF10* [5/26 vs. 1/25], *GDNF* [5/26 vs. 0/25], *GHR* [5/26 vs. 0/25], *IL7R* [6/26 vs. 0/25], *OSMR* [5/26 vs. 0/25], *PRKAA1* [5/26 vs. 0/25], and *PRLR* [5/26 vs. 0/25]; Fisher's exact test mid-*P* values ranged from 0.013 to 0.028) than in *TERT* (3/26 vs. 0/25; Fisher's exact test mid-*P* = 0.12). All five tumors with chr5p gains harbored clonal integration events (Fig. 7a). Importantly, *GHR*, *IL7R*, *OSMR*, and *PRLR* are involved in both PI3K and JAK-STAT signaling pathways, suggesting a potential association between HPV integration and JAK-STAT signaling. In addition, overexpression of *OSMR* in cervical SCC and HNSCC has been associated with epithelial-mesenchymal transition and to worsen the prognosis of cervical SCC⁴⁰, with three of the five tumors with chr5p gains showing poor outcomes. Three clonally-mixed tumors harbored chr17q gains (*CSH1*, *CSHL1*, *GHI*, and *GH2*; 3/26 vs. 0/25; Fisher's exact test mid-*P* = 0.12), related to both PI3K and JAK-STAT signaling. Additionally, five tumors harbored chr19q gains (*IFNL1*, *IFNL2*, and *IFNL3*; 3/26 vs. 2/25; Fisher's exact test mid-*P* = 0.71), also related to JAK-STAT signaling, with three of these showing amplifications.

PCAWG08, a clonally-mixed tumor, showed no clonal SNV/INDELS or CNAs in known cancer-related genes but exhibited whole genome duplication (WGD; tumor ploidy = 3.65). Although carcinogenesis may have been driven by HPV E6/E7 or mutations in unidentified cancer-related genes, the timing of *PIK3CA* gain was considered earlier than genome-wide CNA gains, suggesting that *PIK3CA* gain was the driver in this tumor (Fig. 8d; Supplementary Data 8). In contrast, OC08, a subclonally-mixed tumor, also lacked oncogenic mutations in known cancer-related genes and did not exhibit WGD (tumor ploidy = 1.92).

However, this tumor harbored relatively more heterozygous losses of TSGs, such as *ATM* on chr11q and *FAM107A*, *FOXPI*, and *RBM5* on chr3p, suggesting that, in addition to the potential involvement of unknown cancer-related genes, these TSGs may have contributed to carcinogenesis as well as HPV oncoproteins.

As RNA-seq analysis data for the 18 PCAWG tumors is available in TCGA, we confirmed the expression levels (FPKM) of each proto-oncogene with CNA gain using publicly available datasets (Supplementary Fig. 14). Most proto-oncogenes with CNA gains, including *PIK3CA*, exhibited relatively high expression in HNSCC samples in TCGA.

Discussion

Here, we demonstrate that HPV integration events, once considered essential for carcinogenesis, occur in various forms across different stages of HPV⁺ HNSCC evolution, similar to somatic mutations. This study describes subclonal integration events identified through WGS analysis, an aspect that has been previously underexplored. Subclonal integration is consistent with a previous report where target capture sequencing of primary and recurrent tumors showed an increase in integration events in recurrent tumors⁴¹. We classified four distinct HPV physical states based on the presence of episomes and integrants.

The proportion of tumors with HPV integration events (74.5%) and the unbiased distribution of integration breakpoints in the HPV genome align with a previous report from WGS of 28 HPV⁺ OPSCCs⁴². In this study, we gained insights by considering the clonality of integration breakpoints. Clonal integration frequently involved the disruption of *E1* and the adjacent *E2*, the latter being a negative regulator of the oncogenes *E6/E7*, potentially leading to the dysregulation of *E6/E7* and promoting carcinogenesis. This result is consistent with a long-read sequencing study of cervical cancer, which showed that disruption of *E1* or *E2* is preferred in clonal HPV integration¹⁵. In contrast, the significantly lower frequency of *E6* disruption in subclones implies negative selection, suggesting that *E6* plays a crucial role in the subclonal evolution of cancer. Beyond the intratumor heterogeneity of HPV integration, we emphasize the importance of HPV oncogenes *E6/E7* throughout cancer progression, including subclonal evolution. We observed conservation of *E6/E7* and disruption of *E1* and/or *E2* in integrated-only tumors, although it remains unclear when episomes were eliminated during persistent infection or carcinogenesis, which allows carcinogenesis despite a low HPV copy number. This supports the theory that constitutive *E6/E7* expression is essential for the carcinogenesis of HPV⁺ HNSCC.

Remarkably, HPV integration events were concentrated and co-occurred with SVs, demonstrating focal genomic instability. In genomic regions prone to integration events, large DUPs (> 10 kb) often co-occur, leading not only to subclonal integration events but also to subclonal DELs. A previous study reported HPV integration contributing to the amplification of *MYC* and *CD274*¹³, suggesting that genomic instability surrounding HPV integration may influence cancer-related genes.

HPV integration events are correlated with an increased number of genome-wide somatic mutations⁴. Tumors with HPV integration

harbored a higher number of SNVs/INDELS and a larger proportion of APOBEC signature mutations, suggesting an association between HPV integration and APOBEC⁴³. The fact that even subclonally-mixed tumors tended to harbor more clonal APOBEC signature mutations than episomal-only tumors indicates that APOBEC has already induced a relatively large number of mutations in the subclonally-mixed tumors before integration events occur. These findings suggest that integration events are more likely to occur in tumors where APOBEC has caused a substantial number of SNVs/INDELS.

We identified that haploinsufficiency of *ATM* due to heterozygous loss may contribute to HPV-associated head and neck carcinogenesis. *ATM*, a member of the PI3K-related kinase family, plays a key role in the DNA damage response to double-strand breaks, such as homologous recombination and non-homologous end joining⁴⁴. In cervical cancer, HPV constitutively activates *ATM* and *ATR* (also a PI3K-related kinase) to promote viral genome amplification⁴⁵, and *ATM* knockdown significantly reduces viral episome levels⁴⁶. *BIRC2* and *BIRC3*, members of the human inhibitors of apoptosis proteins family⁴⁷, are frequently amplified in cervical cancer and HNSCC^{3,48}. In the latter study, HPV⁻ HNSCC was predominant. However, scRNA-seq data in this study revealed that *ATM* and *BIRC2* are not activated in HPV⁺ OPSCC, suggesting that *ATM* and *BIRC2* may have distinct roles in HPV infection and carcinogenesis in the cervix versus the oropharynx. This is consistent with a previous study, which noted that pCHK2, the central kinase of ATM pathway, was not preferentially activated in HPV-positive lesions⁴⁹. Interestingly, both *BIRC2* and *BIRC3*, located at chr11q22.2, underwent heterozygous loss in many tumors. However, unlike *BIRC2*, *BIRC3* was upregulated in HPV⁺ OPSCCs compared to both non-malignant epithelial cells and HPV⁻ OPSCCs. These differential roles of *BIRC2* and *BIRC3* in HPV-driven head and neck carcinogenesis warrant further investigation. It is important to note that this study does not rule out the possibility that the remaining allele may be silenced by epigenetic mechanisms⁵⁰. Given the heterozygous losses observed in many TSGs, further analysis of their relationship with gene expression, including potential epigenetic mechanisms, is warranted.

