Diverse tumorigenic consequences of human papillomavirus integration in primary oropharyngeal cancers

David E. Symer,[1] Keiko Akagi,[2] Heather M. Geiger,[3] Yang Song,[2] Gaiyun Li,[2] Anne-Katrin Emde,[3] Weihong Xiao,[2] Bo Jiang,[2] André Corvelo,[3] Nora C. Toussaint,[3] Jingfeng Li,[4] Amit Agrawal,[5] Enver Ozer,[5] Adel K. El-Naggar,[6] Zoe Du,[1] Jitesh B. Shewale,[2] Birgit Stache-Crain,[7] Mark Zucker,[8] Nicolas Robine,[3] Kevin R. Coombes,[8] and Maura L. Gillison[2]

[1]*Department of Lymphoma and Myeloma, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA;* [2]*Department of Thoracic/Head and Neck Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA;* [3]*New York Genome Center, New York, New York 10013, USA;* [4]*Division of Medical Oncology, Department of Internal Medicine, Ohio State University, Columbus, Ohio 43210, USA;* [5]*Department of Otolaryngology – Head and Neck Surgery, Ohio State University Comprehensive Cancer Center, Columbus, Ohio 43210, USA;* [6]*Division of Pathology, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA;* [7]*Complete Genomics, San Jose, California 95134, USA;* [8]*Department of Biomedical Informatics, Ohio State University Comprehensive Cancer Center, Columbus, Ohio 43210, USA*

Human papillomavirus (HPV) causes 5% of all cancers and frequently integrates into host chromosomes. The HPV oncoproteins E6 and E7 are necessary but insufficient for cancer formation, indicating that additional secondary genetic events are required. Here, we investigate potential oncogenic impacts of virus integration. Analysis of 105 HPV-positive oropharyngeal cancers by whole-genome sequencing detects virus integration in 77%, revealing five statistically significant sites of recurrent integration near genes that regulate epithelial stem cell maintenance (i.e., *SOX2, TP63, FGFR, MYC*) and immune evasion (i.e., *CD274*). Genomic copy number hyperamplification is enriched 16-fold near HPV integrants, and the extent of focal host genomic instability increases with their local density. The frequency of genes expressed at extreme outlier levels is increased 86-fold within ±150 kb of integrants. Across 95% of tumors with integration, host gene transcription is disrupted via intragenic integrants, chimeric transcription, outlier expression, gene breaking, and/or de novo expression of noncoding or imprinted genes. We conclude that virus integration can contribute to carcinogenesis in a large majority of HPV-positive oropharyngeal cancers by inducing extensive disruption of host genome structure and gene expression.

[Supplemental material is available for this article.]

Human papillomavirus (HPV) infection causes ~5% of all human cancers, resulting in 650,000 cases worldwide each year (de Martel et al. 2017). These include anogenital cancers and a subset of oropharyngeal cancers that is increasing markedly in incidence (Tota et al. 2019). The HPV oncoproteins E6 and E7 promote host genomic instability in multiple ways including degradation of tumor protein p53 (TP53) and RB transcriptional corepressor 1 (RB1), respectively. E6 and E7 expression is necessary but not sufficient for HPV-associated carcinogenesis. Secondary genetic events such as host gene mutations also are required (The Cancer Genome Atlas Research Network 2017; Gillison et al. 2019).

HPV integration into cervical cancer genomes was first reported over 30 years ago (Dürst et al. 1987). Subsequent studies were unable to reveal the full extent of HPV integration because they utilized biased, insensitive, and/or nonspecific laboratory techniques to detect and map integrants (e.g., Southern blotting, PCR, whole-exome sequencing [WES], RNA sequencing [RNA-seq], amplification of papilloma virus oncogene transcripts [APOTs] assay, and others) (The Cancer Genome Atlas Research Network 2017) (for review, see Bodelon et al. 2016). More recently, a hybrid capture-based method was used to report recurrent hotspots of HPV integration in cervical dysplasias and cancers (Hu et al. 2015). However, in that study, numerous virus–host breakpoints were reported to be identical at a single nucleotide level across multiple samples, raising questions about artifacts and undermining those data and conclusions (Dyer et al. 2016). Recent TCGA genomic studies of oropharyngeal and cervical cancers mostly used RNA-seq to map integration sites indirectly, and sample subsets studied with whole-genome sequencing (WGS) were statistically underpowered to identify recurrent hotspots (Ojesina et al. 2014; Parfenov et al. 2014; The Cancer Genome Atlas Research Network 2017). In sum, the impacts of HPV integration on host genome structures and gene expression have yet to be defined and

Corresponding authors: mgillison@mdanderson.org, desymer@mdanderson.org

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.275911.121. Freely available online through the *Genome Research* Open Access option.

quantified comprehensively across an adequately powered collection of HPV-positive tumors such as oropharyngeal cancers.

Using WGS to study relatively small numbers of HPV-positive cancers and cell lines, we and others have shown that HPV integrants recurrently flank or bridge focal regions of extensive host genomic instability, including copy number variation (CNV) and structural variation (SV) (Akagi et al. 2014; Parfenov et al. 2014; The Cancer Genome Atlas Research Network 2017). We proposed a mechanistic looping model, by which replication of transient, circular, virus–host intermediate structures (using the HPV origin of replication) is followed by recombination and repair, leading to integrated HPV-host concatemers and extensive genomic structural variation (Akagi et al. 2014). Our looping model anticipated that HPV integrants in genomic DNA would be replicated by rolling circle amplification, regardless of their maintenance in chromosomes or potentially in extrachromosomal circular DNA (ecDNA) (Akagi et al. 2014; Parfenov et al. 2014; deCarvalho et al. 2018).

We hypothesize that host genomic alterations associated with HPV integration are critical contributors in the pathogenesis of a majority of HPV-positive primary cancers. Here, we report a comprehensive analysis of virus integration and its impacts in 105 HPV-positive oropharyngeal squamous cell carcinomas (OPSCCs). This is the first study of HPV-positive cancers that is statistically powered to detect significant, recurrent hotspots of HPV integration, using precise and unbiased genomics methods including WGS and RNA-seq. These findings demonstrate that our HPV integrant-mediated looping model, initially formulated based primarily on cell lines, also is supported in primary human oropharyngeal cancers.

Results

Genomic sites of HPV integration reveal clustering in individual tumors

Analysis of WGS data from tumor and normal blood leukocyte (T/N) pairs from 105 patients with newly diagnosed HPV-positive oropharyngeal cancer (Supplemental Table S1.1) reveals that a majority (n = 92, 88%) harbor HPV16 (Gillison et al. 2019). Other viral types detected in remaining samples are HPV18 (n = 2, 2%), HPV33 (n = 6, 6%), HPV35 (n = 3, 3%), HPV59 (n = 1, 1%), and HPV69 (n = 1, 1%) (Gillison et al. 2019; Supplemental Table S1.2). Overall, 874 virus–host breakpoints (i.e., sites of HPV integration with host DNA) are identified in genomic DNA of 81 (77%) tumors, based on detection of split and/or discordant WGS sequencing reads mapping both to HPV and the reference human genome (Supplemental Fig. S1.1). As we previously observed in cultured HPV-positive cancer cells (Akagi et al. 2014), virus–host breakpoint junctions in primary tumors are uniformly distributed across the entire ~8-kb viral genome and thus are not enriched or depleted preferentially in any viral gene (Fig. 1A; Supplemental Fig. S1.2).

We find no enrichment of insertional breakpoints in the HPV E2 gene (Fig. 1A; Supplemental Fig. S1.2; Supplemental Tables S1.3–S1.5), in contrast to an accepted paradigm regarding a selective advantage from disruption of this viral transcriptional regulator. Genetic disruption of E2 has been associated with up-regulation of E6 and E7 expression, which in turn can promote carcinogenesis (Romanczuk and Howley 1992; McBride and Warburton 2017). However, we detect transcripts with coding potential for E2 in 97.3% of HPV16-positive tumors, that is, both with and without viral integrants. We observe no significant difference in expression of polycistronic transcripts encoding HPV E2, E6 or E7 across tumors with or without virus–host breakpoints (Supplemental Figs. S1.3, S1.4; Supplemental Table S1.6). We conclude that E2 disruption or changes in E2 transcript levels are poor proxies for HPV integration (Supplemental Figs. S1.2–S1.4; Supplemental Tables S1.3–S1.6).

Of the 874 virus–host breakpoints, 756 (86.5%) map uniquely in the host genomes of 77 (73%) tumors, whereas remaining breakpoints align either to repetitive elements (n = 95) or unassigned contigs (n = 23) (Supplemental Tables S1.3, S1.7). No breakpoints are detected in 24 (23%) tumors, consistent with the presence of episomal HPV only. We acknowledge that some breakpoints may not be readily detected from WGS data, such as in genomic regions including high or low G:C content or homopolymeric tracts. Nevertheless, our identification of numerous HPV-positive OPSCCs that lack any detectable breakpoints strongly suggests that virus integration is not mandatory for HPV-associated cancer development. In contrast with prior reports (Koneva et al. 2018; Pinatti et al. 2021), we find no significant association between integration status and patients' demographic characteristics or clinical outcomes (Supplemental Fig. S1.5; Supplemental Tables S1.1, S1.3–S1.5).

We used independent methods to validate subsets of uniquely mapped breakpoints. HPV capture-seq, utilizing custom HPV16 baits, confirms 92% of breakpoints detected by WGS in three HPV-positive tumors (Supplemental Table S1.8, and below). We previously used Sanger sequencing to confirm 95% of breakpoints detected in cell lines (Akagi et al. 2014), so we randomly selected ~10% of breakpoints here for similar validation. Sanger sequencing of custom PCR amplicons confirms 100% of WGS predictions and reveals that microhomology between virus and host sequences occurs significantly more than expected by chance at breakpoints (Supplemental Fig. S1.6; Supplemental Table S1.9; Symer et al. 2002). We also detect insertions of heterologous or untemplated DNA sequences but find no identical breakpoint sequences across samples, contradicting a previous report (Hu et al. 2015; Dyer et al. 2016). These data support contributory but nonessential roles for nonhomologous end joining (NHEJ) and/or microhomology-mediated end joining (MMEJ) as DNA repair mechanisms in papillomavirus integration (Hu et al. 2015; Leeman et al. 2019; Yu et al. 2021).

Across all 105 tumors, uniquely mapped breakpoints are distributed broadly across the human genome (Fig. 1B; Supplemental Table S1.7), and we observe breakpoint clustering within individual tumors (Fig. 1B). We define a "cluster" as a grouping of three or more unique breakpoints within a 500-kb genomic segment in an individual tumor (Fig. 1B,C). We and others have reported comparable breakpoint clusters in individual HPV-positive cell lines derived from cervical (i.e., HeLa) and oropharyngeal (e.g., UPCI:SCC090 and others) cancers (Adey et al. 2013; Akagi et al. 2014). Breakpoint counts vary widely per tumor (mean 9, median 4, range 1–99) (Fig. 1C). Of 756 uniquely mapped breakpoints, 70% are located within a cluster in an individual tumor (Fig. 1D), with a median of one cluster per tumor (range 0 to 9). In contrast, when considering all 500 kb loci harboring one or more breakpoints, a majority contain "simple" integrants identified by two or fewer breakpoints, rather than clusters (Fig. 1E).

Integration hotspots target genes involved in epithelial stem cell maintenance and immune evasion

To account for the lack of independence of virus–host breakpoints within a cluster, we identified a single "representative breakpoint"

