Intratumoral heterogeneity and clonal evolution induced by HPV integration

Keiko Akagi[1,*], David E. Symer[2,*], Medhat Mahmoud[3], Bo Jiang[1], Sara Goodwin[4], Darawalee Wangsa[5], Zhengke Li[1,#], Weihong Xiao[1], Joe Dan Dunn[1], 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, University of Texas MD Anderson Cancer Center, Houston, TX. [2]Department of Lymphoma & Myeloma, University of Texas 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. [+]Department of Population Health Science, Medical College of Georgia, Augusta University, Augusta, GA.

Running title: HPV integration drives heterocateny

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

Financial support: Supported by CPRIT (MLG), the Oral Cancer Foundation (MLG), and the University of Texas MD Anderson Cancer Center (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, Unit 432, Houston, TX 77030
email: mgillison@mdanderson.org
telephone: 713-792-6363

Conflict of interest statement: F.S. receives research support from Illumina, PacBio and Oxford Nanopore. M.L.G. has consulted for LLX Solutions, LLC, 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 contribution: 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, DES, MLG; Funding Acquisition, MLG.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

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 identified a previously undescribed form of structural variation, “heterocateny,” characterized by diverse, interrelated, and repetitive patterns of concatemerized virus and host DNA segments within a cancer. Unique breakpoints shared across structural variants facilitated stepwise reconstruction of their evolution from a common molecular ancestor. This analysis revealed that virus and virus-host concatemers are unstable and, upon insertion into and excision from chromosomes, facilitate capture, amplification, and recombination of host DNA and chromosomal rearrangements. Evidence of heterocateny was detected in extrachromosomal and intrachromosomal DNA. These findings indicate that heterocateny 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.

SIGNIFICANCE

Long-read sequencing of HPV-positive cancers revealed “heterocateny,” a previously unreported form of genomic structural variation characterized by heterogeneous, interrelated, and repetitive genomic rearrangements within a tumor. Heterocateny is driven by unstable concatemerized HPV genomes, which facilitate capture, rearrangement, and amplification of host DNA, and promotes intratumoral heterogeneity and clonal evolution.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

INTRODUCTION

Human papillomavirus (HPV) causes more than 630,000 cancers worldwide each year, including anogenital and oropharyngeal squamous cell carcinomas (1). Early in infection, the viral genome is maintained as an ~8-kilobase pair (kb) extrachromosomal circular DNA (ecDNA), i.e., an episome. In a majority of subsequent cancers, HPV DNA has integrated into the host genome, connecting viral and cellular DNAs at breakpoints (2-5) in intrachromosomal DNA (icDNA) and/or ecDNA forms (6). HPV integration promotes tumorigenesis by increasing expression and stability of transcripts encoding the E6 and E7 oncoproteins (7), which target tumor suppressors p53 and pRb for degradation, respectively (8,9). 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) (2,4,5,10). Diverse genetic consequences of HPV integration have been identified, including dysregulated host gene expression near integrants (2-5).

An improved understanding of the mechanisms by which HPV integration leads to SVs, CNVs, and aberrant host gene expression depends upon enhanced resolution of genomic sequence variants and their connectivity. To resolve the structures of genomic rearrangements flanking HPV integration sites, we conducted continuous long-read sequencing (LR-seq) of HPV-positive oropharyngeal cancers and human cell lines. This analysis revealed a form of genomic structural variation, which we named “heterocateny” (for “variable chain”). Heterocateny is characterized by diverse, interrelated, and repetitive patterns of concatemerized virus and host DNA segments within a tumor. Evolutionary models based on LR-seq data explained heterocateny as the consequence of aberrant host DNA replication and recombination induced by HPV integration. We conclude that HPV integration promotes intratumoral heterogeneity and clonal evolution.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

RESULTS

Our analysis of 105 HPV-positive oropharyngeal cancers by WGS identified HPV-host 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, here we used Illumina WGS and two forms of LR-seq, PacBio HiFi and Oxford Nanopore Technologies (ONT). These methods were chosen because they yield high resolution reads with few errors at a single nucleotide level (WGS, PacBio) or longer, continuous reads that can span across genomic features, including repetitive elements, for tens of kilobases (ONT). Given the expected lengths of ONT reads, we selected five HPV-positive primary oropharyngeal cancers and four cell lines, each with virus-host breakpoints that were mapped by WGS to CNV target regions with copy number (CN) ≥4n and/or SVs with breakpoints <60 kb apart. Of the 105 oropharyngeal tumors studied by WGS, CNVs with CN ≥4n were observed within 1 megabasepair (Mbp) of HPV insertional breakpoint(s) in 45% of tumors with integrated virus (5). Cell lines included 93-VU-147T (hereafter VU147) (11), GUMC-395 (12), HeLa (13), and HTEC (14). Details about distributions of read lengths and depths of sequencing coverage are provided in Supplementary Data (Supplementary Figs. S1.1-S1.4 and Supplementary Tables S1.1 and S1.2).

Extensive concatemerization and variation of HPV genomic DNA

After initial infection, HPV is maintained as an ~7.9-kb ecDNA episome in cell nuclei. Therefore, we evaluated the technical ability of both LR-seq approaches to capture and identify small circular DNA molecules, using the endogenous, ~16.5-kb circular mitochondrial (mt)DNA genome as a proxy. Histograms of ONT mtDNA reads displayed frequency peaks at 16.5 and 33 kb (Supplementary Fig. S2.1A-J and Supplementary Table S1.1). Plots of the distance between the 5’ and 3’ ends of mapped reads were within 100 base pairs (bp) in the reference mtDNA genome, indicating predominantly one- and two-unit circular mtDNA genomes. This

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

analysis confirmed the ability of LR-seq to detect ecDNAs, determine their lengths, and identify head-to-tail tandem repeats.

Comparable analysis of ONT reads mapped to the HPV reference genome revealed read lengths frequently exceeding ~7.9 kb (Fig. 1A-D, *top*, and Supplementary Fig. S2.2A and B, *top*). Plots of the distance between the 5' and 3' ends of mapped reads (Fig. 1E) indicated a predominance of single-unit HPV episomes in one primary cancer (Fig. 1A, *bottom,* Tumor 1) and multi-unit, head-to-tail virus-virus concatemers in others (Fig. 1A-D, *bottom*, Fig. 1F, and Supplementary Fig. S2.2A and B, *bottom*), consistent with recent reports (15) (bioRxiv 2021.10.22.465367).

In contrast to mtDNA, ONT reads of HPV genomes deviated more frequently from the expected unit lengths (i.e., multiples of ~7.9 kb; Fig. 1A-D, *bottom*), revealing rearrangements in virus DNA. All unique virus-virus breakpoints were confirmed by at least two and almost always by all three sequencing platforms, arguing against technical artifacts (16) (Supplementary Tables S2.1, S3.1, S3.3, S3.5, S4.1, S5.1, S5.3, and S5.5). Alignment of ONT reads from VU147, Tumor 2 and Tumor 3 against an HPV16 template model revealed rearrangements including tandem duplications, deletions, and inversions (Fig. 1G and Supplementary Fig. S2.3A and B). The seven unique virus-virus breakpoints detected in VU147 were each assigned a numerical identifier (i.e., 1-7). To facilitate pattern recognition, DNA segments in LR-seq reads were visualized using block diagrams, and breakpoints were visualized using breakpoint plots (Fig. 1H). The HPV reference genome coordinate of 0 was depicted as black or red vertical lines in these and all subsequent visualizations (Figs. 1-6). Numerous, diverse rearrangements in HPV DNA were apparent in VU147 (Fig. 1H), indicating genetic instability of the concatemerized virus genomes.

Identification of heterocateny, a unique form of structural variation

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

Extending our analysis beyond virus-only LR-seq reads, we found that Tumor 4 harbored a total of 22 unique breakpoints flanking regions of CNVs and SVs on Chrs. 5p13, 5q14, and Xp22, including 14 HPV-host, five host-host, and three virus-virus breakpoints (Fig. 2A and Supplementary Table S2.1). Rearrangements included two chromosomal translocations -- t(5;X)(p13;p22) and t(5;X)(q14;p22). To facilitate resolution of genomic structural rearrangements as covered by ONT reads, the breakpoints that were best-supported by discordant or split WGS and/or LR-seq reads were selected as segment-defining breakpoints. This allowed us to delineate host or virus DNA segments based on the reference human and HPV type-specific genomes (Supplementary Table S2.2).

Tumor 4 harbored ~171 HPV16 genome copies per haploid genome, as estimated from WGS. Virus-virus concatemers comprising up to 6 tandem HPV16 genomes were detected (Fig. 2B, group A1), but HPV nucleotides 5144-7906 plus 1-776 were deleted intermittently from adjoining viral genome units, forming a unique, recurring virus-virus breakpoint (i.e., breakpoint 20; Fig. 2B, group A2, and Supplementary Table S2.1). ONT reads with lengths ≥20 kb (N = 178) revealed rearranged virus-host structures in which distinct segments of Chr. X (e.g., XB, XD) and/or Chr. 5 (e.g., 5B, 5E, 5G) were inserted precisely where viral genome segments were deleted (Fig. 2B, groups A3-10). Individual molecules displayed specific patterns of virus and/or host DNA segments and breakpoints, which in some cases were repeated in series (Fig. 2B, groups A6, 9, 10). These diverse patterns were analogous to those observed in the virus-only ONT reads of VU147 (Fig. 1H). We clustered Tumor 4 reads into groups based on key distinct breakpoints (Fig. 2B, groups A1-10). Both within and between these read groups, breakpoint patterns diverged markedly, demonstrating extensive intermolecular heterogeneity. Distinct patterns of breakpoints and segments defining a group were occasionally linked with other group patterns in individual molecules. For example, breakpoint 13 in group A3 was also connected to breakpoints 12 and 14 in group A5 (Fig. 2B).

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

We used the unique breakpoints and patterns shared across heterogeneous structures in Tumor 4 as molecular barcodes to reconstruct genomic structural evolution from a common molecular ancestor (Methods). According to the resulting mechanistic model, insertion of concatemerized HPV genomes initially occurred at the host DNA segment XC deletion site on Chr. X (Fig. 3). During their subsequent excision from Chr. X, these concatemerized HPV genomes captured host DNA and formed ecDNA, which then inserted into Chr. 5p and 5q (Fig. 3). Shared virus and host DNA segments and breakpoints were linked in series in recurrent patterns but lacked single nucleotide variants or small insertions/deletions (Supplementary Fig. S3.1A-C), consistent with a formation mechanism involving homologous recombination and intermittent high-fidelity amplification by rolling-circle replication or recombination-dependent replication (RDR) (17-20).

