



# SearchHPV: A Novel Approach to Identify and Assemble Human Papillomavirus–Host Genomic Integration Events in Cancer

Lisa M. Pinatti, BS, PhD<sup>1,2</sup>; Wenjin Gu, MS<sup>3</sup>; Yifan Wang, PhD<sup>4</sup>; Ahmed Elhossiny, BS<sup>3</sup>; Apurva D. Bhangale, MS<sup>2</sup>; Collin V. Brummel, BA<sup>2</sup>; Thomas E. Carey, PhD<sup>1,2,5,6</sup>; Ryan E. Mills, PhD <sup>3,4</sup>; and J. Chad Brenner, PhD <sup>2,5,6</sup>

**BACKGROUND:** Human papillomavirus (HPV) is a well-established driver of malignant transformation at 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 discovering viral-host breakpoints from low-read-coverage sequencing data. **METHODS:** To overcome this limitation, the authors developed SearchHPV, a new HPV detection pipeline based on targeted capture technology, and applied the algorithm to targeted capture data. They performed an integrated analysis of SearchHPV-defined breakpoints with genome-wide linked-read sequencing to identify potential HPV-related structural variations. **RESULTS:** Through an analysis of HPV+ models, the authors showed that SearchHPV detected HPV-host integration sites with a higher sensitivity and specificity than 2 other commonly used HPV detection callers. SearchHPV uncovered HPV integration sites adjacent to known cancer-related genes, including *TP63*, *MYC*, and *TRAF2*, and near regions of large structural variation. The authors further validated the junction contig assembly feature of SearchHPV, which helped to accurately identify viral-host junction breakpoint sequences. They found that viral integration occurred through a variety of DNA repair mechanisms, including nonhomologous end joining, alternative end joining, and microhomology-mediated repair. **CONCLUSIONS:** In summary, SearchHPV is a new optimized tool for the accurate detection of HPV-human integration sites from targeted capture DNA sequencing data. *Cancer* 2021;127:3531–3540. © 2021 American Cancer Society.

**KEYWORDS:** bioinformatics, DNA sequence analysis, genomics, papillomavirus infections, squamous cell carcinoma, virus integration.

## INTRODUCTION

Human papillomavirus (HPV) is a well-established driver of malignant transformation in a number of cancers, including head and neck squamous cell carcinoma (HNSCC). Although HPV genomic integration is not a normal event in the lifecycle of HPV, it is frequently reported in HPV+ cancers,<sup>1,4</sup> and it may be a contributor to oncogenesis. In cervical cancer, HPV integration increases in incidence during progression from the stages of cervical intraepithelial neoplasia I/II, cervical intraepithelial neoplasia III, and invasive cancer development.<sup>5</sup> 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, and this leads to an increase in genetic instability.<sup>6</sup> HPV integration occurs within/near cellular genes more often than expected by chance<sup>7</sup> and has been reported to be associated with structural variations.<sup>8</sup> 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 the generation of hybrid viral-host fusion transcripts.<sup>9</sup>

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<sup>10</sup> and amplification of papillomavirus oncogene transcripts,<sup>11</sup> 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 technologies overcome these limitations. Previous groups have assessed HPV integration within HNSCC tumors in The Cancer Genome Atlas and within cell lines by whole genome sequencing (WGS).<sup>2,3,8</sup> A variety of viral integration detection tools, such as VirusFinder<sup>12,13</sup> and VirusSeq,<sup>14</sup>

**Corresponding Author:** J. Chad Brenner, PhD, University of Michigan, 9301B MSR33, 1150 W Medical Center Dr, Ann Arbor, MI 48109 (chadbren@umich.edu).

<sup>1</sup>Cancer Biology Program, Program in the Biomedical Sciences, Rackham Graduate School, University of Michigan, Ann Arbor, Michigan; <sup>2</sup>Department of Otolaryngology/Head and Neck Surgery, University of Michigan, Ann Arbor, Michigan; <sup>3</sup>Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan; <sup>4</sup>Department of Human Genetics, University of Michigan, Ann Arbor, Michigan; <sup>5</sup>Rogel Cancer Center, Michigan Medicine, Ann Arbor, Michigan; <sup>6</sup>Department of Pharmacology, University of Michigan, Ann Arbor, Michigan

The first 2 authors contributed equally to this article.

The last 2 authors contributed equally to this article.

We thank the University of Michigan Advanced Genomics Core for performing the targeted capture sequencing and 10X linked-read sequencing. We also thank Dr. Tom Wilson for discussions of the data.

Additional supporting information may be found in the online version of this article.

