Intratumoral heterogeneity and clonal evolution induced by HPV integration

Keiko Akagi[1,*], David E. Symer[2,*], Medhat Mahmoud[3], Bo Jiang[1], Sarah Goodwin[4], Darawalee Wangsa[5], Zhengke Li[1,#], Weihong Xiao[1], Joe Dan Dunn[1], Richard McCombie[4], Thomas Ried[5], Kevin R. Coombes[6,+], Fritz J. Sedlazeck[3,7] , and Maura L. Gillison[1,**]

Affiliations: [1]Department of Thoracic / Head and Neck Medical Oncology, MD Anderson Cancer Center, Houston, TX. [2]Department of Lymphoma & Myeloma, MD Anderson Cancer Center, Houston, TX. [3]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX. [4]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY. [5]National Cancer Institute, Bethesda, MD. [6]The Ohio State University, Columbus, OH. [7]Department of Computer Science, Rice University, Houston, TX.
* co-first authors

present address: [#] Boundless Bio, La Jolla, CA. [+]University of Georgia, Athens, GA.

Running title: HPV integration drives intratumoral heterogeneity

Keywords: human papillomavirus, head and neck squamous cell carcinoma, long-read sequencing, genomic structural variation, circacatena, integration, oropharynx

Financial support: Supported by CPRIT (MLG), the Oral Cancer Foundation (MLG), an R50 award from the National Cancer Institute (NIH) (KA), University of Texas MD Anderson Cancer Center (DES, MLG). Dr. Maura L. Gillison is a CPRIT Scholar in Cancer Research.

Corresponding author:
Maura L. Gillison, MD, PhD
Department of Thoracic Head and Neck Medical Oncology
Division of Cancer Medicine
MD Anderson Cancer Center
1515 Holcombe Blvd, Houston, TX 77030
Email: mgillison@mdanderson.org

Conflict of interest statement: F.S. receives research support from Illumina, PacBio and Oxford Nanopore. **M.L.G.** serves as consulting, advisory board roles for LLX Solutions, LLC (Pending), Sensei, Mirati Therapeutics, BioNTech AG, Shattuck Labs Inc., EMD Serono Inc., Debiopharm, Kura Oncology, Merck Co., Ipsen Biopharmaceuticals Inc., Bristol-Myers Squibb, Bicara Therapeutics, Bayer HealthCare Pharmaceuticals, Roche, Roche Diagnostics GmbH, Genocea Biosciences, Inc., NewLink Genetics Corporation, Aspyrian Therapeutics, TRM Oncology, Amgen Inc., AstraZeneca Pharmaceuticals, and Celgene Corp.; and research funding from Genocea, BMS, Kura, Cullinan, Genentech, BioNtech, and Gilead.

Author contributions: Conceptualization, DES, MLG; Methodology, KA, DES, MM, BJ, SG, DW, ZL, WX, RM, TR, FJS; Formal Analysis, KA, DES, MM, KRC, FJS, MLG; Investigation, BJ, SG, DW, ZL, WX; Resources, MLG; Data Curation, KA, MM; Writing-Original Draft, MLG; Writing-Review & Editing, KA, DES, JDD, MLG; Supervision, MLG; Funding Acquisition, MLG.

ABSTRACT

The human papillomavirus (HPV) genome is integrated into host DNA in most HPV-positive cancers, but the consequences for chromosomal integrity are unknown. Continuous long-read sequencing of oropharyngeal cancers and cancer cell lines revealed interrelated but heterogeneous genomic structures comprising repetitive patterns of concatemerized virus and host segments. Evidence of this novel form of structural variation – termed “circacatena” here -- was detected in extrachromosomal and intrachromosomal DNA and at chromosomal rearrangements. Unique breakpoint sequences shared across structurally diverse virus-host concatemers facilitated stepwise reconstruction of their evolution from common molecular ancestors. This analysis revealed that concatemerized HPV genomes are unstable and, upon insertion into and excision from chromosomes, facilitate capture, rearrangement, and rolling-circle amplification of host DNA. The data indicate that circacatena is driven by the dynamic, aberrant replication and recombination of an oncogenic DNA virus, thereby extending known consequences of HPV integration to include promotion of intratumoral heterogeneity and clonal evolution.

INTRODUCTION

Human papillomavirus (HPV) causes more than 630,000 cancers worldwide each year, including anogenital and oropharyngeal squamous cell carcinomas (1). The viral genome is maintained early in infection as an ~8-kilobase-pair (kb) extrachromosomal circular DNA (ecDNA) episome. In a majority of subsequent cancers, the viral genome has integrated into the host genome (2-5). HPV integration promotes tumorigenesis by increasing expression and stability of transcripts encoding the E6 and E7 oncoproteins (6), which target tumor suppressors p53 and pRb for degradation, respectively (7,8). Recent whole genome sequencing (WGS) analyses of cervical and oropharyngeal cancers revealed that HPV integrants are enriched in genomic regions with structural variants (SVs) and copy number variants (CNVs) (2,4,5,9). Diverse genetic consequences of HPV integration have been identified, including dysregulated host gene expression near integrants (2-5).

A greater understanding of the mechanisms by which HPV integration leads to SVs, CNVs, and aberrant host genome expression depends upon improved resolution of variant structures flanking the integration sites. WGS short reads are insufficient to resolve these complex rearrangements (10,11) and cannot adequately distinguish HPV integrants present in ecDNA from intrachromosomal insertions. To resolve the structures of genomic rearrangements flanking HPV integration sites, we conducted continuous long-read sequencing (LR-seq) of genomic DNA from primary HPV-positive oropharyngeal cancers and human cell lines using the PacBio HiFi and Oxford Nanopore Technologies (ONT) platforms. In virtually all cases, we detected a unique form of structural variation, which we named “circacatena” (from the Latin for “circular chain”). Circacatena is characterized by highly diverse and yet interrelated variant structures comprised of repetitive patterns of concatemerized virus and host DNA segments. Evidence of circacatena was detected in ecDNA, chromosomes, or both in individual tumors. This analysis extends the host genomic consequences of HPV integration to include promotion of intratumoral heterogeneity and clonal evolution.

RESULTS

Our analysis of 105 HPV-positive oropharyngeal cancers by WGS identified HPV-host insertional breakpoints that directly flank, bridge, or map within host genomic regions enriched with SVs and CNVs (5). To resolve genomic rearrangements at sites of HPV integration, we selected five HPV-positive primary oropharyngeal cancers and four HPV-positive human cell lines, 93-VU-147T (hereafter VU147) (12), HeLa (13), GUMC-395 (14), and HTEC (15), for WGS using Illumina paired-end sequencing and LR-seq using the PacBio and ONT platforms. Details about distributions of read lengths and depths of sequencing coverage are provided in Supplemental Information (**Figs. S1.1-2, Tables S1.1-2**). WGS depth of sequencing coverage for autosomes across the samples ranged from 50.9 to 115.5x, ONT depth of coverage ranged from 15 to 31.3x, and PacBio depth of coverage ranged from 7.1 to 15.8x.

Continuous long-read sequencing reveals unstable HPV genome concatemers

HPV is thought to be maintained after initial infection as an ~8-kb ecDNA episome in cell nuclei and to persist in a fraction of subsequent cancers as ecDNA. Therefore, we evaluated the technical ability to capture small circular DNA molecules in LR-seq libraries and to identify them in resulting reads. We used the ~16.5-kb circular mitochondrial (mt)DNA genome as a proxy. Histograms of ONT mtDNA read lengths displayed frequency peaks at 16.5 kb (**Fig. S1.3, Table S1.1**). Plots of the distance between the coordinates of the 5’ and 3’ ends of each ONT read mapped against the reference mtDNA genome also revealed a high frequency of reads with ends mapping within 100 base pairs (bp) of each other, i.e., at the unit length of 16.5 kb. These findings revealed the predominance of single-unit circular mtDNA genomes. In one primary cancer, two-unit mtDNA genomes were also abundant (**Fig. S1.3)**. Read lengths exceeding the single mtDNA unit length of 16.5 kb were identified in almost all cancers and cell lines. Overall,

this mtDNA analysis confirms the ability of LR-seq to detect ecDNA and determine its length and sequences accurately.

Comparable analysis of ONT reads mapped to the HPV reference genome revealed read lengths frequently exceeding the single-unit virus length ~8 kb (**Fig. 1A-D** *top*, **Table S1.2, Fig. S1.4**). Plots of the distance between the 5' and 3' ends of mapped reads, as modeled in **Fig. 1E**, indicated the predominance of single-unit HPV episomes in one cancer (**Fig. 1A,** *bottom,* Tumor 1) and the presence of HPV concatemers in the others (**Fig. 1B-D**, *bottom*). The HPV genome was frequently observed in multi-unit, head-to-tail virus-virus concatemers among the primary cancers and a representative cell line (**Fig. 1F**), confirming recent reports (16,17). Variable counts of HPV units per concatemer were detected across the cancers. In contrast to mtDNA, ONT reads of HPV genomes deviated more frequently from the expected unit lengths of multiples of 8 kb (**Fig. 1A-D**, *bottom*), indicating rearrangements within the viral genome had occurred.

Our WGS, PacBio, and ONT data independently confirm unique rearrangements of the HPV genome in individual cancers, strongly arguing against technical artifacts (**Tables S2-S6**). These rearrangements include deletions, tandem duplications, and inversions, as revealed by alignment of ONT reads against template models visualized in dotplots, as well as WGS reads aligned against the reference HPV genome visualized in Circos plots (**Fig. 1G**; **Figs. S1.5-S1.6**). This extensive structural variation suggests that the virus concatemers are genetically unstable.

Identification of circacatena, a unique form of structural variation

Among all primary cancers and cell lines studied here, a recurrent pattern was observed in the structural rearrangements flanking HPV integration sites. In one example, Tumor 4 was estimated by WGS to have ~86 HPV16 genome copies per cell. We identified 14 HPV-host breakpoints, five host-host breakpoints, and three virus-virus breakpoints flanking regions of CNVs and SVs on Chrs. 5p13, 5q14, and Xp22, including two chromosomal translocations, t(5;X)(p13;p22) and t(5;X)(q14;p22) (**Fig. 2A, Table S2**). All 22 breakpoints were confirmed by WGS, PacBio, and/or ONT data at single nucleotide resolution and were assigned a unique numerical identifier (**Table S2**). To facilitate resolution of structures based on ONT reads, the 10 breakpoints that directly flanked CNVs were selected as segment-defining breakpoints and were used to delineate DNA segments in the reference human genome (**Table S2**).

In this tumor, virus-virus concatemers comprising up to 6 tandem unit-length HPV16 genomes were detected. HPV nucleotides 776 - 5144 were deleted intermittently from single units of the viral genome, forming a unique virus-virus breakpoint (designated as #20, **Fig. 2B, Table S2**). ONT sequence reads (N = 178, filtered for reads ≥20 kb) also revealed rearranged virus-host structures in which distinct segments of Chr. X (e.g., XB, XD) and/or Chr. 5 (e.g., 5D, 5E, 5G) were inserted precisely where viral genome segments were deleted (**Fig. 2B,** rearrangement groups A3-10). Individual molecules were detected in which specific patterns of virus and/or host DNA segments and breakpoints were repeated in series (**Fig. 2B,** e.g., groups A6, A9)**.** Such patterns, shared among individual molecules sequenced in the same tumor, also diverged in pattern and/or structure, revealing extensive intermolecular heterogeneity. Additionally, distinct patterns of breakpoints and segments detected in different molecules were occasionally linked together in others (**Fig. 2C**).