Counts of reads supporting integration of rearranged virus-host concatemers into flanking chromosomal DNA were very low in Tumor 4. Therefore, we inferred that the numerous virus-host concatemers observed in this tumor mostly occurred as ecDNA. Notably, predictions for ecDNA made by AmpliconArchitect software (21) were oversimplified and inaccurate compared to the virus-host concatemers we detected by LR-seq (Fig. 2B and Supplementary Fig. S3.2A-H), likely due to inherent limitations of short-read WGS data.

In sum, Tumor 4 LR-seq data revealed a striking degree of genomic structural variation flanking virus-host breakpoints, characterized by diverse, interrelated, and repetitive patterns of virus and host DNA segments and breakpoints. We named this form of genomic structural variation “heterocateny." Multiple independent lines of evidence for heterocateny were observed in all cancers and cell lines studied, as described below.

Tumor 2 harbored a total of 23 breakpoints, including 14 virus-host, four host-host, and five virus-virus, at the ~60 kb *EP300* locus on Chr. 22q13.2 (Supplementary Table S3.1). *EP300* is frequently inactivated by somatic mutation in HPV-positive oropharyngeal cancers (22). Of the breakpoints, 14 were chosen as segment-defining breakpoints (Fig. 4A and Supplementary

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

Table S3.2). As done for Tumor 4, we used key breakpoints to sort ONT reads into groups (Fig. 4B). ONT reads (N = 154) supported concatemers with multiple tandem full-length HPV genome units interspersed with tandems lacking nucleotides 7065-7906 and 1-2312 (breakpoint 17; Fig. 4B, group B2). Virus concatemers containing breakpoint 17 were detected in series with *EP300* segments (Fig. 4B, groups B3-10). Structural heterogeneity within and between read groups, analogous to that in Tumor 4 (Fig. 2B), provided further evidence of heterocateny. Per the model based on LR-seq data, these structures evolved from a clonal ancestor by sequential events, including insertion of concatemerized HPV genomes, ecDNA excision, copy number amplifications, and additional rearrangements such as serial deletions (Supplementary Fig. S4.1). No WGS or LR-seq reads supported integration of virus-host structures into Chr. 22q13.2, suggesting that virus-host concatemers containing *EP300* fragments occurred predominantly or exclusively as ecDNA.

Interestingly, ONT reads in Tumor 2 independently identified integration of virus-host concatemers into flanking host sequences at Chr. 4p15.31 (Fig. 4C and Supplementary Table S3.1). Detection of the same breakpoint 17 both in Chr. 4 (Fig. 4D) and at the Chr. 22 *EP300* locus indicated that the virus-host concatemers at these two distinct sites were clonally related. This example demonstrated that concatemers can coexist as ecDNA and as icDNA integrants within the same tumor and that the same breakpoint can be found in both forms of genomic DNA.

Virus-host concatemers detected in Tumor 5 mapped near or included the cancer driver gene *MYC* on Chr. 8q24.21, a hotspot for HPV integration in oropharyngeal (5) and cervical cancers (23). We identified six breakpoints, three virus-host and three host-host (Supplementary Table S3.3), which were selected to delineate host segments A through J at *MYC* (Supplementary Fig. S4.2A and Supplementary Table S3.4). While HPV concatemers were not detected, a deletion at HPV nucleotides 1803-2170 was identified. We detected 110 ONT reads (each ≥20 kb) defining SVs at the *MYC* locus. Of these, 98 (88%) supported a genomic

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

rearrangement in which *MYC* was duplicated at least twice in tandem (segment E**,** Supplementary Fig. S4.2B). Less common but related SVs were derived from this ancestral molecule by recombination events. Because no reads supported integration of virus-host concatemers into adjacent chromosomal DNA, they likely existed predominantly in ecDNA form. As this tumor harbored ~20 HPV16 genome copies per haploid genome, each cell may have contained a range of ecDNA molecules with lengths commensurate with numbers of linked HPV units (Supplementary Fig. S4.2C). The relative homogeneity of ecDNA structure in Tumor 5, relative to Tumors 4 and 2 above, is consistent with a selective, clonal growth advantage conferred by the captured *MYC* oncogene.

In Tumor 3, HPV16 episomes ranging from one- to six-genome units predominated (Fig. 1C and F). Five virus-host breakpoints mapped to a gene-rich region on Chr. 3q27.1 (Supplementary Fig. S4.2D and Supplementary Tables S3.5 and S3.6), and LR-seq data supported insertion of a virus concatemer at this locus (Supplementary Fig. S4.2E and F). Relatively low read counts and few derivative rearrangements (Supplementary Fig. S4.2E, groups E3-5) suggested that ecDNA excision and recombination likely occurred in a subclonal cell population. Tumors 5 and 3 thus comprised dominant clonal or subclonal cell populations, respectively, harboring ecDNAs induced by HPV integration.

Heterocateny in cancer cell lines

The GUMC-395 cell line was derived from a liver metastasis of an aggressive cervical neuroendocrine carcinoma (12). GUMC-395 cells harbored 13 breakpoints, including five virus-host and seven host-host, clustered within an ~200-kb region of extreme hyper-amplification (up to ~225n) and structural rearrangements at the *MYC* locus (Fig. 5A and Supplementary Table S4.1). Eight of the breakpoints defined sequential host DNA segments A-L (Supplementary Table S4.2). Segments B and C encompassed *MYC*. This cell line had ~112 HPV copies per haploid genome.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

Analogous to our observations in primary cancers, marked heterogeneity in patterns of virus and host DNA segments and breakpoints was observed in ONT reads from GUMC-395 cells (Fig. 5B). Insertion of a virus concatemer was detected in Chr. 8 between segments E and F (Fig. 5A), defining breakpoints 6 and 7. Interestingly, no sequence data supported a normal allele connecting host DNA segments E to F (Supplementary Table S4.1), indicating loss of heterozygosity. Most virus-host concatemers identified in ONT reads (N = 774, ≥20 kb) contained breakpoint 7, nominating this insertion as an early, likely tumorigenic event. Moreover, many structural variants shared the same V-F-B-C pattern containing *MYC* and deletion of host segments D and E (Fig. 5B), consistent with evolution from a common molecular ancestor. In our evolutionary model for GUMC-395, ecDNAs were generated from concatemerized HPV genomes integrated at the *MYC* locus and then underwent subsequent amplification and recombination (Fig. 5C). These HPV-host concatemers continued to evolve via secondary recombination and deletion events (Fig. 5C) and ultimately gave rise to diverse but related variant structures indicative of heterocateny (Fig. 5B). The model provided a potential explanation for the step changes in CNVs identified from WGS data at several segment junctions, including F to G, H to I, J to K, and K to L (Fig. 5A). We concluded that HPV integration was responsible for hyper-amplification of *MYC* in GUMC-395, a seminal event which likely promoted the development and growth of that lethal cancer.

Chromosomal translocations mediated by virus-host concatemers

Fluorescence *in situ* hybridization (FISH) analysis of GUMC-395 cells using an HPV16 probe localized the virus to two copies of Chr. 8q and two copies of Chr. 21 in all metaphase spreads examined, due to a t(8;21)(q24.21;q11.2) translocation involving the centromere of Chr. 21 (Supplementary Fig. S5.1A-E). Consistent with this observation, LR-seq data showed virus-host concatemers integrated adjacent to host segment E on Chr. 8q24.21 (Fig. 5B, group D1) and into a second site joining host segment E to the centromere of Chr. 8 (Fig. 5B, group D9). In

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

addition, numerous ONT reads that joined centromeric repeats of Chrs. 8 and 21 over several kb were detected (Supplementary Table S4.1). We inferred 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).

Numerous HeLa ONT reads supported virus-host concatemers integrated upstream of *MYC* on Chr. 8q24.21 (Fig. 6A and B and Supplementary Tables S5.1 and S5.2), corroborating previous analyses (24-26). A chromosomal translocation in HeLa at t(8;22)(q24;q13) was initially identified by spectral karyotyping (26) but was not detected using WGS or haplotype-resolved data (24,25). Its relationship with HPV integration, if any, was not reported previously. Our LR-seq data uniquely confirmed and resolved this translocation. We identified virus-host concatemers with breakpoints identical to those integrated in Chr. 8 but connecting the 5' end of HeLa genomic segment C with a 2-kb segment of repeated telomeric sequences (i.e., 5'-TTAGGG) on Chr. 22, forming breakpoint 2 (Fig. 6C). Consistent with ONT data, HPV18 FISH probes hybridized to two of three copies of Chr. 8, a t(8;22)(q24;q13) translocation, and a complex der(5)t(5;22;8)(q11;q11q13;q24) rearrangement (Supplementary Fig. S5.1B) (26). WGS data indicated that four of the five copies of Chr. 8q extended from the HPV integration site to a telomere. Thus, we inferred 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 Chr. 22 centromere to the Chr. 5 centromere (Fig. 6D).

Collectively, our combined analysis of HeLa and GUMC-395 cells revealed that integrated virus-host concatemers are unstable and can induce chromosomal translocations and other forms of genomic structural variation.

HPV integrants in cell line icDNA and ecDNA

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

The GUMC-395 and HeLa ONT data supported integration of virus-host concatemers into icDNA. In contrast, VU147 ONT data revealed virus-host concatemers containing Chr. 17q12 segments in ecDNA form and virus-host concatemers anchored into icDNA at Chr. Xp21.1 (Supplementary Fig. S5.2A-E and Supplementary Tables S5.3 and S5.4). To evaluate the possible occurrence of virus-host concatemers in ecDNA form in GUMC-395, HeLa, and VU147, we performed metaphase FISH with HPV probes and Circle-seq. Both methods identified HPV-containing ecDNA in subsets of all cell lines examined (Supplementary Figs. S5.3A-C and S5.4A-D). GUMC-395 and HeLa Circle-seq data aligned well to their respective amplified regions at the *MYC* locus on Chr. 8q24.21, supporting ecDNA, and comparable data also were observed for VU147. This analysis confirmed structurally similar virus-host concatemers in ecDNA and icDNA forms in cell lines, indicating excision from and insertion into chromosomes.