Activation of PI3K signaling by SNVs and CNAs in *PIK3CA* is a critical hallmark of HPV⁺ HNSCC and occurs in early carcinogenesis^{5,6}, consistent with our findings. In addition to mutations in PI3K signaling and genes involved in epidermal differentiation, we identified mutations in cancer-related genes that further enhance HPV oncogene function, including loss of *RBI* and gains of *CCND1*, *CCNE1*, *PIGU*, and *BIRC5*. We have demonstrated that JAK-STAT signaling may be involved in tumors with clonal integration, while NF- κ B signaling appears to play a role in carcinogenesis in tumors without clonal integration.

A relatively high percentage (20%) of patients with HPV⁺ OPSCC who began smoking at a young age were found to have *TP53* mutations and a poor prognosis—chemicals in tobacco smoke induce C > A substitutions in the genome³¹. However, no *TP53* mutations and almost no tobacco-related mutational signatures were found in our study, which included smokers with >50 pack-years. Meta-analyses of risk factors for HNSCC have shown that smoking is a negative prognostic factor for overall survival in HPV⁺ HNSCC⁵², possibly due to smoking-related comorbidities, such as lung disease, and a decreased response to chemoradiotherapy. Our findings suggest that smoking has little impact on genomic mutations in HPV⁺ HNSCC.

Typically, the *E2/E6* ratio is evaluated by qPCR to detect integration events; however, this approach assumes that HPV integration always disrupts *E2* and increases *E6* expression. As a result, it is likely that only certain integration events, primarily those in integrated-only tumors, have been detected. WGS and RNA-seq are suitable for understanding complex SVs associated with genome-wide integration events. However, short-read sequencing has limitations, such as difficulties in identifying tandem repeats of integrated HPV genomes and ecDNA carrying integrated HPV genomes. Long-read sequencing^{15,53} is

expected to enhance our understanding of HPV integration and its correlation with genomic instability. Although we have presented four physical states of the HPV genome, a more detailed classification may be possible by considering tandem repeats of HPV integrants or ecDNAs that carry HPV integrants. Notably, it remains unclear whether subclonal SVs in subclonal-only subregions were found in the same cancer cells where subclonal integration events occurred, which is a limitation of bulk data generated by WGS. Additionally, a limitation of the computational methods is that the CNA estimation methods used in this study, like many others, did not account for subclonal CNAs, considering only clonal segments to estimate a mutation's multiplicity (the number of mutated chromosome copies) in CCF estimation methods⁵⁴.

HPV integration exhibits two key characteristics: 1) it results from broad genomic instability and 2) it leads to further focal genomic instability. APOBEC signature mutations may be associated with integration and occur at various stages of cancer evolution, while PI3K activation typically occurs early in oncogenesis. In our study, at least 49.0% of tumors progressed to cancer in a state of episomal-only maintenance infection, with distinct mutational patterns, such as NF- κ B activation rather than PI3K activation, being associated with carcinogenesis. This underscores the importance of accumulating data not only on HPV genome states but also on somatic mutations and mutational signatures beyond HPV integration events. While our findings support the previously reported association between *UR11* and *OMSR* gains and poor prognosis in HPV⁺ HNSCC, as observed in cervical cancer, we did not identify any somatic mutations that were clear indicators of poor prognosis. Further analyzes, including long-read sequencing and comparisons with distant metastasis tumors, are warranted to deepen our understanding.

Methods

Ethics

We have complied with all relevant ethical regulations. This study was approved by the ethical committee of Osaka University. Informed written consent was obtained from all participants, and no compensation was provided to participants.

Data collection and sequencing

We hypothesize that HPV integration events and accompanying somatic mutations occur more frequently in HPV⁺ OPSCCs with poor prognosis. Indeed, an RNA-seq study of HPV⁺ OPSCCs, including 84 cases from the University of Michigan and TCGA, suggested integration hotspots in the human genome and demonstrated that HPV⁺ OPSCCs without integration had longer survival than those with integration and HPV⁻ OPSCCs⁵⁵. We collected samples from 14 Japanese patients with HPV16⁺ OPSCC, all of whom tested positive for p16 IHC and HPV16 DNA by qPCR. The patients were treated at Osaka University between 2016 and 2021, either receiving standard of care or a clinical trial involving radiation therapy alone⁵⁶. The cohort primarily consisted of patients with poor prognosis, experiencing metastasis or recurrence.

Matched tumor and normal genomic DNA were extracted from untreated primary tumor tissues and whole blood. Sequencing libraries were prepared using reagents from Illumina TruSeq DNA PCR-free Library Prep Kit. We performed WGS for tumor (60 \times) and normal (30 \times) tissue using MGI DNBSSEQ T7, generating paired-end reads with a length of 150-bp.

We used WGS datasets of 18 HPV⁺ HNSCCs in the PCAWG and 19 HPV⁺ HNSCCs in the Ohio cohort⁵, which were generated using Illumina HiSeq with 101-bp and 151-bp paired-end reads. FASTQ records were extracted from sequence alignments in BAM format.

Patients with factors associated with poor prognosis were considered at poor risk in downstream analyzes. However, for PCAWG08, the vital status was recorded as “dead” while the tumor status was

“tumor free,” making it unclear whether the patient should be classified as poor risk.

Detection of HPV genome and HPV integration

SurVirus⁵⁷ and ViFi⁵⁸ were used to identify HPV genotypes and detect integration breakpoints in the tumor WGS data. Since the HPV genome is circular DNA and reads spanning both ends of the reference genome have a lower mapping rate, SurVirus used the HPV genome from RefSeq (NCBI Reference Sequence Database)⁵⁹ as a reference, along with the same genome with a different start position (+ 4000 bp). The combined results from SurVirus and ViFi were used for each tumor. The repetitive region list was downloaded from the UCSC genome browser, and the total length of these regions accounted for 49.7% of the hg38 genome.