We used the unique breakpoint sequences shared across heterogeneous structures as molecular barcodes to reconstruct the evolution of related heterogeneous structures from a common molecular ancestor. According to the resulting model, insertion of concatemerized HPV genomes initially occurred at the site of deletion of host DNA segment XC on Chr. X (**Fig. 2D**). This model also supports capture of host DNA by concatemerized HPV genomes via excision of ecDNA from Chr. X, followed by insertion into Chr. 5p and 5q (**Fig. 2D**). The shared virus and host DNA segments and breakpoints are linked in series in recurrent patterns, consistent with

intermittent amplification by rolling-circle replication. Providing additional support for this mechanism of amplification (18,19), we detected no insertions/ deletions (indels) or single nucleotide variants in 170 PacBio reads at breakpoints 14 and 19, which connect host DNA segment XD to HPV16 (**Fig. S2.1**).

In sum, these LR-seq data reveal a unique form of structural variation at HPV integration sites in this cancer, characterized by highly diverse, yet interrelated, variant structures comprised of repetitive patterns of virus and host DNA segments (**Fig. 2D**). We named this novel form of structural variation "circacatena" (from the Latin for "circular chain"). We detected evidence of circacatena in all primary HPV-positive cancers and cell lines evaluated.

In the same Tumor 4, the counts of reads that support integration of rearranged virus-host chains into flanking chromosomal DNA were very low. Therefore, we infer that the numerous virus-host concatemers observed in this tumor mostly occur in the form of ecDNA. Compared to our model derived from LR-seq reads, the structural predictions for ecDNA made by AmpliconArchitect software, commonly used to predict such molecular structures (20), are oversimplified and inaccurate, likely due to the inherent limitations of short-read WGS data (**Fig. S2.2**).

Circular DNAs contribute to circacatena development

In another tumor (Tumor 2), harboring ~83 HPV16 genome copies per cell, HPV integrants disrupted *EP300* on Chr. 22q13.2. Our previous analysis detected frequent inactivation of this gene by somatic mutations among HPV-positive oropharyngeal cancers (21). A total of 23 breakpoints, including 14 virus-host, four host-host, and five virus-virus were detected, and of these, 14 were chosen as segment-defining breakpoints across 60 kb of the *EP300* locus (**Fig. 3A, Table S3.1**). ONT sequence reads (N = 154; **Figs. S3.1-2**) supported structures comprised of multiple tandem full-length HPV genome units interspersed with intermittent units lacking nucleotides 2312 to 7065.This deletion is annotated in this tumor as breakpoint 20 (**Fig. 3B**). Tandem virus concatemers containing breakpoint 20 were detected in series with host DNA segments of *EP300* (**Fig. 3B, segments B to M**). As in Tumor 4 (**Fig. 2D**), the observed heterogeneity together with shared breakpoints seen in the rearranged genomic structures are characteristic of circacatena, as they suggest evolution from a clonal ancestor by serial events. These events involve insertion of concatemerized HPV genomes, ecDNA excision, rolling-circle replication, and additional serial deletions (**Fig. 3B**). We identified no LR-seq and no WGS reads that support integration of virus-host structures into flanking host sequences at Chr. 22q13.2. This result strongly suggests that these virus-host concatemers occur predominantly or exclusively as ecDNA forms in this primary cancer.

Interestingly, in the same cancer, virus-host concatemers were integrated into flanking host sequences on Chr. 4p19.1, as shown by ONT reads (**Fig. S3.3**). Our detection of virus-virus recombination breakpoint 20 integrated into Chr. 4, identical to the breakpoint 20 integrated into the Chr. 22 *EP300* locus, indicated that the virus-host concatemers at these two distinct sites are clonally related, i.e., that they arose from a common molecular ancestor. This example also demonstrated that concatemers can coexist as ecDNA and as intrachromosomal integrants within the same tumor.

In several of the primary cancers and cell lines, virus-host concatemers mapped near or included the cancer driver gene *MYC* on Chr. 8q24.21, an established hotspot for HPV integration in oropharyngeal (5) and cervical cancers (22). In Tumor 5, we identified six breakpoints, three virus-host and three host-host, which were selected to delineate host segments A through J at *MYC* (**Fig. 3C, Table S3.2**). While HPV concatemers were not detected, a deletion in the viral genome at nucleotides 1803-2170 was identified. We detected 110 ONT reads (each ≥20 kb) defining SVs at this locus. Of them, 97 (88%) supported a repeated genomic structure involving the *MYC* locus (segment E**, Fig. 3D**) occurring at least twice in tandem. Less common but related SVs also were derived from this ancestral molecule

by recombination events including deletions in host and/or virus sequences (**Fig. 3D**). Since no reads supported integration of virus-host concatemers into adjacent chromosomal DNA, they likely existed predominantly or exclusively in ecDNA form. As this tumor harbored ~20 HPV16 genome copies per cell, each cell may have contained, for example, nine ecDNA molecules of two linked units or perhaps one ecDNA molecule of 18 linked units (**Fig. 3E**).

In contrast to primary cancers, LR-seq data from all cell lines indicated that HPV-host concatemers were integrated into flanking chromosomal DNA. For example, in HeLa cervical cancer cells, four virus-host breakpoints and one host-host breakpoint flanked or mapped upstream of *MYC* within a 10.9-kb region on Chr. 8q24.21 with CNVs (3-35n) and SVs that involved *CASC8* (**Fig. 4A, Table S4.1**). These breakpoints defined HeLa genomic DNA segments A to F. Numerous ONT reads showed virus-host concatemers directly connected with upstream and downstream chromosomal genomic segments (**Fig. 4B**). These integrated virus-host concatemers showed considerably less structural heterogeneity than those detected in ecDNA in Tumors 4 and 5. These data corroborate identification of intrachromosomal HPV integrants by WGS, haplotype-resolved sequencing (23,24), and spectral karyotyping and fluorescence in situ hybridization (FISH) (25). These complex SVs and intermolecular heterogeneity were not resolved or predicted accurately by AmpliconArchitect from short-read WGS (**Fig. S4.1)** (20).

A chromosomal translocation in HeLa at t(8;22)(q24;q13) was initially identified by spectral karyotyping (25) but was not detected using WGS or haplotype-resolved data (23,24). Its relationship with HPV integration, if any, has not been described. Our LR-seq data uniquely confirmed and resolved this translocation (**Fig. 4C**). We identified virus-host concatemers with breakpoints identical to those integrated in Chr. 8 but that connected the 5' end of HeLa genomic segment C with a 2-kb segment of repeated telomeric sequence (i.e., 5'-TTAGGG) on Chr. 22, forming breakpoint 2 (**Fig. 4C**). We noted that the repetitive patterns of virus and host DNA segments in Chr.22 were slightly different from those integrated into Chr. 8 (**Fig. 4B**). Consistent with ONT data, HPV18 FISH probes hybridized to two of three copies of Chr. 8, an (8:22)(q24;q13) translocation, and a complex der(5)t(5;22;8)(q11;q11q13:q24) rearrangement (**Fig. 4D**), as observed previously (25). Our WGS data indicated 5 copies of Chr. 8q extended from the HPV integration site to a telomere. Thus, we infer that virus-host concatemers first integrated into Chr. 8, followed by Chr. 8 duplication, translocation to the telomere of Chr.22, and translocation from the centromere of Chr. 22 to the centromere of Chr. 5 (**Fig. 4E**).

To address the possible presence of HPV-containing ecDNA in HeLa, we prepared and analyzed Circle-seq libraries from sequential cell passages. Indeed, ecDNA could be recovered from HeLa cells (**Fig. S4.2**). Virus-host and host-host breakpoints in ecDNA were identical to those in LR-seq and WGS reads, indicating ecDNA was excised from the same region of copy number amplification on Chr. 8q24.21 upstream of *MYC*. However, the Circle-seq reads were identified at very low counts (**Fig. S4.2**), an observation likely explained by infrequent excision from Chrs. 8, 22, and/or 5 by recombination in rare subclones (**Fig. 4F**).

We confirmed our technical ability to detect HPV-containing ecDNA with fluorescence in situ hybridization (FISH) microscopy, by transfecting 293T cells with supercoiled plasmid DNA containing full-length HPV sequences followed by FISH using virus-specific probe (**Fig. S4.3**). We then identified rare HPV FISH probe-positive candidate ecDNAs in a small fraction of HeLa metaphase spreads, consistent with our Circle-seq findings in HeLa (**Figs. S4.3-4, Table S4.2**).

The GUMC-395 cell line was derived from a liver metastasis of a treatment-refractory, rapidly fatal, large cell, cervical neuroendocrine carcinoma (14). Based on WGS data, HPV16 copy number in these cells was estimated as ~56 copies per cell. We detected 13 breakpoints, including five virus-host and seven host-host, clustered within an ~200-kb region at the *MYC* locus that contained extreme hyper-amplification (up to 203n) and structural rearrangements (**Fig. 5A, Table S5**). Nine of the breakpoints delineated sequential host DNA segments A - L. Segments B and C encompass the *MYC* gene.

WGS and LR-seq data from two cell lines studied, HeLa and HTEC, confirmed the presence of a normal allele without structural variation or viral insertions at this genomic site, as expected (**Tables S4.1, S6**). In contrast, no WGS, PacBio, or ONT reads supported a normal allele connecting host DNA segments E to F on Chr. 8 in GUMC-395 cells (**Fig. 5, Table S5**). This result indicated loss of heterozygosity (LOH) of the normal allele at this locus. However, insertion of a virus concatemer was detected at this locus (**Fig. 5A; Fig. S5**), defining breakpoints 6 and 7. Nearly all virus-host concatemers seen in ONT data (N = 774, ≥20kb) contained breakpoint 7, thus nominating this insertion as an early event in this cancer. Moreover, structural variants all shared the same deletion of host segments D and E and the V-F-B-C structural variant containing the *MYC* gene (**Fig. S5**), again consistent with evolution from a common molecular ancestor. These observations provide strong support for the evolutionary model shown in **Fig. 5B**. In this model, ecDNAs that were generated from concatemerized HPV genomes integrated at the *MYC* locus underwent subsequent rolling-circle replication and recombination. These HPV-host concatemers would continue to evolve with secondary recombination and deletion events that give rise to all subsequent variants detected (**Fig. 5B**). This model explains the recurrent patterns of virus and host DNA segments and breakpoints shared among heterogeneous variant structures as detected in LR-seq data. These diverse but related patterns are characteristic of circacatena (**Fig. 5C).** The model also explains step changes in CNVs observed in WGS data at several segment junctions, including F to G, H to I, J to K, and K to L (**Fig. 5A**).

In FISH analysis of GUMC-395 metaphase spreads, HPV16 probes localized to two copies of Chr. 8q and two copies of Chr. 21 in all cells examined, due to a t(8;21)(q24.21:q11.2) translocation into its centromere (**Fig. S5**). Consistent with this observation, LR-seq data showed virus-host concatemers were integrated adjacent to host segment E on Chr. 8q.24.21 (**Fig. S5;** group D1) and into a second site joining host segment E to the centromeres of Chr. 8 and Chr. 21. We infer that these concatemers (likely as ecDNA) were inserted by homologous recombination at the *MYC* locus, followed by Chr. 8 duplication, intrachromosomal Chr. 8q inversion, t(8;21)(q24.21:q11.2) translocation, and duplication of this translocation (**Fig. 5D**). Our analysis of HeLa and GUMC-395 cells revealed that integrated virus-host concatemers are unstable and thus can induce chromosomal translocations and other forms of genomic structural variation.