HPV integration at the *MYC* locus *in vitro*

The human tonsillar epithelial cell line (HTEC) was created upon transfection of primary cells with HPV16 episomal DNA *in vitro*, followed by clonal selection in cell culture (14). Virus integration and formation of HPV-host concatemers occurred solely during cell culture *in vitro*. LR-seq data revealed striking similarities between HPV integration sites and genomic rearrangements at the *MYC* locus in HTEC and those in both GUMC-395 and HeLa 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. 6E and Supplementary Tables S5.5 and S5.6), analogous to HeLa in their location (Fig. 6A) and to GUMC-395 in their structural variation (Fig. 5A). ONT reads demonstrated integrated virus-host concatemers that displayed homology to host DNA segments captured in the concatemer (Fig. 6F), supporting a mechanism of insertion induced by homologous recombination comparable to that of HeLa (Fig. 6B and Supplementary Fig. S5.1C and D). In serially passaged HTEC cells, Circle-seq reads aligning at this locus showed

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

structural variation and additional discordant rearrangements, suggesting the instability of intrachromosomal insertions had resulted in occasional excision of ecDNA from this site (Supplementary Fig. S5.4E). HPV16 FISH probes localized to Chr. 8q and to both ends of isochromosome i(8q) in all metaphase spreads examined (Supplementary Fig. S5.1E), indicating viral integration preceded formation of this chromosomal abnormality as these epithelial cells evolved *in vitro* (Fig. 6G).

HPV genomic structures and transcripts in the context of heterocateny

Virtually all primary tumor and cell line ONT reads containing HPV sequences included at least one copy of the viral origin of replication (HPV16 nucleotides 7838-7906 and 1 to 100) and the region encoding E6 and E7 (nucleotides 83 to 858), even when other HPV genomic sequences were deleted (or not observed). RNA-Seq analysis revealed high levels of E6 and E7 transcripts in all cases (22) (Supplementary Fig. S6). Except for Tumor 5, in which E1 was deleted, the primary tumors with a predominance of virus-host concatemers in ecDNA form contained full-length HPV genomes and expressed E1 and E2 transcripts. In contrast, the cell lines with a predominance of virus-host concatemers in icDNA, i.e., HeLa, GUMC-395, and HTEC, had deletions in E1 and/or E2, and the corresponding transcripts were poorly expressed. Thus, E6 and E7 were expressed regardless of whether E2 had been disrupted.

DISCUSSION

Here we identified heterocateny, a striking form of genomic structural variation induced by HPV integration in human cancers, characterized by highly diverse, interrelated, and repetitive patterns of virus and host DNA segments and breakpoints that coexist within a tumor. We detected strong evidence of heterocateny in HPV-containing ecDNA, icDNA, or both across all cancers and cell lines evaluated. Evolutionary models based on LR-seq data explained heterocateny as the consequence of aberrant host DNA replication and recombination, induced

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

by HPV integration and frequently involving concatemerized, circularized DNA. We inferred that virus-virus and virus-host genomic structural rearrangements characteristic of heterocateny are unstable, whether present in ecDNA or icDNA, leading to further intratumoral heterogeneity and clonal evolution. For this reason, we also use the term heterocateny to describe the stepwise process by which HPV integration induces this form of genomic heterogeneity.

Our previous WGS analyses of cell lines (2) and primary tumors (5) prompted us to develop a mechanistic “looping” model to explain extensive genomic structural variation observed at HPV integration sites (2). Our HPV looping model proposed that double-stranded breaks in HPV DNA facilitate capture of host DNA, resulting in insertional breakpoints followed by amplification, recombination, repair, and integration of virus-host concatemers in icDNA (2). However, short WGS reads limited our ability to connect genomic segments and breakpoints over longer genomic distances. The new insights gained here have enabled expansion of this HPV looping model to include generation and insertion of unstable, concatemerized HPV genomes into icDNA; capture and rearrangement of host DNA during excision of HPV ecDNAs from icDNA and their insertion back into icDNA; HPV-host segment amplification by rolling-circle replication or RDR; recombination between repetitive or homologous segments, likely by homology-directed repair or template-switching during replication, resulting in novel combinations of breakpoint and segment patterns; and formation of chromosomal inversions and translocations between repeats in concatemers and telomeres and/or centromeres (Fig. 7).

Steps 1-5 in our evolutionary model (Fig. 7) are consistent with the existing literature. For example, Southern blotting and two-dimensional electrophoresis provided low-resolution evidence of integrated and/or episomal HPV concatemers in cervical cancers and cell lines (27,28). Excision of HPV integrants from icDNA after unlicensed replication from the HPV origin was proposed after analysis of short-read WGS data from TCGA (29). Unlicensed DNA replication and genome instability can be induced at HPV integration sites in cervical cancer cell

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

lines upon binding of the HPV E1-E2 complex to the viral origin of replication (30,31). However, here we observed heterocateny in tumors and cell lines that did not have detectable E1 and E2 expression (e.g., Tumor 5, GUMC, and VU147) in addition to others that expressed E1 and E2. Although virus-host concatemers (2) and hybrid episomes (29) have been described, the discovery and characteristics of heterocateny as illustrated in steps 6-10 of our model (Fig. 7) have not been reported previously, to the best of our knowledge.

We note both similarities and differences between heterocateny and other causes of cancer genomic structural variation, including chromothripsis, chromoplexy, breakage fusion bridge cycles (BFBC) and seismic amplification. Both heterocateny and chromothripsis are associated with formation of focal host CNVs, SVs, and ecDNAs. While chromothripsis is characterized by random rearrangements of shattered chromosomal segments (32), virus and host genomic segments in heterocateny are joined in organized, repetitive patterns. Formation of chromothriptic ecDNA involves a single catastrophic event, whereas heterocateny occurs sequentially in an orderly way, frequently involving recombination events that cause 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 (33). Virus-host concatemers inserted at icDNA sites share identical host DNA segments captured by virus genomes, implying that homologous recombination mediates their integration. In contrast, chromothriptic ecDNAs preferentially integrate near telomeres (34). The chromosomal translocations observed here are more ordered in structure when compared to chromoplexy (35), in which random fragments from multiple chromosomes are linked in series. Large-scale inversions occur directly within telomeres in heterocateny, whereas BFBC events are attributable to absent telomeres (36). Similar to seismic amplification, HPV concatemers and rearrangements in heterocateny are associated with CNV step changes and increased expression of host genes such as *MYC* (5,37). However, CNV changes in seismic amplification

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

404 are attributable largely to recombination whereas our models indicate serial deletion events
405 predominate in heterocateny. Recombination likely also contributes.

406 Cancer evolution involves two essential processes: genetic variation and clonal selection
407 (38). Comparisons of LR-Seq reads, as visualized in breakpoint plots in Figs. 2B, 4B, and 5B,
408 for example, demonstrated extensive intratumoral genomic structural variation directly linked
409 with HPV integrants. Our evolutionary models implicated these HPV integrants as the inducers
410 of heterocateny across individual cells and subclones in each tumor. Collectively, our data and
411 resulting models describe the selection of DNA segments containing a host oncogene or its
412 regulatory elements by HPV integrants, e.g., *MYC* in Tumor 5 and cell lines GUMC-395 and
413 HeLa, in addition to the viral oncogenes expressed in all HPV-positive cancers. Similarities
414 between the structural variants observed at the *MYC* locus in HTEC, which was immortalized
415 and clonally selected upon transfection with HPV16 *in vitro*, and those in Tumor 5, HeLa, and
416 GUMC-395 provide compelling experimental evidence implicating heterocateny as a driver
417 event in the evolution of human tumors.

418 Overall, virus-virus and virus-host concatemers in ecDNA form showed more extensive
419 heterocateny compared with those anchored into icDNA, implicating circular forms of ecDNA as
420 active agents or substrates in heterocateny. Across primary tumors, several chromosomal loci
421 affected by HPV integration lacked LR-seq support for anchoring of integrants into icDNA,
422 suggesting they harbored ecDNA predominantly. In contrast, FISH analysis of cell lines
423 demonstrated HPV integration in chromosomal DNA in every cell examined here and previously
424 (2). Nevertheless, FISH, LR-seq and Circle-seq data from cell lines consistently suggested that
425 integrated virus-virus and virus-host concatemers also occasionally undergo excision, forming
426 HPV ecDNAs. Distinctions observed between primary cancers and cultured cells may be
427 attributable to differences in numbers of ecDNAs maintained in different cellular contexts. For
428 example, essential factors required to replicate and maintain HPV ecDNA may be

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

429 downregulated or lost upon derivation and growth of cell lines *in vitro*. Alternatively, primary
430 cancer subclones harboring icDNA HPV integrants may benefit from selective growth
431 advantages during cell line derivation.

432 We note both similarities and differences between HPV-containing ecDNAs and HPV-
433 negative ecDNAs observed in neuroblastoma (39), glioma (40), and other cancers. The latter
434 ecDNAs comprise very large (>1 megabase pair) circles (41) with unknown mechanisms of
435 replication (42). Like HPV-containing ecDNAs, they frequently contain cellular oncogenes (e.g.,
436 *MYC*, *EGFRVIII*) (40,43). Such ecDNAs can increase intratumoral heterogeneity and facilitate
437 rapid adaptation to selective environmental pressures, attributed to unequal replication and
438 segregation of ecDNAs in daughter cells during mitosis (39,40,43,44). In contrast, HPV-
439 containing ecDNAs have the viral origin of replication and encode viral proteins including
440 oncoproteins E6 and E7. These features may increase their stable maintenance as ecDNAs by
441 facilitating replication, segregation, and tethering onto chromosomes during mitosis (45,46).
442 Loss of HPV-containing ecDNAs would likely undergo strong negative selection because
443 expression of E6 and E7 is necessary for the malignant phenotype.

444 HPV undergoes two predominant modes of replication that depend upon the
445 differentiation status of the infected cell (47-49). Maintenance replication in the basal epithelium
446 occurs in S phase by bidirectional theta replication initiated from the viral origin and depends
447 upon HPV E1 helicase and E2 transcriptional regulatory proteins. In contrast, rolling-circle
448 replication and RDR occur in the G2/M phase, are less dependent on the viral origin, and are
449 unidirectional (17,47,48). The latter two modes of replication depend upon E7- or E1-induced
450 activation of the ATM-mediated DNA repair pathway (50). The virus-virus and virus-host
451 concatemers observed here, which lacked SNVs or indels at the unit junctions (Supplementary
452 Fig. S3.1A-C), likely resulted from E6/E7 expression, abrogation of the G1-S checkpoint,
453 prolonged stalling of the cell cycle in G2, and rolling-circle replication or RDR (17).