HPV16 sublineage assignment

The HPV16 genomes of HPV16⁺ tumors were reconstructed from BAM files generated by ViFi. We called short variants using GATK4 HaplotypeCaller⁶⁰ with the -ploidy 1 option and created consensus sequences by applying these variants to the HPV16 reference sequence, masking deletions with ‘-’ and regions with coverage < 4 with ‘N.’ These sequences and 212 HPV16 sequences, including 16 sublineage references (A1–4, B1–4, C1–4, D1–4; Supplementary Data 10) published in GenBank⁶¹, were aligned using MAFFT v7.487 L-INS-i⁶². A phylogenetic tree was constructed using RAxML-NG v1.2.1⁶³ with the maximum likelihood method (GTR+I+G4 model) and a bootstrap value of 1,000 repetitions. HPV16 lineage and sublineage assignments were based on the phylogenetic tree.

Preprocessing of WGS data

Sequencing data (FASTQ files) of matched tumor and normal samples were aligned to the human genome reference build GRCh38 using the BWA-MEM algorithm v0.7.17-r1188⁶⁴. Following GATK Best Practices v4.1.9.0⁶⁰, we performed MarkDuplicates and Base Quality Score Recalibration and created BAM files.

Detection of somatic mutations

Somatic SNVs were called using GATK Mutect2 and Strelka2 v2.9.10⁶⁵ in matched tumor and normal BAM files for each case, and two variant callsets were combined and filtered to obtain high-confidence consensus SNVs (Supplementary Fig. 15). Mutect2 used Panel of Normals created for each cohort (14 matched normal samples for OU tumors, 94 normal samples of the PCAWG-TCGA under age 40 for PCAWG tumors, and 42 normal samples of the Ohio cohort including each matched normal sample for OC tumors; Supplementary Table 15). Somatic SVs, including INS, DEL, DUP, INV, and TRA (≥ 50 bp), were called using GRIDSS v2.10.1⁶⁶ with GRIPSS filtering, Manta v1.6.0⁶⁷, and DELLY2 v0.8.6⁶⁸. DELLY2 used normal samples from OU01–14, PCAWG01–13, PCAWG14–18, and OC01–19 as a panel of control samples for OU01–14, PCAWG01–13, PCAWG14–18, and OC01–19, respectively. SVs called by ≥ 2 of the three were considered consensus SVs merged using the StructuralVariantAnnotation package in R (maxgap = 100, sizemargin = 0.25, restrictMarginToSizeMultiple = 0.5, countOnlyBest = TRUE)⁶⁹. For short insertion/deletions (INDELS; < 50 bp), mutations called by Mutect2 and ≥ 1 of Strelka2, GRIDSS2, and DELLY2 were considered consensus INDELS and merged using the StructuralVariantAnnotation package in R (maxgap = 5, sizemargin = 0.25, restrictMarginToSizeMultiple = 0.05, countOnlyBest = TRUE).

Inference of tumor purity and absolute copy numbers

PURPLE v3.2⁷⁰, cnv_facets v0.15.0^{71,72}, and Sclust v1.1⁷³ were used to estimate tumor purity, ploidy, and absolute copy numbers to detect somatic CNAs. First, the consensus SVs were input into PURPLE to estimate segmentation, with BAM read-depths calculated using COBALT v1.11 and B allele frequencies calculated using AMBER v3.5.

Second, we included SVs recovered by PURPLE and some SVs inferred by PURPLE, which were consistent with the filtered variants of the three SV callers (in terms of position, strand, and VAF), to consensus SVs to make the final consensus SVs. These final consensus SVs were then input into PURPLE for further analysis. Similarly, Sclust used the final consensus SVs. For tumor purity and ploidy, we adopted the results of the tool whose purity output was the median of the three tools. For absolute TCNs and MCNs, we adopted the median of the three tool’s results for each 100-bp segment.

We assessed positive selection in CNA gain and loss events. CNA gain was defined as genomic regions with $TCN > \text{tumor ploidy} \times 2^{0.8}$. CNA loss was defined as regions with $TCN < \text{tumor ploidy} \times 2^{-0.8}$. We compared the observed number of tumors with CNA gain and loss in each 100-kb bin with the expected number after 100 rounds of whole-genome randomization using the G-test of goodness of fit, implemented with the R package RVAideMemoire v0.9.81.2⁷⁴. Multiple-testing corrections were applied using the BH procedure ($\alpha = 0.05$)⁵³.

To identify chromothripsis-like events using TCN and SV data, we used ShatterSeek v1.1 and two thresholds: high-confidence calls display oscillations between two states in at least seven adjacent segments, and low-confidence calls involve between four and six segments⁷⁵.

VAF estimation of the HPV integration breakpoints

VAFs of the integration breakpoints were calculated using four methods using read counts output by SurVirus and coverages adjacent to the breakpoints in the human genome.

$$\text{VAF} = \text{Variant} / (\text{Variant} + \text{Normal}) \quad (1)$$

Variant count was the number of reads involved in the integration event at the breakpoint in the human genome. Normal count was the number of reads not involved in the integration event at the breakpoint in the human genome.

$$\text{Normal} = \text{Coverage}_{5\text{bp}_{\text{Median}}} - \text{Split}_{5\text{bp}} \quad (2)$$

$$\text{Split}_{5\text{bp}} = \text{Split}_{5\text{bp}_{\text{Max}}} - \text{Split}_{5\text{bp}_{\text{Min}}} \quad (3)$$

Coverage_{5bp_{Median}} count was median coverage in 5 bp adjacent to the breakpoint (upstream if strand was +, downstream if strand was –) in the human genome. Split_{5bp} count was the number of split reads, which are single reads composed of the human and HPV genomes, in 5 bp adjacent to the breakpoint. Split_{5bp_{Max}} count was the maximum coverage for reads not properly paired in 5 bp adjacent to the breakpoint (upstream if strand was +, downstream if strand was –) and Split_{5bp_{Min}} count was the minimum coverage for reads not properly paired in 5 bp adjacent to the breakpoint (downstream if strand was +, upstream if strand was –). Coverage_{5bp_{Median}}, Split_{5bp_{Max}}, Split_{5bp_{Min}} were calculated from retained-pairs.remapped.cs.bam output by SurVirus using pysamstats v1.1.2⁷⁶ with the parameters, type = coverage, coverage_ext, or coverage_ext, respectively. Then, four estimated VAFs were defined using different Variant counts:

- (1) Split Reads VAF

$$\text{Variant} = \text{SPLIT}_{\text{READS}} \quad (4)$$

SPLIT_{READS} is the number of split reads reported by SurVirus.