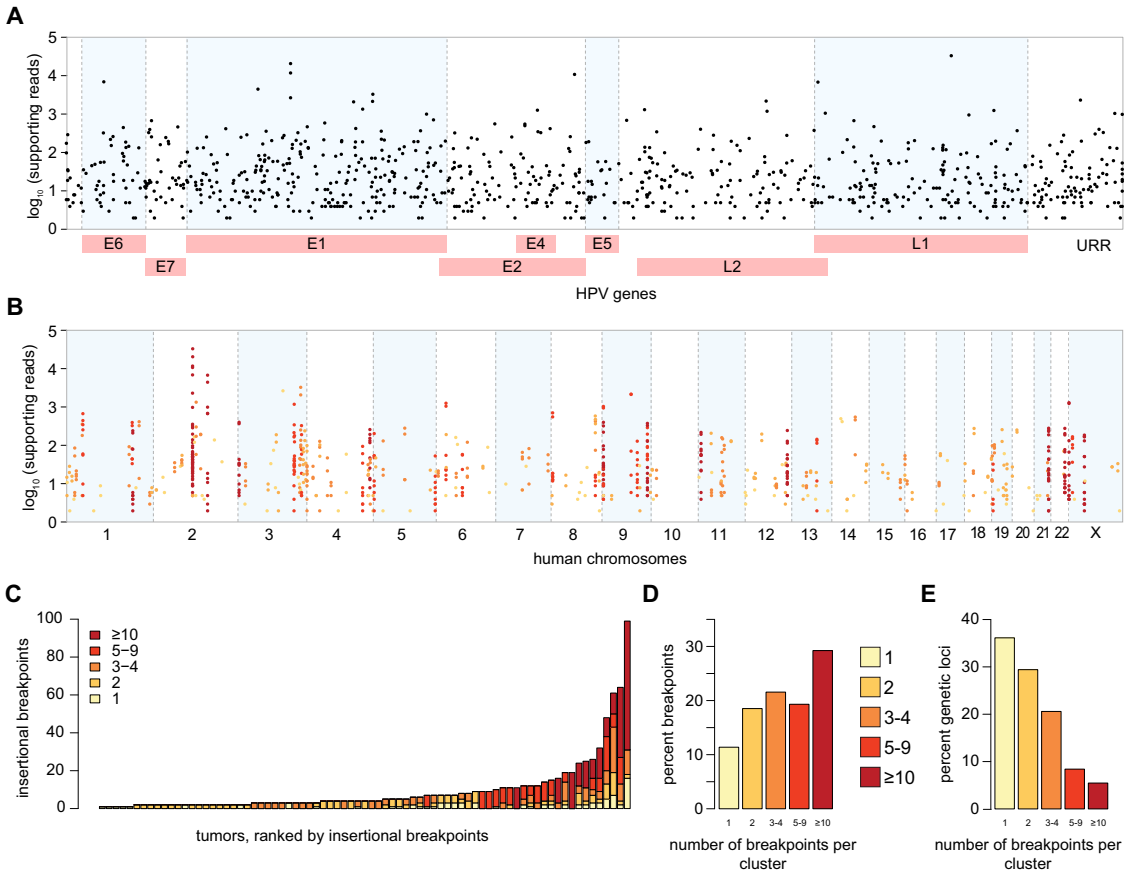


Figure 1. Frequent clustering of virus–host breakpoints in individual tumors. (*A*) Breakpoints (dots, n = 874) identified in 105 HPV-positive OPSCCs mapped to the HPV16 genome (*x*-axis). Non-HPV16 breakpoint (n = 50) coordinates are approximated; *y*-axis, $\log_{10}$ of WGS reads supporting each breakpoint. (*B*) Breakpoints uniquely mapped to the human genome (*x*-axis, n = 756, hg19) are clustered within 500-kb windows; *y*-axis as in *A*. Colors: breakpoints in individual tumors in clusters, as per key in panel *C*. (*C*) Counts of uniquely mapped breakpoints (*y*-axis), ranked by frequency in individual HPV-positive OPSCCs (*x*-axis). Colors: breakpoint counts per cluster. (*D*) Overall frequencies of breakpoints in clusters across all tumors (*x*-axis; colors, breakpoint counts per cluster). (*E*) Overall frequencies of distinct genetic loci harboring clusters of various breakpoint counts within 500 kb (*x*-axis) in individual tumors. See also Supplemental Figures S1.1–S1.6 and Supplemental Tables S1.1–S1.9.

for each cluster, defined as that breakpoint in the cluster that is supported by the highest number of WGS reads. We performed null hypothesis testing in which we simulated random distributions of all 238 defined representative breakpoints in non-overlapping genomic segments, each of one megabase pair (Mbp) length, across all 105 tumors. Our analysis identifies five independent genomic sites showing significantly recurrent HPV integration (i.e., "hotspots"), each targeted across at least three independent tumors, including *SOX2, TP63, FGFR3, MYC*, and *CD274* (empirical probability, $P = 1 \times 10^{-6}$) (Fig. 2A; Supplemental Table S2.1). Each of these genes has well-established roles in epithelial stem cell maintenance or antitumor immunity. A separate, gene-centric bioinformatics approach confirms that the same five genes are recurrently targeted hotspots (Supplemental Table S2.2). Both the *MYC* and *TP63* loci also were identified as sites of recurrent integration in independent cervical cancers (Bodelon et al. 2016).

Within the 1-Mbp genomic segment containing *MYC*, breakpoints are identified in four tumors and are associated with flanking CNVs and SVs (Fig. 2B). In one case, breakpoints directly flank an approximately ninefold amplification of *MYC,* and in another they flank a threefold amplification upstream of *MYC*. Both are associated with elevated *MYC* transcript levels (Fig. 2C; Supplemental Figs. S2.1, S2.2), consistent with impacts of amplification of the gene or adjacent super-enhancers (Zhang et al. 2016). HPV integration near *MYC* previously was reported in cervical cancers and derived cell lines (Peter et al. 2006; Yuan et al. 2017).

Additional examples of integration hotspots targeting "stemness" genes involve *FGFR3*, *SOX2*, and *TP63*. In four tumors, breakpoints map to a hotspot containing *FGFR3* (Fig. 2C; Supplemental Fig. S2.1). We previously documented recurrent FGFR3 p.249S > C activating mutations in 11% of HPV-positive OPSCCs (Gillison et al. 2019). Cancer stem cell renewal is sustained by FGF pathway activation with downstream regulation of several key transcription factors (Mossahebi-Mohammadi et al. 2020). Another hotspot involves *SOX2*, encoding a stem cell pluripotency factor. In one tumor, insertional breakpoints flank approximately fivefold amplification with outlier transcript up-regulation (Fig. 2C; Supplemental Figs. S2.1, S2.2). Such SOX2 overexpression promotes proliferation and anchorage-independent growth of squamous epithelial cancers (Bass et al. 2009). Hotspot integrants also map near *TP63*, encoding a transcription factor that regulates squamous epithelial stem cell maintenance, differentiation, and proliferation (Senoo et al. 2007). Our prior analysis identified inactivating mutations in genes promoting epithelial differentiation

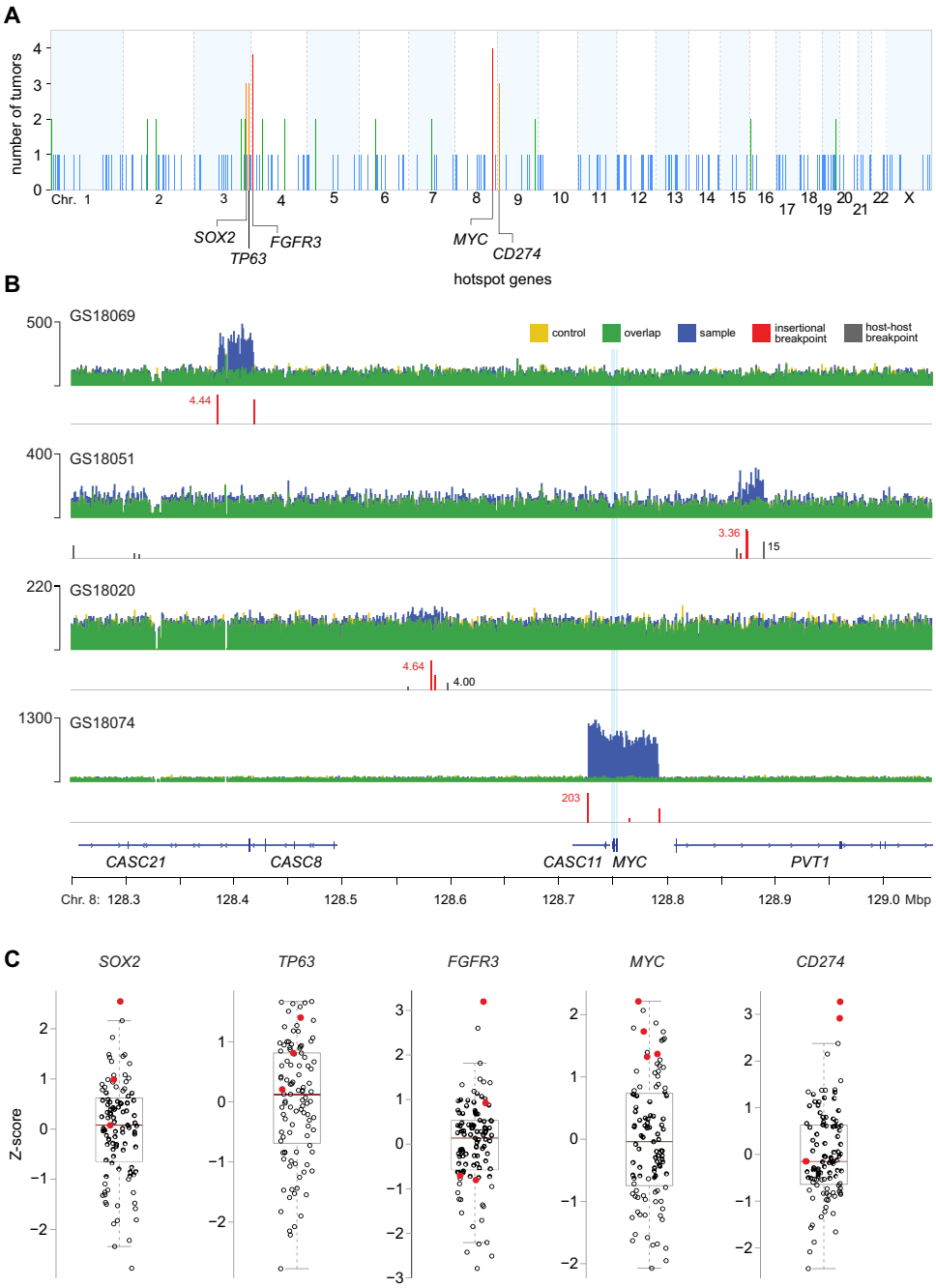


Figure 2. Genomic hotspots of integration near genes involved in epithelial stem cell maintenance and immune evasion. (*A*) Counts of independent tumors harboring ≥1 virus–host breakpoints (*y*-axis) in 1-Mbp genomic segments across the human genome (*x*-axis). Recurrent hotspots are identified at five segments containing *SOX2, TP63*, *FGFR3*, *MYC*, and *CD274*, each across at least three tumors (orange, n = 3 tumors; red, n = 4; empirical probability, $P = 1 \times 10^{-6}$). Other integration sites are not statistically significant hotspots in these tumors (blue, n = 1 tumor; green, n = 2). (*B*) IGV browser views of WGS depth of coverage (*y*-axis) and virus–host breakpoints (red) in four independent tumors within a 1-Mbp genomic segment containing *MYC* (*x*-axis; light blue vertical lines, exons). Colors as in key at *top*. *Y*-axis scale, tick mark, range (maximal count) of mapped WGS reads at locus in each tumor. See also Supplemental Figure S2.1. (*C*) Transcript levels of *SOX2*, *TP63, FGFR3*, *MYC*, and *CD274* (*y*-axis, *Z*-score of $\log_2$ TPM value), in tumors quantified by RNA-seq (circles, n = 103). Red fill: tumors with breakpoints near the hotspot genes (panels *A*,*B*). Box and whiskers, median (brown horizontal line), quartiles (light gray box). See also Supplemental Figures S2.1, S2.2 and Supplemental Tables S2.1, S2.2.

(i.e., *ZNF750, KMT2D*, *RIPK4*, and *TGFB1*) in 37% of HPV-positive tumors (Gillison et al. 2019). We conclude that disruption of differentiation and maintenance of epithelial stemness are important in the pathogenesis of these cancers.

A recurrent integration hotspot in three tumors involves the immune checkpoint ligand gene *CD274* (also known as programmed cell death 1 ligand 1 [PD-L1]). HPV integration near this gene has been reported previously (The Cancer Genome

Atlas Research Network 2017; Koneva et al. 2018). In two of the cases studied here, HPV integrants are associated with five- to 10-fold amplification and outlier expression of *CD274* (Fig. 2C; Supplemental Figs. S2.1, S2.2).

In a tumor with 63 breakpoints distributed across the viral genome, insertional breakpoint clusters are identified on Chr 4, 5, 9, 10, 19, and 22, in direct proximity to CNVs affecting specific host genes (e.g., *CD274* and *EP300*) (Fig. 3A; Supplemental Fig. S3.1; Supplemental Table S3.1). This demonstrates that diverse insertional breakpoints and/or intact copies of virus genes can co-occur in the same tumor (Akagi et al. 2014; Parfenov et al. 2014). Inspection of WGS depth of coverage around a cluster of 18 breakpoints on Chr 9p.24.1 reveals extensive CNV and SV, including ~11-fold amplification of *CD274* (Fig. 3A). Linked-read sequencing (10x Genomics) demonstrates that the HPV integrants co-occur with genomic deletions and amplifications on the same haplotype (Fig. 3B; Supplemental Figs. S3.2, S3.3). Barcodes of linked reads mapping to the CNVs or SVs are shared at high frequencies with HPV16, establishing direct connectivity between integrants and flanking structural variants. In these cases, HPV integration near *CD274* likely promoted immune escape and tumor development.

Cancer-driving genes (Sondka et al. 2018) are enriched in genetic loci neighboring HPV insertional breakpoints over loci that lack such breakpoints (Fig. 3C; Supplemental Table S3.2). Ontology analysis of genes in these loci reveals enrichment of genes involved in regulation of activated T cell proliferation, somatic stem cell maintenance, and mitochondrial apoptotic signaling, among others (Supplemental Table S3.3). These findings further confirm enrichment of HPV integrants near genes and pathways involved in cancers and strongly implicate viral integration as a driver of carcinogenesis by clonal selection.

HPV capture-seq data from additional tumors identify another hotspot involving epithelial stemness genes

To confirm and extend the recurrent HPV integration hotspots identified from WGS data (Fig. 2), we analyzed additional HPV-positive OPSCC with HPV capture-seq, a targeted sequencing method (Warburton et al. 2018). First, we compared insertional breakpoints identified from HPV capture-seq versus WGS from the same tumor. Upon normalization of sequencing depth of coverage, numbers of supporting reads are closely correlated for breakpoints detected by both methods (Fig. 3D).