We used Circle-seq to identify HPV-positive ecDNA in serial passages of GUMC-395. The resulting reads consistently identified ecDNA from the *MYC* locus (**Fig. 5A, Fig. S6.1**). To enhance visualization of ecDNA, HPV16 FISH was repeated with sensitive YOYO-1 DNA staining. In 85 metaphase spreads examined, the HPV16 probes localized consistently to Chr. 8 and to the t(8;21) translocation. The HPV16 probe also localized to ecDNA as visualized by YOYO-1 dye in 22% (19 of 85) of metaphase spreads (**Table S4.2, Figs. S4.4, S5**). This result indicates that virus-host concatemers were stably integrated into Chr. 8 and 21 but in subclones were either maintained as ecDNA or generated by excision from chromosomes. Comparable results were observed in VU147, HeLa, and HTEC cell lines (**Table S4.2, Fig. S4.3**). We conclude that HPV integration was responsible for hyper-amplification of *MYC* in GUMC-395 and thereby promoted growth of that lethal cancer.

The human tonsillar epithelial cell line (HTEC) was created *in vitro* upon transfection of primary cells with HPV16 episomal DNA, followed by clonal selection (15). Virus integration and formation of HPV-host concatemers occurred during cell culture *in vitro*. We observed striking similarities between HPV integration sites and genomic rearrangements at the *MYC* locus in HTEC and those in both HeLa and GUMC-395 cells. In HTEC, two virus-host breakpoints flanked the 5' ends of two amplified genomic loci (i.e.,16-19n), ~350 kb upstream of *MYC* (**Fig. S6.2**), which is analogous to GUMC-395. ONT reads demonstrated integrated virus-host concatemers that displayed homology to host DNA segments captured in the concatemer, supporting a mechanism of insertion induced by homologous recombination, comparable to

HeLa (**Fig. 4F, Fig. S6.2**). In serially passaged HTEC cells, Circle-seq reads that aligned at low counts to this locus showed structural variation and additional discordant rearrangements, suggesting instability of intrachromosomal insertions and ecDNA occasionally generated from this site (**Fig. S6.1**). Similar findings were observed in VU147 and GUMC-395 (**Fig. S6.1**). In HTEC, HPV16 FISH probes localized to Chr. 8q and to both ends of an isochromosome i(8q) in all metaphase spreads examined, indicating viral integration preceded formation of a chromosomal abnormality as the epithelial cells evolved *in vitro* (**Fig. S6.2**).

DISCUSSION

Here we define circacatena as a unique form of genomic structural variation induced by HPV integration in human cancers. Circacatena is characterized by diverse and yet interrelated variant structures harboring repetitive patterns of concatemerized virus and host DNA segments. Our evolutionary models, supported by LR-seq data, explain circacatena as the result of aberrant host DNA replication and recombination induced by HPV integration. We infer that the virus-virus and virus-host genomic structural rearrangements characteristic of circacatena are genetically unstable, leading to further intratumoral heterogeneity and clonal evolution. Evidence of circacatena is present in HPV-containing ecDNA, intrachromosomal integrants, or both, in all cancers evaluated.

Our previous WGS analyses of cell lines (2) and primary tumors (5) prompted us to develop a mechanistic looping model explaining the extensive host genomic structural variation we observed at HPV integration sites (2). This HPV looping model proposed that double-strand DNA breaks in viral and host genomes lead to viral capture of host DNA, forming insertional breakpoints, followed by rolling-circle replication, recombination, and repair, resulting in intrachromosomal virus-host concatemers (2). Here, LR-seq data provide considerable additional support for this original looping model. FISH analysis of cell lines demonstrated HPV integration in chromosomal DNA in every cell examined (2). However, very few, if any, LR-seq reads that support integration of virus-host concatemers into chromosomes were identified among several primary cancers, suggesting they harbor HPV-containing ecDNA predominantly or exclusively (**Figs. 3C and D**). This discrepancy between cultured cells and primary cancers may be attributable to differences in numbers of ecDNAs maintained in different cellular contexts. For example, essential factors required to replicate and maintain HPV ecDNA may be downregulated or lost with derivation and growth of cell lines *in vitro*. Alternatively, subclones of primary cancers harboring intrachromosomal HPV integrants may have growth advantages during development of cell lines.

New insights gained from LR-seq and Circle-seq prompted us to expand upon our original looping model (2) to include generation and insertion of unstable concatemerized HPV genomes into chromosomes; capture and rearrangement of host DNA during excision of HPV ecDNAs from chromosomes and their insertion into chromosomes; ecDNA amplification by rolling-circle replication; recombination, likely by homology-directed repair, formation of chromosomal inversions and translocations between concatemers, telomeres and centromeres; and development of circacatena (**Fig. 6**).

Our model of the molecular evolution of HPV integrants, supported by WGS and LR-seq data, indicates that circacatena results from ongoing, aberrant DNA replication and recombination facilitated by HPV integration. HPV has two predominant modes of viral replication that depend upon the differentiated state of the infected cell (26-28). Maintenance replication in the basal epithelium occurs in S phase by bidirectional theta replication initiated from the viral origin. This mode is dependent upon the viral E1 helicase and the E2 transcriptional regulatory proteins. By contrast, productive viral replication by rolling-circle replication occurs in the G2/M phase, is less dependent on the viral origin, and is unidirectional (26,27,29). This mode of viral replication is dependent upon E7- or E1-induced activation of the

ATM-mediated DNA repair pathway (30). The virus-virus and virus-host concatemers observed here that lacked SNVs or indels at the unit junctions (**Fig. S2.1**) are likely the consequence of E6/E7 expression, abrogation of the G1-S checkpoint, prolonged stalling in the G2 phase of the cell cycle, and rolling-circle replication.

HPV-containing ecDNAs as identified here are analogous to those observed in neuroblastoma (31), glioma (32) and other cancers. The ecDNAs in HPV-negative cancer types of course lack HPV sequences and frequently comprise very large (>1 megabase pair) circles. Unlike HPV ecDNAs, which typically harbor the viral origin of replication, the mechanism of replication of ecDNAs in other cancer types remains mostly unknown (33). HPV ecDNAs typically encode the viral oncogenes E6 and E7 and also can encompass host oncogenes such as *MYC*. Among other cancer types, host oncogenes (e.g., *MYC*, *EGFRVIII*) also can be encoded in ecDNAs (32,34), and their copy numbers can be amplified. Such ecDNAs can increase intratumoral heterogeneity and facilitate rapid adaptation to selective environmental pressures. This is attributed to unequal replication and segregation of ecDNAs in daughter cells during mitosis (31,32,34,35). Such ecDNA presence has been linked with poorer survival (36). By contrast, reported associations between HPV integration and survival in oropharyngeal cancer are inconsistent (5,37,38).

HPV-containing concatemers detected here contain the viral promoter, origin of replication and HPV oncogenes. These features may increase their stable maintenance as ecDNAs by facilitating replication, segregation, and tethering onto chromosomes during mitosis, even in the absence of E1 and E2 protein expression (39,40). Moreover, since expression of E6 and E7 oncoproteins is necessary for the malignant phenotype, we infer that loss of HPV-containing ecDNA would undergo strong negative selection. Primary tumors with heterogeneous virus-virus and virus-host concatemers in ecDNA form showed more extensive circacatena than did cell lines, underscoring that HPV-host ecDNAs likely are more unstable than are intrachromosomal integrants. Nevertheless, FISH, LR-seq and Circle-seq data from cell lines strongly suggest that integrated virus-virus and virus-host concatemers occasionally undergo excision, forming HPV ecDNAs.

We identified both similarities and differences between circacatena and other forms of genomic structural variation in human tumors, including chromothripsis, chromoplexy, breakage fusion bridge cycles (BFBC), and seismic amplification. Both circacatena and chromothripsis are associated with focal host CNVs and SVs as well as formation of ecDNAs. Whereas chromothripsis is characterized by random rearrangements of shattered chromosomal segments (41), virus and host genomic segments in circacatena are joined in organized, repetitive patterns. While evolution of chromothriptic ecDNA involves a single catastrophic event, we infer that the molecular evolution of circacatena occurs sequentially in an orderly way, frequently involving recombination causing serial deletions and insertions. This difference may be due to tethering of HPV-containing ecDNA to mitotic chromosomes, whereas other ecDNAs are subject to mitotic micronuclear expulsion (42). We note virus-host concatemers are inserted at intrachromosomal loci bearing sequence homology to the host DNA segments captured by virus genomes, implying that homologous recombination mediates integration. In contrast, chromothriptic ecDNAs preferentially integrate near telomeres (43).

We have identified associations between integration of virus-host concatemers and chromosomal translocations or inversions, including concatenation of host DNA segments from multiple chromosomes and large-scale translocations of chromosomal arms. We documented translocations between virus-host concatemers and other forms of tandem repeats (e.g., telomeric or centromeric repeats), indicating shared susceptibility to DNA double-strand breaks. The chromosomal translocations we observed are more ordered in structure when compared to chromoplexy (44), in which random fragments from multiple chromosomes are linked in series. In contrast to BFBC, large-scale inversions in circacatena occur directly within telomeres, whereas BFBC are attributable to absent telomeres (45). Similar to seismic amplification, HPV

concatemers and rearrangements are associated with step-changes in CNVs and increased expression of host genes such as *MYC* (5,46). However, in contrast to seismic amplification, we infer that step-changes in CNVs at sites with circacatena are due to serial deletion events, although recombination between tandem repeats in ecDNA molecules may also play a role.

Hotspots of recurrent HPV integration near *MYC*, *ERBB2*, and *RAD51* in cervical cancers and near *MYC*, *SOX2*, *FGFR3*, *TP63*, *KLF5*, and *CD274* in oropharyngeal cancers indicate that HPV integrants can drive carcinogenesis upon clonal selection (3,5). The similarity between the structural variation we observed at the *MYC* locus in HTEC, a cell line derived from tonsillar epithelial cells that were immortalized *in vitro* upon transfection with HPV16, and those present in primary tumors and cancer cell lines (e.g., HeLa and GUMC) provides additional experimental evidence to support HPV integration as a critical event in the development of a majority of human tumors. However, a question about whether or not the diversity or extent of circacatena identified in individual cancers has clinical significance has not been answered yet.

Each primary cancer and cell line analyzed here provided a snapshot in time to inform our expanded looping model. We acknowledge a lack of longitudinally collected cancers and data to validate the time course of events. We have not demonstrated here that HPV ecDNA-mediated amplification of host oncogenes contributes directly to cancer formation or progression; we are addressing this question in our ongoing studies (5). Furthermore, despite their many advantages over WGS data, including longer read length distributions and continuous sequences, LR-seq data still cannot distinguish whether the heterogeneous, repetitive virus-host concatemerized structures detected here were linked within the same, very long (>100 kb) molecules, co-existed in the same cells and/or were segregated between distinct subclones. While multiple independent methods supported the findings reported here across the cancers and cell lines, discordant observations were made in a few instances. Among these, Circle-seq reads in HTEC cells at Chr. 8q24 and in VU147 at Chr. 15q13.3 did not match breakpoints or CNVs identified in WGS or LR-seq data. We infer that these discrepancies could result from occasional rearrangements in subclones, suggesting instability of intrachromosomal insertions and of HPV ecDNAs.