**DOI:** 10.1002/cncr.33691, **Received:** January 21, 2021; **Revised:** April 15, 2021; **Accepted:** May 3, 2021, **Published online** June 23, 2021 in Wiley Online Library (wileyonlinelibrary.com)

have been developed for WGS data. However, these strategies are designed for a broad range of virus types and require whole genomes to be sequenced at uniform coverage, and this 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.<sup>5,15-18</sup> This strategy allows for better coverage of integration sites than an untargeted approach such as WGS but requires sensitive and accurate viral-human fusion detection bioinformatic tools, which the field has been lacking. In our laboratory, 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, and this 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 HPV-16+ UM-SCC-47 cell line,<sup>19</sup> patient-derived xenograft 294R (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.<sup>20</sup> Targeted capture was performed with a custom-designed probe panel with high-density coverage of the HPV-16 genome, the HPV-18/33/35 L2/L1 regions, and more than 200 HNSCC-related genes, which have been detailed by Heft Neal et al.<sup>21</sup> After library preparation and capture, the samples were sequenced on an Illumina NovaSEQ6000 or HiSEQ4000, respectively, with a 300-nt paired-end run. Data were demultiplexed, and FastQ files were generated.

### Novel Integration Caller (SearchHPV)

The pipeline of SearchHPV has 4 main steps, which are detailed next: 1) alignment, 2) genome fusion point calling, 3) assembly, and 4) HPV fusion point calling (Fig. 1). The package is available on GitHub (<https://github.com/mills-lab/SearchHPV>).

### Alignment

The customized reference genome used for alignment was constructed by catenation of the HPV-16 genome (from the Papillomavirus Episteme database<sup>22,23</sup>) 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 with the Burrows-Wheeler Aligner (BWA) mem aligner.<sup>24</sup> Then, we performed an indel realignment with Picard Tools<sup>25</sup> and the Genome Analysis Toolkit.<sup>26</sup> Duplications were marked by the Picard MarkDuplicates tool<sup>25</sup> for the filtering in downstream steps.

### Genome fusion point calling

To identify the fusion points, we extracted reads with regions matched to HPV-16 and filtered those reads to meet these criteria: 1) not secondary alignment, 2) mapping quality greater or equal than 50, and 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 to the human genome) in the surrounding region ( $\pm 300$  bp; Fig. 1A). The cutoff criteria for identifying the fusion points were based on empirical practice. We then clustered the integration sites within 100 bp to avoid duplicated counting of integration events due to the stochastic nature of read mapping and structural variations.

### Assembly

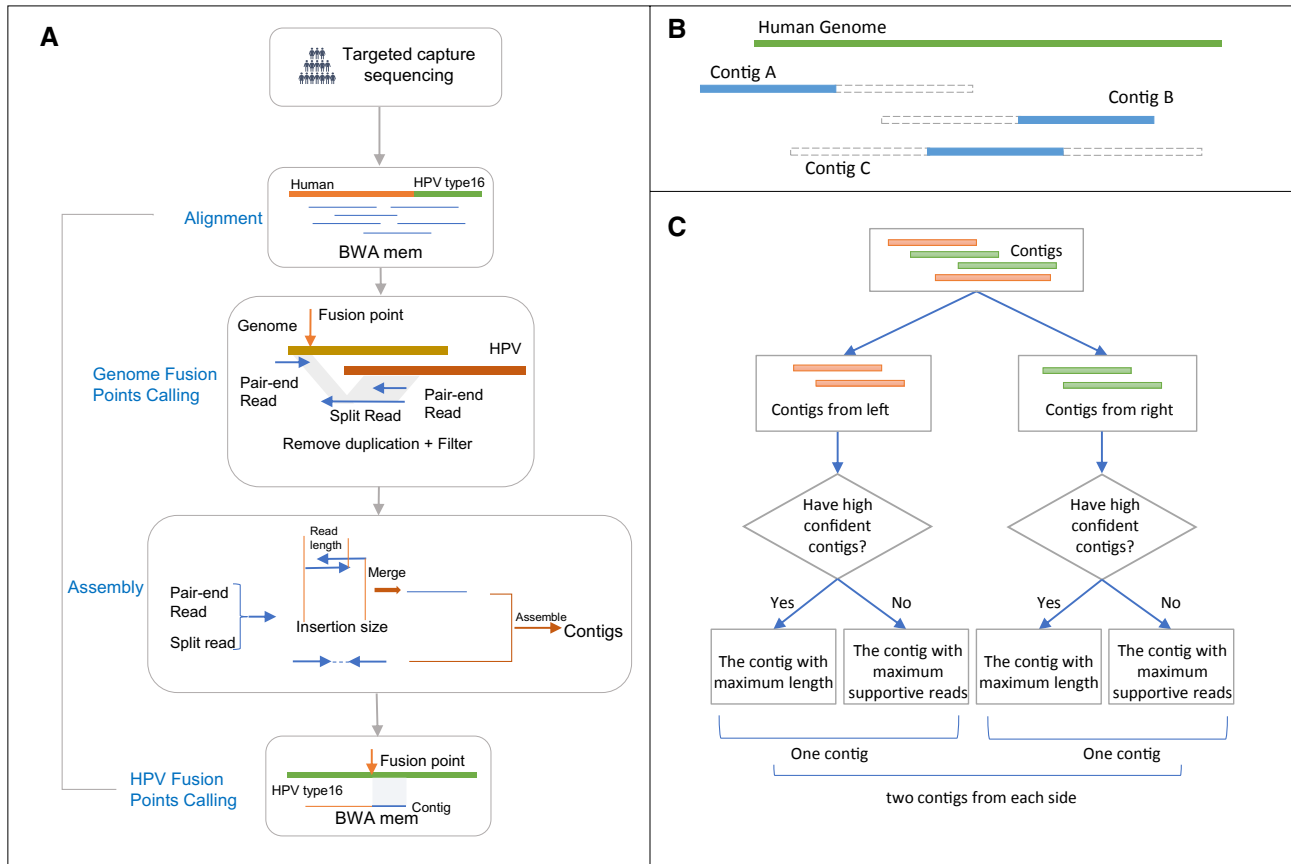
To construct longer sequence contigs from individual reads, we extracted supporting split reads and paired-end reads for local assembly from each integration event. Because of the library preparation methods that we implemented for the targeted capture approach, some reads exhibited an insertion size less than 2 times the read length, and this resulted in overlapping read segments. For such events, we first merged these reads with Paired-End Read Merger<sup>27</sup> and then combined them with other individual reads to perform a local assembly by CAP3<sup>28</sup> (Fig. 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 the HPV fusion point more precisely. First, we aligned the contigs against the human genome and the HPV genome separately by BWA mem. If the contig met the following criteria, we marked it as high confidence:

1. Had at least 10 supportive reads
2.  $10\% < \frac{\text{Matched length of the contig to HPV}}{\text{Length of the contig}} < 95\%$

Then, we separated the contigs that we assembled into 2 classes: from the left side (contig A in Fig. 1B) and from the right side (contig B in Fig. 1B). For each class, if there



**Figure 1.** Workflow of SearchHPV. (A) Paired-end reads from targeted capture sequencing were aligned to a catenated human-HPV reference genome. After removal of the duplication and filter, fusion points were identified by split reads and paired-end reads. Informative reads were extracted for local assembly. Read pairs that had 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 2 classes. A blue solid rectangle demonstrates the matched region of the contig. A gray dashed rectangle 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 points with multiple candidate contigs. For each fusion point, we report at least 1 contig and at most 2 contigs representing 2 directions. BWA, Burrows-Wheeler Aligner; GATK, Genome Analysis Toolkit; HPV indicates human papillomavirus.

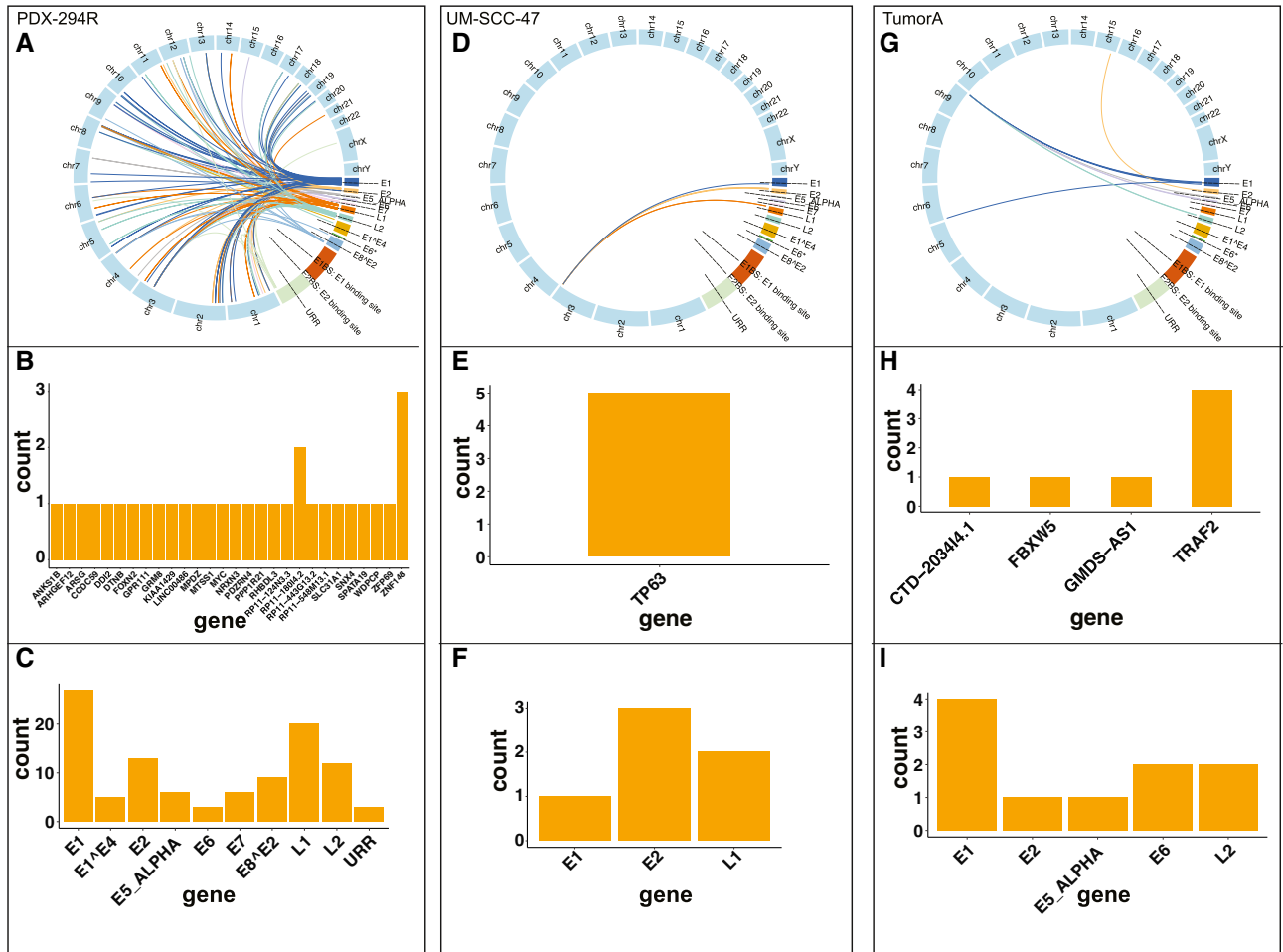
were high-confidence contigs in the class, we selected the contig with the maximum length among them; otherwise, we selected the contig with the most supportive reads. For each insertion event, we reported 1 contig if it had contigs from only 1 side, and we reported 2 contigs if it had contigs from both sides (Fig. 1C). Finally, we identified the fusion points within HPV on the basis of the alignment results of the selected contigs against the HPV genome. The bam/sam file processing in this pipeline was performed with SAMtools,<sup>24</sup> and the analysis was performed with R 3.6.1<sup>29</sup> and Python.<sup>30</sup>