Based on these results, we used HPV capture-seq to identify HPV integrants in 53 additional HPV-positive OPSCCs (Supplemental Tables S3.4, S3.5). These independent tumors harbor breakpoints near the same hotspot genes noted above, adding further support for them (i.e., at *SOX2, MYC, CD274*). By combining these 53 tumors with the 105 WGS tumors, we identify an additional, significant hotspot in three tumors near the zinc-finger transcription factor Krüppel-like factor 5 (*KLF5*) on Chr 13q22.1 (Fig. 3E; Supplemental Tables S3.6, S3.7). *KLF5* is a candidate oncogene that regulates stemness, proliferation, and differentiation of the basal epithelial cell (Ghaleb et al. 2005), the cell specifically infected by HPV. We conclude that HPV integration near genes that regulate epithelial stem cell maintenance is likely to confer a selective growth advantage, promoting tumorigenesis.

Enrichment of HPV integrants in genomic regions with CNVs and SVs

Breakpoint clusters in individual tumors with identified hotspots frequently are associated with CNVs and SVs (Figs. 2, 3).

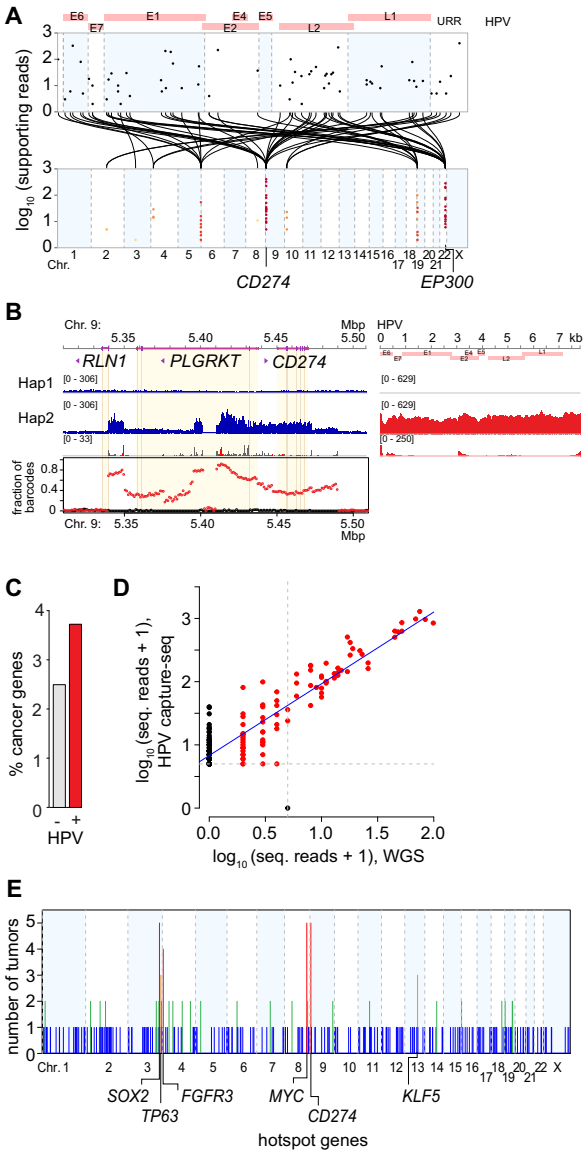


Figure 3. Associations between HPV integrants, CNVs, and SVs in individual tumors. (*A*) Strudel plot shows virus–host breakpoints in a representative OPSCC, GS18047. Breakpoints mapped to the HPV16 genome (*top*, *x*-axis) are connected (black lines) with the host genome (*bottom*, *x*-axis), clustered on Chr 4, 5, 9, 10, 19, and 22 (colored dots, as per key in Fig. 1C). Disrupted genes include *CD274* and *EP300*. (*B*) Haplotype-resolved linked reads (blue, host depths of sequencing coverage; red, HPV) connect HPV16 sequences (*right*), virus–host breakpoints (red peaks) and host–host breakpoints (gray) on one allele (haplotype 2) at the *CD274* locus on Chr 9p24.1. Graph (*bottom*), shared linked-read barcodes connect HPV16 exclusively to haplotype 2 (red) but not haplotype 1 (black). (*C*) Fraction of genes with (red) or without (gray) HPV breakpoints within ± 500 kb that are annotated cancer genes (*y*-axis) (Sondka et al. 2018). Fisher's exact test, $P = 6.3 \times 10^{-5}$ (Supplemental Table S3.2). (*D*) Scatterplot shows strong correlation between read counts supporting individual breakpoints (red dots) identified with HPV capture-seq (*y*-axis, n = 164) versus WGS (*x*-axis, n = 86) in the same tumor (r = 0.91; $P = 1.8 \times 10^{-63}$). (*E*) Adding 53 tumors studied by HPV capture-seq to 105 tumors studied by WGS (Fig. 1A), tumors harboring ≥1 virus–host breakpoints (*y*-axis) in 1-Mbp genomic segments were recounted across the human genome (*x*-axis). Statistically significant, recurrent hotspots (orange, n = 3 tumors; red, n = 4 or 5) are detected at segments containing *SOX2*, *TP63*, *FGFR3*, *MYC*, *CD274*, and *KLF5* (empirical probability, $P = 7 \times 10^{-6}$). See also Figure 2, Supplemental Figures S3.1–S3.3, and Supplemental Tables S3.1–S3.7.

Therefore, we investigated associations between HPV integration and CNVs and SVs across all tumors studied by WGS. The frequency distribution of CNVs is markedly different in comparing 100-kb host genomic segments with and without virus–host breakpoints (Fig. 4A). Quantile-quantile (Q-Q) plots comparing the distribution of genomic copy numbers in the presence versus absence of HPV breakpoints demonstrate unequivocally that viral insertions are strongly associated with copy number alterations ($P < 2.2 \times 10^{-16}$, Kolmogorov–Smirnov test) (Fig. 4B). Breakpoints are highly enriched in segments containing CNVs, particularly in hyperamplified segments with estimated copy number ≥ 4 n (16.3-fold enrichment, binomial test, adj. $P = 2.06 \times 10^{-20}$) (Supplemental Table S4). Breakpoints directly flank CNV regions with amplification up to 15-fold and/or lengths exceeding 5 Mbp.

Genomic SVs including deletions, insertions, inversions, and chromosomal translocations are enriched in genomic segments with HPV breakpoints compared to those without (44.7% vs. 0.47%) (Fig. 4C). Copy number transitions (CTs, i.e., step changes in copy number >0.5 n) also are enriched in segments with breakpoints compared to those without (37.6% vs. 0.88%) (Fig. 4C). Across all tumors, breakpoints map within 10 kb of inversions in 21%, duplications in 34%, deletions in 20%, and chromosomal translocations in 6%. The larger the number of breakpoints within a cluster, the more frequent the concomitant SV counts and step-changes in CNVs (Fig. 4D,E), supporting direct involvement of HPV integration in generation of host genomic rearrangements.

Association between HPV integrants and outlier expression of neighboring host genes including cancer genes

Z-scores were calculated for all genes' transcript expression levels across 103 OPSCC tumors with available RNA-seq data. To investigate impacts of HPV integration on host gene expression, we identified neighboring genes within ±500 kb of breakpoints in affected tumors and used Q-Q plots to compare the distribution of expression levels for these genes near a breakpoint versus those without a corresponding breakpoint. In many cases with breakpoint(s) present, neighboring genes are significantly overexpressed (and less frequently underexpressed) in the affected tumor compared to controls, as shown by significant deviation of many data points away from the line of identity (Fig. 5A).

Genes with statistical outlier expression levels (*Z*-scores ≥ 2) are disproportionately higher in frequency within ±500 kb of HPV breakpoints compared with those lacking breakpoints, and this difference is greater among cancer genes (Sondka et al. 2018) than noncancer genes (Fig. 5B). Among the 2898 genes neighboring individual breakpoints across the samples, 220 have outlier expression and 108 are cancer genes (Supplemental Tables S5.1, S5.2). Of these, 16 are cancer genes that neighbor an HPV breakpoint and display outlier expression (Fig. 5C; Supplemental Fig. S4.1). Thus, known cancer genes displaying outlier expression in individual tumors are enriched in proximity to HPV integrants, implicating a selective growth advantage imparted by this proximity.

Further analysis of the relationship between a gene's proximity to an insertional breakpoint and its outlier expression status reveals that integrants' impacts on gene expression are greatest at distances <±150 kb (Fig. 5D). Outlier levels of gene expression (*Z*-score ≥ 2) are sixfold more likely (binomial test, adj. *P*-value 1.43×10^{-63}), and extreme outlier levels of expression (*Z*-score ≥ 4) are 86-fold more likely (binomial test, adj. *P*-value 1.23×10^{-86}), for genes within ±150 kb of an HPV breakpoint in affected tumors, compared to the same genes in tumors without those integrants.

Even after accounting for the strong association between HPV integration and local genomic copy number as noted above, outlier expression is significantly enriched in regions with HPV integrants (Fig. 5E). This demonstrates that proximal host gene expression is affected by HPV integrants, even after accounting for effects of differences in local copy number.

To analyze HPV insertions' impacts on regional gene expression levels, we defined HPV-linked host genomic rearrangements as the broader chromosomal regions containing breakpoints (or clusters), flanked by ±1-Mbp margins (Supplemental Table S5.3). We compiled 238 HPV-linked rearrangements across the 105 tumors and compared the

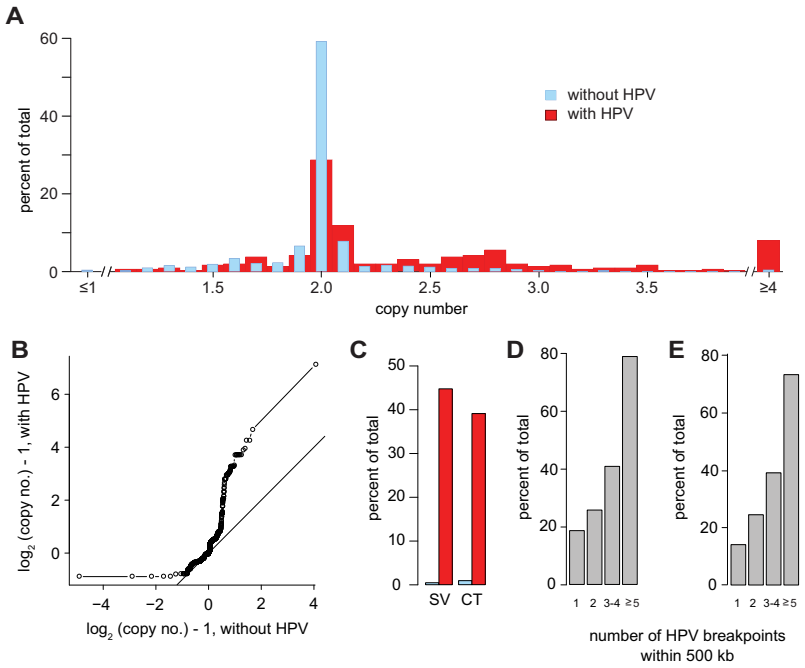
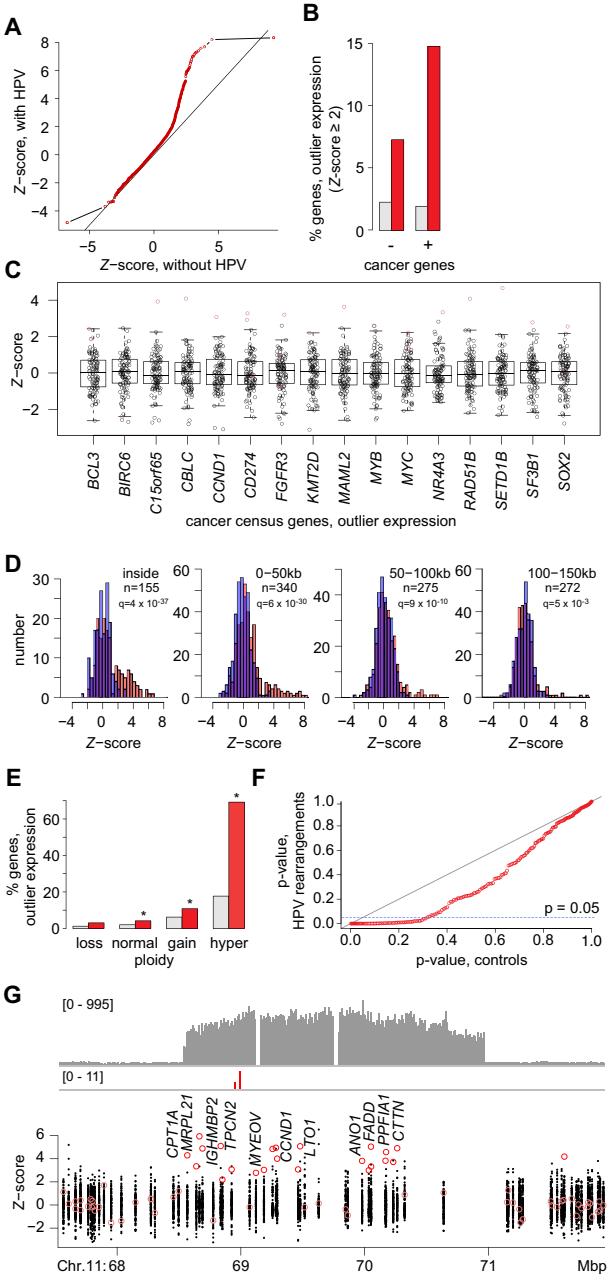


Figure 4. HPV integrants are associated with CNVs and SVs across tumors. (*A*) Shown are distinct frequency distributions (*y*-axis) of copy numbers (*x*-axis) of 100-kb genomic segments with (red) versus without (blue) virus–host breakpoints across 105 HPV-positive OPSCC (χ^2, $P = 1.8 \times 10^{-18}$). (*B*) Quantile-quantile (Q-Q) plot confirms differences in copy numbers of genomic segments with (*y*-axis) and without (*x*-axis) breakpoints, deviating significantly from the line of identity ($P < 2.2 \times 10^{-16}$, Kolmogorov–Smirnov test). (*C*) Frequencies (*y*-axis) of structural variation (SV, *left*) and step-changes in copy number (copy number transition [CT] ±0.5 n, *right*) are significantly greater in 100-kb segments with a breakpoint (red) versus without (gray) (SV; binomial test, one-tailed, $P = 3.3 \times 10^{-224}$; CT, $P = 2.39 \times 10^{-14}$). (*D*,*E*) Among 500-kb genomic segments with ≥1 breakpoints, frequencies (*y*-axis) of (*D*) SVs and (*E*) CTs increase with breakpoint counts in the cluster. See also Supplemental Table S4.

cumulative distributions of involved genes' expression levels in tumors with versus without such rearrangements in a Q-Q plot, by calculating the sum of the square of their *Z*-scores. The results show that expression levels of the genes in ~30% of the rearrangements are significantly different from those in the same regions in control tumors without breakpoints (Fig. 5F). A representative HPV-linked rearrangement showing expression levels of involved genes reveals that HPV integrants flank and bridge host CNVs and SVs and induce significant outlier expression of numerous regional genes, in this case including cyclin D1 (*CCND1*) (Fig. 5G).

