# SearcHPV: a novel approach to identify and assemble human papillomavirus-host genomic integration events in cancer

**Lisa M. Pinatti, B.S.**[1,2,*], **Wenjin Gu, M.S.**[3,*], **Yifan Wang, PhD**[4], **Ahmed Elhossiny**[3], **Apurva D. Bhangale, M.S.**[2], **Collin V. Brummel, B.A.**[2], **Thomas E. Carey, PhD**[1,2,5,6], **Ryan E. Mills, PhD**[3,4,†], **J. Chad Brenner, PhD**[2,5,6,†]

[1]Cancer Biology Program, Program in the Biomedical Sciences, Rackham Graduate School, University of Michigan, Ann Arbor, MI

[2]Department of Otolaryngology/Head and Neck Surgery, University of Michigan, Ann Arbor, MI

[3]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI

[4]Department of Human Genetics, University of Michigan, Ann Arbor, MI

[5]Rogel Cancer Center, Michigan Medicine, Ann Arbor, MI

[6]Department of Pharmacology, University of Michigan, Ann Arbor, MI

## Abstract

**Background:** Human papillomavirus (HPV) is a well-established driver of malignant transformation in a number of sites including head and neck, cervical, vulvar, anorectal and penile squamous cell carcinomas; however, the impact of HPV integration into the host human genome on this process remains largely unresolved. This is due to the technical challenge of identifying HPV integration sites, which includes limitations of existing informatics approaches to discover viral-host breakpoints from low read coverage sequencing data.

**Methods:** To overcome this limitation, we developed a new HPV detection pipeline called SearcHPV based on targeted capture technology and applied the algorithm to targeted capture

**CORRESPONDING AUTHOR:** J. Chad Brenner, 9301B MSRB3, 1150 W. Medical Center Drive, Ann Arbor, MI 48109, 734-763-2761, chadbren@umich.edu.

*L.M. Pinatti and W. Gu should be considered joint first author

†R.E. Mills and J.C. Brenner should be considered joint senior author

data. We performed an integrated analysis of SearcHPV-defined breakpoints with genome-wide linked read sequencing to identify potential HPV-related structural variations.

**Results:** Through analysis of HPV+ models, we show that SearcHPV detects HPV-host integration sites with a higher sensitivity and specificity than two other commonly used HPV detection callers. SearcHPV uncovered HPV integration sites adjacent to known cancer-related genes including *TP63*, *MYC* and *TRAF2*, as well as near regions of large structural variation. We further validated the junction contig assembly feature of SearcHPV, which helped to accurately identify viral-host junction breakpoint sequences. We found that viral integration occurred through a variety of DNA repair mechanisms including non-homologous end joining, alternative end joining and microhomology mediated repair.

**Conclusions:** In summary, we show that SearcHPV is a new optimized tool for the accurate detection of HPV-human integration sites from targeted capture DNA sequencing data.

## PRECIS:

To overcome technical challenges of detecting viral integrations in human papillomavirus-related cancers, we optimized a new pipeline called SearcHPV. Using this tool, we found frequent integration near genes and areas of large structural rearrangements in HPV+ models.

### Keywords

genomics; bioinformatics; papillomavirus infections; virus integration; squamous cell carcinoma; DNA sequence analysis

## INTRODUCTION:

Human papillomavirus (HPV) is a well-established driver of malignant transformation in a number of cancers, including head and neck squamous cell carcinomas (HNSCC). Although HPV genomic integration is not a normal event in the lifecycle of HPV, it is frequently reported in HPV+ cancers[1–4] and it may be a contributor to oncogenesis. In cervical cancer, HPV integration increases in incidence during progression from stages of cervical intraepithelial neoplasia (CIN) I/II, CIN III and invasive cancer development.[5] This process has a variety of impacts on both the HPV and cellular genomes, including disruption of E2, the transcriptional repressor of the HPV oncoproteins, leading to an increase in genetic instability.[6] HPV integration occurs within/near cellular genes more often than expected by chance[7] and has been reported to be associated with structural variations[8]. Recent studies in HNSCCs have also suggested that additional oncogenic mechanisms of HPV integration may exist through direct effects on cancer-related gene expression and generation of hybrid viral-host fusion transcripts.[9]

A wide array of methods has been previously used for the detection of HPV integration. Polymerase chain reaction (PCR)-based methods, such as Detection of Integrated Papillomavirus Sequences PCR (DIPS-PCR)[10] and Amplification of Papillomavirus Oncogene Transcripts (APOT)[11], are low sensitivity assays and are limited in their ability to detect the broad spectrum of genomic changes resulting from this process. Next-generation sequencing (NGS) technologies overcome these limitations. Previous groups have assessed

HPV integration within HNSCC tumors in The Cancer Genome Atlas (TCGA) and cell lines by whole-genome sequencing (WGS).[2, 3, 8] There are a variety of viral integration detection tools developed for WGS data, such as VirusFinder2[12, 13] and VirusSeq[14]. However, these strategies are designed for a broad range of virus types and require whole genomes to be sequenced at uniform coverage, which can result in a lower sensitivity of detection for specific types of rare viral integration events.

To overcome this issue, others have begun to use HPV targeted capture sequencing.[5, 15–18] This strategy allows for better coverage of integration sites than an untargeted approach like WGS but requires sensitive and accurate viral-human fusion detection bioinformatic tools, of which the field has been lacking. In our lab, we have found the previously available viral integration callers to have a relatively low validation rate and limitations on the structural information surrounding the fusion sites, which impairs mechanistic studies. Therefore, we set out to generate a novel pipeline specifically for targeted capture sequencing data to serve as a new gold standard in the field.