- (2) Supporting Pairs VAF

$$\text{Variant} = \text{SUPPORTING}_{\text{PAIRS}} \times (2 \times \text{Read Length} / \text{Insert Size}) \quad (5)$$

SUPPORTING_{PAIRS} is the number of read pairs that support the integration event reported by SurVirus.

(3) Discordant Pairs VAF

$$\text{Variant} = \text{DISCORDANT}_{\text{PAIRS}} \times \left\{ \frac{2 \times \text{Read Length}}{\text{Insert Size} - 2 \times \text{Read Length}} \right\} \quad (6)$$

$$\text{DISCORDANT}_{\text{PAIRS}} = \text{SUPPORTING}_{\text{PAIRS}} - \text{SPLIT}_{\text{READS}} \quad (7)$$

DISCORDANT_{PAIRS} is the number of read pairs with one read aligning to the human genome and the other to the HPV genome.

(4) Split 5bp VAF

$$\text{Variant} = \text{Split5bp} \quad (8)$$

To validate the accuracy of each calculation, we used the reads simulator wgsim v.0.3.1-r13⁷⁷ to simulate 2×101 -bp and 2×150 -bp paired-end FASTQs of three regions, with or without HPV integration (chr1:100k–120k with or without the HPV16 genome integrated at chr1:110k, chr1:110k–130k at chr1:120k, and chr1:120k–140k at chr1:130k). We concatenated FASTQs with and without integration to get 60× coverage and VAF from 1% to 100% and ran SurVirus. Then, the four estimated VAFs were calculated and evaluated (Supplementary Fig. 16; Supplementary Table 16). We mostly used Split Reads VAF, because it is the best estimation method. If no SPLIT_{READS} or SUPPORTING_{PAIRS} were detected, we used Split 5bp VAF. If only SPLIT_{READS} were not detected, we used the median value of the three estimated VAFs other than Split Reads VAF (median VAF). If Split 5bp VAF and median VAF were >1.5 times Split Reads VAF, we used median VAF.

Clonality analysis of somatic mutations and integration breakpoints

The mutational contexts of SNVs/INDELS extracted using SigProfilerMatrixGenerator v1.1.26⁷⁸ and absolute copy numbers were used to cluster mutations that were attributed to the same mutational processes and had similar CCFs, i.e., the proportions of cancer cells harboring mutations in tumors, simultaneously using CloneSig v1.00⁷⁹ with the integrated matrix of single base substitution and insertion/deletion mutational signatures from the Catalogue of Somatic Mutations in Cancer v3.1¹⁷. The maximum number of clones was 4 and the minimal number of mutations per clone was 5 for running CloneSig. We obtained the CCF of each mutation and the cluster (clone) to which each mutation belonged, as well as the mutational signature of each mutation. Hierarchical clustering was performed based on the relative contribution of clonal mutational signatures in each tumor using Ward D2 clustering and Pearson's distances.

Using SVclone v1.00⁸⁰, we first estimated cluster CCFs of SVs (only SV breakpoint pairs without single breakends) based on absolute copy number. We then reassigned the CCFs of SVs and cluster CCFs of SNVs/INDELS to the joint SV + SNV model, obtaining post-assigned CCFs that reflect both. Finally, we assigned post-assigned CCFs of SVs and VAFs of integration breakpoints to a joint post-assigned SV + integration model and obtained CCFs of integration breakpoints. Integration breakpoints with variant count = 1 were excluded as false positives. Cluster CCFs were used for each mutation in downstream analyzes.

For the clonality of SNVs/INDELS, breakpoints with CCF ≥ 0.9 and CCF < 0.9 were considered clonal and subclonal, respectively, based on distribution (Supplementary Fig. 17). For the clonalities of integration events and SVs, breakpoints with CCF ≥ 0.8 and CCF < 0.8 were considered clonal and subclonal, respectively, based on the

distribution of CCFs of SVs. Integration breakpoints with CCF < 0.1 were excluded from downstream analyzes because they could be false positives.

Estimation of HPV genome copy number

Mean coverage of the HPV genome was calculated using BAM generated by ViFi. HPV genome copy number per cell was estimated using tumor purity, ploidy, and median coverage of the human genome as follows:

$$\text{Corrected tumor coverage} = \frac{\text{Median WGS coverage} \times (\text{Ploidy} \times \text{Tumor purity})}{(\text{Ploidy} \times \text{Tumor purity}) + (2 \times (1 - \text{Tumor purity}))} \quad (9)$$

HPV copies per cell =

$$\text{Mean HPV coverage} / \left(\frac{\text{Corrected tumor coverage}}{\text{Ploidy}} \right) \quad (10)$$

Classification of the HPV genome physical states

First, we classified tumors without integration breakpoints, which retain the entire HPV genome, as episomal-only tumors. Next, tumors with exactly two integration breakpoints and no HPV copy number (coverage) between the breakpoints on the HPV genome were classified as integrated-only tumors. Finally, we classified the remaining tumors as clonally-mixed tumors if they had clonal integration breakpoints, or subclonally-mixed tumors if they did not.

Since subclonally-mixed tumors do not have HPV integration events throughout the cancer cells, we assumed that episomes are present in all cancer cells. Although it is possible that in clonally-mixed tumors, several patterns of truncated integrants combine to cover the entire HPV genome, we considered these tumors to harbor the entire HPV genome as episomes and/or tandemly duplicated genome within their integrants. However, it is challenging to directly verify such physical states of the HPV genome or to accurately estimate integration breakpoint pairs when the number of breakpoints is large using short-read WGS, leaving other methods, such as long-read sequencing and dual-color fluorescence in situ hybridization⁸¹, suitable for further verification.

Rainfall plot

Rainfall plots were generated by plotting integration breakpoints and SV breakpoints using the kpPlotRainfall() function in the R package karyoplotR v1.24.0⁸². Both breakpoints of TRA were plotted, while only the 5' breakpoints of other SVs were plotted. When summarizing the integration and SV breakpoints for each cohort in Supplementary Fig. 4, only the chromosomes containing integration breakpoints from each tumor were extracted and overlaid.

Randomization of integration and SV breakpoint position on the human genome

DELs, DUPs, INSSs, and INVs were randomly located on the human genome (hg38) while retaining their sizes. On the other hand, integration breakpoints, TRA breakpoints, and single breakends were randomly located without considering their pairs (TRA breakpoint pairs were allowed to be located on the same chromosome). This randomization process was repeated 100 times.

The distances of randomly allocated integration and SV breakpoints were compared with the observed distances using the two-sided Wilcoxon signed-rank test. ITG regions and SV regions were defined by randomly allocated integration and SV breakpoints. The percentage of SV breakpoints in the ITG regions was compared with the observed percentage using the two-sided Wilcoxon signed-rank test. Similarly, the percentage of integration breakpoints in the ITG

overlapped subregions and SV breakpoints in the SV overlapped subregions were compared with the observed percentage using two-sided Wilcoxon signed-rank test. These comparisons were performed for each randomization.