To evaluate a possible association between expression of HPV E6 or E7 oncogenes and expression of genes near HPV integrants in individual tumors, we plotted the maximum *Z*-score out of all genes within ±500 kb of viral integrants, for each tumor in the Ohio cohort. Scatterplots revealed no such association between viral gene expression and maximum host gene expression in each tumor (Supplemental Fig. S4.2).



Diverse forms of genetic disruption by HPV integrants

We investigated relationships between HPV integrants and host genomic features such as annotated genes and regulatory elements. Consistent with previous findings (Bodelon et al. 2016), breakpoints are enriched within fragile sites (binomial test, adj. $P = 3.5 \times 10^{-6}$) and DNase I hypersensitive sites (binomial test, adj. $P = 4.4 \times 10^{-7}$), indicating that open chromatin may facilitate HPV integration. Breakpoints are enriched in protein-coding gene promoters (binomial test, adj. *P*-value 5.5×10^{-8}) and exons (binomial test, adj. *P*-value 4.4×10^{-6}), and a slight majority (56.2%) is localized in intragenic regions of annotated genes (Supplemental Tables S6.1, S6.2). Breakpoints are enriched in genomic sites bearing marks of active enhancer elements (e.g., histone 3 lysine 27 acetylation [H3K27ac], binomial test, adj. $P = 6.4 \times 10^{-9}$), as detected in normal human epithelial keratinocytes (NHEK), but conversely also at sites bound by CCCTC-binding factor (CTCF), a transcription factor and insulator protein (binomial test, adj. $P = 2.5 \times 10^{-7}$) (Supplemental Table S6.3). Our findings corroborate associations between HPV integrants and transcriptionally active chromatin bearing H3K27ac marks in HPV-positive cell lines and patient-derived xenografts (Kelley et al. 2017). Moreover, HPV insertions amplified and hijacked a cellular enhancer in subclones derived from W12 cells, a cervical cell line, forming a super-enhancer-like element (Warburton et al. 2018). These data collectively indicate that alterations in host chromatin structure and

Figure 5. HPV integrants are associated with outlier expression of neighboring host genes. (*A*) Q-Q plot compares *Z*-score distributions of expression levels for genes near (±500 kb) virus–host breakpoints (*y*-axis) versus expression of the same genes without nearby breakpoints in all other tumors (*x*-axis; Kolmogorov–Smirnov test, $P < 2.2 \times 10^{-16}$). Line of identity (dark gray). (*B*) Percent of genes with outlier expression that are not cancer genes (–, *left*) or are cancer genes (+, *right*) as per Cancer Gene Census Database, and are not (gray) or are (red) within ±150 kb of an HPV integrant (Fisher's exact test, FDR correction, $P = 2.2 \times 10^{-11}$ [*left*] and $P = 7.2 \times 10^{-55}$ [*right*], respectively). (*C*) Of 220 genes expressed at outlier levels (Z-score ≥ 2 or ≤−2) in ≥1 tumor and within ±500 kb of a breakpoint, 16 are cancer genes as shown. Box and whiskers plot, *Z*-scores (*y*-axis) for cancer genes (*x*-axis) in samples harboring nearby breakpoints (red) versus lacking them (no fill). (*D*) Comparison of gene counts (*y*-axis) expressed at various levels (*Z*-scores, *x*-axis), grouped in 50-kb genomic distances from the nearest breakpoint, in tumors with (red) and without (blue) breakpoints in those segments. Of 194 genes harboring breakpoints, 155 are expressed as per available RNA-seq data. *Left to right*, breakpoints across the tumors inside or outside genes as indicated; n, counts; *q* = adj. *P*-values, binomial test. (*E*) Percentages of genes expressed at outlier levels ($Z \geq 2$) at indicated copy numbers (*x*-axis) in absence (gray) or presence (red) of breakpoints within ±500 kb. Copy number loss, $n < 1.5$; normal, $1.5 \leq n \leq 2.5$; gain $2.5 \leq n \leq 5$; hyper-gain $n > 5$. Asterisks, adj. $P < 1 \times 10^{-4}$, binomial test, one-tailed, adjusted by FDR. (*F*) Q-Q plot of χ^2adjusted *P*-values calculated from comparison of gene expression *Z*-score distributions (i.e., sum of the square of *Z*-scores) at chromosomal loci with HPV-mediated rearrangements (*y*-axis) versus matched loci without rearrangements (*x*-axis). (*G*, *top*) Depth of sequencing coverage (*y*-axis); (*bottom*) *Z*-scores of $\log_2$ TPM (*y*-axis) for genes in a tumor with an HPV-linked rearrangement on Chr 11q13.3 (*x*-axis); genes with outlier expression (red fill). Breakpoints (red vertical lines) mapping within a 2.4-Mbp region with eightfold amplification result in outlier expression (Z-score ≥ 2) for 22 (67%) of 33 genes including cyclin D1 (*CCND1*). Tumor with HPV-linked rearrangement: Z-score < 2 (pink), outlier Z-score ≥ 2 (red); all other tumors without detectable local HPV insertions (black). See also Supplemental Figures S4.1, S4.2 and Supplemental Tables S5.1–S5.3.

concomitant changes in gene regulation can serve as a target for and/or result from HPV integration.

Previous analysis of 35 HPV-positive OPSCCs identified intragenic HPV integrants which disrupted *RAD51* and *ETS2* (Parfenov et al. 2014). Here, breakpoints map within annotated genes in 79% of tumors with HPV integration (n = 194, mean 3.2, median 2, range 1–20 genes per tumor) (Supplemental Fig. S5.1; Supplemental Table S6.4). Intragenic HPV integration coincides with several forms of gene disruption in 71% of these tumors, including CNVs (53%), SVs (41%), and/or virus–host chimeric transcripts (35%). Expression of an additional 79 host genes without detectable intragenic breakpoints is altered by chimeric transcript expression, indicating creation of fusions from nearby HPV insertions. Overall, ~92% of tumors harboring HPV integrants have one or more genes (mean 3.97, median 3, range 1–23) disrupted by intragenic insertions and/or chimeric transcript expression.

In 85% of the tumors with HPV integration, a broad range of distinct virus–host chimeric transcripts is detected in RNA-seq data (mean 10.5, median 8, range 1–56 unique transcripts per tumor) (Supplemental Table S6.5). We aligned chimeric transcripts expressed in HPV16-positive tumors to the HPV16 reference genome (Fig. 6A). A majority of transcripts initiated from viral promoters utilize established viral splice donor sites (ranked by frequency at HPV16 nt. 880 > 226 > 1302 > 3632) (Fig. 6A; Supplemental Fig. S5.2) spliced to diverse host splice acceptor sites. When aligned to the reference human genome (Fig. 6B), chimeric transcripts within the same tumor frequently map in close proximity to one another, consistent with alternative splicing and/or transcription from clustered HPV integrants. Splice junctions in virus–host chimeric transcripts detected in RNA-seq data frequently map at considerable (i.e., ≥1 Mbp) distances from the nearest virus–host DNA breakpoint (Fig. 6C; Supplemental Fig. S5.3), indicating that analysis of chimeric transcripts can mislocate the actual HPV integration site in the host genome. We detect no chimeric transcripts from ~50% of inserted virus sequences as represented by DNA breakpoints. Approximately half of the host genes from which chimeric transcripts are expressed lack intragenic breakpoints. These results show that chimeric transcripts detected by RNA-seq or other RNA-based mapping methods are poor surrogates for detection of all virus–host DNA breakpoints. Because only 56% of virus–host breakpoints are intragenic, whole-exome sequencing also would be a poor proxy for detection of all breakpoints.

Chimeric transcripts are expressed at 147 genes across the tumors, inducing outlier expression of 35% of them (Supplemental Tables S6.5, S6.6). In contrast to canonical splicing of transcripts expressed from amplified host genes (Fig. 5G), chimeric transcripts display markedly altered structures when expressed at outlier levels. Chimeric transcripts disrupt the 147 host genes in part via readthrough expression and/or usage of host splice donor, splice acceptor, and/or cryptic splice sites (Fig. 6D). For example, 31 distinct chimeric transcripts were detected at *CASC8* in a single tumor (Supplemental Table S6.6).

Several forms of genetic disruption are manifested in chimeric transcripts, including expression of novel exons, premature transcriptional termination, and gene breaking. Exon-by-exon expression analysis of all host genes affected by fusion transcripts helps visualize these diverse forms of genetic disruption (Supplemental Fig. S5.4). An example of gene breaking (Wheelan et al. 2005), induced by intragenic HPV integrants, is identified at mastermind like transcriptional coactivator 2 (*MAML2*), a transcriptional coactivator of Notch (Fig. 6E; Wu et al. 2002). Premature transcriptional termination of *MAML2* after exon 2 and de novo initiation of downstream transcripts from HPV were observed. Although *NOTCH1* itself is mutated in ~17% of head and neck cancers, Notch pathway signaling is disrupted by other mechanisms, including rare driver gene mutations in an estimated 67% (Loganathan et al. 2020). Gene breaking by HPV integrants at *MAML2* uncovers a novel additional mechanism for Notch pathway disruption. Comparable instances of genes harboring intragenic insertional breakpoints are 130-fold more likely to show extreme outlier expression (i.e., *Z*-scores ≥ 4) compared to the same genes in control tumors lacking such breakpoints (8.4 vs. 0.06%, binomial test, adj. $P = 1.3 \times 10^{-22}$). Overexpression of disruptive, chimeric transcripts drives this effect in most cases. In another example, an intragenic HPV integrant within a nuclear importin gene, *IPO8*, results in deletion of 3′ exons 23 to 25 (Fig. 6F). Genetic disruption by an intergenic HPV16 integrant upstream of insulin induced gene 2, *INSIG2*, encoding a negative regulator of cholesterol biosynthesis, is shown in Figure 6G. Here, numerous chimeric transcripts initiated from HPV promoters are spliced to a novel exon, a novel splice acceptor site, and exons 1, 2, and 3. Disruption of this HIF1A-inducible gene may facilitate cell growth in a hypoxic tumor microenvironment (Hwang et al. 2017).

A simple HPV*69* integrant induces high expression of an imprinted oncogene, *RTL1*

On an infrequent basis, involving ~7% of genes with chimeric transcripts, intragenic HPV integrants markedly up-regulate otherwise completely unexpressed genes. Most of these cases affect noncoding RNAs such as *LINC0001* (Zapatka et al. 2020), supporting their putative interactions with HPV (Supplemental Table S7; Sharma and Munger 2020). In a tumor with a low burden of somatic mutations, CNVs, and SVs (Fig. 7A–C), we detected a simple insertion of a segment of HPV69 (nt. 2802–2266) in the *DLK1-D103* imprinted domain on Chr 14q32. HPV E6, E6*I, and E7 transcripts are spliced to exon 1 of retrotransposon Gag like 1 (*RTL1*), resulting in extreme outlier levels of expression (396 TPM vs. median = 0.011 TPM in controls) (Fig. 7D,F). In the same tumor, another simple HPV69 insertion on Chr 15q21 results in dramatic up-regulation of E6/E7 chimeric transcripts and outlier expression of *C15orf65* (27.1 TPM vs. median = 3.44) (Fig. 7E,G). These results demonstrate that simple integration even by a rare HPV type with unconfirmed oncogenicity can induce outlier expression levels of candidate driver genes in a primary tumor with a low burden of mutations, CNVs, and SVs. We conclude that these HPV69 integrants likely contributed substantially to the etiology of this tumor, suggesting that this rare viral type may indeed be oncogenic (Bernard et al. 2013).