The model shown in **Fig. 6** informs the mechanisms by which HPV integration causes extensive, focal host CNVs and SVs as well as the evolution of circacatena. We conclude that this structural variation is caused by HPV integration and does not reflect a preference for HPV integration at sites of pre-existing SVs and CNVs. Our data shift the current paradigm of HPV integration in human cancers to include capture, replication, and recombination of host DNA within single- or multi-unit HPV episomes. Our analysis also extends the consequences of HPV integration to include promotion of intratumoral heterogeneity and clonal evolution in human cancers.

METHODS

Cancer cell lines and primary tumors

HeLa, 93-VU-147T (VU147), GUMC-395 and HTEC cell lines were obtained from ATCC and kindly provided by Drs. RD Steenbergen, Richard Schlegel, and John Lee, respectively. Primary oropharyngeal cancer specimens were obtained with informed consent from human subjects enrolled in a genomics study at Ohio State University and studied under approved Institutional Review Board protocols (OSU, MDACC) as described (5,21).

Sequencing libraries and data generation

Genomic DNA was extracted from cancer samples as previously described (21). For WGS, all samples were prepared for 2 x 150 bp paired end WGS libraries for sequencing on the Illumina platform (5). See **Tables S1.1-2** for more details.

For LR-seq libraries, molecular weight distributions of genomic DNA samples were evaluated using a Femto Pulse pulse-field capillary electrophoresis system (Agilent). To prepare

PacBio libraries, first genomic DNA was sheared with a Megaruptor (Diagenode) or Covaris g-tube to obtain >15 kb – 25 kb fragments. Resulting sheared DNA fragments were re-assessed using the Femto Pulse. Up to 5 µg DNA was used to prepare a SMRTbell library with a PacBio SMRTbell Express Template prep kit 2.0 (Pacific Biosciences of California). Briefly, single-stranded DNA overhangs were removed, DNA damage was repaired by end-repair and A-tailing, PB adapters were ligated, desired size fragments were purified using AMpure PB beads, and resulting CCS HiFi libraries were sized-selected in the 10-50 kb fragment range using a Blue Pippin system (Sage Science). LR-seq data were generated on one SMRT cell 8M with v2.0/v2.0 chemistry on a PacBio Sequel II instrument (Pacific Biosciences) with movie length of 30 hours. Circular consensus sequence (CCS) data files and high accuracy subreads were generated using SMRTLink software, v. 9.0.0 to 10.1.0. If yield was < 10 x fold coverage, additional library aliquots were re-sequenced.

For ONT libraries, samples containing high molecular-weight DNA fragments were sheared by passage 2-5 times (depending on starting material size distribution) through a 26.5-gauge needle. DNA size distributions were assessed again with Femto Pulse. Five µg of DNA was used to prepare each ONT library with an Oxford Nanopore SQK-LSK-110 Kit. Libraries were size-selected to remove shorter fragments using the Circulomics Short Read Eliminator (SRE) kit (Circulomics). Sized libraries were sequenced on a PromethION 24 cell PROM0002 instrument for 3 days, including a nuclease flush performed at 24 h to increase yield. Basecalling, trimming of adapters and quality checking were performed using Guppy (Oxford Nanopore), resulting in FASTQ files. See **Tables S1.1-2** for more details.

To prepare Circle-seq libraries from cultured cancer cells, we followed a published protocol (47). Briefly, 5 ug genomic DNA was purified from serial passages of each cell line by proteinase K digestion and phenol/chloroform extraction. DNA was treated with 0.2 units/ul Plasmid-Safe ATP-Dependent DNase (Epicentre) for 5 days at 37 degrees. A SYBR Green quantitative (q)PCR (Thermo Fisher Scientific) assay of a 173bp amplicon of *HBB* and TaqMan qPCR (Life Technologies) assay of a 153bp amplicon of *ERV3* were used to confirm degradation of linear chromosomal DNA (i.e. Cycle Threshold value >35). Remaining circular DNA was amplified by Multiple Displacement Amplification using φ29 DNA polymerase and random hexamer primers using the Qiagen REPLI-g Mini Kit (Qiagen). Magnetic bead-based purification was used to remove the polymerase and primers. Amplified circular DNA was sheared with ten cycles (on/off, 30/30) using a Bioruptor Pico with a cooler (Diagenode). Sequencing libraries were prepared using a NEBNext DNA Library prep kit (New England Biolabs) resulting in a target insert size of 250 bp as confirmed by TapeStation (Agilent). Resulting DNA libraries were pooled at 10 nM and sequenced in 2 x 76-bp format (Illumina), resulting in >35 million read pairs.

Bioinformatics analysis of sequencing data

Global sequence alignment and analysis: WGS data (Illumina) were aligned against a hybrid human-HPV reference genome comprised of GRCh37 + 15 high-risk HPV type genomes (GRCh37 + HPV) as previously described (5). SVs and breakpoints were detected using Lumpy (48). See **Tables S1.1-2** for more details.

PacBio and ONT reads were aligned globally against a hybrid GRCh37 + HPV16 reference using Minimap2 version 2.17 (49), as part of PRINCESS version 1.0 (50). We selected default options appropriate to each sequencing platform (-x map-pb and -x map-ont, respectively). For HeLa cell analysis, we used a hybrid GRCh37 + HPV18 reference. Resulting alignments were compared against those from LRA version 1.3.2 (51), based on the same hybrid reference genomes indexed using the commands lra global, with lra align and option -CCC for PacBio HiFi data and with -ONT for ONT data. Comparable results were observed. SVs were identified from these global alignments using Sniffles v1.0.12 (52) with or without a

VCF file generated by Lumpy analysis of WGS short reads (option –lvcf). This step identified target regions of interest containing clustered HPV breakpoints (**Tables S2-S6**).

Local realignments and analysis: Because virus-host and host-host SVs are occasionally missed in global alignments by different platforms, we compared SVs called from short reads and long reads to determine the union of SVs called across the sequencing platforms. SV breakpoints that directly flank regions of CNV were selected as segment-defining breakpoints. Non-segment defining breakpoints included those that did not flank CNV and were <1 kb from a segment-defining breakpoint due to alignment constraints (**Tables S2-S6**). We selected breakpoints detected with high number of supporting reads (i.e., ≥20 Illumina short reads, ≥5 PacBio reads and/or ≥5 ONT reads, called by at least by two or more platforms) to select genomic segments to be realigned locally. Breakpoints called from lower numbers of supporting reads were not analyzed further (data not shown).

For local realignments, we extracted long reads that aligned in part or in total to the target regions plus flanking SVs. Additional long reads that initially aligned to an additional +/- 50 kb outside of these target regions also were extracted. Target regions for local realignments were extended by adding 1 Mbp of reference sequences up- and downstream of the initial regions. We created a local reference sequence model for each sample locus as template for local re-alignment. Genomic coordinates of segments used for local realignments are listed in **Tables S2-S6**.

Realignments of extracted long reads against extended target regions were performed using Minimap2 (49). Reads with at least one segmental alignment > 1 kbp were included for further analysis. SVs in the realigned long reads were confirmed using Sniffles by alignment with these custom local sequence models (**Tables S2-S6**). Further local realignments were evaluated using a custom script to count numbers of long reads supporting individual segments and/or breakpoint junctions. Local realignments and qualities were visualized in alignment dotplots generated using pafR package (https://github.com/dwinter/pafr) and then further evaluated.

Reconstructing clonal evolution of virus-host concatemers and rearrangements: LR-seq reads identifying virus-virus and virus-host concatemers and rearrangements were grouped based on patterns of genomic segments and breakpoints that they support. Analysis of ONT reads was restricted to those with length ≥20kb that harbor HPV segments and/or host-host breakpoints. All breakpoints in each read were identified and annotated, resulting in a list of breakpoint patterns. Segments in LR-seq reads were visualized using block diagrams to facilitate manual curation. We defined preliminary groups by identifying overlaps in these patterns, requiring three or more reads supporting a given pattern of segments and breakpoints. Reads were then sorted into different prototypes based on their breakpoint connections.

Grouping of reads was started by identifying a group having the most abundant reads and simplest connection structures that support a given pattern of breakpoints and/or segments. We then moved on to a derivative group which is related to the previous structure but adds the lowest number of additional breakpoints and/or segments.

Upon identification of read groups, we traced the lineage of evolution from a common ancestral structure, by taking the minimum number of steps needed to progress from one group to the next based on sharing of breakpoint patterns. We assumed that unique individual breakpoints occurred only once in time, and would remain in downstream genomic structures unless they were deleted. Such a deletion would result in a novel breakpoint, allowing us to trace its molecular lineage. We applied this examination within and across groups of reads.

Bioinformatics analysis for ecDNA detection using Circle-seq data: To increase the accuracy of structural variant (SV) detection, we merged paired-end reads having ≥15 nt overlap between them to form longer, continuous single reads using BBMAP (https://sourceforge.net/projects/bbmap/) before alignment. Resulting merged reads were aligned to human reference genome GHCh37 + HPV16/18 genome by BWA v0.7.17 (53). SVs

including duplications were called by Lumpy v 0.3.0 (48). Candidate circular DNAs were detected by the following criteria: SVs (duplications as a marker of circular DNA) with ≥2 supporting reads; 95% coverage of regions flanked by SVs; and the mean depth of sequencing coverage in the amplified SV region was greater than that in the flanking region of the same length (54).

Prediction of ecDNA and rearrangement structures by AmpliconArchitect (AA): We used 20x coverage Illumina paired-end WGS data as input for AmpliconArchitect (v1.2) (20). First, reads were aligned against human GRCh37 + HPV reference genome using BWA, and highly amplified regions were selected using amplified_intervals script (option --gain: 4n, --cnsize: 1000 bp). We ran AmpliconArchitect using both EXPLORE mode and VIRAL mode and compared these results to determine virus-associated amplicons. For those virus-associated amplicons, we also ran AmpliconArchitect on virus-associated amplification regions using VIRAL_CLUSTERED mode for further resolution. We annotated amplicon types using AmpliconClassifier (v0.3.8) and visualized those amplicons predicted as ecDNA-like circular structure using CycleViz (0.1.1).

Fluorescence in situ hybridization (FISH)

Metaphase chromosomes were prepared from cultured cells by incubating them in 0.02 mg/ml Colcemid (Invitrogen; Grand Island, NY) for ~2 h. Cells then were incubated in hypotonic (0.075M) KCl solution and fixed in methanol/acetic acid (3:1). Slides were incubated at 37°C before FISH was performed. Biotinylated HPV probes were purchased from Enzo Life Sciences (Enzo Biochem, Farmingdale, NY). Whole chromosome paint probes were generated in-house using PCR labeling techniques (55). To increase the signal of the HPV probe, the Tyramide SuperBoost kit (ThermoFisher Scientific, Waltham, MA) was used during detection. Slides were imaged on a Leica DM-RXA fluorescence microscope (Leica; Wetzlar, Germany) equipped with appropriate optical filters (Chroma, Bellows Falls, VT) and a 63X fluorescence objective. Slides then were counterstained with 4′,6-diamidino-2-phenylindole (DAPI) or with YOYO-1 (ThermoFisher). When HPV probe signal co-localized with YOYO-1 signal detecting DNA at 63x magnification, HPV-containing ecDNA was counted. In a proof-of-principle experiment, 293T cells were transfected with a pGEM-T vector containing or lacking full-length HPV16. Colocalized HPV and YOYO-1 DNA signals were observed only when cells were transfected with HPV, but not when empty vector DNA was used in the transfection (see **Fig. S4.3**).