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

Each primary cancer and cell line analyzed here provided a snapshot in time to inform our model for heterocateny (Fig. 7) (2). We acknowledge a lack of longitudinally collected cancers and data to validate the sequence of events. To date, we have not demonstrated that HPV ecDNA-mediated amplification of host oncogenes contributes directly to cancer formation or progression. Furthermore, despite many advantages over WGS data, including longer read length distributions and continuous sequences, LR-seq data still cannot determine whether the heterogeneous, repetitive virus-host concatemerized structures detected here were linked within the same, very long (>100 kb) molecules, co-existed within the same cells, and/or were segregated among distinct subclones. Our rigorous requirement for validation by multiple supporting ONT LR-seq reads may have underestimated the extent of molecular heterogeneity in each cancer. Although we observed evidence of heterocateny in all samples studied here, a larger sample size would be required to estimate the proportion of HPV-positive cancers with heterocateny.

The model shown in Fig. 7 proposes mechanisms by which HPV integration induces formation of CNVs and SVs, extensive diversity, and heterocateny. 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. These data extend our understanding of the consequences of HPV integration to include promotion of intratumoral heterogeneity and clonal evolution in human cancers. In addition, these findings may have broader implications for cancers caused by other DNA tumor viruses that integrate into host genomic DNA, including Merkel cell polyomavirus and hepatitis B virus (51-53). To our knowledge, neither the genomic structures of these cancers nor the potential of these viruses to induce heterocateny has been investigated using LR-seq to date. We speculate that aberrant firing of origins of replication endogenous to human chromosomes (54) also could induce various forms of genomic instability, potentially including heterocateny.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

METHODS

Cell lines and primary tumors

HeLa (13) and 293T cell lines were obtained from ATCC. 93-VU-147T (VU147), GUMC-395, and HTEC were obtained from Drs. RD Steenbergen (11), Richard Schlegel (12), and John Lee (14), respectively. The cell lines were authenticated using short tandem repeat DNA profiling at The University of Texas MD Anderson Cancer Center cytogenetics and cell authentication core, and were tested periodically for mycoplasma using the MycoAlert PLUS mycoplasma detection kit (Lonza, LT07-703). Primary oropharyngeal cancer specimens were obtained with written informed consent from human subjects enrolled in a genomics study at The Ohio State University conducted in accordance with the Declaration of Helsinki and studied under approved Institutional Review Board protocols (OSU, MDACC) as described (5,22).

Sequencing libraries and data generation

Genomic DNA was extracted from cancer samples as previously described (22). For WGS, all samples were prepared for 2 x 150 bp paired end libraries for Illumina WGS sequencing (5).

For LR-seq libraries, molecular weight distributions of genomic DNA samples were evaluated using a Femto Pulse pulse-field capillary electrophoresis system (Agilent; RRID:SCR_019498). To prepare PacBio libraries, genomic DNA was sheared with a Megaruptor (Diagenode) or Covaris g-tube to obtain >15-25 kb fragments. Resulting sheared DNA fragments were re-assessed using the Femto Pulse. Up to 5 µgs of 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 BluePippin system (Sage Science; RRID:SCR_020505). LR-seq

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

data were generated on one SMRT cell 8M with v2.0/v2.0 chemistry on a PacBio Sequel II instrument (Pacific Biosciences; RRID:SCR_017990) with movie length of 30 hours. Circular consensus sequence (CCS) data files and high accuracy subreads were generated using SMRT Link software, v. 9.0.0 to 10.1.0 (RRID:SCR_021174). If yield was < 10x 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 were used to prepare each ONT library with an Oxford Nanopore SQK-LSK-110 kit. Libraries were size-selected to remove shorter fragments using a Short Read Eliminator (SRE) kit (Circulomics). Sized libraries were sequenced on a PromethION 24 cell PROM0002 instrument (ONT; RRID:SCR_017987) for 3 days, including a nuclease flush performed at 24 h to increase yield. Base-calling, trimming of adapters and quality checking were performed using Guppy (Oxford Nanopore), resulting in FASTQ files.

We prepared Circle-seq libraries from cultured cancer cells as described (https://doi.org/10.1038/protex.2019.006). Briefly, 5 µg of 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°C. A SYBR Green quantitative (q)PCR (Thermo Fisher Scientific) assay of a 173bp *HBB* amplicon and TaqMan qPCR (Life Technologies) assay of a 153bp *ERV3* amplicon were used to confirm degradation of linear chromosomal DNA (i.e., expected cycle threshold values >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, 150023). 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

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

England Biolabs, E7805S) resulting in a target insert size of 250 bp as confirmed by TapeStation 4200 (Agilent; RRID:SCR_018435). Resulting DNA libraries were pooled at 10 nM and sequenced in 2 x 76-bp format (Illumina), resulting in >35 million read pairs per library.

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) (5). CNVs, SVs and breakpoints were detected as described (5,55). We previously validated our WGS pipeline for virus-host breakpoint calls with Sanger sequencing, which confirmed ~100% (2,5).

PacBio and ONT reads were aligned globally against a hybrid GRCh37 + HPV16 reference using Minimap2 version 2.17 (RRID:SCR_018550) (56), as part of PRINCESS version 1.0 (57). 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 (58), 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 (RRID:SCR_017619) (59) with or without a VCF file generated by Lumpy analysis of WGS short reads (option –lvcf; RRID:SCR_003253). This step identified reads covering target regions of interest including clustered HPV insertional breakpoints (Supplementary Tables S2-S5).

Local realignments and analysis: Breakpoints (i.e., virus-virus, virus-host, or host-host) that were detected with ≥20 Illumina short reads, ≥5 PacBio, and/or ≥5 ONT reads, and called by two or more of these platforms, were selected for further analysis. We defined boundaries of genomic segments by identifying sites of copy number transitions or discontinuous read alignments. Particular breakpoints that were best-supported by discordant or split WGS and/or

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

LR-seq reads were selected as segment-defining breakpoints to delineate host or virus DNA segments based on the reference human and HPV type-specific genomes. By contrast, other breakpoints included those that did not flank a copy number transition site or were <1 kb from a segment-defining breakpoint due to alignment constraints (Supplementary Tables S2-S5).

Target regions of interest were defined by alignment of virus-host breakpoints against the human reference genome, and then we added +/- 50 kb of flanking genomic sequences. For local realignments, we extracted long reads that aligned in part or in total to the target regions. To facilitate local alignment, target regions of interest were extended by adding 1 Mbp of reference sequences up- and downstream (referred to as "pad" in Supplementary Tables S2-S5). We used these coordinates to create a local reference sequence model for each sample locus as template for local re-alignments. Genomic coordinates of segments used for local realignments are listed in Supplementary Tables S2-S5.

Realignments of extracted long reads against extended target regions were performed using Minimap2 (56). 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 (Supplementary Tables 2-5). 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 (e.g., Fig. 1) generated using pafR package (https://github.com/dwinter/pafr; RRID:SCR_023151).

Reconstructing clonal evolution of virus-host concatemers and rearrangements: This analysis was restricted to informative ONT reads ≥20 kb in length that contained HPV and host DNA segments and breakpoints in a target region of interest. All breakpoints and segments in each read and their order in sequence were annotated. Further analysis was restricted to annotated patterns supported by three or more reads. To facilitate manual curation, DNA segments in LR-seq reads were visualized using block diagrams and breakpoints were

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

visualized using breakpoint plots. LR-seq reads were then sorted into groups based upon differences in annotated patterns of segments and breakpoints.

To elucidate how annotated patterns in grouped LR-Seq reads from target regions of interest may be interrelated, grouped LR-Seq reads were serially ordered based upon the minimal number of additional DNA segments or breakpoints present when compared to the previous and subsequent group. The analysis was repeated until all LR-Seq groups from target regions of interested were included.

After the grouped LR-Seq reads were ordered, differences in annotated patterns and genomic coordinates between groups were manually inspected at single-base pair resolution, using breakpoints as molecular barcodes, to infer a mechanism by which one group could be derived from the previously ordered group with the minimal number of events, including deletion, insertion, inversion, ecDNA excision, amplification by rolling-circle or recombination-dependent replication, recombination, or translocation. We applied this examination within and across ordered groups of LR-Seq reads. This analysis was predicated upon a reasonable statistical assumption that a unique individual breakpoint 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, prompting us to trace its molecular lineage. For some models, hypothetical intermediate structures were proposed to explain stepwise evolution of breakpoint patterns observed in LR-Seq reads. The sequence of inferred, ordered events were then used to create evolutionary models for each tumor or cell line.

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 (60). SVs including duplications were called by Lumpy v 0.3.0 (RRID:SCR_003253) (55). Candidate

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

609 circular DNAs were detected by the following criteria: SVs (duplications as a marker of circular
610 DNA) with ≥2 supporting reads; 95% coverage of regions flanked by SVs; and the mean depth
611 of sequencing coverage in the amplified SV region was greater than that in the flanking region
612 of the same length (61).
613 *Prediction of ecDNA and rearrangement structures by AmpliconArchitect (AA):* We used
614 20x coverage Illumina paired-end WGS data as input for AmpliconArchitect (v1.2) (21)
615 (RRID:SCR_023150). First, reads were aligned against human GRCh37 + HPV reference
616 genome using BWA, and highly amplified regions were selected using amplified_intervals script
617 (option --gain: 4n, --cnsize: 1000 bp). We ran AmpliconArchitect using both EXPLORE mode
618 and VIRAL mode and comparatively predicted virus-associated amplicons. We also ran
619 AmpliconArchitect on virus-associated amplification regions using VIRAL_CLUSTERED mode
620 for further resolution. Amplicon types were annotated using AmpliconClassifier (v0.3.8) and
621 amplicons predicted as ecDNA-like circular structures were visualized using CycleViz (v. 0.1.1;
622 RRID:SCR_023149).
623
624 **RNA-seq analysis**
625 Total RNA was extracted and strand-specific RNA-seq libraries were prepared and sequenced
626 as previously described (22). RNA-seq reads (2 x 150 nt) were aligned against a custom, hybrid
627 genome comprised of human GRCh37 reference with 13 appended HPV type genome
628 sequences (2) using STAR aligner version 2.7.2 (RRID:SCR_004463) (62). For HPV transcript
629 analysis, we calculated mean depth of coverage every 10 bp along the HPV16 or -18 reference
630 genomes (NC_001526.3 and NC_001357.1) and normalized against total aligned read count
631 per million.
632
633 **Fluorescence in situ hybridization (FISH)**

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

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. Whole chromosome paint probes were generated in-house using PCR labeling techniques (63). To increase the signal of the HPV probe, the Tyramide SuperBoost kit (ThermoFisher Scientific) was used during detection. Slides were imaged on a Leica DM-RXA fluorescence microscope equipped with appropriate optical filters (Chroma) and a 63X fluorescence objective. Slides then were counterstained with 4′,6-diamidino-2-phenylindole (DAPI) or with YOYO-1 (ThermoFisher Y3601). 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 (Promega A362A) engineered to contain or lack full-length HPV16, and processed as described above.