Annotation of somatic mutations and driver discovery

PCGR v0.9.2⁸³ was used to annotate somatic SNVs/INDELs and CNAs. To calculate the TMB per megabase, PCGR divided the total number of somatic SNVs/INDELs in the protein-coding regions, including synonymous alterations, by the size of the coding regions.

We focused on cancer-related genes (annotated as TSGs or oncogenes by PCGR) and signaling-related genes (hsa04010, hsa04012, hsa04014, hsa04015, hsa04020, hsa04022, hsa04024, hsa04064, hsa04066, hsa04068, hsa04070, hsa04071, hsa04072, hsa04150, hsa04151, hsa04152, hsa04310, hsa04330, hsa04340, hsa04350, hsa04370, hsa04371, hsa04390, hsa04630, and hsa04668 in KEGG database²⁷) with somatic SNVs/INDELs and those covered (i.e., transcript overlapping $\geq 50\%$) by genomic segments subject to CNAs. We defined $TCN > \text{tumor ploidy} \times 2^{2.3}$ as amplification ($TCN > \text{approximately } 10$ when tumor ploidy = 2), $TCN > \text{tumor ploidy} \times 2^{1.6}$ as high copy number gain ($TCN > \text{approximately } 6$ when ploidy = 2), $TCN > \text{tumor ploidy} \times 2^{0.8}$ as low copy number gain ($TCN > \text{approximately } 3.5$ when ploidy = 2), $TCN < 0.5$ as homozygous deletion, $TCN < 1.5$ as heterozygous deletion, and $TCN > 1.5$ and $MCN < 0.5$ as copy-neutral loss of heterozygosity. The dNdScv v0.1.0 package in R⁸⁴ was used to identify cancer driver genes undergoing positive selection. The expression levels of proto-oncogenes with CNA gains in 13 PCAWG tumors were examined using TCGA RNA-seq data.

Timing estimation of CNA gains

The MutationTimeR v1.00.1 package in R²⁶ was used to calculate the relative timing of CNA gain.

Statistical analysis

All analyzes were conducted using R v4.0.5. Binomial tests, Wilcoxon rank sum tests and Wilcoxon signed-rank tests with the R package exactRankTests v0.8.35⁸⁵, and Fisher's test with exact2x2 v1.6.6⁸⁶ were two-sided. For spearman rank correlations, due to presence of ties, the conditional null distribution of the tests was approximated via Monte Carlo resampling by *spearman.test(integration_count ~ sv_count, distribution = approximate(nresample = 1,000,000,000))* implemented in the R package coin v1.4.2⁸⁷ to get exact *P*-values.

RNA-seq

mRNA was extracted from primary tumors of 19 Japanese patients with HPV⁺ OPSCC and 17 Japanese patients with HPV⁻ OPSCC who underwent standard of care or the clinical trial of radiation therapy alone⁵⁶, using p16 IHC and HPV DNA detection by qPCR at Osaka University. In total, 1/19 patient with HPV⁺ OPSCC was included in WGS analysis. Sequencing libraries were prepared using Illumina TruSeq Stranded mRNA Library Prep Kit. We performed RNA-seq using Illumina HiSeq. FASTQs were aligned to human (GRCh38, GENCODE v36) and HPV genome references using STAR v2.6.1b⁸⁸ and quantified using RSEM v1.3.3⁸⁹.

HPV integration was detected using CTAT-Virus Integration Finder v0.1.0⁹⁰. CTAT-VIF used the HPV genome from RefSeq and the same genome with a different start position (+4000 bp) as in WGS analysis. We defined integration positivity as the total number of split reads and spanning read-pairs ≥ 5 .

DE analyzes were performed between HPV⁺ OPSCC with HPV integration and HPV⁺ OPSCC without HPV integration, as well as HPV⁺ OPSCC and HPV⁻ OPSCC, using the DESeq2 v1.30.1 package in R⁹¹ with LFC shrinkage through apeglm⁹². Genes with an absolute value of shrunken \log_2 fold change > 1 and a BH-adjusted *P*-value < 0.05 were defined as significantly upregulated or downregulated.

Publicly available scRNA-seq

We collected published scRNA-seq Cell Ranger outputs of 12 tumors from 11 patients with HPV⁺ OPSCC and three normal tissues adjacent to each tumor from three patients with HPV⁺ OPSCC²¹. Cell doublets were removed by scds v1.10.0⁹³ for each sample. Quality control was performed to remove low-quality cells with < 200 genes expressed, > 5000 genes expressed, > 6000 UMIs, $< 5\%$ of reads from ribosomal genes, or $> 10\%$ of reads from hemoglobin genes. We obtained 27,473 high-quality cells from HPV⁺ OPSCCs and 3,813 high-quality cells from normal tissues. We extracted epithelial cells using the R package Azimuth v0.4.6⁹⁴. The R package Seurat v4.3.0⁹⁴ was used for data scaling, transformation, clustering, dimensionality reduction, and visualization. Data were scaled and transformed using the SCTransform() function with parameters method = "glmGamPoi", ncells = 2000, n_genes = 2000, and do.correct.umi = FALSE. In this process, we used the Azimuth human tonsil v2 reference⁹⁵. Anchors were identified between query data and Azimuth reference data using FindTransferAnchors() with parameters k.filter = NA, normalization.method = "SCT", dims = 1:50, n.trees = 20, and mapping.score.k = 100. Cell type labels were transferred with TransferData() using parameters dims = 1:50 and n.trees = 20, and embeddings were calculated on the reference supervised PCA (IntegrateEmbeddings() with the default options). Neighbors were identified using FindNeighbors() with l2.norm = TRUE. We transformed an NN index with NNTransform() (default parameters) and projected the query data onto the reference UMAP using RunUMAP() (default parameters).

We estimated CNAs in epithelial cells from HPV⁺ OPSCCs and normal tissues using inferCNV v1.10.0⁹⁶ to categorize subclusters based on CNAs. Epithelial cells in HPV⁺ OPSCC subcluster with a small number of genes affected by CNAs and a low proportion of *HPVon* cells in each tumor were considered suspected non-malignant cells. Principal component analysis was performed, and subclusters with outliers in the principal components were excluded. We performed pseudobulk DE analysis between 40 malignant cell subclusters from HPV⁺ OPSCCs (8490 cells) and 17 non-malignant epithelial cell subclusters from both HPV⁺ OPSCCs and normal tissues (2475 cells) using DESeq2 v1.34.0⁹¹, excluding genes with fewer than 17 subclusters with normalized counts ≥ 10 . Additionally, we performed a similar pseudobulk DE analysis between *HPVon* cells (malignant cells with HPV expression; 40 subclusters, 5846 cells) and *HPVoff* cells (malignant cells without HPV expression; 40 subclusters, 2644 cells). Genes with an absolute value of shrunken \log_2 fold change > 1 and a BH-adjusted *P*-value < 0.05 were defined as significantly upregulated or downregulated.