In sum, HPV disrupts host transcription in 95% of the tumors with integration, via intragenic integrants, chimeric transcription, outlier expression, gene breaking, and/or de novo expression of noncoding and imprinted genes (Supplemental Table S6.4).

Discussion

In this analysis, we show that HPV integrants disrupt host genomic structure and expression in almost all tumors with viral integration. This prospective study was powered to detect significantly recurrent hotspots for integration, and we identified six hotspots near genes that regulate specific biological processes of epithelial stem cell maintenance (i.e., *MYC*, *FGFR3*, *SOX2*, *TP63*, *KLF5*) and immune cell function (i.e., *CD274*). The

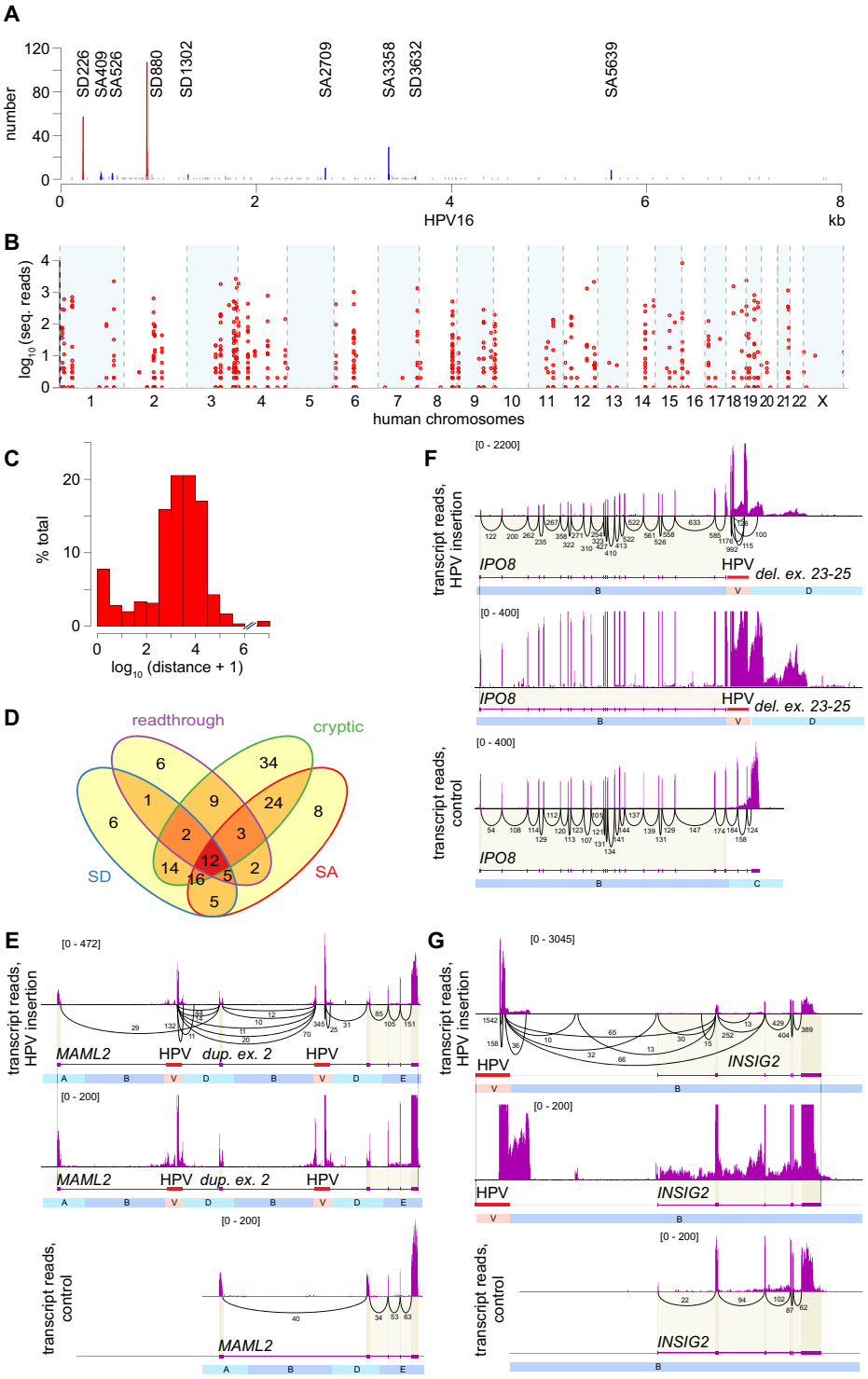


Figure 6. HPV integrants induce various forms of genetic disruption including gene breakage and chimeric transcription. (*A*) Counts of virus–host chimeric transcript junctions (*y*-axis) in 91 HPV16-positive tumors, aligned to the HPV16 genome (*x*-axis) with known splice donor (SD coordinate, red), splice acceptor (SA coordinate, blue), and other (gray) sites. (*B*) Counts of split or discordant RNA-seq reads (*y*-axis) in 103 HPV-positive tumors supporting chimeric transcript junctions (n = 673), aligned to the human genome (*x*-axis). (*C*) Frequency distribution (*y*-axis, percent total) of $\log_{10}$-transformed genomic distances (*x*-axis) between virus–host junctions from RNA-seq versus nearest DNA breakpoint (n = 604). (*D*) Venn diagram counts chimeric transcripts expressed at 147 genes, via host splice donor (SD, blue, n = 61); splice acceptor (SA, red, n = 75); readthrough transcription (purple, n = 40), and/or cryptic splice sites (green, n = 114). (*E–G*) Sashimi plots depict counts of mapped RNA-seq reads at genes with HPV integrants in affected tumor (*top*, *center* panels) versus without integrants in control tumor (*bottom*). Intron sequences not shown to scale. *Center*, *bottom* panels, identical scale of reads (*y*-axis, brackets). Black arcs, numbers, read counts connecting spliced exons. (*E*) Intragenic HPV integrants in *MAML2* (red) flank a ~75-kb duplication including exon 2, and delete small intronic segment C. Gene breaking involves premature transcriptional termination of *MAML2* after exon 2 and de novo initiation of downstream transcripts from HPV. Segment B is truncated for visualization. (*F*) Intragenic HPV integrants in *IPO8* (red) delete distal exons 23–25, disrupting 3′ transcripts and up-regulating upstream exons. (*G*) Intergenic HPV integrants both upstream of and downstream from *INSIG2* flank a ~665-kb duplication on Chr 2q14, extending from Chr 2:118.826 to 119.492 Mbp. Numerous chimeric transcripts originating from an upstream, intergenic HPV16 integrant are spliced to a novel exon, novel splice acceptor site, and exons 1, 2, and 3, causing gene disruption. See also Supplemental Figures S5.1–S5.4 and Supplemental Tables S6.1–S6.6.

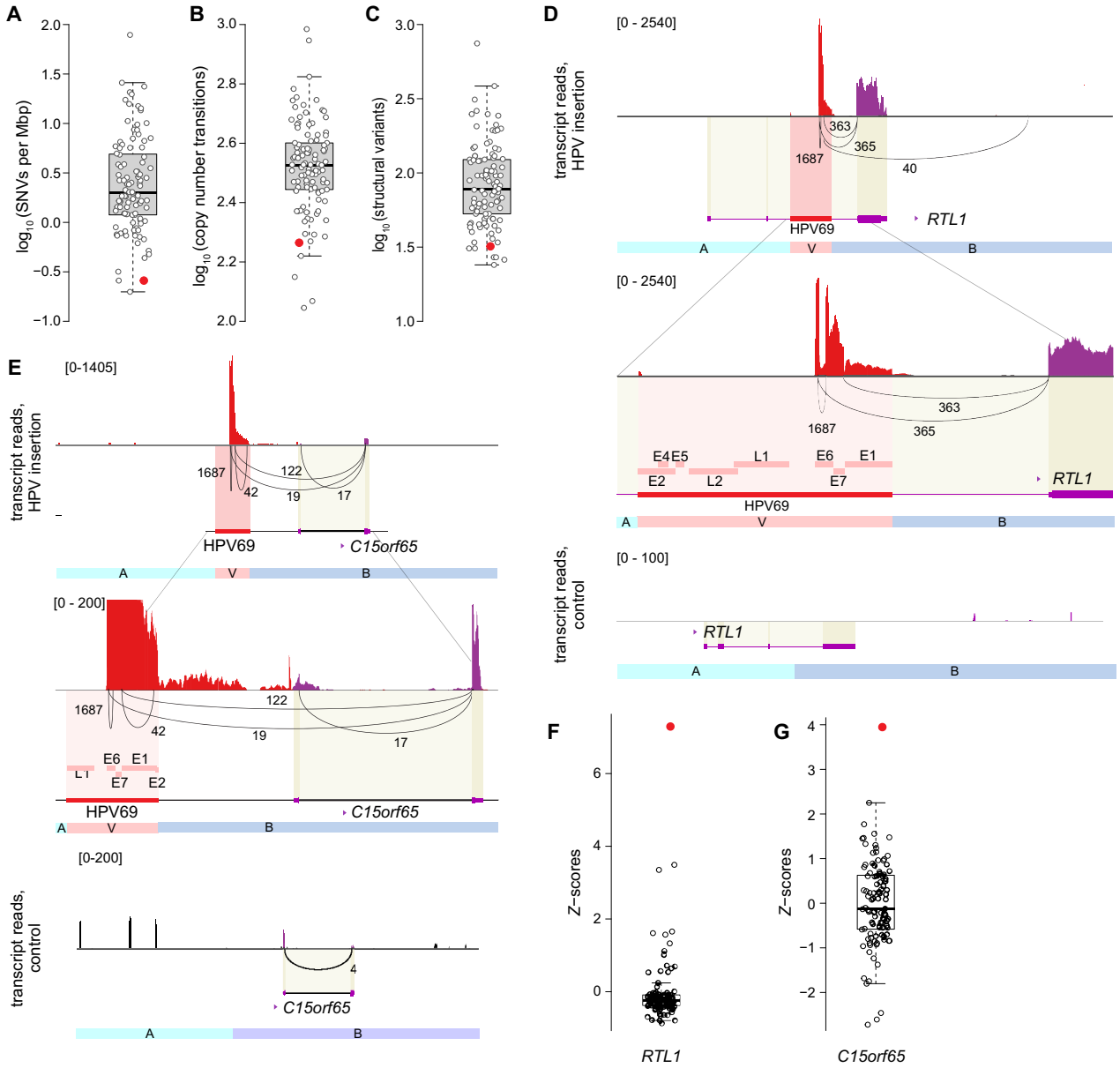


Figure 7. Simple HPV69 integrants induce high expression of an imprinted oncogene, *RTL1*, and of *C15orf65*. (*A*–*C*) Box and whisker plots depict $\log_{10}$-transformed counts of (*A*) SNVs or small indels per megabase pair, (*B*) copy number step-changes (±0.5 n), and (*C*) SV breakpoints in 105 OPSCC (circles): red, tumor with HPV69 integrants; no fill, all others. (*D*,*E*) Sashimi plots of (*top*) chimeric transcripts initiated in HPV69, spliced to exons of (*D*) *RTL1* and (*E*) *C15orf65*, leading to extremely high expression relative to controls. Black line, numbers, read counts connecting spliced exons; *bottom*, read counts of conventional transcripts in control tumor. Some *RTL1* fusion transcripts extend past the 3′ transcription termination signal. (*F*,*G*) Box and whiskers plots of expression levels of (*F*) *RTL1* and (*G*) *C15orf65* in 103 HPV-positive OPSCC. Red, HPV69-positive tumor; gray, all others. See also Supplemental Table S7.

pathophysiological significance of this finding is underscored by known roles for these genes in cancer development. Among these are transcription factors known to mediate self-renewal, proliferation, and epithelial differentiation and the CD274 (also known as PD-L1) immune checkpoint ligand that facilitates tumor escape from immune surveillance. Second, the numbers of virus–host breakpoints increase directly with the frequency of regional host genomic rearrangements, supporting a direct role for HPV integrants in SV formation. Third, even after accounting for genomic copy number changes, we find significant enrichment of integrants within 150 kb of genes expressed at outlier levels, thereby implicating HPV integration as the direct cause of genetic disruption. Fourth, essentially all (~95%) tumors with HPV integration have one or more genes disrupted by intragenic insertions and/or chimeric RNA transcription. For example, in a tumor with low burden of mutations, that is, lacking frequent SNVs, CNVs, and/or SVs, we discover simple HPV integrants inducing outlier levels of chimeric transcripts involving candidate oncogenes. Taken together, our findings indicate that HPV integrant-mediated

genetic alterations are among the secondary genetic events that promote carcinogenesis.

Three of the recurrent hotspot genes that we find across our large collection of OPSCC samples, that is, *MYC, TP63*, and *KLF5*, also have been identified across cervical cancers in a meta-analysis (Bodelon et al. 2016), supporting a common HPV-mediated tumor biology across distinct anatomic sites. The three additional hotspot sites in OPSCC identified here, involving *SOX2*, *FGFR3*, and *CD274*, are previously unreported as hotspots in HPV-positive cancers, although additional cases of HPV integration reported near *CD274* (Koneva et al. 2018) support our findings. Differences by anatomic site could be attributable to limitations in sample sizes or methodologies of prior studies, random fluctuations between sample collections, or specific biological differences across the tissue sites of HPV infection. The latter explanation may underlie why we did not find the *RAD51B* or other hotspots in OPSCC as reported in cervical cancers (Ojesina et al. 2014; Bodelon et al. 2016; The Cancer Genome Atlas Research Network 2017). We note that studies utilizing WES miss half of HPV integration breakpoints. RNA-seq can detect chimeric transcripts, but it mismaps the corresponding integrant templates by up to several megabase pairs. Moreover, RNA-seq altogether misses the half of genomic integrants that we find do not generate chimeric transcripts. In the 874 integration breakpoints identified here, we find no instances of identical breakpoints shared across any two independent tumors (Hu et al. 2015; Dyer et al. 2016), highlighting the robustness of our approach.