Date availability

Illumina WGS data and LR-seq data from all cancer samples and cell lines (with the exception of HeLa) were deposited at European Genome Archive (EGA; https://ega-archive.org/). The accession numbers are EGAD00001009630, EGAD00001009631, and EGAD00001009632. WGS and LR-seq data from HeLa cells were deposited at the database of Genotypes and Phenotypes (dbGaP) as a substudy under accession number phs000640.

ACKNOWLEDGMENTS

The authors thank the cancer patients who enrolled in our genomics study. We also thank 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. Where indicated, some genome

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

sequences as analyzed in this study were derived from a HeLa cell line. Henrietta Lacks, and the HeLa cell line that was established from her tumor cells without her knowledge or consent in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

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. Pang J, Nguyen N, Luebeck J, Ball L, Finegersh A, Ren S*, et al.* Extrachromosomal DNA in HPV-Mediated Oropharyngeal Cancer Drives Diverse Oncogene Transcription. Clin Cancer Res **2021**;27(24):6772-86 doi 10.1158/1078-0432.Ccr-21-2484.
7. 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.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

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. 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.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

15. 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.

16. Dyer N, Young L, Ott S. Artifacts in the data of Hu et al. Nat Genet **2016**;48(1):2-4 doi 10.1038/ng.3392.

17. 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.

18. McBride AA. Mechanisms and strategies of papillomavirus replication. Biol Chem **2017**;398(8):919-27 doi 10.1515/hsz-2017-0113.

19. Liblekas L, Piirsoo A, Laanemets A, Tombak EM, Laaneväli A, Ustav E*, et al.* Analysis of the Replication Mechanisms of the Human Papillomavirus Genomes. Front Microbiol **2021**;12:738125 doi 10.3389/fmicb.2021.738125.

20. Chang HHY, Pannunzio NR, Adachi N, Lieber MR. Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nat Rev Mol Cell Biol **2017**;18(8):495-506 doi 10.1038/nrm.2017.48.

21. 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.

22. 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.

23. 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.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

24. 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.

25. 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.

26. 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.

27. Dürst M, Croce CM, Gissmann L, Schwarz E, Huebner K. Papillomavirus sequences integrate near cellular oncogenes in some cervical carcinomas. Proc Natl Acad Sci U S A **1987**;84(4):1070-4 doi 10.1073/pnas.84.4.1070.

28. Kristiansen E, Jenkins A, Holm R. Coexistence of episomal and integrated HPV16 DNA in squamous cell carcinoma of the cervix. J Clin Pathol **1994**;47(3):253-6 doi 10.1136/jcp.47.3.253.

29. Nulton TJ, Olex AL, Dozmorov M, Morgan IM, Windle B. Analysis of The Cancer Genome Atlas sequencing data reveals novel properties of the human papillomavirus 16 genome in head and neck squamous cell carcinoma. Oncotarget **2017**;8(11):17684-99 doi 10.18632/oncotarget.15179.

30. Kadaja M, Sumerina A, Verst T, Ojarand M, Ustav E, Ustav M. Genomic instability of the host cell induced by the human papillomavirus replication machinery. EMBO J **2007**;26(8):2180-91 doi 10.1038/sj.emboj.7601665.

31. Kadaja M, Isok-Paas H, Laos T, Ustav E, Ustav M. Mechanism of genomic instability in cells infected with the high-risk human papillomaviruses. PLoS Pathog **2009**;5(4):e1000397 doi 10.1371/journal.ppat.1000397.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

32. 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.

33. 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.

34. 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.

35. 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.

36. 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.

37. 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.

38. Greaves M, Maley CC. Clonal evolution in cancer. Nature **2012**;481(7381):306-13 doi 10.1038/nature10762.

39. 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.

40. deCarvalho AC, Kim H, Poisson LM, Winn ME, Mueller C, Cherba D*, et al.* Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

dynamic disease evolution in glioblastoma. Nat Genet **2018**;50(5):708-17 doi 10.1038/s41588-018-0105-0.

41. Wu S, Turner KM, Nguyen N, Raviram R, Erb M, Santini J*, et al.* Circular ecDNA promotes accessible chromatin and high oncogene expression. Nature **2019**;575(7784):699-703 doi 10.1038/s41586-019-1763-5.

42. 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.

43. 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.

44. 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.

45. McBride AA. Replication and partitioning of papillomavirus genomes. Adv Virus Res **2008**;72:155-205 doi 10.1016/S0065-3527(08)00404-1.

46. 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.

47. 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.

48. 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.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

49. 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.

50. 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.

51. Jiang Z, Jhunjhunwala S, Liu J, Haverty PM, Kennemer MI, Guan Y*, et al.* The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. Genome Res **2012**;22(4):593-601 doi 10.1101/gr.133926.111.

52. Starrett GJ, Marcelus C, Cantalupo PG, Katz JP, Cheng J, Akagi K*, et al.* Merkel Cell Polyomavirus Exhibits Dominant Control of the Tumor Genome and Transcriptome in Virus-Associated Merkel Cell Carcinoma. mBio **2017**;8(1) doi 10.1128/mBio.02079-16.

53. Feng H, Shuda M, Chang Y, Moore PS. Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science **2008**;319(5866):1096-100 doi 10.1126/science.1152586.

54. Prioleau MN, MacAlpine DM. DNA replication origins-where do we begin? Genes Dev **2016**;30(15):1683-97 doi 10.1101/gad.285114.116.

55. 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.

56. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **2018**;34(18):3094-100 doi 10.1093/bioinformatics/bty191.

57. 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.

58. 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.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

59. 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.

60. 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.

61. 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.

62. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S*, et al.* STAR: ultrafast universal RNA-seq aligner. Bioinformatics **2013**;29(1):15-21 doi 10.1093/bioinformatics/bts635.

63. 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.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

FIGURE LEGENDS

Figure 1. LR-seq reads containing only HPV sequences revealed frequent HPV concatemers with and without structural variants 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 ONT 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 bp ÷ 2). *Grey*, *top and bottom*, two ONT reads aligned against (*red*) a one-unit circle of the HPV16 genome. **F,** Representative ONT reads from samples in panels A-D aligned against concatemeric HPV genomes. *X-axis*, *dashed lines,* ~7.9-kb HPV genome unit length; *black arrows*, orientation of HPV genome from coordinate 1 to 7906. **G,** Dotplots depict (*light gray*) alignments of *(x-axis)* representative ONT reads from VU147 cells of variable lengths against *(y-axis, arrow)* one ~7.9-kb HPV genome unit. *DUP*, duplications; *DEL*, deletions; *INV*, inversions; *colored circles*, sites of discordant or split reads supporting a breakpoint. **H,** Virus-only VU147 ONT reads shown as (*top*) block diagrams and (*bottom*) breakpoint plots, grouped by presence of unique virus-virus breakpoints. *Red lines*, HPV genome (*vertical black ticks*, HPV reference coordinate 0; *vertical white ticks*, HPV rearrangement); *colored dots, numbers*, *inset key,* breakpoints; *numbers below block diagrams*, group-defining breakpoints. See also Supplementary Figs. S2.2 and S2.3.

Figure 2. HPV integration induced intratumoral heterogeneity and clonal evolution. 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, estimated copy number, and breakpoints at HPV integration sites at (*left to right*) Chrs. 5p, 5q, and Xp and in the HPV16 genome, as indicated. *Top*, IGV browser display of (*y-axis, blue*) WGS coverage;

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

middle (*red*) virus-host and (*gray*) host-host or virus-virus breakpoints at chromosomal coordinates. *Bracketed numbers*, range of aligned sequence read counts; *numbers above WGS coverage*, estimated copy number; *circles, numbers*, identifiers of each (*top*) segment-defining and (*bottom*) segment non-defining breakpoint (see Supplementary Table S2.1). *Bottom, left*, genomic segments defined by breakpoints (see Supplementary Table S2.2); *right*, HPV genes. **B,** ONT reads ≥20 kb are shown as (*top*) block diagrams and (*bottom*) breakpoint plots. Groups A1-A10 are defined by shared breakpoint patterns based on breakpoint IDs specified below block diagrams. *Red blocks*, HPV genome; *vertical black lines*, HPV reference coordinate 0; *white vertical lines*, HPV rearrangement; *colored blocks*, host genome segment as indicated in panel A. Breakpoint plots within groups also display further heterogeneity characteristic of heterocateny. *Red lines*, HPV genome; *vertical red ticks*, HPV reference coordinate 0; *gray lines*, host DNA segments; *colored dots, numbers*, *inset key,* breakpoints. *Numbers in parentheses*, counts of reads in group, from which representative reads were selected for presentation.

Figure 3. A model of heterocateny depicts how groups of structural variants could evolve from a common molecular ancestor. *Block diagrams* (e.g., A1, A2, A3), representative ONT reads as in Fig. 2B; *brackets*, hypothetical intermediate structures; *gray*, deletions; *green*, insertions; *tan*, ecDNA excisions; *dashed lines*, circularized segments; *circular arrow*, amplification; *block colors*, segments defined in Fig. 2A and B.