## RESULTS

### SearchHPV Pipeline

To overcome the limitations of viral integration detection in WGS for detecting rare events, we performed HPV

targeted capture sequencing, which allows for deeper investigation of these events. The bioinformatics pipelines currently 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 *SearchHPV*. Two HPV-16+ HNSCC models, UM-SCC-47 and patient-derived xenograft 294R (PDX-294R), as well as an HPV-16+ HNSCC tumor, TumorA, were subjected to targeted capture-based Illumina sequencing using a custom panel of probes spanning the entire HPV-16 genome. The paired-end reads then went through the 4 steps of analysis of SearchHPV: alignment to a custom reference genome, genome fusion point calling, local assembly, and HPV fusion point calling (Fig. 1). An analysis of the integration sites in the models using our pipeline SearchHPV showed a high frequency of HPV-16



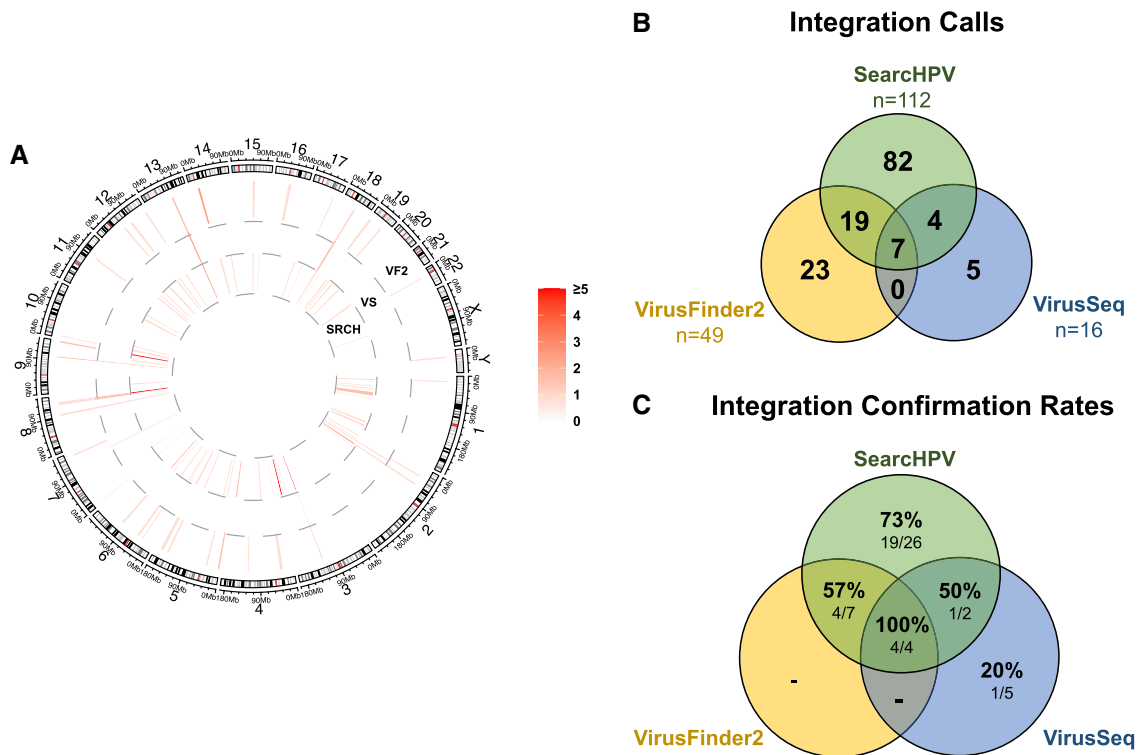
**Figure 2.** Distribution of breakpoints in the human and HPV genomes called by SearchHPV. (A-C) Results for PDX-294R: (A) links of breakpoints in the human and HPV-16 genomes for PDX-294R, (B) quantification of breakpoint calls in human genes for PDX-294R, and (C) quantification of breakpoint calls in the HPV-16 genes for PDX-294R. (D-F) As described in panels A-C for UM-SCC-47. (G-I) As described in panels A-C for 4840 TumorA. HPV indicates human papillomavirus; PDX-294R, patient-derived xenograft 294R.