ACKNOWLEDGMENTS

The authors thank the cancer patients who enrolled on our genomics study, and members of the Gillison and Symer laboratories for helpful comments. The authors acknowledge computational resources from the High Performance Computing for Research facility at the University of Texas MD Anderson Cancer Center.

REFERENCES

1. Forman D, de Martel C, Lacey CJ, Soerjomataram I, Lortet-Tieulent J, Bruni L*, et al.* Global burden of human papillomavirus and related diseases. Vaccine **2012**;30 Suppl 5(2):F12-23 doi 10.1016/j.vaccine.2012.07.055.
2. Akagi K, Li J, Broutian TR, Padilla-Nash H, Xiao W, Jiang B*, et al.* Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. Genome Res **2014**;24(2):185-99 doi 10.1101/gr.164806.113.
3. Cancer Genome Atlas Research N, Albert Einstein College of M, Analytical Biological S, Barretos Cancer H, Baylor College of M, Beckman Research Institute of City of H*, et al.* Integrated genomic and molecular characterization of cervical cancer. Nature **2017**;543(7645):378-84 doi 10.1038/nature21386.
4. Parfenov M, Pedamallu CS, Gehlenborg N, Freeman SS, Danilova L, Bristow CA*, et al.* Characterization of HPV and host genome interactions in primary head and neck cancers. Proc Natl Acad Sci U S A **2014**;111(43):15544-9 doi 10.1073/pnas.1416074111.
5. Symer DE, Akagi K, Geiger HM, Song Y, Li G, Emde AK*, et al.* Diverse tumorigenic consequences of human papillomavirus integration in primary oropharyngeal cancers. Genome Res **2022**;32(1):55-70 doi 10.1101/gr.275911.121.
6. Jeon S, Lambert PF. Integration of human papillomavirus type 16 DNA into the human genome leads to increased stability of E6 and E7 mRNAs: implications for cervical carcinogenesis. Proc Natl Acad Sci U S A **1995**;92(5):1654-8 doi 10.1073/pnas.92.5.1654.
7. Crook T, Tidy JA, Vousden KH. Degradation of p53 can be targeted by HPV E6 sequences distinct from those required for p53 binding and trans-activation. Cell **1991**;67(3):547-56 doi 10.1016/0092-8674(91)90529-8.
8. Gonzalez SL, Stremlau M, He X, Basile JR, Münger K. Degradation of the retinoblastoma tumor suppressor by the human papillomavirus type 16 E7 oncoprotein is important for functional inactivation and is separable from proteasomal degradation of E7. J Virol **2001**;75(16):7583-91 doi 10.1128/jvi.75.16.7583-7591.2001.
9. Ojesina AI, Lichtenstein L, Freeman SS, Pedamallu CS, Imaz-Rosshandler I, Pugh TJ*, et al.* Landscape of genomic alterations in cervical carcinomas. Nature **2014**;506(7488):371-5 doi 10.1038/nature12881.
10. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol **2019**;20(1):246 doi 10.1186/s13059-019-1828-7.
11. De Coster W, Weissensteiner MH, Sedlazeck FJ. Towards population-scale long-read sequencing. Nat Rev Genet **2021**;22(9):572-87 doi 10.1038/s41576-021-00367-3.
12. Steenbergen RD, Hermsen MA, Walboomers JM, Joenje H, Arwert F, Meijer CJ*, et al.* Integrated human papillomavirus type 16 and loss of heterozygosity at 11q22 and 18q21 in an oral carcinoma and its derivative cell line. Cancer Res **1995**;55(22):5465-71 doi 10.1038/srep45617.
13. Schneider-Gadicke A, Schwarz E. Different human cervical carcinoma cell lines show similar transcription patterns of human papillomavirus type 18 early genes. EMBO J **1986**;5(9):2285-92 doi 10.1002/j.1460-2075.1986.tb04496.x.
14. Yuan H, Krawczyk E, Blancato J, Albanese C, Zhou D, Wang N*, et al.* HPV positive neuroendocrine cervical cancer cells are dependent on Myc but not E6/E7 viral oncogenes. Sci Rep **2017**;7:45617 doi 10.1038/srep45617.
15. Lace MJ, Anson JR, Klingelhutz AJ, Lee JH, Bossler AD, Haugen TH*, et al.* Human papillomavirus (HPV) type 18 induces extended growth in primary human cervical,

tonsillar, or foreskin keratinocytes more effectively than other high-risk mucosal HPVs. J Virol **2009**;83(22):11784-94 doi 10.1128/jvi.01370-09.
16. Zhou L, Qiu Q, Zhou Q, Li J, Yu M, Li K*, et al.* Long-read sequencing unveils high-resolution HPV integration and its oncogenic progression in cervical cancer. Nat Commun **2022**;13(1):2563 doi 10.1038/s41467-022-30190-1.
17. Rossi NM, Dai J, Xie Y, Lou H, Boland JF, Yeager M*, et al.* Extrachromosomal Amplification of Human Papillomavirus Episomes as a Mechanism of Cervical Carcinogenesis. bioRxiv **2021**:2021.10.22.465367 doi 10.1101/2021.10.22.465367.
18. Carvalho CM, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet **2016**;17(4):224-38 doi 10.1038/nrg.2015.25.
19. Skryabin BV, Kummerfeld DM, Gubar L, Seeger B, Kaiser H, Stegemann A*, et al.* Pervasive head-to-tail insertions of DNA templates mask desired CRISPR-Cas9-mediated genome editing events. Sci Adv **2020**;6(7):eaax2941 doi 10.1126/sciadv.aax2941.
20. Deshpande V, Luebeck J, Nguyen ND, Bakhtiari M, Turner KM, Schwab R*, et al.* Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. Nat Commun **2019**;10(1):392 doi 10.1038/s41467-018-08200-y.
21. Gillison ML, Akagi K, Xiao W, Jiang B, Pickard RKL, Li J*, et al.* Human papillomavirus and the landscape of secondary genetic alterations in oral cancers. Genome Res **2019**;29(1):1-17 doi 10.1101/gr.241141.118.
22. Bodelon C, Untereiner ME, Machiela MJ, Vinokurova S, Wentzensen N. Genomic characterization of viral integration sites in HPV-related cancers. Int J Cancer **2016**;139(9):2001-11 doi 10.1002/ijc.30243.
23. Adey A, Burton JN, Kitzman JO, Hiatt JB, Lewis AP, Martin BK*, et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. Nature **2013**;500(7461):207-11 doi 10.1038/nature12064.
24. Landry JJ, Pyl PT, Rausch T, Zichner T, Tekkedil MM, Stutz AM*, et al.* The genomic and transcriptomic landscape of a HeLa cell line. G3 (Bethesda) **2013**;3(8):1213-24 doi 10.1534/g3.113.005777.
25. Macville M, Schrock E, Padilla-Nash H, Keck C, Ghadimi BM, Zimonjic D*, et al.* Comprehensive and definitive molecular cytogenetic characterization of HeLa cells by spectral karyotyping. Cancer Res **1999**;59(1):141-50 doi 10.1534/g3.113.005777.
26. Flores ER, Lambert PF. Evidence for a switch in the mode of human papillomavirus type 16 DNA replication during the viral life cycle. J Virol **1997**;71(10):7167-79 doi 10.1128/JVI.71.10.7167-7179.1997.
27. Orav M, Geimanen J, Sepp EM, Henno L, Ustav E, Ustav M. Initial amplification of the HPV18 genome proceeds via two distinct replication mechanisms. Sci Rep **2015**;5:15952 doi 10.1038/srep15952.
28. Hoffmann R, Hirt B, Bechtold V, Beard P, Raj K. Different modes of human papillomavirus DNA replication during maintenance. J Virol **2006**;80(9):4431-9 doi 10.1128/JVI.80.9.4431-4439.2006.
29. Sakakibara N, Chen D, McBride AA. Papillomaviruses use recombination-dependent replication to vegetatively amplify their genomes in differentiated cells. PLoS Pathog **2013**;9(7):e1003321 doi 10.1371/journal.ppat.1003321.
30. Moody CA, Laimins LA. Human papillomaviruses activate the ATM DNA damage pathway for viral genome amplification upon differentiation. PLoS Pathog **2009**;5(10):e1000605 doi 10.1371/journal.ppat.1000605.
31. Koche RP, Rodriguez-Fos E, Helmsauer K, Burkert M, MacArthur IC, Maag J*, et al.* Publisher Correction: Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. Nat Genet **2020**;52(4):464 doi 10.1038/s41588-020-0598-1.

32. deCarvalho AC, Kim H, Poisson LM, Winn ME, Mueller C, Cherba D*, et al.* Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. Nat Genet **2018**;50(5):708-17 doi 10.1038/s41588-018-0105-0.
33. Bailey C, Shoura MJ, Mischel PS, Swanton C. Extrachromosomal DNA-relieving heredity constraints, accelerating tumour evolution. Ann Oncol **2020**;31(7):884-93 doi 10.1016/j.annonc.2020.03.303.
34. Nathanson DA, Gini B, Mottahedeh J, Visnyei K, Koga T, Gomez G*, et al.* Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. Science **2014**;343(6166):72-6 doi 10.1126/science.1241328.
35. Verhaak RGW, Bafna V, Mischel PS. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. Nat Rev Cancer **2019**;19(5):283-8 doi 10.1038/s41568-019-0128-6.
36. Kim H, Nguyen NP, Turner K, Wu S, Gujar AD, Luebeck J*, et al.* Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. Nat Genet **2020**;52(9):891-7 doi 10.1038/s41588-020-0678-2.
37. Koneva LA, Zhang Y, Virani S, Hall PB, McHugh JB, Chepeha DB*, et al.* HPV Integration in HNSCC Correlates with Survival Outcomes, Immune Response Signatures, and Candidate Drivers. Mol Cancer Res **2018**;16(1):90-102 doi 10.1158/1541-7786.Mcr-17-0153.
38. Pinatti LM, Sinha HN, Brummel CV, Goudsmit CM, Geddes TJ, Wilson GD*, et al.* Association of human papillomavirus integration with better patient outcomes in oropharyngeal squamous cell carcinoma. Head Neck **2021**;43(2):544-57 doi 10.1002/hed.26501.
39. McBride AA. Replication and partitioning of papillomavirus genomes. Adv Virus Res **2008**;72:155-205 doi 10.1016/S0065-3527(08)00404-1.
40. Pittayakhajonwut D, Angeletti PC. Analysis of cis-elements that facilitate extrachromosomal persistence of human papillomavirus genomes. Virology **2008**;374(2):304-14 doi 10.1016/j.virol.2008.01.013.
41. Cortes-Ciriano I, Lee JJ, Xi R, Jain D, Jung YL, Yang L*, et al.* Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nat Genet **2020**;52(3):331-41 doi 10.1038/s41588-019-0576-7.
42. van Leen E, Brückner L, Henssen AG. The genomic and spatial mobility of extrachromosomal DNA and its implications for cancer therapy. Nat Genet **2022**;54(2):107-14 doi 10.1038/s41588-021-01000-z.
43. Shoshani O, Brunner SF, Yaeger R, Ly P, Nechemia-Arbely Y, Kim DH*, et al.* Chromothripsis drives the evolution of gene amplification in cancer. Nature **2021**;591(7848):137-41 doi 10.1038/s41586-020-03064-z.
44. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y*, et al.* Punctuated evolution of prostate cancer genomes. Cell **2013**;153(3):666-77 doi 10.1016/j.cell.2013.03.021.
45. Gisselsson D, Pettersson L, Hoglund M, Heidenblad M, Gorunova L, Wiegant J*, et al.* Chromosomal breakage-fusion-bridge events cause genetic intratumor heterogeneity. Proc Natl Acad Sci U S A **2000**;97(10):5357-62 doi 10.1073/pnas.090013497.
46. Rosswog C, Bartenhagen C, Welte A, Kahlert Y, Hemstedt N, Lorenz W*, et al.* Chromothripsis followed by circular recombination drives oncogene amplification in human cancer. Nat Genet **2021**;53(12):1673-85 doi 10.1038/s41588-021-00951-7.
47. Henssen A, MacArthur I, Koche R, Dorado García H. Purification and Sequencing of Large Circular DNA from Human Cells. Protocol Exchange **2019** doi 10.1038/protex.2019.006.

48. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol **2014**;15(6):R84 doi 10.1186/gb-2014-15-6-r84.
49. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **2018**;34(18):3094-100 doi 10.1093/bioinformatics/bty191.
50. Mahmoud M, Doddapaneni H, Timp W, Sedlazeck FJ. PRINCESS: comprehensive detection of haplotype resolved SNVs, SVs, and methylation. Genome Biol **2021**;22(1):268 doi 10.1186/s13059-021-02486-w.
51. Ren J, Chaisson MJP. lra: A long read aligner for sequences and contigs. PLoS Comput Biol **2021**;17(6):e1009078 doi 10.1371/journal.pcbi.1009078.
52. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A*, et al.* Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods **2018**;15(6):461-8 doi 10.1038/s41592-018-0001-7.
53. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **2009**;25(14):1754-60 doi 10.1093/bioinformatics/btp324.
54. Møller HD, Mohiyuddin M, Prada-Luengo I, Sailani MR, Halling JF, Plomgaard P*, et al.* Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. Nat Commun **2018**;9(1):1069 doi 10.1038/s41467-018-03369-8.
55. Schrock E, du Manoir S, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA*, et al.* Multicolor spectral karyotyping of human chromosomes. Science **1996**;273(5274):494-7 doi 10.1126/science.273.5274.494.

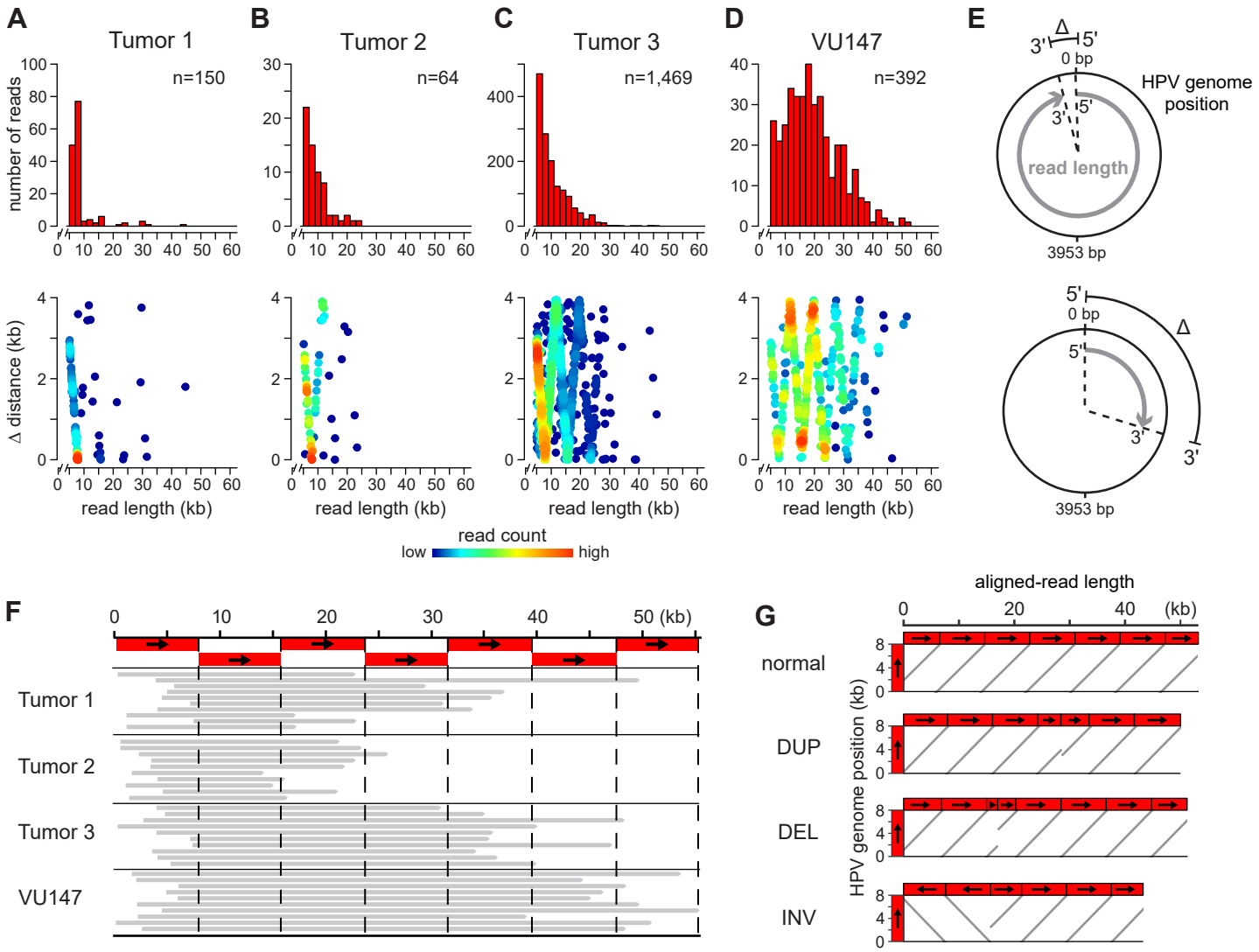

Fig. 1 - Akagi, et al.

Figure 1. Detection of HPV concatemers and SVs by LR-seq reads. LR-seq (ONT) reads containing only HPV sequences revealed frequent HPV concatemers with and without structural variants (SVs) in multiple cancers and cell lines. (A-D) Shown are (*top, y-axis*) read count histograms and (*bottom, y-axis*) plots of the distance (Δ) between 5' and 3' mapped coordinates when HPV-only reads were aligned against the HPV16 reference genome for (A) Tumor 1, (B) Tumor 2, (C) Tumor 3, and (D) VU147 cell line. *X-axis*, *top and bottom panels,* ONT read lengths in kilobase pairs (kb); *n*, number of aligned ONT reads. *Bottom, heatmap*, read counts. (E) Schematic depicting distance Δ between read 5' and 3' ends (based on half-maximal genome unit circumference, 7906 ÷ 2 bp). *Red*, *top and bottom*, two ONT reads aligned against (*black*) one unit circle of the HPV16 genome. (F) Representative ONT reads from samples in panels A-D, aligned against (*x-axis*, *dashed lines,* ~7.9-kb HPV genome unit length) concatemers of HPV genome. *Black arrows*, orientation of HPV genome from coordinate 1 to 7905. (G) Dotplots depict (*light gray*) alignments of *(x-axis)* representative ONT reads of variable lengths against *(y-axis, arrow)* one ~7.9-kb unit of HPV genome from VU147 cells. *DUP*, duplications; *DEL*, deletions; *INV*, inversions. See **Fig. S1.5**.

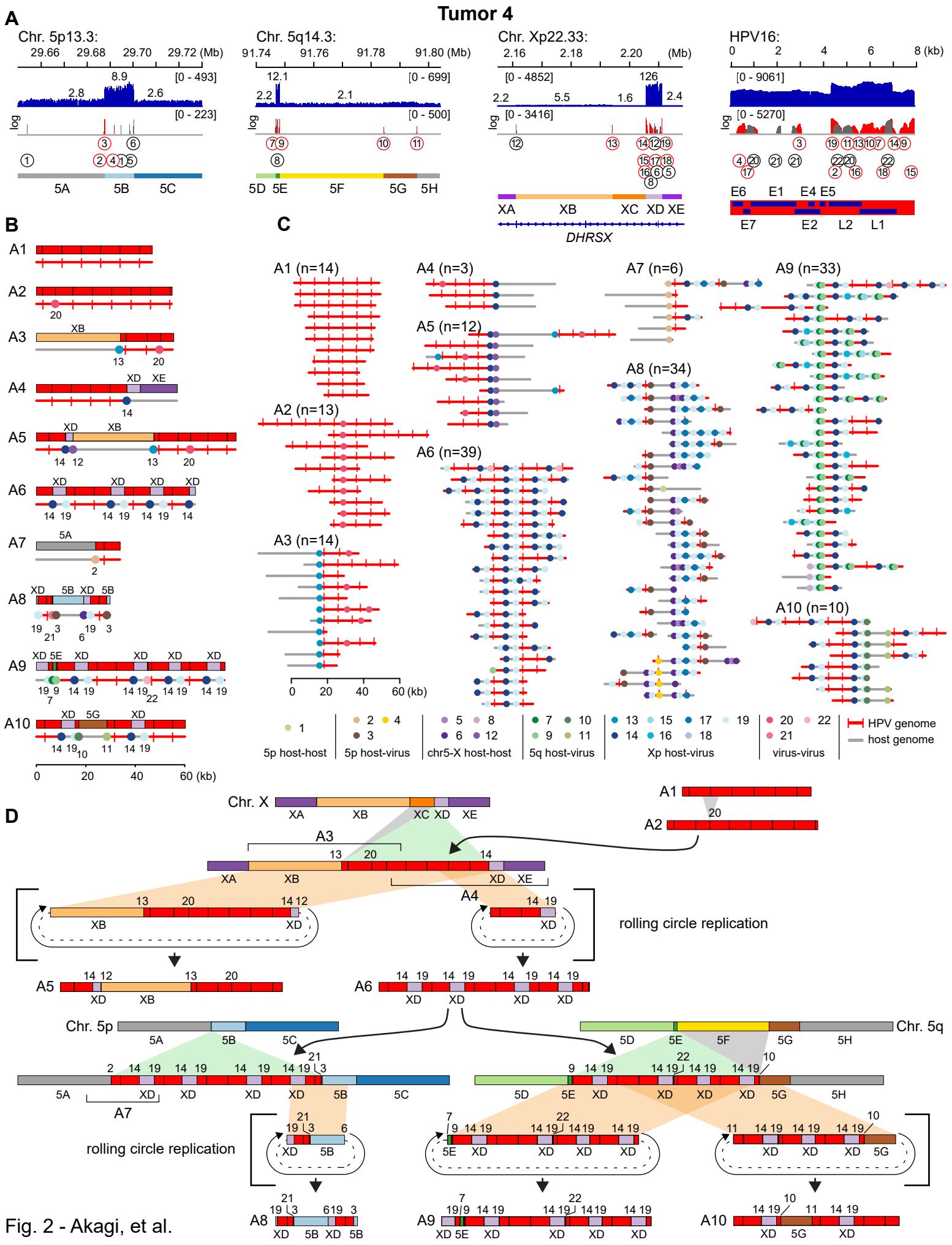

Fig. 2 - Akagi, et al.