## MATERIALS AND METHODS:

### Targeted Capture Sequencing:

DNA from the HPV16-positive UM-SCC-47 cell line[45], a Patient derived xenograft (PDX)-294R (National Cancer Institute Identifier: PDX-932174–294-R) and a frozen HPV+ sample, TumorA, were submitted to the University of Michigan Advanced Genomics Core for targeted capture sequencing. The patient donating TumorA was consented for next generation sequencing under a previously described protocol approved by the University of Michigan Institutional Review Board [41]. Targeted capture was performed using a custom designed probe panel with high density coverage of the HPV16 genome, the HPV18/33/35 L2/L1 regions, and over 200 HNSCC-related genes, which are detailed in Heft Neal et. al 2020.[19] Following library preparation and capture, the samples were sequenced on an Illumina NovaSEQ6000 or HiSEQ4000, respectively, with 300nt paired end run. Data was de-multiplexed and FastQ files were generated.

### Novel Integration Caller (SearcHPV):

The pipeline of SearcHPV has four main steps which are detailed below: (1) Alignment; (2) Genome fusion point calling; (3) Assembly; (4) HPV fusion point calling (Figure 1). The package is available on Github: https://github.com/mills-lab/SearcHPV.

**Alignment**—The customized reference genome used for alignment was constructed by catenating the HPV16 genome (from Papillomavirus Episteme (PAVE) database[20, 21]) and the human genome reference (1000 Genomes Reference Genome Sequence, hs37d5). We aligned paired-end reads from targeted capture sequencing against the customized reference genome using BWA mem aligner.[22] Then we performed an indel realignment by Picard Tools[23] and GATK[24]. Duplications were marked by Picard MarkDuplicates Tool[23] for the filtering in downstream steps.

**Genome Fusion Points Calling—**To identify the fusion points, we extracted reads with regions matched to HPV16 and filtered those reads to meet these criteria: (1) not secondary alignment; (2) mapping quality greater or equal than 50; (3) not duplicated. Genome fusion points were called by split reads (reads spanning both the human and HPV genomes) and the paired-end reads (reads with one end matched to HPV and the other matched the human genome) at the surrounding region (+/−300bp) (Figure 1A). The cut-off criteria for identifying the fusion points were based on empirical practice. We then clustered the integration sites within 100bp to avoid duplicated counting of integration events due to the stochastic nature of read mapping and structural variations.

**Assembly—**To construct longer sequence contigs from individual reals, we extracted supporting split reads and paired-end reads for local assembly from each integration event. Due to the library preparation methods we implemented for the targeted capture approach, some reads exhibited an insertion size less than $2 \times$ read length, resulting in overlapping read segments. For such events, we first merged these reads using PEAR[25] and then combined them with other individual reads to perform a local assembly by CAP3[26] (Figure 1).

**HPV Fusion Point Calling—**For each integration event, the assembly algorithm was able to report multiple contigs. We developed a procedure to evaluate and select contigs for each integration event to call HPV fusion point more precisely. First, we aligned the contigs against the human genome and HPV genome separately by BWA mem. If the contig met the following criteria, we marked it as high confidence:

1.   Has at least 10 supportive reads

2.   $10\% < \frac{matched\ length\ of\ the\ contig\ to\ HPV}{length\ of\ contig} < 95\%$

Then we separated the contigs we assembled into two classes: from left side (Contig A in Fig 1B) and from right side (Contig B in Fig 1B). For each class, if there were high confidence contigs in the class, we selected the contig with maximum length among them, otherwise we selected the contig with most supportive reads. For each insertion event, we reported one contig if it only had contigs from one side and we reported two contigs if it had contigs from both sides (Figure 1C). Finally, we identified the fusion points within HPV based on the alignment results of the selected contigs against the HPV genome. The bam/sam file processing in this pipeline was done by Samtools[22] and the analysis was performed with R 3.6.1[27] and Python.[28]

## RESULTS:

### SearcHPV pipeline:

To overcome the limitations of viral integration detection in WGS of detecting rare events, we performed HPV targeted capture sequencing which allows for deeper investigation of these events. Current bioinformatics pipelines available are not designed for this type of data so we developed a novel HPV integration detection tool for targeted capture sequencing data, which we termed "SearcHPV". Two HPV16+ HNSCC models, UM-SCC-47 and Patient derived xenograft (PDX)-294R as well as an HPV16+ HNSCC tumor, TumorA, were

subjected to targeted-capture based Illumina sequencing using a custom panel of probes spanning the entire HPV16 genome. The paired end reads then went through the four steps of analysis of SearcHPV: alignment to custom reference genome, genome fusion points calling, local assembly and HPV fusion point calling (Figure 1). Analysis of the integration sites in the models using our pipeline SearcHPV showed a high frequency of HPV16 integration with a total of six events in UM-SCC-47, ninety-eight in PDX-294R and eight in TumorA (Figure 2, Figure S3, Table S1–S3).

**Comparison to other integration callers and confirmation of integration sites:**

In addition to using SearcHPV, we used two previously developed integration callers, VirusFinder2 and VirusSeq to independently call integration events in UM-SCC-47, PDX-294R and TumorA (Figure 3, Tables S4–S6). We found that SearcHPV called HPV integration events at a much higher rate than either previous caller (Figure 3B). There were a large number of sites that were only identified by SearcHPV (n=82). In order to assess the accuracy of each caller, we performed PCR for PDX-294R and UM-SCC-47 on source genomic DNA followed by Sanger sequencing with primers spanning the HPV-human junction sites predicted by the callers. We tested all integration sites with sufficient sequence complexity for primer design (n=43), twenty-five of which were unique to SearcHPV and five of which were unique to VirusSeq. VirusFinder2 does not allow for local assembly of the integration junctions which rendered us unable to test these sites. UM-SCC-47 was also subjected to Oxford Nanopore GridION sequencing to provide additional supportive evidence of integration sites. We combined the information from PCR and Nanopore sequencing to interrogate a total of 44 integration sites and compared the conformation rates for each caller. (Figure 3C. S1, Table S7, S17). Sites unique to SearcHPV had a confirmation rate of 19/26 (73%). The confirmation rate of high confidence SearcHPV sites was higher than that for low confidence sites (25/32 (78%) versus 4/7 (57%)). In contrast, only 1/5 (20%) sites unique to VirusSeq could be confirmed.