integration with a total of 6 events in UM-SCC-47, 98 in PDX-294R, and 8 in TumorA (Fig. 2, Supporting Fig. 3, and Supporting Tables 1-3).

**Comparison With Other Integration Callers and Confirmation of Integration Sites**

In addition to using SearchHPV, we used 2 previously developed integration callers, VirusFinder2 and VirusSeq, to independently call integration events in UM-SCC-47, PDX-294R, and TumorA (Fig. 3 and Supporting Tables 4-6, and 18). We found that SearchHPV called HPV integration events at a much higher rate than either previous caller (Fig. 3B). There were a large number of sites that were identified only by SearchHPV (n = 82). To assess the accuracy of each caller, we performed PCR for PDX-294R and UM-SCC-47 on source genomic DNA,

which was followed by Sanger sequencing with primers spanning the HPV-human junction sites predicted by the callers (Supporting Table 9). We tested all integration sites with sufficient sequence complexity for primer design (n = 43), 25 of which were unique to SearchHPV and 5 of which were unique to VirusSeq. VirusFinder2 does not allow for local assembly of the integration junctions, and this 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 (Fig. 3C, Supporting Fig. 1, and Supporting Tables 7 and 17). Sites unique to SearchHPV had a confirmation rate of 73% (19 of 26).



**Figure 3.** Comparison of integration sites called by SearchHPV, VirusSeq, and VirusFinder2 in 3 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 100 bp to be kept consistent with SearchHPV. (C) PCR and Nanopore confirmation rate for a subset of panel B that was chosen to assess accuracy with both PCR and Nanopore sequencing where available. If there was at least 1 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. PCR indicates polymerase chain reaction; SRCH, SearchHPV; VF2, VirusFinder2; VS, VirusSeq.

The confirmation rate of high-confidence SearchHPV sites was higher than that of low-confidence sites (25 of 32 [78%] vs 4 of 7 [57%]). In contrast, only 1 of 5 sites (20%) unique to VirusSeq could be confirmed.

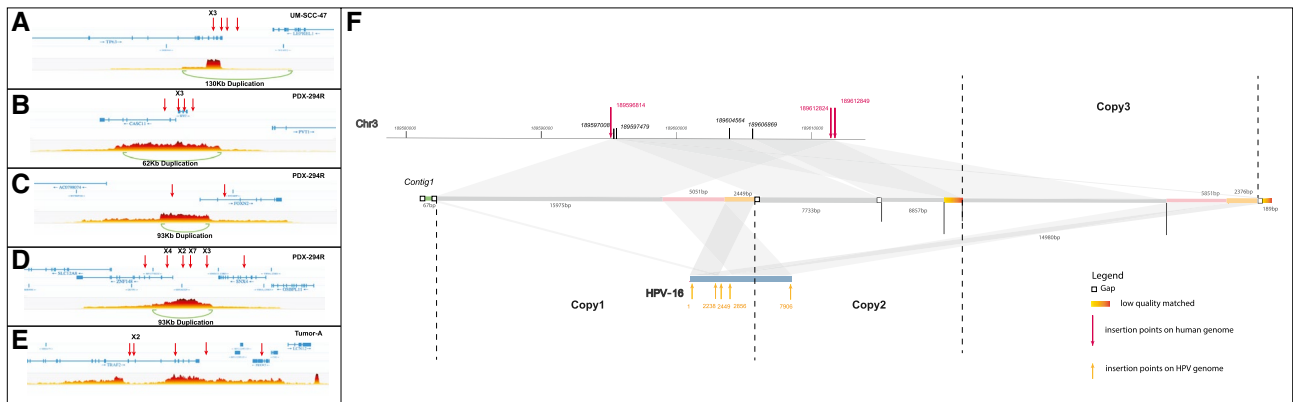
To further compare the performance of SearchHPV and the other 2 callers, we expanded the sequencing requirements by applying them to whole exome sequencing (WES) data for UM-SCC-47 and PDX-294R, which either were previously generated by our laboratory<sup>20,31</sup> or were publicly available, respectively. VirusSeq did not report any integration results for either sample from the WES data. For UM-SCC-47, SearchHPV and VirusFinder2 both called 1 integration site. This site was reported by SearchHPV from targeted capture data. For PDX-294R, SearchHPV identified 3 integration sites, whereas VirusFinder2 did not identify any sites. Two of the 3 integration sites were also called by SearchHPV from targeted capture data, and the other one was not covered in the targeted region of our targeted capture technology (Supporting Tables 10-13).