Figure 2. Intratumoral heterogeneity and clonal evolution induced by HPV integration. Analysis of LR-seq reads from Tumor 4 revealed shared breakpoint patterns and extensive heterogeneity in virus-virus and virus-host DNA structures. (A) Depths of sequencing coverage and breakpoints at HPV integration sites at (*left to right*) Chrs. 5p, 5q and Xp and in the HPV genome, as indicated. *Top*, IGV browser display of (*y-axis, blue*) WGS coverage; *middle* (*red*) virus-host and (*gray*) host-host or virus-virus breakpoints. *Circles, numbers*, identifiers of each breakpoint defining boundaries of segments (see **Table S2**). *Bottom, left*, genomic segments defined by breakpoints; *right*, HPV genes. (B) Pairs of schematics display (*top*) block diagrams depicting representative ONT reads and (*bottom*) breakpoint plots, supporting groups of various HPV-host structures sharing patterns of breakpoints and genomic segments. Groups A1-A10 are defined by the breakpoint patterns represented. *Red lines*, HPV; *gray lines*, host DNA segments; *colored dots, numbers*, breakpoints (*see inset key, panel C*). (C) Breakpoint plots depicting (*x-axis*) representative ONT reads of length >20 kb grouped by recurrent breakpoint patterns also display heterogeneity and incorporation of other patterns. *Parentheses*, count of reads in group. (D) A model depicts how various groups of structural variants evolved from a common molecular ancestor. *Block diagrams* (e.g., A1, A2, A3, in panel B), representative ONT reads; *brackets*, hypothetical intermediate structures; *gray*, deletions; *green*, insertions; *tan*, ecDNA excisions; *dashed lines*, circularized segments; *circular arrow*, rolling-circle replication; *block colors*, segments defined in panels A and B.

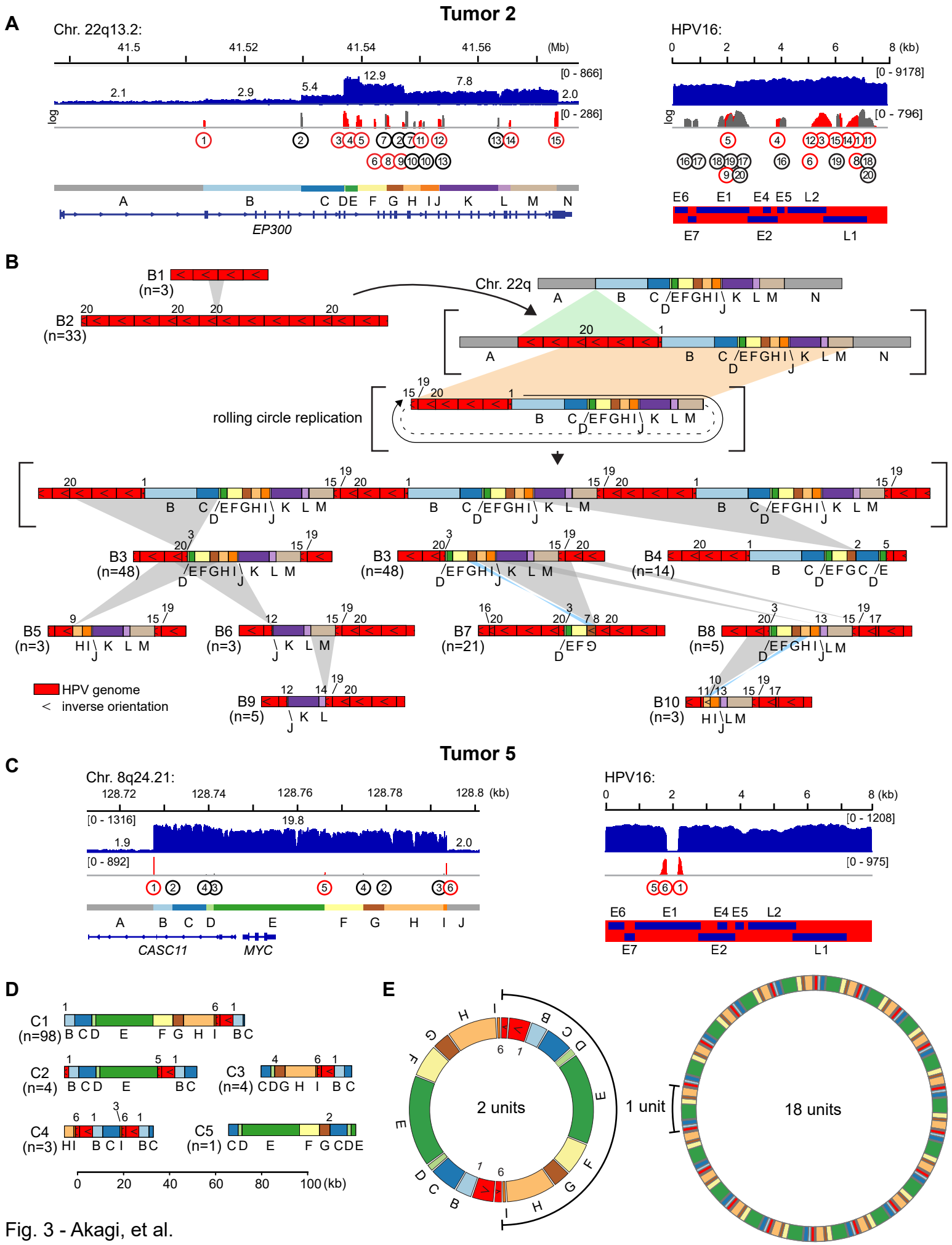

Fig. 3 - Akagi, et al.

Figure 3. Models of circacatena evolution in primary HPV-positive cancers. (A-B) *EP300* gene locus in Tumor 2; (C-E) *MYC* gene locus in Tumor 5. (A) In Tumor 2, depths of sequencing coverage and HPV insertional breakpoints at (*left to right*) the *EP300* gene locus at Chr. 22q13.2 and in the HPV genome, as indicated. See legend of Figure 2, panel A and **Table S3.1** for more details. (B) Schematic depicts potential evolution of groups of structural variants from a common molecular ancestor. *Block diagrams* (e.g., B1, B2, B3), representative ONT reads. See legend of Figure 2, panel B. (C) In Tumor 5, depths of sequencing coverage and breakpoints at HPV integration sites at (*left to right*) the *MYC* locus on Chr. 8q24.21, as in panel A. See **Table S3.2**. (D) Schematic depicts potential clonal evolution at *MYC* locus in Tumor 5, with annotations as in legend, panel B. *Block diagrams* (e.g., C1, C2, C3), representative ONT reads. *Bottom*, relative length of reads, kb. (E) Model shows potential numbers of unit chains in eccDNAs, ranging from 2 (*left*) to 18 (*right*) linked units per molecule. Many combinations per concatemer are possible but cannot be distinguished using current methods.

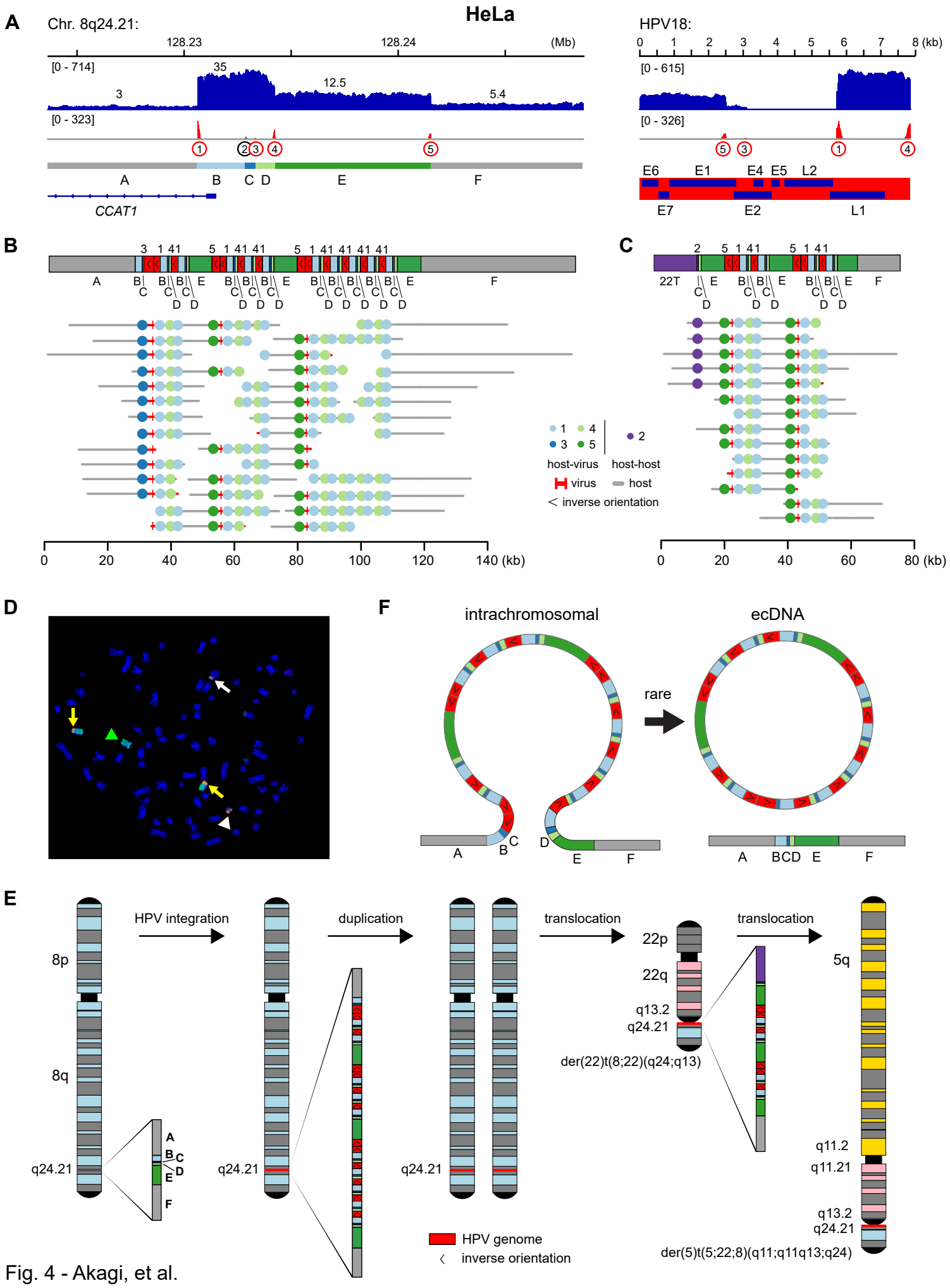

Fig. 4 - Akagi, et al.