Figure 4. Heterocateny disrupted the *EP300* locus and Chr. 4p15 in Tumor 2. A, Depths of sequencing coverage, estimated copy number, and HPV insertional breakpoints at (*left to right*) the *EP300* gene locus at Chr. 22q13.2 and in the HPV16 genome as indicated (see legend of Figure 2, panel A and Supplementary Tables S3.1 and S3.2 for more details). **B,** ONT reads of length ≥20 kb shown as (*top*) block diagrams or (*bottom*) breakpoint plots. Groups B1-B10 are

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

defined by the breakpoint patterns per breakpoint IDs specified below block diagrams. *Red blocks*, HPV genome; *vertical black lines*, HPV reference coordinate 0; *white vertical lines*, HPV rearrangement; *arrowhead*, inverse orientation; *colored blocks*, host genome segment as indicated in panel A. Breakpoint plots within groups display further heterogeneity characteristic of heterocateny. *Red lines*, HPV genome; *vertical red ticks*, HPV reference coordinate 0; *gray lines*, host DNA segments; *colored dots, numbers*, *inset key,* breakpoints. *Numbers in parentheses,* counts of reads in group, from which representative reads were selected for presentation. **C,** Depths of sequencing coverage, estimated copy number, and virus-host breakpoints at Chr. 4p15 in Tumor 2 as per panel A. **D,** Block diagram (*top*) for a virus-host concatemer in icDNA in Chr. 4 supported by (*bottom*) representative LR-seq reads ≥20 kb depicted as breakpoint plots. Breakpoint 17 is shared by concatemers at both chromosomal loci.

Figure 5. Intratumoral heterogeneity and clonal evolution are observed in LR-seq reads at *MYC* in GUMC-395 cells. **A,** Depths of sequencing coverage, estimated copy number, 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 Supplementary Tables S4.1 and S4.2 for more details). **B,** ONT reads of length ≥20 kb shown as (*top*) block diagrams or (*bottom*) breakpoint plots. Structural variant groups D1-D9 are defined by the breakpoint patterns per breakpoint IDs specified below block diagrams. *Red blocks*, HPV genome; *vertical black lines*, HPV reference coordinate 0; *colored blocks*, host genome segment as indicated in panel A. Breakpoint plots within groups display further heterogeneity characteristic of heterocateny. *Red lines*, HPV genome; *vertical red ticks*, HPV reference coordinate 0; *gray lines*, host DNA segments; *colored dots, numbers*, *inset key,* breakpoints. *Numbers in parentheses,* counts of reads in group, from which representative reads were selected for presentation. **C,** Schematic depicts potential evolution of structural variant groups in panel B from a common molecular ancestor. *Black X*, site of potential homologous recombination; *brackets*, hypothetical

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

intermediate structures; *gray*, deletions; *green*, insertions; *tan*, ecDNA excisions; *dashed lines*, circularized segments; *circular arrow*, amplification; *block colors*, segments defined in panel A. **D,** Schematic supported by LR-seq reads depicts a stepwise model by which 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. *White arrowhead*, inverse orientation.

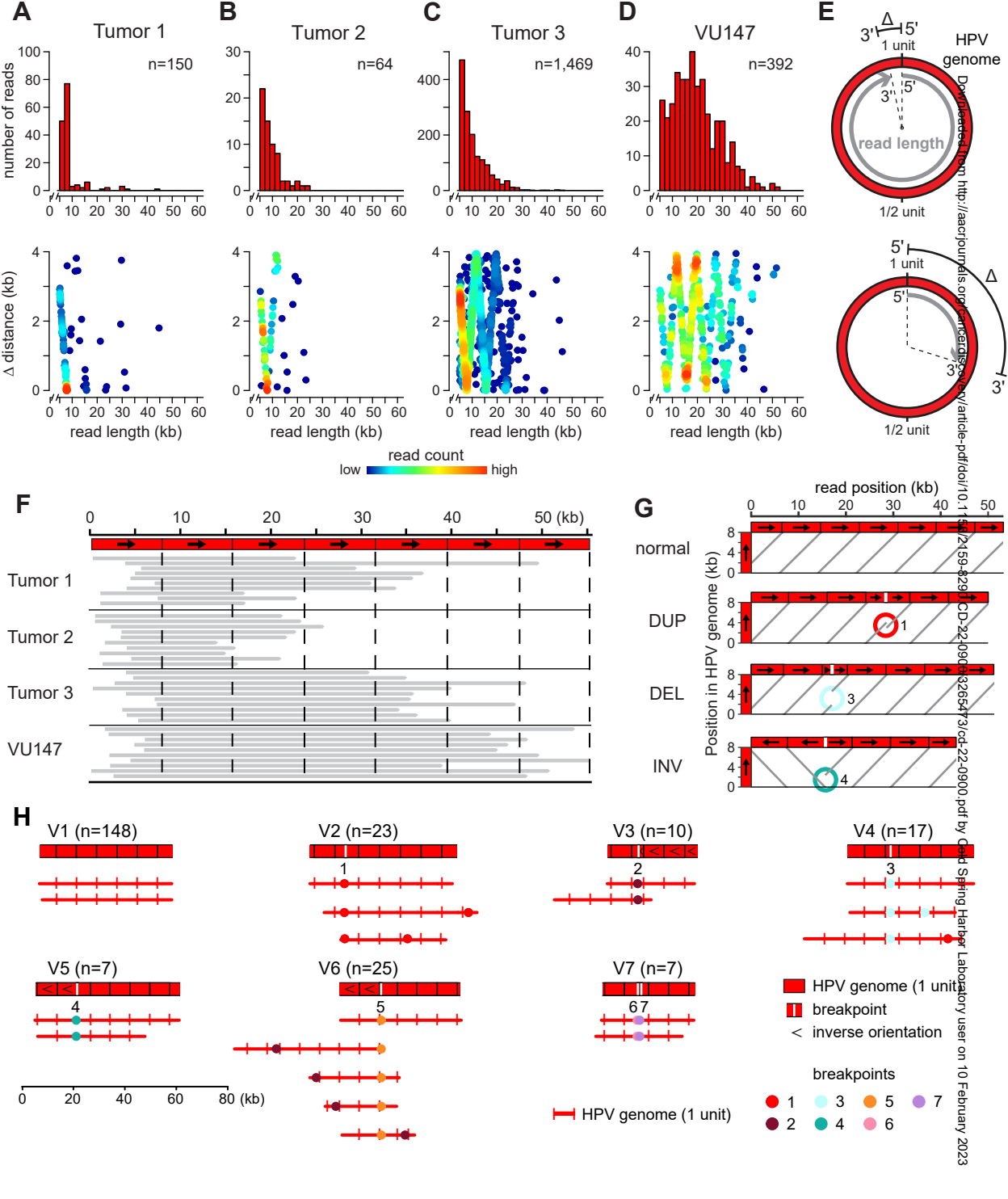
Figure 6. HPV integration in HeLa cells and human tonsillar epithelial cells (HTEC) induced CNV, SV, and intrachromosomal rearrangements. Virus-host concatemers in icDNA lead to chromosomal instability in HeLa (**A-D**) and HTEC (**E-G**) cells. **A,** Depths of sequencing coverage, estimated copy number, and breakpoints at HPV integration sites in HeLa at (*left to right*) Chr. 8q24.21 (upstream of *MYC*) and in the HPV18 genome, as indicated (see legend of Fig. 2A and Supplementary Tables S5.1 and S5.2 for more details). **B** and **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 translocation. *Red blocks*, HPV genome; *vertical black lines*, HPV reference coordinate 0; *arrowhead*, inverse orientation; *colored blocks*, host genome segment as indicated in panel A. *Bottom*, breakpoint plots of representative ONT reads ≥20 kb supporting each block diagram. Many of the ONT reads demonstrate intrachromosomal integration as they directly connect concatemers with flanking host DNA segments A (*left*) and F (*right*). *Red lines*, HPV genome; *vertical red ticks*, HPV reference coordinate 0; *gray lines*, host DNA segments; *colored dots, numbers*, *inset key,* breakpoints. **D,** 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. **E,** Depths of sequencing coverage, estimated copy number, and breakpoints at HPV integration sites in HTEC at (*left to right*) Chr. 8q24.13 (upstream of *MYC*)

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

and in the HPV16 genome, as indicated (see legend of Fig. 2A and Supplementary Tables S5.5 and S5.6 for more details. **F,** ONT reads (*bottom, breakpoint plots*) supporting integration of a virus-host concatemer in icDNA at Chr. 8q24.13 (*top, block diagram*). **G,** *Left to right*, stepwise model depicting molecular evolution of Chr. 8 in HTEC *in vitro*, starting with insertion of a virus-host concatemer (inset) into Chr. 8q24.13, likely by homologous recombination, followed by chromosomal duplication and development of isochromosome 8.