By examining the location of integration sites called from targeted capture sequencing for these 2 samples, we found that most (102 of 104) fell outside the targeted region of WES, and this resulted in lower coverage of reads and insufficient evidence to identify the integration events (Supporting Tables 14-16). Given this limitation of WES for capturing genome-wide HPV integration events, our approach was still more applicable for identifying HPV integration events than VirusSeq and VirusFinder2.

#### Localization of Integration Sites

We next examined the integration sites detected by SearchHPV. The 6 integration sites discovered in UM-SCC-47 were clustered on chromosome 3q28 within/near the cellular gene *TP63* and had breakpoints within the HPV-16 gene 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.





**Figure 4.** Genomic duplications associated with HPV integration: (A) UM-SCC-47, (B-D) PDX-294R, and (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 from Nanopore sequencing data. The scaffold mapped to different regions is marked by different colors: gray for a match to the human genome reference and green, pink, and yellow for a match to the HPV genome. Potential duplications are marked by the same color. HPV indicates human papillomavirus; PDX-294R, patient-derived xenograft 294R.

For TumorA, 6 of 8 integration sites were clustered on chromosome 9q34 within/near gene *TRAF2*; this included 1 integration site that fell within *FBXW5*, which was 15.8 kb downstream of *TRAF2*. Among them, 3 integration sites fell within intron 5 of *TRAF2*, and 1 mapped to intron 8.