To further compare the performance of SearcHPV and the other two callers, we expanded the sequencing requirements by applying them on whole exome sequencing data (WES) for UM-SCC-47 and PDX-294R, which were either previously generated by our lab [41,42] or were publicly available, respectively. VirusSeq did not report any integration results in either sample from the WES data. For UM-SCC-47, SearcHPV and VirusFinder2 both called one integration site. This site was reported by SearcHPV from targeted capture data. For PDX-294R, SearcHPV identified three integration sites while VirusFinder2 did not identify any sites. Two of three integration sites were also called by SearcHPV from targeted capture data and the other one was not covered in the targeted region of our targeted capture technology (Table S10–13). By examining the location of integration sites called from targeted capture sequencing for these two samples, we found that most (102/104) fell outside of the targeted region of WES, resulting in lower coverage of reads and insufficient evidence to identify the integration events (Table S14–16). Given this limitation of WES on capturing genome-wide HPV integration events, our approach was still more applicable on identifying HPV integration events than VirusSeq and VirusFinder2.

## Localization of integration sites:

We next examined the integration sites detected by SearcHPV. The six integration sites discovered in UM-SCC-47 were clustered on chromosome 3q28 within/near the cellular gene *TP63* and either had breakpoints within the HPV16 genes E1, E2 or L1. The integration sites fell within intron 10, intron 12 and exon 14. One additional integration site was 8.6 kb downstream of the *TP63* coding region.

For TumorA, six of eight integration sites were clustered on chromosome 9q34 within/near gene *TRAF2,* including one integration site that fell within *FBXW5* which was 15.8kb downstream of *TRAF2.* Among them, three integration sites fell within intron 5 of *TRAF2* and one mapped to intron 8.

Within PDX-294R, HPV16 integration sites were identified across 21 different chromosomes, occurring most frequently on chromosome 3. For the 98 integration events of PDX-294R, we identified 142 breakpoints in the HPV genome. The most frequently involved HPV genes were E1 (45/142 (32%)) and L1 (31/142 (22%)). Most of the integration sites mapped to within/near (<50 kb) a known cellular gene (89/98 (91%)). Of the sites that fell within a gene, the majority of integrations took place within an intronic region (3¾2 (78%)). Although the integration sites were scattered throughout the human genome, we saw examples of closely clustered sites around cancer-relevant genes, including *ZNF148* and *SNX4* on chromosome 3q21.2, *MYC* on chromosome 8q24.21 and *FOXN2* on chromosome 2p16.3.

**Association of integration sites and large-scale duplications—**We predicted that the complex integration sites we discovered in UM-SCC-47, PDX-294R and TumorA would be associated with large-scale structural alterations of the genome, such as rearrangements, deletions and duplications. To identify these alterations, we subjected UM-SCC-47, PDX-294R and TumorA to 10X linked-read sequencing. We generated over 1 billion reads for each sample (Table S8), with phase blocks (contiguous blocks of DNA from the same allele) of up to 28.9M, 3.8M and 15.3M bases in length for UM-SCC-47, PDX-294R and TumorA, respectively (Figure S2). This led to the identification of 444 high confidence large structural events in UM-SCC-47, 126 events in the PDX-294R model and 49 events in TumorA. We then performed integrated analysis with our SearcHPV results. There was a 130 kb duplication surrounding the integration events in *TP63* in UM-SCC-47 (Figure 4A). In PDX-294R, 38/98 (39%) integration sites were within a region that contained a large-scale duplication, while the other 50 integration events fell outside regions of large structural variation. This suggested that in this PDX model, 38/126 (30%) large structural events were potentially induced during HPV integration. For example, the clusters of integration events surrounding *ZNF148* and *SNX4*, *MYC*, as well as *FOXN2* were also associated with large genomic duplications (Figure 4B–C). For TumorA, large duplications were not observed within the surrounding region of the eight integration events (Figure 4E).

To further resolve the structure around the clusters of integration sites, we performed local assembly for UM-SCC-47 using Nanopore sequencing data (See Supplementary File, Figure 4F). The 60K-bp scaffold indicated a 15K-bp, twice amplified segment that matched against the human genome and a 7.5K-bp, twice amplified segment matched against HPV genome.

These segments were potentially amplified from a large 22.5K-bp focal genomic segment that has both human and HPV genomic components (Figure 4F, copy1–3) and then parts of one duplication were deleted resulting in the shorter segment in the middle (Figure 4F, copy2). These human segments and HPV segments were all bounded by identical or very near breakpoints. The integration sites on the human genome shown by the local assembly kept consistent with results from SearcHPV. Notably, within the focal HPV segments, an HPV-HPV junction structure was also identified showing an HPV internal rearrangement structure (Figure 4F, pink and yellow parts). This HPV internal rearrangement occurred twice and resulted in additional breakpoints on the HPV genome. The focal amplification structure resolved by local assembly from Nanopore sequencing confirmed the duplications predicted by 10X linked-read sequencing and indicated the association of HPV integrations and large-scaled duplications.

### Microhomology at junction sites:

Finally, to evaluate possible mechanisms of DNA repair-mediated integration, we examined the degree of sequence overlap between the genomes at junction sites that were covered by contigs. We saw three types of junction points: those with a gap of unmapped sequence between the human and HPV genomes, those that had a clean breakpoint between the genomes, and those with sequence that could be mapped to both genomes (Figure 5A). The majority (59%) of junction sites in the three samples had at least some degree of microhomology (Figure 5B–D). Integration sites with clean breaks (0 bp overlap) and 3 bp of overlap were the most frequently seen junctions in PDX-294R, but there was a wide range of levels seen. There was also a large number of junctions with gaps between the human and HPV genomes ranging from 1 – 54 bp long.