Here, we show that HPV integration disrupts host genes by mechanisms including genomic amplifications, deletions, inversions, translocations and other rearrangements, intragenic insertion, initiation of chimeric transcripts from heterologous viral promoters, and introduction of promoter, splice acceptor, splice donor, and transcriptional terminator elements. Each of these disruptive mechanisms can lead to alterations in gene expression, structure, and function. HPV integrants amplified oncogenes including *MYC, CCND1,* and *SOX2*, and an immune checkpoint ligand *CD274*; disrupted known tumor suppressors including *MAML2, EP300*, and *INSIG2*, and induced aberrant overexpression of the imprinted oncogene *RTL1*. Our data indicate that HPV-mediated insertional mutagenesis is at least as disruptive of host genomic homeostasis as are SNVs but would not be detected or resolved by widely utilized platforms such as WES or targeted cancer gene panels (Chung et al. 2015).

We note that the primary HPV-positive OPSCCs studied here could not be further manipulated experimentally. However, prior studies of unrelated HPV-positive cancer cell lines demonstrated that their proliferation and viability depend on genetic disruptions caused by HPV integrants (Akagi et al. 2014; Shen et al. 2017; Warburton et al. 2018; Broutian et al. 2020). The best-studied of these is the HeLa cell line, established in culture from Henrietta Lacks' aggressive cervical cancer. A likely instigator of tumorigenesis in this case consists of the HPV-mediated amplifications, rearrangements, and long-range interactions with enhancer elements that induce massive up-regulation of *MYC* (Adey et al. 2013; Shen et al. 2017). Experimental deletion of these HPV integrants resulted in marked reductions in *MYC* expression (Shen et al. 2017). Similarly, we demonstrated collaboratively that *MYC* overexpression drives proliferation of the cervical neuroendocrine cell line GUMC-395, in which HPV integrants directly flank a ~40-fold amplification of *MYC* (Yuan et al. 2017). *MYC* overexpression drives tumorigenesis upon *TP53* loss (Zindy et al. 1998). Documented cooperative interactions between HPV E6 and MYC in inducing telomerase (*TERT*) expression and keratinocyte immortalization further support the pathophysiological significance of viral integration at this hotspot (Zhang et al. 2017). Moreover, we have shown that genetic knockdown or small molecule inhibition of PIM1 induces cell death in UPCI:SCC090, a head and neck cancer cell line in which HPV integrants directly flank a 16-fold amplification of *PIM1* (Broutian et al. 2020). These examples illustrate the contributions of HPV integrant-mediated transcriptional alterations to the malignant phenotype.

In a tumor with very low burdens of somatic SNVs, CNVs, and SVs, two simple HPV69 integrants were identified on Chr 14 and 15 (Fig. 7). HPV69 has not been classified as a high-risk, oncogenic virus type, likely because of its low prevalence in tumors (Bernard et al. 2013). Highly expressed chimeric transcripts are initiated from both integrated HPV promoters and include the viral E6/E7 genes. The Chr 14 transcripts are spliced to exons of *RTL1,* an imprinted oncogene, whereas Chr 15 transcripts are spliced to *C15orf65* exons, encoding a protein of unknown function (Fig. 7). Mouse models of hepatocellular carcinogenesis involving an engineered Sleeping Beauty transposon showed highly analogous, frequent activating insertions upstream of *RTL1*, resulting in overexpression of transposon-*RTL1* chimeric transcripts (Riordan et al. 2013). Overexpression of *RTL1* in adult mouse liver led to tumor formation in 86%, revealing that it is a potent oncogene. Aberrant expression of *RTL1* also has been reported in melanoma (Fan et al. 2017). By analogy, we conclude that marked up-regulation of these two genes due to simple HPV integrants (without associated CNVs or SVs) also could drive OPSCC tumorigenesis directly, drawing parallels to retrovirus and activating transposon integrants causing cancer in numerous contexts (Kawakami et al. 2017; Bushman 2020).

Persistent HPV infection comprises stable maintenance of HPV episomes, initiated years to decades prior to cancer diagnosis, but the timing of viral integration in tumorigenesis remains uncertain. HPV integration increases in frequency with severity of cervical dysplasia and is present in the majority of cervical cancers (Bodelon et al. 2016), suggesting that it drives clonal selection and carcinogenesis. In vitro models demonstrated that HPV integration confers a selective growth advantage, attributed to increased expression and stabilization of viral transcripts encoding oncoproteins (Jeon et al. 1995).

Our working model holds that initial HPV integration occurs randomly but preferentially in genomic regions of open chromatin. Simple HPV integration may occur at sites of individual DNA double-strand breaks. In contrast, at sites of multiple such breaks, HPV may capture transient free ends of intervening host DNA, generate virus–host concatemers via rolling circle amplification typically initiated from the viral origin of replication, and after recombination and repair, result in clusters of integrants, CNVs, and SVs (Akagi et al. 2014). Resulting viral-host concatenated arrays at insertion sites occur because viral replication depends upon host DNA damage response pathways (Gillespie et al. 2012; Reinson et al. 2013). Individual cell clones with growth advantages imparted by HPV integration are positively selected, resulting in associations between insertional breakpoints and CNVs, SVs, cancer genes, and outlier levels of expression specific to each emerging cancer. We formulated this mechanistic model, termed the HPV-mediated looping model, to explain the formation and impacts of HPV integrants in cultured cancer cells (Akagi et al. 2014). Here, we have provided comprehensive additional evidence from primary cancers to support and extend this model.

Limitations to our approach include a lack of analysis of HPV insertion-mediated effects on small RNAs, long noncoding RNAs, or epigenetic controls such as DNA methylation or chromatin changes. We also have not fully resolved the genomic architecture of some of the most complex HPV-mediated SVs, including presence or absence of ecDNA, nor the structures of spliced, polycistronic transcripts; therefore, currently we are utilizing complementary methods including long-read DNA and RNA sequencing to address these points. We plan to investigate tumor heterogeneity of HPV integration-mediated SVs, CNVs, and chimeric transcription at subclonal or single cell levels.

We conclude that host genomic alterations and gene disruptions induced by virus integration are necessary driver events in a high proportion of HPV-positive oropharyngeal cancers. We acknowledge that ~23% of these tumors lack detectable insertional breakpoints, suggesting that they may harbor episomal HPV DNA only (Gray et al. 2010). In such tumors, other forms of secondary genetic events (e.g., aneuploidy or mutations in specific genes) promote cancer development (Gray et al. 2010; McBride and Warburton 2017; Gillison et al. 2019).

In summary, we find HPV integrants to be significantly associated with recurrent hotspots, CNVs, SVs, and numerous additional forms of genetic disruption, supporting our looping model (Akagi et al. 2014). We and others have reported similar impacts in tumors caused by Merkel cell polyomavirus (Starrett et al. 2020) and hepatitis B virus (Jiang et al. 2012). Thus, growing evidence supports impacts of virus–mediated insertional mutagenesis involving genomic instability and genetic disruption—particularly at cancer genes—as a common hallmark of the ~10% of cancers caused by DNA viruses.

Methods

Study population

Patients with newly diagnosed, HPV-positive OPSCC, presenting at Ohio State University Comprehensive Cancer Center from 2011 to 2016, provided written, informed consent for genomics studies and prospective collection of clinical data (Gillison et al. 2019). This study was approved by Institutional Review Boards at Ohio State University (OSU) and the University of Texas MD Anderson Cancer Center (MDACC). The overall study population included 105 patients with HPV-positive OPSCC, including 86 from OSU and 19 studied by The Cancer Genome Atlas project (TCGA, https://gdc.cancer.gov/). Clinical characteristics were reported previously (Gillison et al. 2019). A statistical power calculation indicated that at least 96 tumors would be required to provide 90% power to detect recurrent integration events within the same 1-Mbp window in 4% of tumors. An additional 53 de-identified, fresh-frozen OPSCCs were obtained from an MDACC head and neck cancer specimen bank, to increase genomic DNA sample numbers to a total of 158 HPV-positive OPSCC for detection of additional recurrent integration hotspots. All tumors were confirmed p16-positive by immunohistochemistry (p16, also known as INK4, is a protein isoform encoded by *CDKN2A*) and HPV-positive by quantitative PCR (Gillison et al. 2019).

Genomic DNA WGS analysis

OPSCC tumors were snap-frozen and microdissected to ensure >70% tumor content. Genomic DNA was isolated from paired tumors and matched normal blood leukocytes. Tumors' HPV status and genomic DNA sample quality were measured (Gillison et al. 2019). In the Ohio cohort tumors, 34 HPV-positive OPSCC T/N pairs were sequenced at ~90× mean depth of coverage by Complete Genomics (CGI) WGS (Gillison et al. 2019). The CGI aligner was used to map paired-end WGS reads against the human reference genome GRCh37 (hg19) (Carnevali et al. 2012). Illumina WGS data from 52 Ohio cohort HPV-positive OPSCC T/N pairs were generated at the New York Genome Center (NYGC), including 40× mean depth of coverage for normal samples and ~90× for tumors. Illumina WGS data for HPV-positive oropharyngeal cancers with at least 40× coverage (n = 19) were downloaded from TCGA. GATK v.3 was used to identify duplicate reads, realign reads surrounding indels, and recalculate alignment quality scores (McKenna et al. 2010). To estimate HPV copy numbers per sample, mean depth of sequencing coverage in the viral genome was divided by mean depth of autosomal coverage. HPV copy number was compared in tumors with and without HPV integrants by *t*-test. WGS data from the Ohio cohort (Gillison et al. 2019) have been deposited at the European Genome-phenome Archive (EGA; http://www.ebi.ac.uk/ega/), under accession numbers EGAS00001002393 and EGAS00001003228.

HPV insertional breakpoint detection

To detect virus–host breakpoints, raw Illumina sequence reads were aligned by BWA-MEM (v0.7.15) (Li and Durbin 2009) against a hybrid reference assembly combining human (hg19) and 15 high-risk HPV genomes (hereafter named hg19 + HPV). To harmonize breakpoint calls with detection of other variants (Gillison et al. 2019), we continued use of hg19 in this study. Breakpoint calls based on GRCh37 (hg19) and GRCh38 (hg38) reference sequences were similar, with the exception of one tumor in which several breakpoint sites (Chr X) were masked in hg38. The virus types and NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/) accession numbers are HPV16 [NC_001526.2]; HPV18 [NC_001357.1]; HPV31 [HQ537687.1]; HPV33 [HQ537707.1]; HPV35 [M74117.1]; HPV39 [M62849.1]; HPV45 [X74479.1]; HPV51 [M62877.1]; HPV52 [X74481.1]; HPV56 [X74483.1]; HPV58 [D90400.1]; HPV59 [X77858.1]; HPV66 [U31794.1]; HPV68 [DQ080079.1], and HPV69 [AB027020.1]. Discordant read pairs and split reads supporting the presence of virus–host breakpoints were extracted using three breakpoint callers, that is, Hydra (Quinlan et al. 2010), DELLY (Rausch et al. 2012), and SplazerS (Emde et al. 2012). HPV breakpoints supported by at least two independent discordant or split read pairs by each of at least two callers, or by four or more pairs by at least one caller, were analyzed further. Discordant pairs were extracted from CGI data using proprietary scripts (https://www.completegenomics.com/customer-support/tool-repository/) with hybrid genome assemblies. Virus–host breakpoints were called based on ≥5 supporting pairs.

Confirmation of HPV insertional breakpoints by HPV capture-seq

Custom Agilent SureSelect baits were designed to capture HPV (Warburton et al. 2018). Genomic DNA of selected tumors (i.e., GS18109, GS18041, GS18006) was hybridized and libraries were prepared following the manufacturer's directions. We generated 20 to 50 million read pairs per sample. HPV breakpoints were detected using published methods (Quinlan et al. 2010). In a separate approach, ~10% of breakpoints identified from WGS data were randomly selected for confirmation by Sanger sequencing. Custom PCR primers were designed to amplify across each breakpoint (Akagi et al. 2014). PCR products were sequenced using conventional Sanger sequencing. If unsuccessful, PCR products were cloned using the TOPO TA cloning kit (Invitrogen) and sequenced bidirectionally using M13 forward (–20) and M13 reverse primers.

Statistical analysis of demographic characteristics and survival outcomes

Details are available in the Methods section of the Supplemental Materials.

Hotspot detection

Identification of recurrent hotspots of HPV integration across tumors could be confounded by breakpoints clustered in individual tumors, which are not mutually independent. Therefore, breakpoint clusters within 500-kb genomic segments were identified in each tumor using mergeBed function of BEDTools (Quinlan and Hall 2010), and particular breakpoints supported by the highest read counts were identified as representative breakpoints. We identified 238 representative breakpoints in 105 OPSCCs studied by WGS, and 80 more in 53 OPSCCs studied by HPV capture-seq (totaling 318) (Supplemental Table S3.6). To identify expected breakpoints, we performed in silico simulations to calculate the probability of observing loci with recurrent integration.