ATM Immunohistochemistry

IHC was performed on formalin-fixed, paraffin-embedded samples from three HPV⁺ OPSCCs (OU04, OU13, and OU08) and a healthy tonsil sample from a patient without malignant disease, who underwent tonsillectomy at Osaka University. A rabbit monoclonal anti-ATM antibody (1:1000; ab32420, Abcam, UK) was used for staining. The stained sections were imaged using a cellSense microscope (Olympus).

Reporting summary

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

Data availability

WGS data of tumor and matched normal samples from 14 Japanese patients with HPV⁺ OPSCC generated in this study is publicly deposited at the Japanese Genotype-phenotype Archive (JGA) with accession codes [JGAS000751/JGAD000890](https://www.genotype-phenotype.org/entry/JGAS000751/JGAD000890). RNA-seq data of primary tumor samples from 19 Japanese patients with HPV⁺ OPSCC and 17 Japanese patients with HPV⁻ OPSCC is also publicly deposited at JGA with accession codes [JGAS000751/JGAD000890](https://www.genotype-phenotype.org/entry/JGAS000751/JGAD000890). The WGS and RNA-Seq data are available under controlled access to ensure patient

confidentiality. Access for research purposes without limitation of period can be obtained by submitting a request to the National Bioscience Database Center (NBDC) Human Database (<https://humandbs.dbcls.jp/hum0197-latest>). WGS datasets of tumor and matched normal samples from 18 PCAWG cases with HPV⁺ HNSCC and normalized transcriptomic expression data of HNSCCs were derived from the TCGA Research Network. An example of a dataset link for PCAWG01 (DOI14264, TCGA-BB-4225) is provided here: <https://portal.gdc.cancer.gov/cases/c1f50a22-38df-41cc-alf4-f7985504a7ac>. We downloaded TCGA clinical data through cBioPortal and extracted reported clinical variables of 18 cases analyzed in PCAWG from `hncsc_tcg_pan_cancer_atlas_2018_clinical_data.tsv`^{97–99}. WGS datasets of 19 tumor and 42 normal samples from patients with HPV⁺ HNSCC in the Ohio cohort were derived from the European Genome-phenome Archive (EGA) under an accession number [EGAS00001003228](https://ega-archive.org/studies/EGAS00001003228). scRNA-seq datasets of 12 HPV⁺ OPSCCs (OP4-CD45-, OP5, OP6-CD45-, OP9-CD45-, OP13, OP14-CD45-, OP17, OP20, OP33-tumor, OP34-tumor, OP35-tumor, OP35-LN) and three normal tissues adjacent to each tumor (OP33-normal, OP34-normal, OP35-normal) were derived from the Gene Expression Omnibus (GEO) under an accession number [GSE182227](https://.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182227). All remaining data can be found in the Article, Supplementary and Source data files. Source data are provided with this paper.

Code availability

The codes generated for the purpose of this study, including WGS data preprocessing, variant merging, VAF calculation of integration breakpoints, and scRNA-seq analysis are available on Zenodo (<https://doi.org/10.5281/zenodo.14199195>¹⁰⁰).

References

- Rusan, M., Li, Y. Y. & Hammerman, P. S. Genomic landscape of human papillomavirus-associated cancers. *Clin Cancer Res.* **21**, 2009–2019 (2015).
- Romanczuk, H. & Howley, P. M. Disruption of either the E1 or the E2 regulatory gene of human papillomavirus type 16 increases viral immortalization capacity. *Proc Natl Acad Sci USA* **89**, 3159–3163 (1992).
- Cancer Genome Atlas Research Network. et al. Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**, 378–384 (2017).
- Labarge, B. et al. Human Papillomavirus Integration Strictly Correlates with Global Genome Instability in Head and Neck Cancer. *Mol Cancer Res.* **20**, 1420–1428 (2022).
- Gillison, M. L. et al. Human papillomavirus and the landscape of secondary genetic alterations in oral cancers. *Genome Res* **29**, 1–17 (2019).
- Leshchiner, I. et al. Inferring early genetic progression in cancers with unobtainable premalignant disease. *Nat Cancer* **4**, 550–563 (2023).
- Verlaet, W. et al. Somatic mutation in PIK3CA is a late event in cervical carcinogenesis. *J Pathol Clin Res.* **1**, 207–211 (2015).
- Aaltonen, L. A. et al. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
- Hirose, Y. et al. Whole-genome analysis of human papillomavirus type 16 prevalent in japanesewomen with or without cervical lesions. *Viruses* **11**, 350 (2019).
- Zu, Y. et al. Genetic characteristics of human papillomavirus type 16, 18, 52 and 58 in southern China. *Genomics* **113**, 3895–3906 (2021).
- Yoshida, T. et al. Investigation of the diversity of human papillomavirus 16 variants and L1 antigenic regions relevant for the prevention of human papillomavirus-related oropharyngeal cancer in Japan. *Auris Nasus Larynx* **49**, 1033–1041 (2022).
- Lang Kuhs, K. A. et al. Genetic variation within the human papillomavirus type 16 genome is associated with oropharyngeal cancer prognosis. *Ann Oncol* **33**, 638–648 (2022).
- Symer, D. E. et al. Diverse tumorigenic consequences of human papillomavirus integration in primary oropharyngeal cancers. *Genome Res.* **32**, 55–70 (2022).
- Akagi, K. et al. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res.* **24**, 185–199 (2014).
- Zhou, L. et al. Long-read sequencing unveils high-resolution HPV integration and its oncogenic progression in cervical cancer. *Nat Commun* **13**, 2563 (2022).
- Vitale, I., Shema, E., Loi, S. & Galluzzi, L. Intratumoral heterogeneity in cancer progression and response to immunotherapy. *Nat Med.* **27**, 212–224 (2021).
- Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
- Kim, K. T., Kim, B.-S. & Kim, J. H. Association between FAT1 mutation and overall survival in patients with human papillomavirus-negative head and neck squamous cell carcinoma. *Head Neck* **38**, E2021–E2029 (2016).
- Jaiswal, N., Nandi, D., Cheema, P. S. & Nag, A. The anaphase-promoting complex/cyclosome co-activator, Cdh1, is a novel target of human papillomavirus 16 E7 oncoprotein in cervical oncogenesis. *Carcinogenesis* **43**, 988–1001 (2022).
- Srivastava, S. et al. Downregulated Expression of WWOX in Cervical Carcinoma: A Case-Control Study. *Int J Mol Cell Med.* **9**, 273–288 (2020).
- Puram, S. V. et al. Cellular states are coupled to genomic and viral heterogeneity in HPV-related oropharyngeal carcinoma. *Nat Genet* **55**, 640–650 (2023).
- Shihab, H. A., Rogers, M. F., Campbell, C. & Gaunt, T. R. HIPred: an integrative approach to predicting haploinsufficient genes. *Bioinformatics* **33**, 1751–1757 (2017).
- Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* **6**, e1001154 (2010).
- Qiu, L. et al. Characterization of the genomic landscape in cervical cancer by next generation sequencing. *Genes (Basel)* **13**, 287 (2022).
- Martínez-Jiménez, F. et al. A compendium of mutational cancer driver genes. *Nat Rev Cancer* **20**, 555–572 (2020).
- Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
- Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
- Schaefer, T. & Lengerke, C. SOX2 protein biochemistry in stemness, reprogramming, and cancer: the PI3K/AKT/SOX2 axis and beyond. *Oncogene* **39**, 278–292 (2020).
- Clement, E., Inuzuka, H., Nihira, N. T., Wei, W. & Toker, A. Skp2-dependent reactivation of AKT drives resistance to PI3K inhibitors. *Sci Signal* **11**, eaao3810 (2018).
- Krig, S. R. et al. ZNF217, a candidate breast cancer oncogene amplified at 20q13, regulates expression of the ErbB3 receptor tyrosine kinase in breast cancer cells. *Oncogene* **29**, 5500–5510 (2010).
- Aguayo, F., Perez-Dominguez, F., Osorio, J. C., Oliva, C. & Calaf, G. M. PI3K/AKT/mTOR signaling pathway in HPV-driven head and neck carcinogenesis: therapeutic implications. *Biology (Basel)* **12**, 672 (2023).
- Hajek, M. et al. TRAF3/CYLD mutations identify a distinct subset of human papillomavirus-associated head and neck squamous cell carcinoma. *Cancer* **123**, 1778–1790 (2017).
- Barr, J. A. et al. Long non-coding RNA FAM83H-AS1 is regulated by human papillomavirus 16 E6 independently of p53 in cervical cancer cells. *Sci Rep* **9**, 3662 (2019).