## Discussion

We developed a novel bioinformatics pipeline that we termed "SearcHPV" and show that it operated in a more accurate and efficient manner than existing pipelines on targeted capture sequencing data. The software also has the advantage of performing local contig assembly around the junction sites, which simplifies downstream confirmation experiments. We used our new caller to interrogate the integration sites found in two HNSCC models and one frozen HNSCC HPV+ sample, in order to compare the accuracy of our caller to the existing pipelines. We then evaluated the genomic effects of these integrations on a larger scale by 10X linked-reads sequencing and Oxford Nanopore sequencing to identify the role of HPV integration in driving structural variation in the tumor genome.

Using SearcHPV, we were able to investigate the HPV-human integration events present in UM-SCC-47, PDX-294R and TumorA. Importantly, UM-SCC-47 has been previously assessed for HPV integration by a variety of methods[8, 29–32], which we leveraged as ground truth knowledge to validate our integration caller. All previous studies were in agreement that HPV16 is integrated within the cellular gene *TP63*, although the exact number of sites and locations within the gene varied by study. In this study, SearcHPV also called HPV integration sites within *TP63*. We found integrations of E1, E2 and L1 within *TP63* intron 10, L1 within intron 12 and E2 within *TP63* exon 14. These integration sites were also

detected using DIPS-PCR[32] and/or WGS[8] with the exception of E1 into intron 10, which was unique to our caller and confirmed by direct PCR. It is possible that the integration sites detected in this sample represent multiple fragments of one larger integration site. There were additional sites called by other WGS studies that we did not detect (intron 9[8] and exon 7[31]), although it is possible that alternate clonal populations grew out due to different selective pressures in different laboratories. Nonetheless, the analysis clearly demonstrated that SearcHPV was able to detect a well-established HPV insertion site.

In contrast to UM-SCC-47, to our knowledge TumorA and PDX-294R have not been previously analyzed for viral-host integration sites and therefore represented a true discovery case. For TumorA, we identified a cluster of HPV integration sites within/near *TRAF2.* Interestingly, *TRAF2* was previously identified as a potential downstream effector of E6/E7 [43,44], and due to the role of *TRAF2* in regulating innate immunity, this gene may have a larger role in HPV16-mediated biology than previously recognized.

For PDX-294R, we identified widespread HPV integration sites throughout the host genome and also observed that 66% of integration sites were found within or near genes. This aligns with previous reports that integrations are detected in host genes more frequently than expected by chance.[2, 3, 7, 33] One particularly interesting cluster of integration events surrounded the cellular proto-oncogene *MYC.* Importantly, *MYC* has been identified as a potential hotspot for HPV integration[7, 34] and the junctions we detected in/near this gene had 2–4 bp of microhomology, potentially driving this observation. Accordingly, an HPV-integration related promoter duplication event, which may be expected to drive expression, would be consistent with a novel genetic mechanism to drive expression of this oncogene.

*TP63* has also been reported to be a hotspot for HPV integration, as it has been recorded in multiple samples besides UM-SCC-47.[3, 7, 35, 36] There is a high degree of microhomology between HPV16 and this gene. Given the high frequency of molecular alterations in the epidermal differentiation pathway (e.g. *NOTCH1/2*, *TP63* and *ZNF750*) in HPV+ HNSCCs, this data supports HPV integration as a pivotal mechanism of viral-driven oncogenesis in this model.[37]

HPV integration sites have been associated with structural variations in the human genome[3, 8, 37], which supports an additional genetic mechanism as to why HPV integration sites may often be detected adjacent to host cancer-related genes. These structural variation events are thought to be due to the rolling circle amplification that takes place at the integration breakpoint, leading to the formation of amplified segments of genomic sequence flanked by HPV segments.[8, 38] Our data are consistent with these previous reports in that approximately half of the integration events we discovered were associated with a large-scale amplification. It is unclear why only some integration sites were associated with structural variants, but it is possible that an alternative mechanism of integration occurred.[38] Notably, we resolved and identified an HPV-HPV junction that bounded in a large duplication segment and showed the possibility of HPV internal rearrangement to be involved in HPV integration events.

Importantly, this observation that HPV integration events tended to be enriched in cellular genes could result from multiple different mechanisms. Integration could occur preferentially in regions of open chromatin during cell replication and keratinocyte differentiation. Other potential mechanisms are: 1) that HPV integration is directed to specific host genes by homology, or 2) that HPV integration is random, but events that are advantageous for oncogenesis are clonally selected and expanded, implicating non-homology based DNA repair mechanisms. Therefore, to help resolve differences in the mechanism of integration, we assessed microhomology at the HPV-human junction points. The majority of breakpoints had some level of microhomology. The most frequent levels of overlap were 0 and 3 bp, which potentially implicates non-homologous end joining (NHEJ) in repair at these sites, since this pathway most frequently results in 0–5 bp of overlap.[39] There were also a number of junction sites that demonstrated a gap of inserted sequence between the HPV and human genomes. It has been described that during polymerase theta-mediated end joining (TMEJ), stretches of 3–30 bp are frequently inserted at the site of repair, possibly accounting for these sites.[40] However, given the relatively small number of events we examined, we expect that future analysis with our pipeline will help resolve the specific role of each DNA repair pathway in HPV-human fusion breakpoints.