Figure 4. HPV integration in HeLa cells induced intrachromosomal rearrangements, chromosomal translocations, and rare excision of ecDNA. (A) Depths of sequencing coverage and breakpoints at HPV integration sites at (*left to right*) Chr. 8q24.21 (upstream of *MYC*, not shown) and in the HPV18 genome, as indicated. See legend of **Fig. 2A** and **Table S4.1** for more details. (B, C) *Top*, block diagrams depicting concatemerized HPV integrants and rearrangements (B) integrated into flanking intrachromosomal segments at Chr. 8q24, and (C) joining Chr. 22 and Chr. 8 at a novel translocation. *Bottom*, breakpoint plots depicting (*x-axis*) representative ONT reads of length >20 kb, as annotated in legend, **Fig. 2C**. Many of the ONT reads demonstrate intrachromosomal integration as they directly connect concatemers with flanking host segments A (*left*) and F (*right*). *Key, inset,* breakpoints, virus segments as defined in **Table S4.1**. (D) FISH microscopy of representative HeLa metaphase spread. *Red*, HPV18 probed in (*yellow arrow*) two of three copies of Chr. 8 and in translocations between Chr. 8 and (*thin white arrow*) Chr. 22, and (*thick white*) Chr. 5. *Blue*, DAPI probe for DNA. (E) *Left to right*, stepwise model depicting molecular evolution of Chr. 8, starting with insertion of a virus-host concatemer (inset) into Chr. 8q24.21, likely by homologous recombination, followed by chromosomal translocation to the telomere of Chr. 22 and then to the centromere of Chr. 5. (F) Schematic depicting rare excision of ecDNA containing virus-host concatemer. Although the presence of ecDNA containing HPV integrants in HeLa (*top, right*) is supported by low counts of Circle-seq reads (see **Fig. S4.1**), we did not detect aberrant empty target sites (*bottom, right*). A large majority of HPV-containing concatemers appear to be integrated intrachromosomally in HeLa cells.

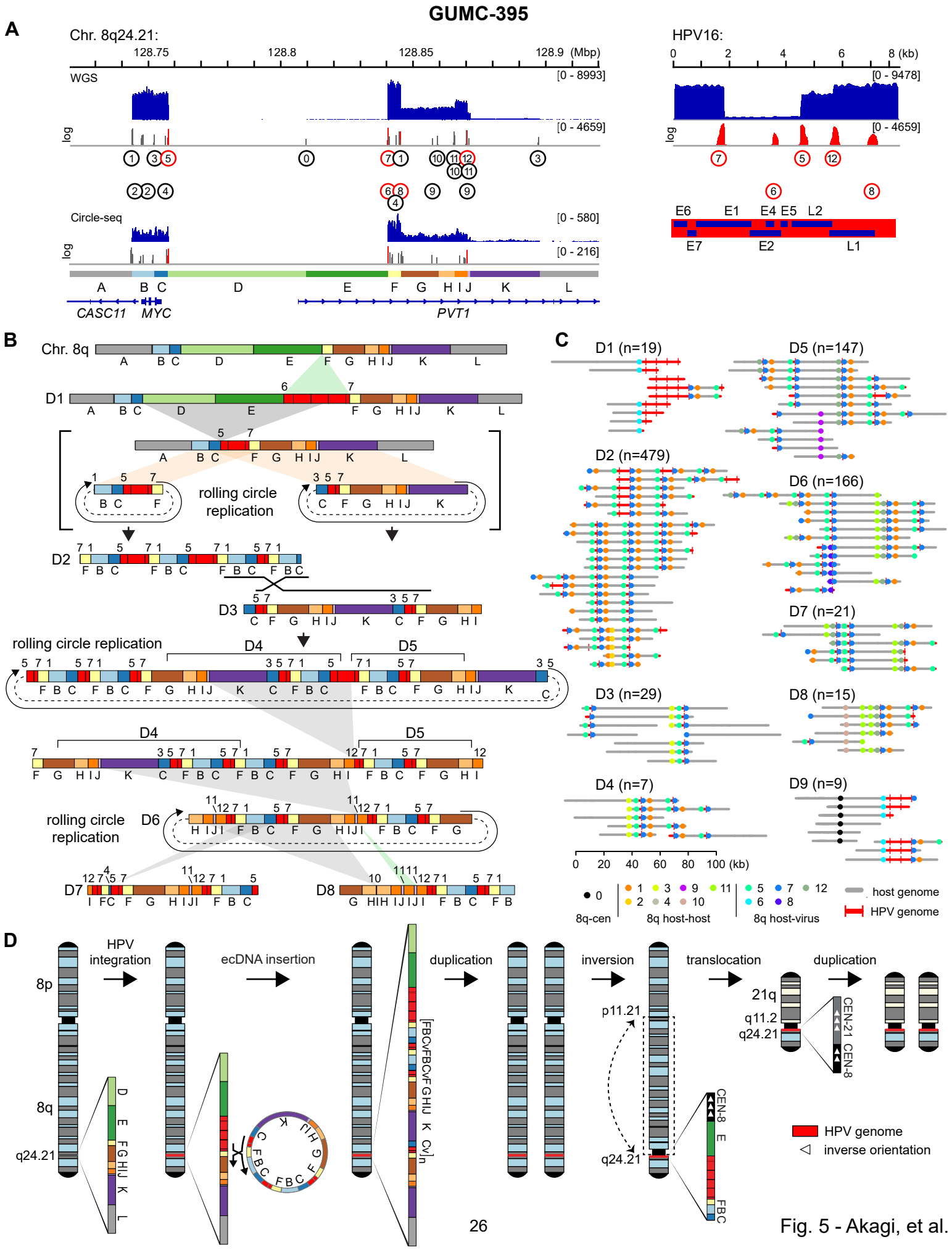

Fig. 5 - Akagi, et al.

Figure 5. Intratumoral heterogeneity and clonal evolution at *MYC* in GUMC-395 cells. (A) Depths of sequencing coverage and breakpoints at HPV integration sites at (*left to right*) Chr. 8q24.21 (*MYC* and *PVT1* genes) and in HPV16, as indicated. See legend of **Fig. 2A**, and **Table S5** for more details. (B) Schematic depicts potential evolution of groups of structural variants from a common molecular ancestor. *Block diagrams* (e.g., D1, D2, D3), representative ONT reads. See legend of **Fig. 2B**. *Black X*, site of potential homologous recombination. (C) Schematics of representative ONT reads of length >20 kb grouped by recurrent breakpoint patterns also display heterogeneity and incorporation of other patterns. *Parentheses*, count of reads in group; *key, inset*, breakpoints and virus segments. (D) Schematic supported by CLR-seq reads depicts a stepwise model by which (*left*) insertion of a virus-host concatemer containing *MYC* is followed by Chr. 8 duplication, inversion of Chr. 8q, chromosomal translocation between centromeres of Chr. 8 and Chr. 21 resulting in t(8;21)(q24;q11), and duplication of this translocation.

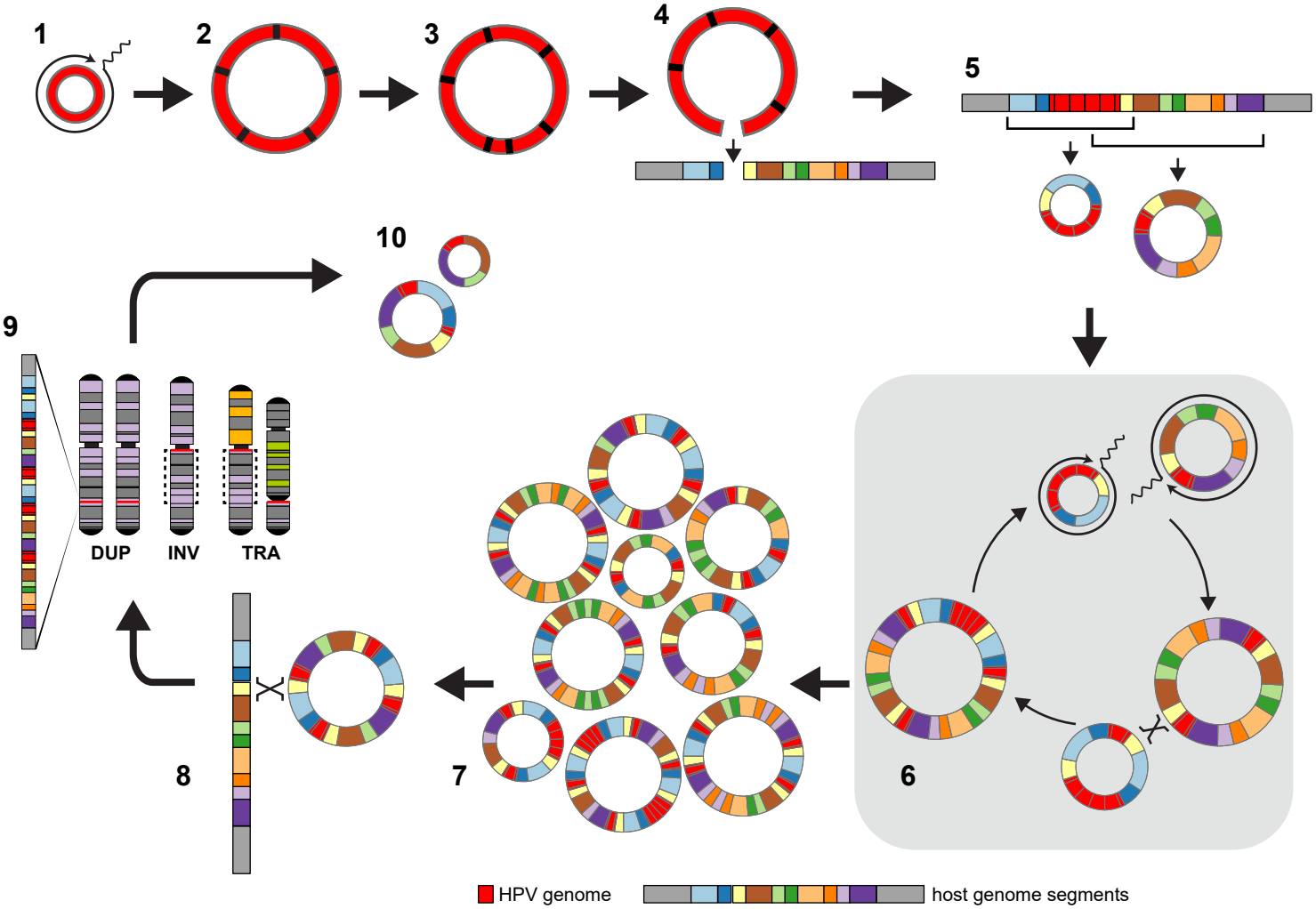

Fig. 6 - Akagi, et al.

Figure 6. Model of HPV circacatena, derived from multiple lines of evidence. A general model for development of HPV circacatena with highly diverse but related genomic rearrangements, including CNV and SV, at HPV integration sites. (1) Rolling-circle replication of HPV episomes results in (2) unstable virus genome ecDNA concatemers that (3) acquire structural rearrangements and (4) integrate into chromosomes at sites of double-strand DNA breaks. (5) Dynamic excision of virus with captured host DNA leads to (6) serial rounds of amplification of ecDNA by rolling-circle replication and recombination events between host and/or HPV segments in the same cells, driving (7) HPV circacatena and thus intratumoral heterogeneity and clonal evolution. (8) Insertion of ecDNA by recombination into chromosomes (likely through homology-directed repair) can induce (9) chromosomal inversions (INV) and translocations (TRA). (10) Occasional additional rounds of excision may produce more diverse HPV ecDNAs.