34. Zhao, H. et al. BCL3 exerts an oncogenic function by regulating STAT3 in human cervical cancer. *Onco Targets Ther* **9**, 6619–6629 (2016).
35. Concha-Benavente, F. et al. Identification of the cell-intrinsic and -extrinsic pathways downstream of EGFR and IFN γ that induce PD-L1 expression in head and neck cancer. *Cancer Res* **76**, 1031–1043 (2016).
36. Beier, U. H., Maune, S., Meyer, J. E. & Görögh, T. Overexpression of p68 mRNA in head and neck squamous cell carcinoma cells. *Anticancer Res* **26**, 1941–1946 (2006).
37. Gu, J. et al. URI expression in cervical cancer cells is associated with higher invasion capacity and resistance to cisplatin. *Am J Cancer Res* **5**, 1353–1367 (2015).
38. Guo, Z. et al. CDC91L1 (PIG-U) is a newly discovered oncogene in human bladder cancer. *Nat Med* **10**, 374–381 (2004).
39. Borbély, Á. A. et al. Effects of human papillomavirus type 16 oncoproteins on survivin gene expression. *J Gen Virol* **87**, 287–294 (2006).
40. Kucia-Tran, J. A. et al. Overexpression of the oncostatin-M receptor in cervical squamous cell carcinoma is associated with epithelial–mesenchymal transition and poor overall survival. *Br J Cancer* **115**, 212–222 (2016).
41. Cao, Y. et al. Early HPV ctDNA kinetics and imaging biomarkers predict therapeutic response in p16+ oropharyngeal squamous cell carcinoma. *Clin Cancer Res* **28**, 350–359 (2022).
42. Gao, G. et al. Whole genome sequencing reveals complexity in both HPV sequences present and HPV integrations in HPV-positive oropharyngeal squamous cell carcinomas. *BMC Cancer* **19**, 352 (2019).
43. Kondo, S. et al. APOBEC3A associates with human papillomavirus genome integration in oropharyngeal cancers. *Oncogene* **36**, 1687–1697 (2017).
44. Ciccia, A. & Elledge, S. J. The DNA damage response: making it safe to play with knives. *Mol Cell* **40**, 179–204 (2010).
45. Moody, C. A. & Laimins, L. A. Human papillomaviruses activate the ATM DNA damage pathway for viral genome amplification upon differentiation. *PLoS Pathog* **5**, e1000605 (2009).
46. Edwards, T. G., Helmus, M. J., Koeller, K., Bashkin, J. K. & Fisher, C. Human papillomavirus episome stability is reduced by aphidicolin and controlled by DNA damage response pathways. *J Virol* **87**, 3979–3989 (2013).
47. Liang, J. et al. Comprehensive molecular characterization of inhibitors of apoptosis proteins (IAPs) for therapeutic targeting in cancer. *BMC Med Genomics* **13**, 7 (2020).
48. LEE, M. K. et al. Prognostic significance of the BIRC2-BIRC3 gene signature in head and neck squamous cell carcinoma. *Cancer Genomics - Proteomics* **19**, 591–605 (2022).
49. Kono, T. et al. Activation of DNA damage repair factors in HPV positive oropharyngeal cancers. *Virology* **547**, 27–34 (2020).
50. Ai, L. et al. Ataxia-telangiectasia-mutated (ATM) gene in head and neck squamous cell carcinoma: promoter hypermethylation with clinical correlation in 100 cases. *Cancer Epidemiol Biomarkers Prev* **13**, 150–156 (2004).
51. Dogan, S. et al. Identification of prognostic molecular biomarkers in 157 HPV-positive and HPV-negative squamous cell carcinomas of the oropharynx. *Int J Cancer* **145**, 3152–3162 (2019).
52. Alotaibi, M. et al. Impact of smoking on the survival of patients with high-risk hpv-positive hnscc: a meta-analysis. *In Vivo* **35**, 1017–1026 (2021).
53. Akagi, K. et al. Intratumoral heterogeneity and clonal evolution induced by HPV integration. *Cancer Discov* **13**, 1–18 (2023).
54. Satas, G., Zaccaria, S., El-Kebir, M. & Raphael, B. J. DeCifering the elusive cancer cell fraction in tumor heterogeneity and evolution. *Cell Syst* **12**, 1004–1018.e10 (2021).
55. Koneva, L. A. et al. HPV integration in HNSCC correlates with survival outcomes, immune response signatures, and candidate drivers. *Mol Cancer Res* **16**, 90–102 (2018).
56. Takemoto, N. et al. Radiation therapy alone for human papillomavirus-related squamous cell carcinoma of the oropharynx: a single-arm, phase 2 study. *Int J Radiat Oncol Biol Phys* **110**, 403–411 (2021).
57. Rajaby, R. et al. SurVirus: a repeat-aware virus integration caller. *Nucleic Acids Res* **49**, e33 (2021).
58. Nguyen, N.-P. D., Deshpande, V., Luebeck, J., Mischel, P. S. & Bafna, V. ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer. *Nucleic Acids Res* **46**, 3309–3325 (2018).
59. O’Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–D745 (2016).
60. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
61. Benson, D. A. et al. GenBank. *Nucleic Acids Res* **41**, D36–D42 (2013).
62. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066 (2002).
63. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
64. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://doi.org/10.48550/arXiv.1303.3997> (2013).
65. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591–594 (2018).
66. Cameron, D. L. et al. GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol* **22**, 202 (2021).
67. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
68. Rausch, T. et al. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
69. Cameron, D. L., Dong, R. & Papenfuss, A. T. StructuralVariantAnnotation: a R/Bioconductor foundation for a caller-agnostic structural variant software ecosystem. *Bioinformatics* **38**, 2046–2048 (2022).
70. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
71. Dario, B. *cnv_facets*: Somatic copy variant caller (CNV) for next generation sequencing. *GitHub Repository* https://github.com/dariober/cnv_facets (2018).
72. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res* **44**, e131 (2016).
73. Cun, Y., Yang, T. P., Achter, V., Lang, U. & Peifer, M. Copy-number analysis and inference of subclonal populations in cancer genomes using ScIust. *Nat Protoc* **13**, 1488–1501 (2018).
74. Herve, M. RVAideMemoire: Testing and Plotting Procedures for Biostatistics. <https://doi.org/10.32614/CRAN.package.RVAideMemoire> (2023).
75. Cortés-Ciriano, I. et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet* **52**, 331–341 (2020).