Overall, our new HPV detection pipeline SearchHPV overcomes a gap in the field of viral-host integration analysis. While the performance of SearcHPV has only been examined on three samples, in the future, we expect that the application of this pipeline in large HPV+ cancer tissue cohorts will help advance our understanding of the potential oncogenic mechanisms associated with viral integration. With the emerging set of tools such as SearcHPV, we believe the field is now primed to make major advances in the understanding of HPV-driven pathogenesis, some of which may lead to the development of novel biomarkers and/or treatment paradigms.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## REFERENCES:

1. Gao G, Wang J, Kasperbauer JL, et al.Whole genome sequencing reveals complexity in both HPV sequences present and HPV integrations in HPV-positive oropharyngeal squamous cell carcinomas. BMC Cancer. 4112019;19(1):352. doi:10.1186/s12885-019-5536-1 [PubMed: 30975103]

2. 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. 3142017;8(11):17684–17699. doi:10.18632/oncotarget.15179 [PubMed: 28187443]

3. Parfenov M, Pedamallu CS, Gehlenborg N, et al.Characterization of HPV and host genome interactions in primary head and neck cancers. Proc Natl Acad Sci U S A. 10282014;111(43):15544–9. doi:10.1073/pnas.1416074111 [PubMed: 25313082]

4. Pinatti LM, Sinha HN, Brummel CV, et al.Association of human papillomavirus integration with better patient outcomes in oropharyngeal squamous cell carcinoma. Head Neck. 10192020;doi:10.1002/hed.26501

5. Tian R, Cui Z, He D, et al.Risk stratification of cervical lesions using capture sequencing and machine learning method based on HPV and human integrated genomic profiles. Carcinogenesis. 10162019;40(10):1220–1228. doi:10.1093/carcin/bgz094 [PubMed: 31102403]

6. McBride AA, Warburton A. The role of integration in oncogenic progression of HPV-associated cancers. PLoS Pathog. 42017;13(4):e1006211. doi:10.1371/journal.ppat.1006211 [PubMed: 28384274]

7. Bodelon C, Untereiner ME, Machiela MJ, Vinokurova S, Wentzensen N. Genomic characterization of viral integration sites in HPV-related cancers. Int J Cancer. 11012016;139(9):2001–11. doi:10.1002/ijc.30243 [PubMed: 27343048]

8. Akagi K, Li J, Broutian TR, et al.Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. Genome Res. 22014;24(2):185–99. doi:10.1101/gr.164806.113 [PubMed: 24201445]

9. Pinatti LM, Walline HM, Carey TE. Human Papillomavirus Genome Integration and Head and Neck Cancer. J Dent Res. 62018;97(6):691–700. doi:10.1177/0022034517744213 [PubMed: 29227715]

10. Luft F, Klaes R, Nees M, et al.Detection of integrated papillomavirus sequences by ligation-mediated PCR (DIPS-PCR) and molecular characterization in cervical cancer cells. Int J Cancer. 4012001;92(1):9–17. [PubMed: 11279600]

11. Klaes R, Woerner SM, Ridder R, et al.Detection of high-risk cervical intraepithelial neoplasia and cervical cancer by amplification of transcripts derived from integrated papillomavirus oncogenes. Cancer Res. Dec 151999;59(24):6132–6.

12. Wang Q, Jia P, Zhao Z. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. Plos One. 2013;8(5):e64465. doi:10.1371/journal.pone.0064465 [PubMed: 23717618]

13. Wang Q, Jia P, Zhao Z. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome Med. 2015;7(1):2. doi:10.1186/s13073-015-0126-6 [PubMed: 25699093]

14. Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics. 1152013;29(2):266–7. doi:10.1093/bioinformatics/bts665 [PubMed: 23162058]

15. Holmes A, Lameiras S, Jeannot E, et al.Mechanistic signatures of HPV insertions in cervical carcinomas. NPJ Genom Med. 2016;1:16004. doi:10.1038/npjgenmed.2016.4 [PubMed: 29263809]

16. Montgomery ND, Parker JS, Eberhard DA, et al.Identification of Human Papillomavirus Infection in Cancer Tissue by Targeted Next-generation Sequencing. Appl Immunohistochem Mol Morphol. 82016;24(7):490–5. doi:10.1097/PAI.0000000000000215 [PubMed: 26371432]

17. Morel A, Neuzillet C, Wack M, et al.Mechanistic Signatures of Human Papillomavirus Insertions in Anal Squamous Cell Carcinomas. Cancers (Basel). 11222019;11(12)doi:10.3390/cancers11121846

18. Nkili-Meyong AA, Moussavou-Boundzanga P, Labouba I, et al.Genome-wide profiling of human papillomavirus DNA integration in liquid-based cytology specimens from a Gabonese female population using HPV capture technology. Sci Rep. 262019;9(1):1504. doi:10.1038/s41598-018-37871-2 [PubMed: 30728408]

19. Heft Neal ME, Bhangale AD, Birkeland AC, et al.Prognostic Significance of Oxidation Pathway Mutations in Recurrent Laryngeal Squamous Cell Carcinoma. Cancers (Basel). 10222020;12(11)doi:10.3390/cancers12113081

20. NIAID. Papillomavirus Episteme. Bioinformatics and Computational Biosciences Branch. 2020. https://pave.niaid.nih.gov/

21. Van Doorslaer K, Li Z, Xirasagar S, et al.The Papillomavirus Episteme: a major update to the papillomavirus sequence database. Nucleic Acids Res. 142017;45(D1):D499–D506. doi:10.1093/nar/gkw879 [PubMed: 28053164]

22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 7152009;25(14):1754–60. doi:10.1093/bioinformatics/btp324 [PubMed: 19451168]

23. Institute BPicard toolkit. Broad Institute GitHub Repository. 2019;

24. McKenna A, Hanna M, Banks E, et al.The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 92010;20(9):1297–303. doi:10.1101/gr.107524.110 [PubMed: 20644199]

25. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. 312014;30(5):614–20. doi:10.1093/bioinformatics/btt593 [PubMed: 24142950]

26. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 91999;9(9):868–77. doi:10.1101/gr.9.9.868 [PubMed: 10508846]

27. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2019;

28. Van Rossum G, Drake FL. Python 3 Reference Manual: Python Documentation Manual Part 2. CreateSpace Independent Publishing Platform. 2009;

29. Khanal S, Shumway BS, Zahin M, et al.Viral DNA integration and methylation of human papillomavirus type 16 in high-grade oral epithelial dysplasia and head and neck squamous cell carcinoma. Oncotarget. 7132018;9(54):30419–30433. doi:10.18632/oncotarget.25754 [PubMed: 30100997]