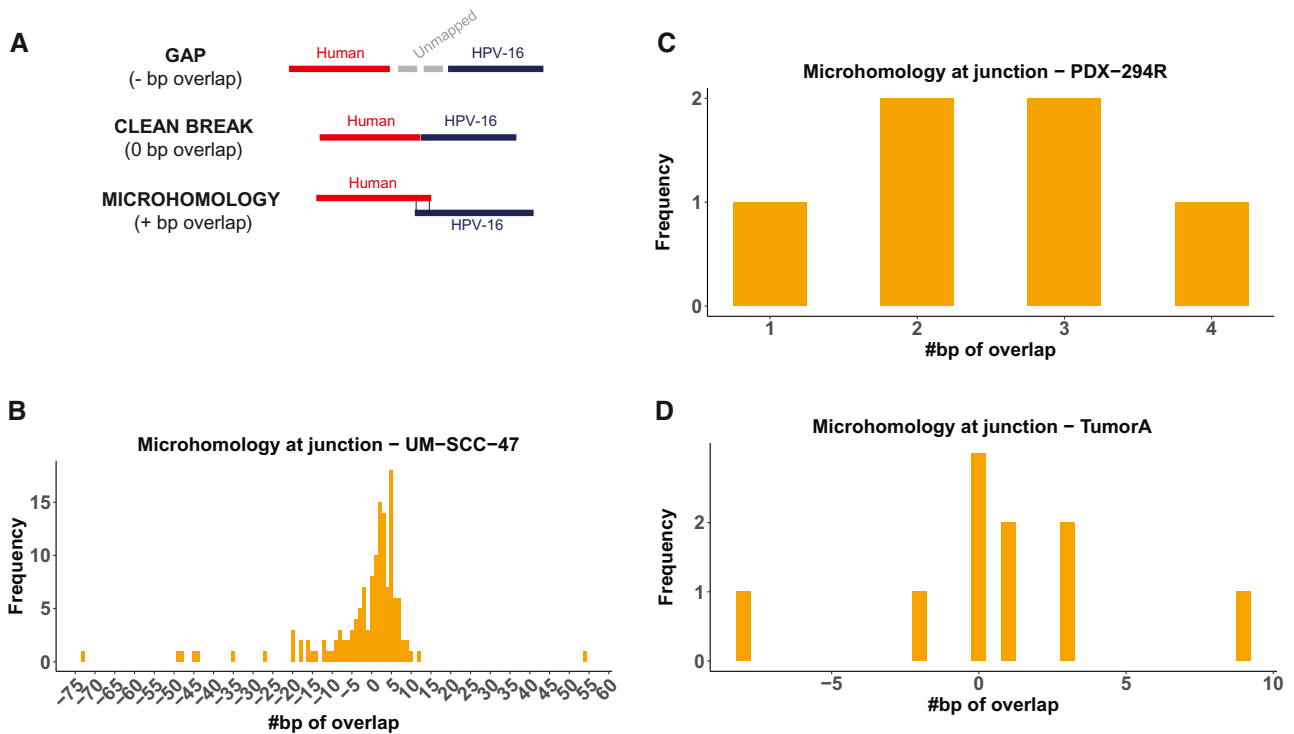
Within PDX-294R, HPV-16 integration sites were identified across 21 different chromosomes, and they occurred most frequently on chromosome 3. For the 98 integration events of PDX-294R, we identified 142 breakpoints in the HPV genome. The most frequently involved HPV genes were E1 (45 of 142 [32%]) and L1 (31 of 142 [22%]). Most of the integration sites mapped to within/near (<50 kb) a known cellular gene (89 of 98 [91%]). For the sites that fell within a gene, the majority of integrations took place within an intronic region (33 of 42 [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 more than 1 billion reads for each sample (Supporting Table 8), with phase blocks (contiguous blocks of DNA from the same allele) up to 28.9M, 3.8M, and 15.3M bases in length for UM-SCC-47, PDX-294R, and TumorA, respectively (Supporting Fig. 2). 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 an integrated analysis with our SearchHPV results. There was a 130-kb duplication surrounding the integration events in *TP63* in UM-SCC-47 (Fig. 4A). In PDX-294R, 38 of 98 integration sites (39%) were within a region that contained a large-scale duplication, whereas the other 50 integration events fell outside regions of large structural variation. This suggested that in this PDX model, 38 of 126 large structural events (30%) were potentially induced during HPV integration. For example, the clusters of integration events surrounding *ZNF148* and *SNX4*, *MYC*, and *FOXN2* were also associated with large genomic duplications (Fig. 4B,C). For TumorA, large duplications were not observed within the surrounding region of the 8 integration events (Fig. 4E).

To further resolve the structure around the clusters of integration sites, we performed local assembly for UM-SCC-47 by using Nanopore sequencing data (see the supporting information and Fig. 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 that matched against the HPV genome. These segments were potentially amplified from



**Figure 5.** Microhomology at junction points. (A) Three types of junction points. (B) Level of microhomology (in base pairs) in UM-SCC-47. (C) Level of microhomology (in base pairs) in PDX-294R. (D) Level of microhomology (in base pairs) in TumorA. Junctions with a gap are shown as negative numbers. HPV indicates human papillomavirus; PDX-294R, patient-derived xenograft 294R.

a large 22.5K-bp focal genomic segment that has both human and HPV genomic components (Fig. 4E, copies 1-3); then, parts of 1 duplication were deleted, and this resulted in the shorter segment in the middle (Fig. 4E, copy 2). 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 were consistent with results from SearchHPV. Notably, within the focal HPV segments, an HPV-HPV junction structure was also identified that showed an HPV internal rearrangement structure (Fig. 4E, 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 3 types of junction points: those with a gap of unmapped sequence between the human and HPV genomes, those with a clean breakpoint between the genomes, and those with a sequence that could be mapped to both genomes (Fig. 5A). The majority of junction sites (59%) in the 3 samples had at least some degree of microhomology (Fig. 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, which ranged in length from 1 to 54 bp.

#### DISCUSSION

We have developed a novel bioinformatics pipeline that we have termed *SearchHPV*, and we have shown that it operates 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 2

HNSCC models and 1 frozen HNSCC HPV+ sample to compare the accuracy of our caller with that of the existing pipelines. We then evaluated the genomic effects of these integrations on a larger scale by 10X linked-read sequencing and Oxford Nanopore sequencing to identify the role of HPV integration in driving structural variation in the tumor genome.