76. Alistair, M. pysamstats: A fast Python and command-line utility for extracting simple statistics against genome positions based on sequence alignments from a SAM or BAM file. *GitHub Repository* <https://github.com/alimanfoo/pysamstats> (2014).
77. Heng, L. wgsim: Reads simulator. *GitHub Repository* <https://github.com/lh3/wgsim> (2011).
78. Bergstrom, E. N. et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).
79. Abécassis, J., Reyal, F. & Vert, J. P. CloneSig can jointly infer intra-tumor heterogeneity and mutational signature activity in bulk tumor sequencing data. *Nat Commun* **12**, 5352 (2021).
80. Cmero, M. et al. Inferring structural variant cancer cell fraction. *Nat Commun* **11**, 730 (2020).
81. Xiong, J. et al. Detection of HPV and human chromosome sites by dual-color fluorescence in situ hybridization reveals recurrent hvp integration sites and heterogeneity in cervical cancer. *Front Oncol* **11**, 734758 (2021).
82. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
83. Nakken, S. et al. Personal cancer genome reporter: variant interpretation report for precision oncology. *Bioinformatics* **34**, 1778–1780 (2018).
84. Martincorena, I. et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, (2017).
85. Hothorn, T. & Hornik, K. exactRankTests: Exact Distributions for Rank and Permutation Tests. <https://CRAN.R-project.org/package=exactRankTests> (2022).
86. Fay, M. P. Confidence intervals that match Fisher's exact or Blaker's exact tests. *Biostatistics* **11**, 373–374 (2010).
87. Hothorn, T., Hornik, K., van de Wiel, M. A. & Zeileis, A. Implementing a Class of Permutation Tests: The coin Package. *J Stat Softw* **28**, (2008).
88. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
89. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
90. CTAT Virus Integration Finder. <https://github.com/broadinstitute/CTAT-VirusIntegrationFinder/wiki>.
91. 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).
92. Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* **35**, 2084–2092 (2019).
93. Bais, A. S. & Kostka, D. scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics* **36**, 1150–1158 (2020).
94. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
95. Massoni-Badosa, R. et al. An atlas of cells in the human tonsil. *Immunity* **57**, 379–399.e18 (2024).
96. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
97. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401–404 (2012).
98. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1 (2013).
99. Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304.e6 (2018).
100. Sasa, N. Scripts for 'Intratumor Heterogeneity of HPV Integration in HPV-associated Head and Neck Cancer'. *Zenodo*. <https://doi.org/10.5281/zenodo.14199195> (2024).
101. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

Acknowledgements

We would like to thank all the participants involved in this study. This study makes use of the data deposited by Drs. Gillison, Symer, and Akagi in the HPV Virome Consortium, formerly at the Ohio State University Comprehensive Cancer Center and then at the University of Texas MD Anderson Cancer Center. Funding and computational support for these data were provided in part by the Ohio State University Comprehensive Cancer Center, the Ohio Supercomputer Center, Cancer Prevention & Research Institute of Texas, and the University of Texas MD Anderson Cancer Center. H.I. was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (16H02681, 19H01062). Y.O. was supported by JSPS KAKENHI (22H00476, 24K02697), AMED (JP22ek0410075, JP23km0405211, JP23km0405217, JP23ek0109594, JP23ek0410113, JP223fa627002, JP223fa627010, JP233fa627011, JP23zf0127008, JP23tm0524002, JP24wm0625504, JP24gm1810011), JST Moonshot R&D (JPMJMS2021, JPMJMS2024), Takeda Science Foundation, Ono Pharmaceutical Foundation for Oncology, Immunology, and Neurology, Bioinformatics Initiative of Osaka University Graduate School of Medicine, Institute for Open and Transdisciplinary Research Initiatives, Center for Infectious Disease Education and Research (CiDER), and Center for Advanced Modality and DDS (CAMaD), Osaka University. This study was supported by JST (JPMJSP2138).

Author contributions

N.S., T.K., and H.I. designed the study. N.S. and Y.O. conducted the data analysis. N.S. wrote the manuscript in consultation with Ma.S., Y.O., and H.I. M.M. and Ma.S. performed immunostaining. R.I., Y.M., H.E., H.T., T.F., Mo.S., Y.T. and K.N. collected and managed the samples. Y.O. and H.I. supervised the study. All authors contributed to the generation of the primary data incorporated in the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-56150-z>.

Correspondence and requests for materials should be addressed to Yukinori Okada or Hidenori Inohara.

Peer review information *Nature Communications* thanks Takashi Kohno, Andrew McPherson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025