30. Myers JE, Guidry JT, Scott ML, et al.Detecting episomal or integrated human papillomavirus 16 DNA using an exonuclease V-qPCR-based assay. Virology. 112019;537:149–156. doi:10.1016/j.virol.2019.08.021 [PubMed: 31493653]

31. Olthof NC, Huebbers CU, Kolligs J, et al.Viral load, gene expression and mapping of viral integration sites in HPV16-associated HNSCC cell lines. Int J Cancer. 312015;136(5):E207–18. doi:10.1002/ijc.29112 [PubMed: 25082736]

32. Walline HM, Goudsmit CM, McHugh JB, et al.Integration of high-risk human papillomavirus into cellular cancer-related genes in head and neck cancer cell lines. Head Neck. 52017;39(5):840–852. doi:10.1002/hed.24729 [PubMed: 28236344]

33. Hu Z, Zhu D, Wang W, et al.Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. Nat Genet. 22015;47(2):158–63. doi:10.1038/ng.3178 [PubMed: 25581428]

34. Ferber MJ, Thorland EC, Brink AA, et al.Preferential integration of human papillomavirus type 18 near the c-myc locus in cervical carcinoma. Oncogene. 10 16 2003;22(46):7233–42. doi:10.1038/sj.onc.12070061207006 [pii] [PubMed: 14562053]

35. Schmitz M, Driesch C, Jansen L, Runnebaum IB, Durst M. Non-random integration of the HPV genome in cervical cancer. Plos One. 2012;7(6):e39632. doi:10.1371/journal.pone.0039632PONE-D-12-09523 [pii] [PubMed: 22761851]

36. Walline HM, Komarck CM, McHugh JB, et al.Genomic Integration of High-Risk HPV Alters Gene Expression in Oropharyngeal Squamous Cell Carcinoma. Mol Cancer Res. 102016;14(10):941–952. doi:10.1158/1541-7786.MCR-16-0105 [PubMed: 27422711]

37. Cancer Genome Atlas N.Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 1292015;517(7536):576–82. doi:10.1038/nature14129 [PubMed: 25631445]

38. Groves IJ, Coleman N. Human papillomavirus genome integration in squamous carcinogenesis: what have next-generation sequencing studies taught us?The Journal of pathology. 52018;245(1):9–18. doi:10.1002/path.5058 [PubMed: 29443391]

39. Pannunzio NR, Li S, Watanabe G, Lieber MR. Non-homologous end joining often uses microhomology: implications for alternative end joining. DNA Repair (Amst). 52014;17:74–80. doi:10.1016/j.dnarep.2014.02.006 [PubMed: 24613510]

40. Carvajal-Garcia J, Cho JE, Carvajal-Garcia P, et al.Mechanistic basis for microhomology identification and genome scarring by polymerase theta. Proc Natl Acad Sci U S A. 4142020;117(15):8476–8485. doi:10.1073/pnas.1921791117 [PubMed: 32234782]

41. Shuman AG, Gornick MC, Brummel C, Kent M, Spector-Bagdady K, Biddle E, et al.Patient and Provider Perspectives Regarding Enrollment in Head and Neck Cancer ResearchOtolaryngology--head and neck surgery: official journal of American Academy of Otolaryngology-Head and Neck Surgery2020;162:73–78. [PubMed: 31818193]

42. Liu J, Pan S, Hsieh MH, Ng N, Sun F, Wang T, et al.Targeting Wnt-driven cancer through the inhibition of Porcupine by LGK974. Proceedings of the National Academy of Sciences of the United States of America2013;110:20224–20229. [PubMed: 24277854]

43. Poirson J, Biquand E, Straub M-L, Cassonnet P, Nominé Y, Jones L, et al.Mapping the interactome of HPV E6 and E7 oncoproteins with the ubiquitin-proteasome system. The FEBS journal2017;284:3171–3201. [PubMed: 28786561]

44. Thompson DA, Zacny V, Belinsky GS, Classon M, Jones DL, Schlegel R, et al.The HPV E7 oncoprotein inhibits tumor necrosis factor alpha-mediated apoptosis in normal human fibroblasts. Oncogene2001;20:3629–3640. [PubMed: 11439326]