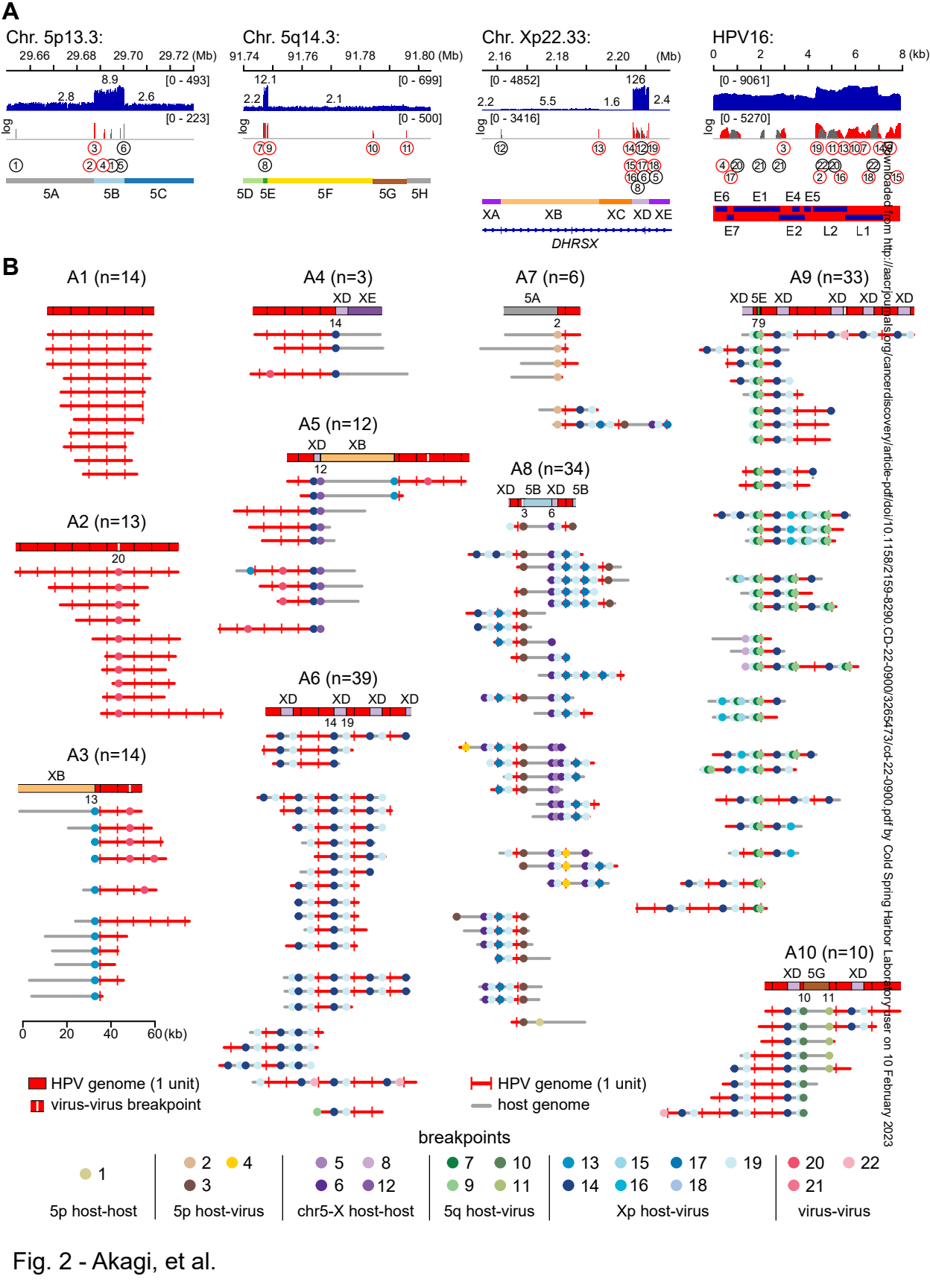
Figure 7. A model of HPV heterocateny development, depicting highly diverse but related genomic rearrangements including CNVs and SVs at HPV integration sites, is derived from multiple lines of evidence. (1) Rolling-circle replication of HPV episomes results in (2) unstable virus genome ecDNA concatemers 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 or recombination-dependent replication and recombination events between host and/or HPV segments in the same cells, driving (7) HPV heterocateny 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.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023



Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

Fig. 1 - Akagi, et al.



Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

Fig. 2 - Akagi, et al.

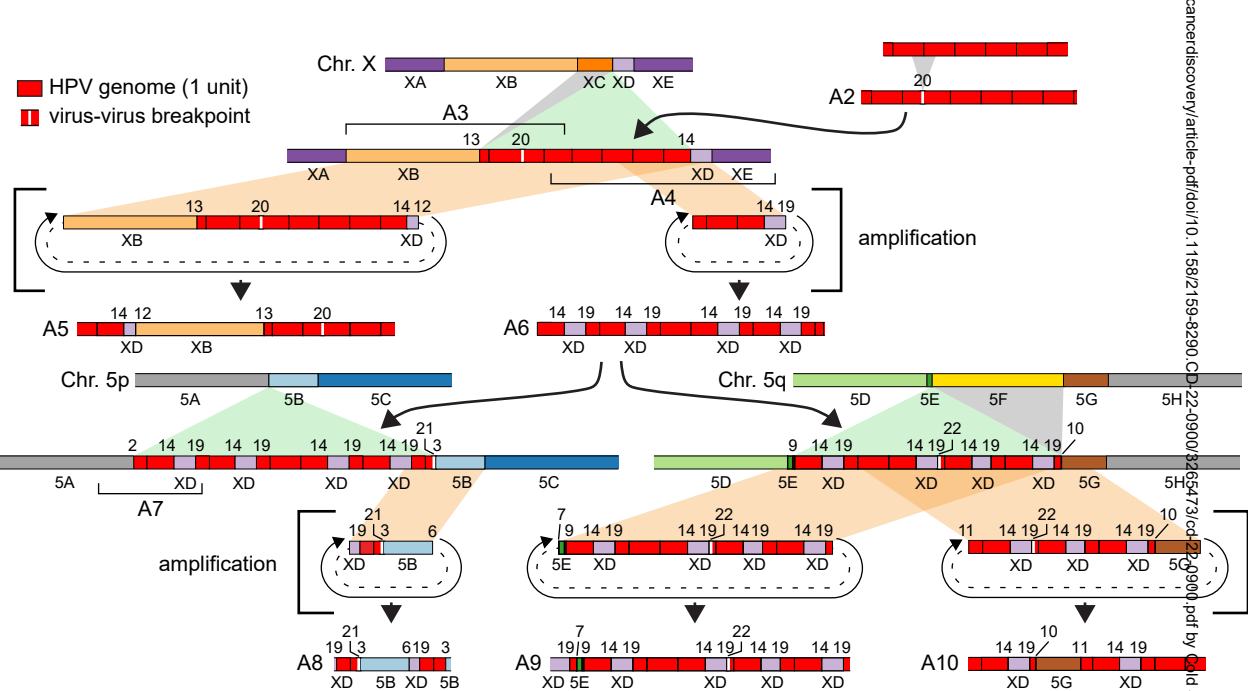
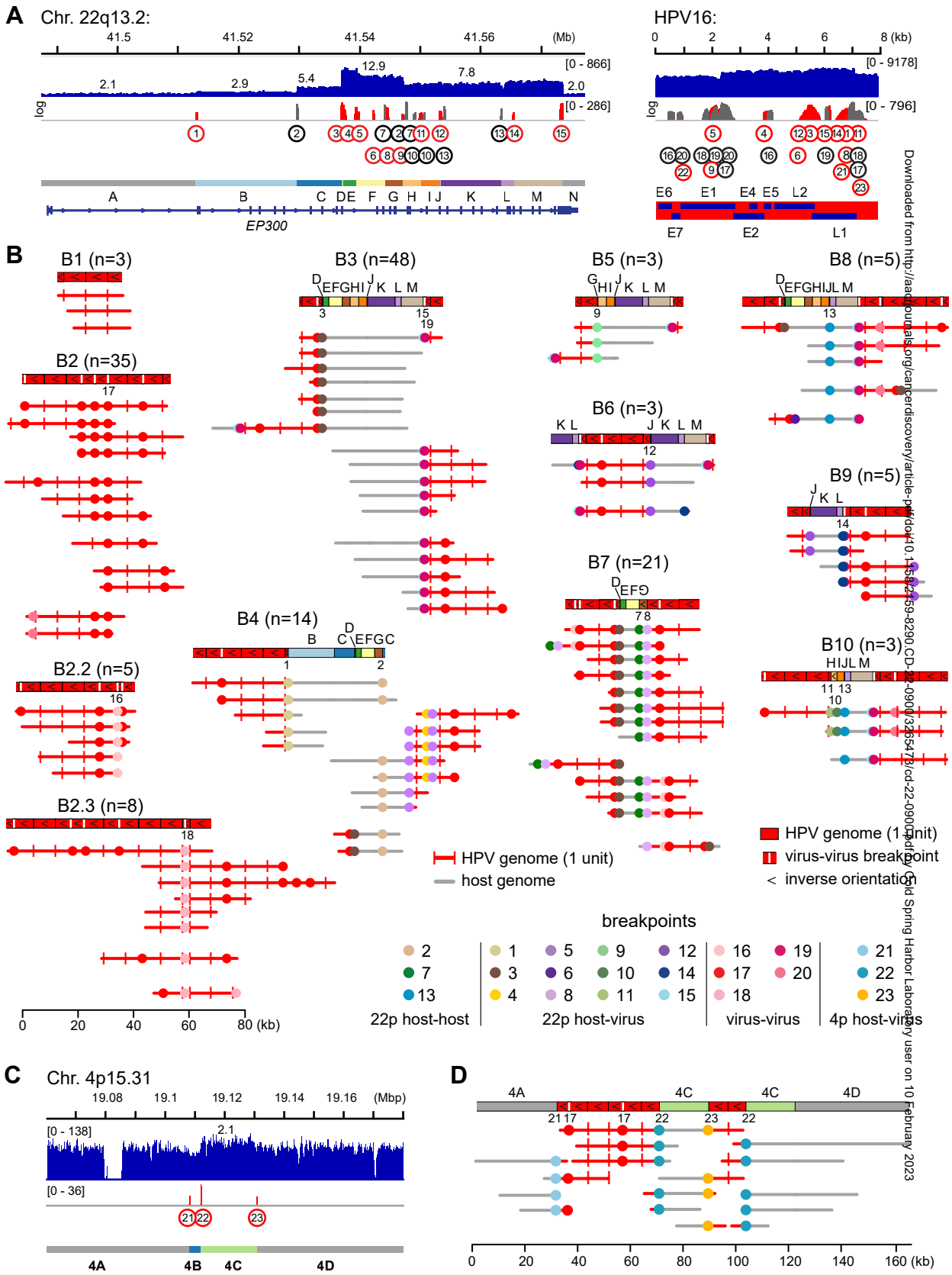