We defined statistically significant genomic sites of recurrent HPV integration, that is, "hotspots," as those genomic sites that harbor integrants across multiple independent tumors at a frequency greater than expected by chance. We segmented the human genome into ~3000 independent, non-overlapping, 1-Mbp "tiles" or segments. We simulated 238 or 318 representative breakpoints randomly hitting these tiles 1 million times. The expected distribution of HPV breakpoints genome-wide and across the tumors would be random, allowing us to calculate empirical probabilities of the observed distribution. For example, observing at least two independent 1-Mbp tiles hit by insertional breakpoints across at least three independent tumors would not be expected by chance ($P = 0.0188$). An observation of at least five independent 1-Mbp genomic segments, each hit by insertional breakpoints across at least three independent tumors, would be even more significant (empirical probability, $P = 1 \times 10^{-6}$).

Details about a second, gene-centric approach to detect integration hotspots are available in the Methods section of the Supplemental Materials.

Gene Ontology

Details are available in the Methods section of the Supplemental Materials.

Detection of small variants, SVs, and CNVs

Illumina WGS reads were aligned against human reference genome hg19 using BWA.aln version 0.7.15 (Li and Durbin 2009). To detect SNVs and small indels, we used variant caller packages Mutect (Cibulskis et al. 2013), LoFreq (Wilm et al. 2012), and Strelka (Saunders et al. 2012). Variants supported by two or more callers were selected for further analysis (Gillison et al. 2019). For CGI WGS data, small variants were called using the CGI Cancer Genomics pipeline at default settings. CNVs were detected by comparing the depths of coverage in WGS data in T/N pairs using CNANorm with smoothening and segmentation (Gusnanto et al. 2012). Ratios of sequencing depth of coverage were calculated for 2-kb bins in each tumor versus its matched normal sample and were smoothed using the DNAcopy algorithm within CNANorm. To visualize sequence alignments, depth of coverage, discordant pairs, and breakpoints, we used Broad Institute's Integrative Genome Viewer (IGV) (Robinson et al. 2011). Structural variants were identified using three callers, that is, Crest (Wang et al. 2011), DELLY (Rausch et al. 2012), and BreakDancer (Chen et al. 2009). SVs identified by two or more callers were selected for further analysis.

Relationships between HPV breakpoints, CNVs, and SVs

To evaluate relationships between HPV breakpoints and CNVs, first we defined 100-kb bins across the human genome. Estimated copy numbers for each bin were determined by identifying the majority copy number segment representing the bin (Gusnanto et al. 2012). Bins with majority copy number $n \geq 2.5$ were defined as copy number amplifications, whereas bins with majority copy number $n \leq 1.5$ were defined as copy number losses. Hyperamplified bins were defined by majority copy number $n \geq 4$ or 5 depending on analysis.

Using CNANorm outputs of segments and ploidies, we identified copy number transition sites as boundaries between 100-kb segments with copy number changes of ≥0.5 n. Genomic copy numbers and copy number transitions were compared between bins with HPV breakpoints versus those without. Associations were evaluated using a one-tailed binomial test. *P*-values were adjusted by FDR multiple testing correction. To evaluate associations between HPV breakpoints and SVs (i.e., chromosomal translocations, deletions, inversions, rearrangements), we conducted a similar analysis comparing bin lengths of 100, 200, and 500 kb genome-wide. Statistical analysis (as described above) was performed using Bioconductor (http://www.bioconductor.org) packages.

Linked-read sequencing

Details are available in the Methods section of the Supplemental Materials.

RNA-seq libraries and analysis

Total RNA was isolated and RNA-seq was performed as described (Gillison et al. 2019). RNA-seq reads from HPV-positive OPSCCs were aligned against the human GRCh37 reference using STAR aligner version 2.4.2a (Dobin et al. 2013). RNA-seq reads from all TCGA samples also were similarly aligned. To determine expression levels of each transcript, transcript structures first were downloaded from GENCODE v.18 (http://www.gencodegenes.org/) as gene models. Aligned reads were counted using featureCounts (Liao et al. 2014). Raw counts were transformed into transcripts per million reads (TPM) using gene lengths defined by the union of all annotated exons. Genes with low expression values (maximum TPM across all samples <2) were excluded from further analysis. A pseudocount of one was added to all TPM values to avoid undefined data upon log transformation. Resulting $\log_2$ TPM values were normalized and batch-corrected using Bioconductor SVA function ComBat (Leek et al. 2012). To study expression of HPV polycistronic transcripts, we analyzed RNA-seq data with Salmon v. 1.3, a highly accurate transcript-level quantification tool (Patro et al. 2017), and transcript splicing models (Zheng and Baker 2006). RNA-seq data from the Ohio cohort (Gillison et al. 2019) have been deposited at the European Genome-phenome Archive (EGA; http://www.ebi.ac.uk/ega/), under accession number EGAS00001003237.

RNA expression analysis (Z-scores)

Gene-level expression data ($\log_2$-transformed TPM values) for 103 samples with available RNA-seq data were analyzed. For each annotated gene, *Z*-scores were calculated by standardizing the data based on the mean and standard deviation across the samples.

Fractions of genes with outlier expression (absolute value of Z-score ≥ 2) were compared in cases with versus without HPV breakpoints nearby (i.e., ±500 kb). The statistical significance of this enrichment in outlier gene expression in genes with HPV breakpoints was assessed using a one-tailed binomial test. P-values were adjusted by FDR multiple testing correction. We also calculated the sum of the square of Z-scores for all genes within HPV-linked rearrangements (HPV breakpoint clusters ±1-Mbp margins). Under the null hypothesis of no effect, the sum of the squared Z-scores would be expected to follow a chi-squared distribution with the number of degrees of freedom equal to the number of genes. Using this chi-squared distribution, these summary measurements were converted to P-values in order to compare regions containing different numbers of genes. We plotted the distribution of the sum of squared Z-scores using Q-Q plots.

Associations between HPV breakpoints, annotated host genes, and various genomic features

Details are available in the Methods section of the Supplemental Materials.

Chimeric HPV-host transcript analysis

RNA-seq reads were aligned against the hg19 + HPV reference assembly using TopHat-Fusion (Kim and Salzberg 2011). Salmon v1.3 was used to evaluate expression of HPV16 spliced transcripts (Zheng and Baker 2006). RNA-seq reads were aligned against custom virus–host chimeric models using GSNAP (Wu et al. 2016). At least one split read and ≥2 discordant pairs were required to identify chimeric HPV-host transcripts, which were visualized using the Sashimi plot function of IGV (Katz et al. 2015). To confirm structures of selected chimeric transcripts, de novo RNA-seq assembly was performed using Trinity (Grabherr et al. 2011).

Detection of HPV-mediated intragenic expression changes

Exon start and stop coordinates were downloaded from GENCODE v18. Sums of read base pair counts mapping to each exon were counted for each sample using SAMtools bedcov function (Li and Durbin 2009). Base pair coverage was normalized by RNA-seq library size. Batch correction was performed using SVA ComBat (Leek et al. 2012). Mean depths of coverage were calculated on an exon-by-exon basis for each sample, that is, for exons before and after HPV chimeric transcript junctions. Ratios calculated from after coverage versus before coverage read counts were $\log_2$-transformed for each sample, and Z-scores were calculated across all samples.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank the patients with oropharyngeal cancers at Ohio State University (OSU) who enrolled in our research study; members of the Gillison and Symer laboratories for insightful comments; Elisa Venturini, Karen Bunting, Benjamin Hubert, and Dayna M. Oschwald for project management at New York Genome Center; the Genomics Shared Resource at OSU Comprehensive Cancer Center (OSUCCC) for DNA and RNA quality assays; and Jordan Pietz (MD Anderson Cancer Center [MDACC]) for help with graphical figures. This study was funded by the Oral Cancer Foundation (M.L.G.); OSUCCC (M.L.G., D.E.S.); University of Texas MDACC (M.L.G., D.E.S.); Ohio Supercomputer Center (PAS0425, D.E.S.); Ohio Cancer Research Associate grant (GRT00024299, K.A.); Cancer Prevention Research Institute of Texas (CPRIT, RR170005, M.L.G.); and National Cancer Institute grant R50CA211533 (K.A.). Dr. Gillison is a CPRIT Scholar in Cancer Research. D.E.S. would like to dedicate this study to the memory of his father, Donald G. Symer, who died of COVID-19 during the preparation of this manuscript.

Author contributions: Conceptualization, M.L.G. and D.E.S.; methodology, M.L.G., D.E.S., K.A., H.M.G., K.R.C., A-K.E., B.S-C., M.Z., J.L., and N.R.; formal analysis, M.L.G., D.E.S., K.A., H.M.G., K.R.C., A-K.E., B.S-C., M.Z., Z.D., A.C., N.C.T., and N.R.; investigation, W.X., B.J., Y.S., and G.L.; resources, A.A., E.O., and A.K.E-N.; data curation, K.A. and J.B.S.; writing—original draft, M.L.G.; writing—review and editing, M.L.G. and D.E.S.; supervision, M.L.G. and D.E.S.; funding acquisition, M.L.G.

References

Adey A, Burton JN, Kitzman JO, Hiatt JB, Lewis AP, Martin BK, Qiu R, Lee C, Shendure J. 2013. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500:** 207–211. doi:10.1038/nature12064

Akagi K, Li J, Broutian TR, Padilla-Nash H, Xiao W, Jiang B, Rocco JW, Teknos TN, Kumar B, Wangsa D, et al. 2014. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res* **24:** 185–199. doi:10.1101/gr.164806.113

Bass AJ, Watanabe H, Mermel CH, Yu S, Perner S, Verhaak RG, Kim SY, Wardwell L, Tamayo P, Gat-Viks I, et al. 2009. *SOX2* is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nat Genet* **41:** 1238–1242. doi:10.1038/ng.465

Bernard E, Pons-Salort M, Favre M, Heard I, Delarocque-Astagneau E, Guillemot D, Thiébaut AC. 2013. Comparing human papillomavirus prevalences in women with normal cytology or invasive cervical cancer to rank genotypes according to their oncogenic potential: a meta-analysis of observational studies. *BMC Infect Dis* **13:** 373. doi:10.1186/1471-2334-13-373

Bodelon C, Untereiner ME, Machiela MJ, Vinokurova S, Wentzensen N. 2016. Genomic characterization of viral integration sites in HPV-related cancers. *Int J Cancer* **139:** 2001–2011. doi:10.1002/ijc.30243

Broutian TR, Jiang B, Li J, Akagi K, Gui S, Zhou Z, Xiao W, Symer DE, Gillison ML. 2020. Human papillomavirus insertions identify the PIM family of serine/threonine kinases as targetable driver genes in head and neck squamous cell carcinoma. *Cancer Lett* **476:** 23–33. doi:10.1016/j.canlet.2020.01.012

Bushman FD. 2020. Retroviral insertional mutagenesis in humans: evidence for four genetic mechanisms promoting expansion of cell clones. *Mol Ther* **28:** 352–356. doi:10.1016/j.ymthe.2019.12.009

The Cancer Genome Atlas Research Network. 2017. Integrated genomic and molecular characterization of cervical cancer. *Nature* **543:** 378–384. doi:10.1038/nature21386

Carnevali P, Baccash J, Halpern AL, Nazarenko I, Nilsen GB, Pant KP, Ebert JC, Brownley A, Morenzoni M, Karpinchyk V, et al. 2012. Computational techniques for human genome resequencing using mated gapped reads. *J Comput Biol* **19:** 279–292. doi:10.1089/cmb.2011.0201

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6:** 677–681. doi:10.1038/nmeth.1363

Chung CH, Guthrie VB, Masica DL, Tokheim C, Kang H, Richmon J, Agrawal N, Fakhry C, Quon H, Subramaniam RM, et al. 2015. Genomic alterations in head and neck squamous cell carcinoma determined by cancer gene-targeted sequencing. *Ann Oncol* **26:** 1216–1223. doi:10.1093/annonc/mdv109

Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31:** 213–219. doi:10.1038/nbt.2514

deCarvalho AC, Kim H, Poisson LM, Winn ME, Mueller C, Cherba D, Koeman J, Seth S, Protopopov A, Felicella M, et al. 2018. Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. *Nat Genet* **50:** 708–717. doi:10.1038/s41588-018-0105-0

de Martel C, Plummer M, Vignat J, Franceschi S. 2017. Worldwide burden of cancer attributable to HPV by site, country and HPV type. *Int J Cancer* **141:** 664–670. doi:10.1002/ijc.30716

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635

Dürst M, Croce CM, Gissmann L, Schwarz E, Huebner K. 1987. Papillomavirus sequences integrate near cellular oncogenes in some cervical carcinomas. *Proc Natl Acad Sci* **84:** 1070–1074. doi:10.1073/pnas.84.4.1070

Dyer N, Young L, Ott S. 2016. Artifacts in the data of Hu et al. *Nat Genet* **48:** 2–3. doi:10.1038/ng.3392

Emde AK, Schulz MH, Weese D, Sun R, Vingron M, Kalscheuer VM, Haas SA, Reinert K. 2012. Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS. *Bioinformatics* **28:** 619–627. doi:10.1093/bioinformatics/bts019