45. Walline HM, Goudsmit CM, McHugh JB, Tang AL, Owen JH, Teh BT, et al.Integration of high-risk human papillomavirus into cellular cancer-related genes in head and neck cancer cell lines. Head & neck2017;39:840–852. [PubMed: 28236344]
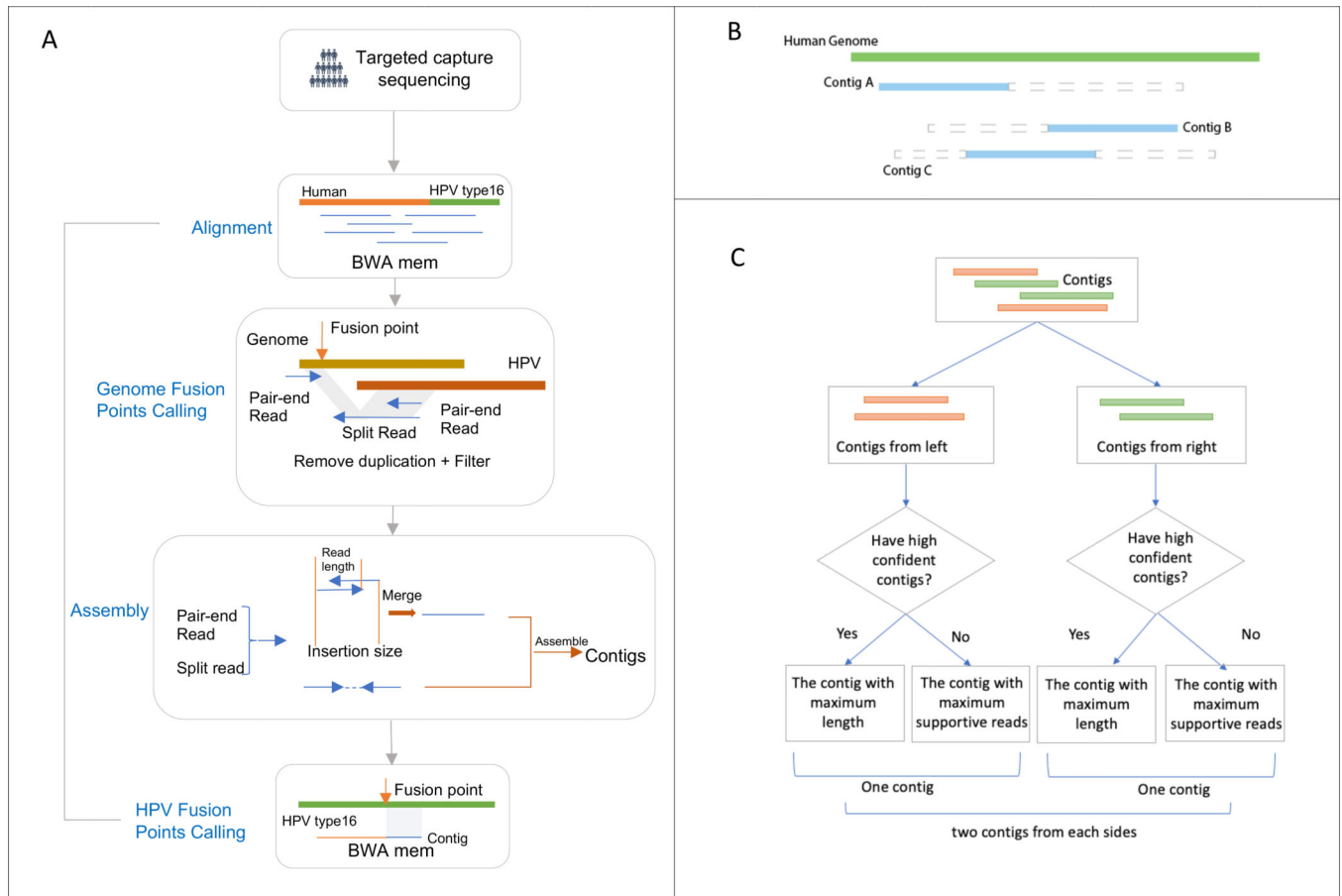
**Figure 1: Workflow of SearcHPV.**
(A) Paired-end reads from targeted capture sequencing were aligned to a catenated Human-HPV reference genome. After removing duplication and filter, fusion points were identified by split reads and pair-end reads. Informative reads were extracted for local assembly. Reads pairs that have overlaps were merged first before assembly. Assembled contigs were aligned to the HPV genome to identify the breakpoints on HPV. (B) Contigs were divided into two classes. Blue solid triangle demonstrates the matched region of the contig. Grey dashed triangle demonstrates the clipped region of the contig. Contig A would be assigned to the left group and Contig B would be assigned to the right group. Contig C would be randomly assigned to the left or right group. (C) Workflow for the contig selection procedures for fusion point with multiple candidates contigs. For each fusion point, we report at least one contig and at most two contigs representing two directions.
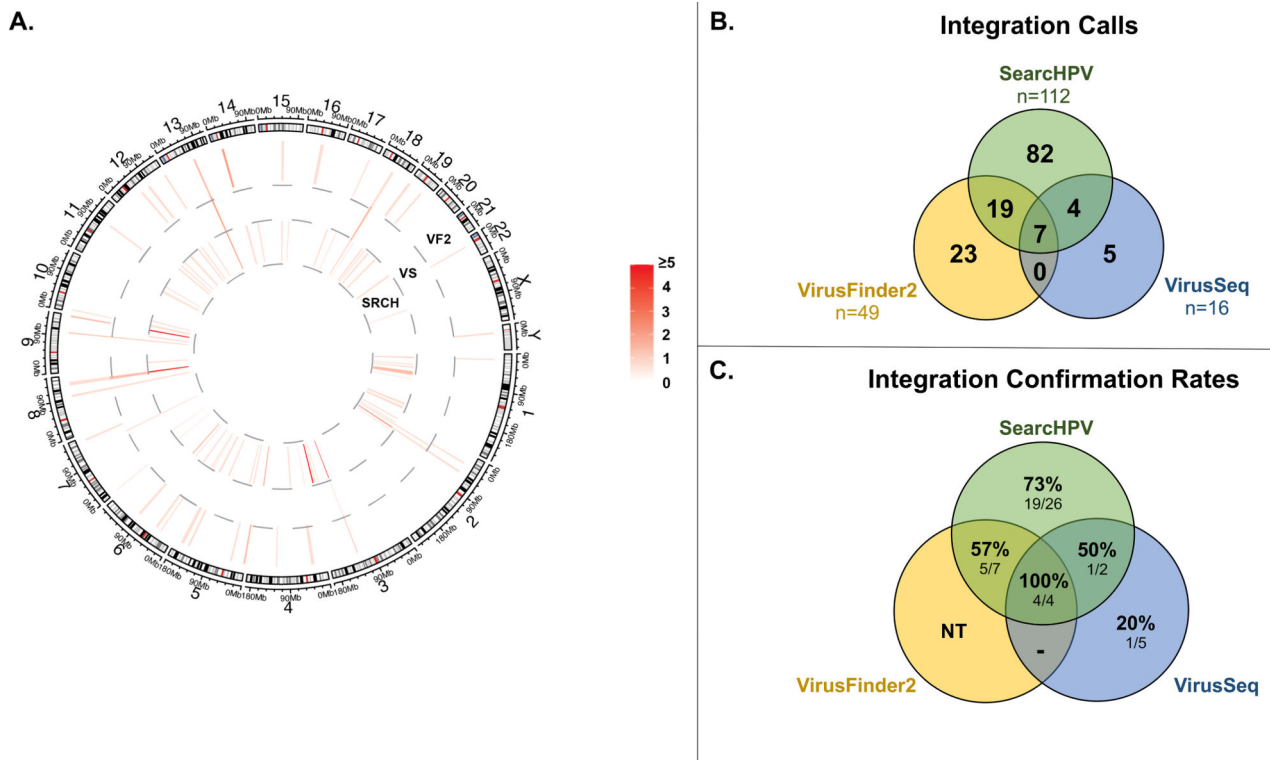
**Figure 2: Distribution of breakpoints in the human and HPV genomes called by SearcHPV.**
(A-C) Results for PDX-294R. (A) Links of breakpoints in the human and HPV16 genomes for PDX-294R. (B) Quantification of breakpoint calls in human genes for PDX-294R. (C) Quantification of breakpoints calls in the HPV16 genes for PDX-294R. (D-F) As described in A-C for UM-SCC-47. (G-I) As described in A-C for 4840 TumorA