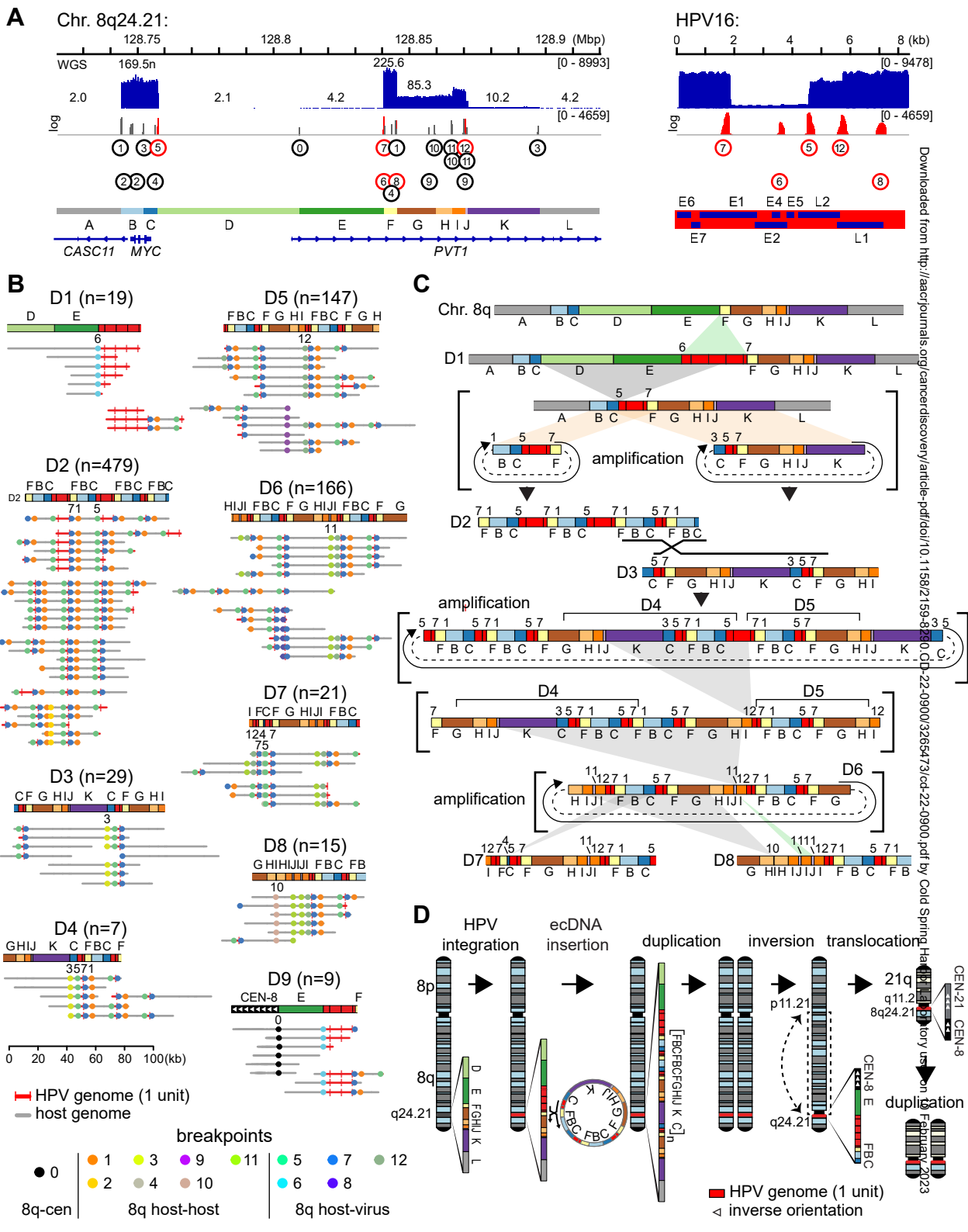
Fig. 3 - Akagi, et al.

cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Har



Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3205473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

Fig. 4 - Akagi, et al.



Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on February 2023

Fig. 5 - Akagi, et al.

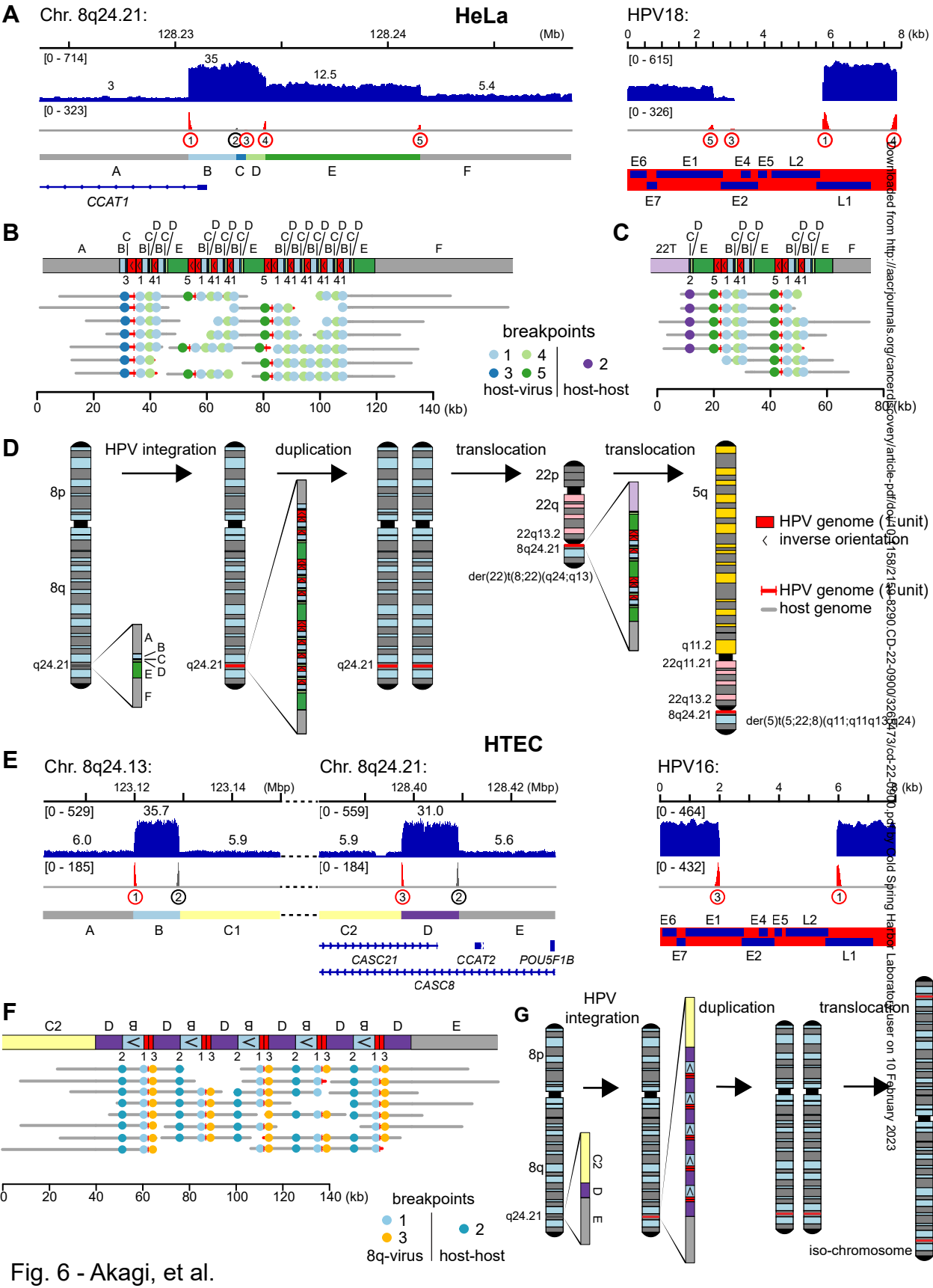


Fig. 6 - Akagi, et al.

Downloaded from http://aacrjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laboratory user on 10 February 2023

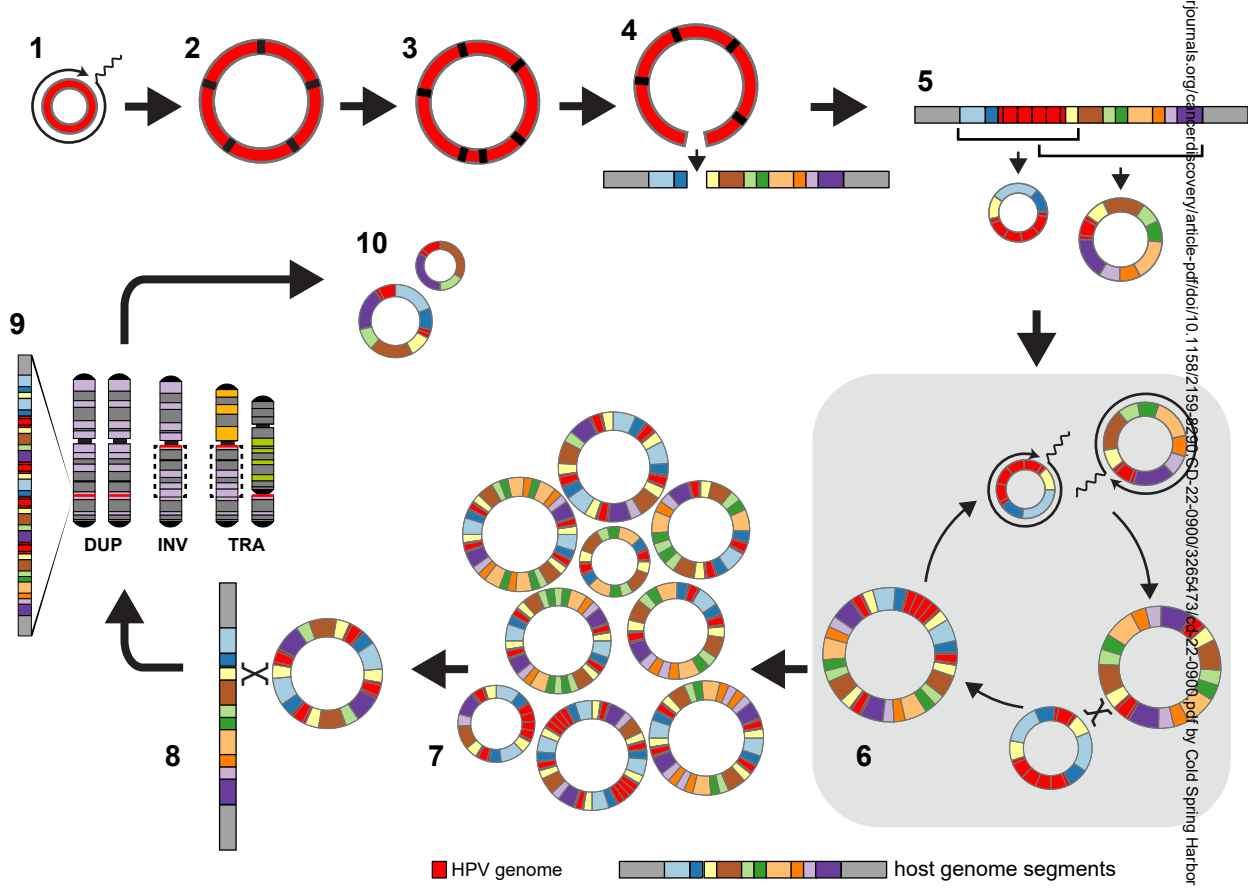


Fig. 7 - Akagi, et al.

rjournals.org/cancerdiscovery/article-pdf/doi/10.1158/2159-8290.CD-22-0900/3265473/cd-22-0900.pdf by Cold Spring Harbor Laborator