Fan G, Ye D, Zhu S, Xi J, Guo X, Qiao J, Wu Y, Jia W, Wang G, Fan G, et al. 2017. RTL1 promotes melanoma proliferation by regulating Wnt/β-catenin signalling. *Oncotarget* **8:** 106026–106037. doi:10.18632/oncotarget.22523

Ghaleb AM, Nandan MO, Chanchevalap S, Dalton WB, Hisamuddin IM, Yang VW. 2005. Krüppel-like factors 4 and 5: the yin and yang regulators of cellular proliferation. *Cell Res* **15:** 92–96. doi:10.1038/sj.cr.7290271

Gillespie KA, Mehta KP, Laimins LA, Moody CA. 2012. Human papillomaviruses recruit cellular DNA repair and homologous recombination factors to viral replication centers. *J Virol* **86:** 9520–9526. doi:10.1128/JVI.00247-12

Gillison ML, Akagi K, Xiao W, Jiang B, Pickard RKL, Li J, Swanson BJ, Agrawal AD, Zucker M, Stache-Crain B, et al. 2019. Human papillomavirus and the landscape of secondary genetic alterations in oral cancers. *Genome Res* **29:** 1–17. doi:10.1101/gr.241141.118

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29:** 644–652. doi:10.1038/nbt.1883

Gray E, Pett MR, Ward D, Winder DM, Stanley MA, Roberts I, Scarpini CG, Coleman N. 2010. *In vitro* progression of human papillomavirus 16 episome-associated cervical neoplasia displays fundamental similarities to integrant-associated carcinogenesis. *Cancer Res* **70:** 4081–4091. doi:10.1158/0008-5472.CAN-09-3335

Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, Berri S. 2012. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics* **28:** 40–47. doi:10.1093/bioinformatics/btr593

Hu Z, Zhu D, Wang W, Li W, Jia W, Zeng X, Ding W, Yu L, Wang X, Wang L, et al. 2015. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat Genet* **47:** 158–163. doi:10.1038/ng.3178

Hwang S, Nguyen AD, Jo Y, Engelking LJ, Brugarolas J, DeBose-Boyd RA. 2017. Hypoxia-inducible factor 1α activates insulin-induced gene 2 (Insig-2) transcription for degradation of 3-hydroxy-3-methylglutaryl (HMG)-CoA reductase in the liver. *J Biol Chem* **292:** 9382–9393. doi:10.1074/jbc.M117.788562

Jeon S, Allen-Hoffmann BL, Lambert PF. 1995. Integration of human papillomavirus type 16 into the human genome correlates with a selective growth advantage of cells. *J Virol* **69:** 2989–2997. doi:10.1128/jvi.69.5.2989-2997.1995

Jiang Z, Jhunjhunwala S, Liu J, Haverty PM, Kennemer MI, Guan Y, Lee W, Carnevali P, Stinson J, Johnson S, et al. 2012. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Res* **22:** 593–601. doi:10.1101/gr.133926.111

Katz Y, Wang ET, Silterra J, Schwartz S, Wong B, Thorvaldsdóttir H, Robinson JT, Mesirov JP, Airoldi EM, Burge CB. 2015. Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics* **31:** 2400–2402. doi:10.1093/bioinformatics/btv034

Kawakami K, Largaespada DA, Ivics Z. 2017. Transposons as tools for functional genomics in vertebrate models. *Trends Genet* **33:** 784–801. doi:10.1016/j.tig.2017.07.006

Kelley DZ, Flam EL, Izumchenko E, Danilova LV, Wulf HA, Guo T, Singman DA, Afsari B, Skaist AM, Considine M, et al. 2017. Integrated analysis of whole-genome ChIP-Seq and RNA-Seq data of primary head and neck tumor samples associates HPV integration sites with open chromatin marks. *Cancer Res* **77:** 6538–6550. doi:10.1158/0008-5472.CAN-17-0833

Kim D, Salzberg SL. 2011. TopHat-fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol* **12:** R72. doi:10.1186/gb-2011-12-8-r72

Koneva LA, Zhang Y, Virani S, Hall PB, McHugh JB, Chepeha DB, Wolf GT, Carey TE, Rozek LS, Sartor MA. 2018. HPV integration in HNSCC correlates with survival outcomes, immune response signatures, and candidate drivers. *Mol Cancer Res* **16:** 90–102. doi:10.1158/1541-7786.MCR-17-0153

Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. 2012. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28:** 882–883. doi:10.1093/bioinformatics/bts034

Leeman JE, Li Y, Bell A, Hussain SS, Majumdar R, Rong-Mullins X, Blecua P, Damerla R, Narang H, Ravindran PT, et al. 2019. Human papillomavirus 16 promotes microhomology-mediated end-joining. *Proc Natl Acad Sci* **116:** 21573–21579. doi:10.1073/pnas.1906120116

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760. doi:10.1093/bioinformatics/btp324

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30:** 923–930. doi:10.1093/bioinformatics/btt656

Loganathan SK, Schleicher K, Malik A, Quevedo R, Langille E, Teng K, Oh RH, Rathod B, Tsai R, Samavarchi-Tehrani P, et al. 2020. Rare driver mutations in head and neck squamous cell carcinomas converge on NOTCH signaling. *Science* **367:** 1264–1269. doi:10.1126/science.aax0902

McBride AA, Warburton A. 2017. The role of integration in oncogenic progression of HPV-associated cancers. *PLoS Pathog* **13:** e1006211. doi:10.1371/journal.ppat.1006211

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20:** 1297–1303. doi:10.1101/gr.107524.110

Mossahebi-Mohammadi M, Quan M, Zhang JS, Li X. 2020. FGF signaling pathway: a key regulator of stem cell pluripotency. *Front Cell Dev Biol* **8:** 79. doi:10.3389/fcell.2020.00079

Ojesina AI, Lichtenstein L, Freeman SS, Pedamallu CS, Imaz-Rosshandler I, Pugh TJ, Cherniack AD, Ambrogio L, Cibulskis K, Bertelsen B, et al. 2014. Landscape of genomic alterations in cervical carcinomas. *Nature* **506:** 371–375. doi:10.1038/nature12881

Parfenov M, Pedamallu CS, Gehlenborg N, Freeman SS, Danilova L, Bristow CA, Lee S, Hadjipanayis AG, Ivanova EV, Wilkerson MD, et al. 2014. Characterization of HPV and host genome interactions in primary head and neck cancers. *Proc Natl Acad Sci* **111:** 15544–15549. doi:10.1073/pnas.1416074111

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14:** 417–419. doi:10.1038/nmeth.4197

Peter M, Rosty C, Couturier J, Radvanyi F, Teshima H, Sastre-Garau X. 2006. *MYC* activation associated with the integration of HPV DNA at the *MYC* locus in genital tumors. *Oncogene* **25:** 5985–5993. doi:10.1038/sj.onc.1209625

Pinatti LM, Sinha HN, Brummel CV, Goudsmit CM, Geddes TJ, Wilson GD, Akervall JA, Brenner CJ, Walline HM, Carey TE. 2021. Association of human papillomavirus integration with better patient outcomes in oropharyngeal squamous cell carcinoma. *Head Neck* **43:** 544–557. doi:10.1002/hed.26501

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, Mell JC, Hall IM. 2010. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* **20:** 623–635. doi:10.1101/gr.102970.109

Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28:** i333–i339. doi:10.1093/bioinformatics/bts378

Reinson T, Toots M, Kadaja M, Pipitch R, Allik M, Ustav E, Ustav M. 2013. Engagement of the ATR-dependent DNA damage response at the human papillomavirus 18 replication centers during the initial amplification. *J Virol* **87:** 951–964. doi:10.1128/JVI.01943-12

Riordan JD, Keng VW, Tschida BR, Scheetz TE, Bell JB, Podetz-Pedersen KM, Moser CD, Copeland NG, Jenkins NA, Roberts LR, et al. 2013. Identification of *Rtl1*, a retrotransposon-derived imprinted gene, as a novel driver of hepatocarcinogenesis. *PLoS Genet* **9:** e1003441. doi:10.1371/journal.pgen.1003441

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29:** 24–26. doi:10.1038/nbt.1754

Romanczuk H, Howley PM. 1992. Disruption of either the E1 or the E2 regulatory gene of human papillomavirus type 16 increases viral immortalization capacity. *Proc Natl Acad Sci* **89:** 3159–3163. doi:10.1073/pnas.89.7.3159

Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. 2012. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28:** 1811–1817. doi:10.1093/bioinformatics/bts271

Senoo M, Pinto F, Crum CP, McKeon F. 2007. p63 is essential for the proliferative potential of stem cells in stratified epithelia. *Cell* **129:** 523–536. doi:10.1016/j.cell.2007.02.045

Sharma S, Munger K. 2020. The role of long noncoding RNAs in human papillomavirus-associated pathogenesis. *Pathogens* **9:** 289. doi:10.3390/pathogens9040289

Shen C, Liu Y, Shi S, Zhang R, Zhang T, Xu Q, Zhu P, Chen X, Lu F. 2017. Long-distance interaction of the integrated HPV fragment with MYC gene and 8q24.22 region upregulating the allele-specific MYC expression in HeLa cells. *Int J Cancer* **141:** 540–548. doi:10.1002/ijc.30763

Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. 2018. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18:** 696–705. doi:10.1038/s41568-018-0060-1

Starrett GJ, Thakuria M, Chen T, Marcelus C, Cheng J, Nomburg J, Thorner AR, Slevin MK, Powers W, Burns RT, et al. 2020. Clinical and molecular characterization of virus-positive and virus-negative Merkel cell carcinoma. *Genome Med* **12:** 30. doi:10.1186/s13073-020-00727-4

Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, Boeke JD. 2002. Human L1 retrotransposition is associated with genetic instability in vivo. *Cell* **110:** 327–338. doi:10.1016/S0092-8674(02)00839-5

Tota JE, Best AF, Zumsteg ZS, Gillison ML, Rosenberg PS, Chaturvedi AK. 2019. Evolution of the oropharynx cancer epidemic in the United States: moderation of increasing incidence in younger individuals and shift in the burden to older individuals. *J Clin Oncol* **37:** 1538–1546. doi:10.1200/JCO.19.00370

Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, Chen K, Harris CC, Ding L, et al. 2011. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods* **8:** 652–654. doi:10.1038/nmeth.1628

Warburton A, Redmond CJ, Dooley KE, Fu H, Gillison ML, Akagi K, Symer DE, Aladjem MI, McBride AA. 2018. HPV integration hijacks and multimerizes a cellular enhancer to generate a viral-cellular super-enhancer that drives high viral oncogene expression. *PLoS Genet* **14:** e1007179. doi:10.1371/journal.pgen.1007179

Wheelan SJ, Aizawa Y, Han JS, Boeke JD. 2005. Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. *Genome Res* **15:** 1073–1078. doi:10.1101/gr.3688905

Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* **40:** 11189–11201. doi:10.1093/nar/gks918

Wu L, Sun T, Kobayashi K, Gao P, Griffin JD. 2002. Identification of a family of mastermind-like transcriptional coactivators for mammalian notch receptors. *Mol Cell Biol* **22:** 7688–7700. doi:10.1128/MCB.22.21.7688-7700.2002

Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. 2016. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods Mol Biol* **1418:** 283–334. doi:10.1007/978-1-4939-3578-9_15

Yu L, Majerciak V, Xue XY, Uberoi A, Lobanov A, Chen X, Cam M, Hughes SH, Lambert PF, Zheng ZM. 2021. Mouse papillomavirus type 1 (MmuPV1) DNA is frequently integrated in benign tumors by microhomology-mediated end-joining. *PLoS Pathog* **17:** e1009812. doi:10.1371/journal.ppat.1009812

Yuan H, Krawczyk E, Blancato J, Albanese C, Zhou D, Wang N, Paul S, Alkhilaiwi F, Palechor-Ceron N, Dakic A, et al. 2017. HPV positive neuroendocrine cervical cancer cells are dependent on Myc but not E6/E7 viral oncogenes. *Sci Rep* **7:** 45617. doi:10.1038/srep45617

Zapatka M, Borozan I, Brewer DS, Iskar M, Grundhoff A, Alawi M, Desai N, Sultmann H, Moch H, Pathogens P, et al. 2020. The landscape of viral associations in human cancers. *Nat Genet* **52:** 320–330. doi:10.1038/s41588-019-0558-9

Zhang X, Choi PS, Francis JM, Imielinski M, Watanabe H, Cherniack AD, Meyerson M. 2016. Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat Genet* **48:** 176–182. doi:10.1038/ng.3470

Zhang Y, Dakic A, Chen R, Dai Y, Schlegel R, Liu X. 2017. Direct HPV E6/Myc interactions induce histone modifications, Pol II phosphorylation, and hTERT promoter activation. *Oncotarget* **8:** 96323–96339. doi:10.18632/oncotarget.22036

Zheng ZM, Baker CC. 2006. Papillomavirus genome structure, expression, and post-transcriptional regulation. *Front Biosci* **11:** 2286–2302. doi:10.2741/1971

Zindy F, Eischen CM, Randle DH, Kamijo T, Cleveland JL, Sherr CJ, Roussel MF. 1998. Myc signaling via the ARF tumor suppressor regulates p53-dependent apoptosis and immortalization. *Genes Dev* **12:** 2424–2433. doi:10.1101/gad.12.15.2424

Received June 20, 2021; accepted in revised form November 10, 2021.