**Figure 3: Comparison of integration sites called by SearcHPV, VirusSeq and VirusFinder2 in three samples.**
(A) Each bar denotes integration sites within the region. The colormap shows the count of the integration sites. (B) Number of integration sites called by each program. Integration sites from VirusSeq and VirusFinder2 were clustered within 100bp to keep consistent with SearcHPV. (C) PCR and Nanopore confirmation rate for a subset of (B) that were chosen to assess accuracy using both PCR and Nanopore sequencing where available.

If there is at least one split read from Nanopore sequencing data supporting an integration site, the integration site was regarded as validated by Nanopore sequencing. An integration site was counted as confirmed if it was validated by PCR or Nanopore sequencing.
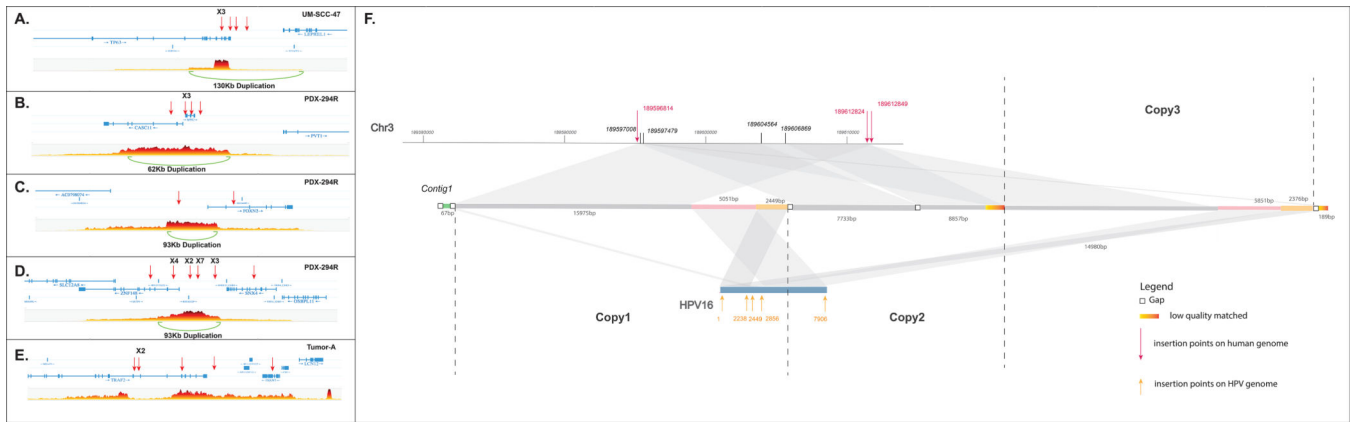
**Figure 4: Genomic duplications associated with HPV integration.**
(A) UM-SCC-47. (B-D) PDX-294R. (E) TumorA. Red arrows indicate integration sites. Each plot shows the number of overlapping barcodes observed in sequencing reads of that region. (F) Local assembly around the HPV integration sites in UM-SCC-47 using Nanopore sequencing data. The scaffold mapped to different region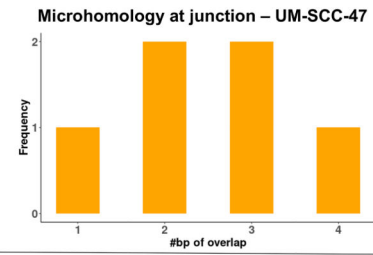s was marked by different colors. Gray: match to human genome reference. Green, pink and yellow: match to HPV genome. Potential duplications were marked by the same color.
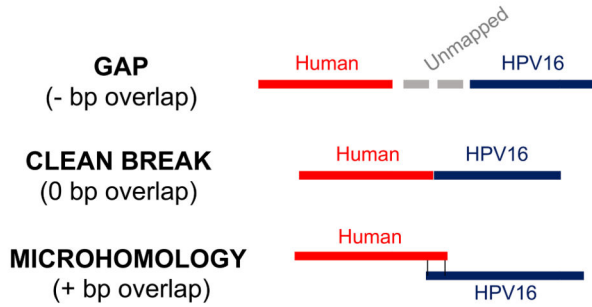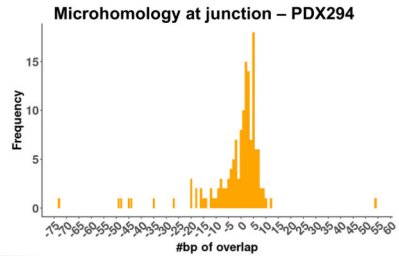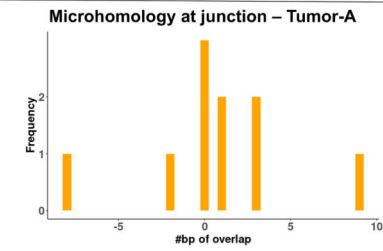
**Figure 5: Microhomology at junction points.**
(A) The three types of junction points. (B) Level of microhomology (in bp) in UM-SCC-47.
(C) Level of microhomology (in bp) in PDX-294R. (D) Level of microhomology (in bp) in
TumorA. Junctions with a gap are shown as negative numbers.