Using SearchHPV, 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,<sup>8,19,32-34</sup> which we leveraged as ground-truth knowledge to validate our integration caller. All previous studies were in agreement that HPV-16 is integrated within the cellular gene *TP63*, although the exact number of sites and locations within the gene varied by study. In this study, SearchHPV also called HPV integration sites within *TP63*. We found integrations of E1, E2, and L1 within *TP63* intron 10, of L1 within intron 12, and of E2 within *TP63* exon 14. These integration sites were also detected via detection of integrated papillomavirus sequences PCR<sup>19</sup> and/or WGS<sup>8</sup> 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 1 larger integration site. There were additional sites called by other WGS studies that we did not detect (intron 9<sup>8</sup> and exon 7<sup>34</sup>), although it is possible that alternate clonal populations grew out because of different selective pressures in different laboratories. Nonetheless, the analysis clearly demonstrated that SearchHPV 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,<sup>35,36</sup> and because of the role of *TRAF2* in regulating innate immunity, this gene may have a larger role in HPV-16-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 showing that integrations are detected in host genes more frequently than expected by chance.<sup>2,3,7,37</sup> 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,<sup>7,38</sup> and the junctions that we detected in/near this gene had 2 to 4 bp of microhomology, which potentially drove this insertion event. 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.<sup>3,7,39,40</sup> There is a high degree of microhomology between HPV-16 and this gene. Because of the high frequency of molecular alterations in the epidermal differentiation pathway (eg, *NOTCH1/2*, *TP63*, and *ZNF750*) in HPV+ HNSCCs, these data support HPV integration as a pivotal mechanism of viral-driven oncogenesis in this model.<sup>41</sup>

HPV integration sites have been associated with structural variations in the human genome,<sup>3,8,41</sup> and this supports an additional genetic mechanism for 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.<sup>8,42</sup> Our data are consistent with these previous reports in that approximately half of the integration events that 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.<sup>42</sup> Notably, we resolved and identified an HPV-HPV junction that bounded in a large duplication segment and showed the possibility of an HPV internal rearrangement being involved in HPV integration events.

Importantly, this observation that HPV integration events tended to be enriched in cellular genes could have resulted from multiple different mechanisms. Integration could occur preferentially in regions of open chromatin during cell replication and keratinocyte differentiation. Other potential mechanisms include the following: 1) HPV integration is directed to specific host genes by homology, and 2) HPV integration is random, but events that are advantageous for oncogenesis are clonally selected and expanded (this implicates non-homology-based DNA repair mechanisms). Therefore, to help to 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; this potentially implicates nonhomologous end joining in repair at these sites because this pathway most frequently results in 0 to 5 bp of overlap.<sup>43</sup> There were also a number of junction sites that demonstrated a gap of inserted sequence between the HPV and human genomes. It has been reported that during polymerase theta-mediated end joining, stretches of 3 to 30 bp are frequently inserted at the site of repair, and this possibly accounts for these sites.<sup>44</sup> However, given the relatively small number of events that we examined, we expect that future analysis with our pipeline will help to 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. Although the performance of SearchHPV has been examined for only 3 samples, in the future, we expect that the application of this pipeline in large HPV+ cancer tissue cohorts will help to advance our understanding of the potential oncogenic mechanisms associated with viral integration. With the emerging set of tools such as SearchHPV, we believe that 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.

## FUNDING SUPPORT

This study was supported by the National Institutes of Health/National Cancer Institute (R01 CA194536 to Thomas E. Carey and J. Chad Brenner) as well as startup discretionary funds to J. Chad Brenner and Ryan E. Mills from the University of Michigan. Lisa M. Pinatti was supported by the National Institutes of Health/National Cancer Institute (R01 CA194536).

## CONFLICT OF INTEREST DISCLOSURES

The University of Michigan has licensed some UM-SCC cell lines to Millipore EMD/Merck; royalty payments have been made to the University of Michigan as well as Thomas E. Carey and J. Chad Brenner. Carey also has received consulting fees for being an advisor to large grant planning committees at the University of Michigan and Massachusetts General Hospital–Harvard University within the past 36 months and travel support and honoraria for invited lectures at various universities. The other authors made no disclosures.

## AUTHOR CONTRIBUTIONS

**Lisa M. Pinatti:** Conceptualization, data curation, formal analysis, investigation, project administration, validation, visualization, writing—original draft, and writing—review and editing. **Wenjin Gu:** Conceptualization, data curation, formal analysis, investigation, project administration, methodology, resources, software, visualization, writing—original draft, and writing—review and editing. **Yifan Wang:** Data curation, formal analysis, investigation, methodology, resources, and software. **Ahmed Elhossiny:** Data curation, investigation, and software. **Apurva D. Bhangale:** Data curation, formal analysis, investigation, methodology, resources, software, and validation. **Collin V. Brummel:** Data curation, investigation, project administration, and resources. **Thomas E. Carey:** Conceptualization, funding acquisition, project administration, resources, supervision, and

writing—review and editing. **Ryan E. Mills:** Conceptualization, funding acquisition, methodology, resources, software, supervision, and writing—review and editing. **J. Chad Brenner:** Conceptualization, funding acquisition, project administration, resources, software, supervision, visualization, and writing—review and editing.

## REFERENCES

- 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*. 2019;19:352. doi:10.1186/s12885-019-5536-1
- 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;8:17684-17699. doi:10.18632/oncotarget.15179
- 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*. 2014;111:15544-15549. doi:10.1073/pnas.1416074111
- Pinatti LM, Sinha HN, Brummel CV, et al. Association of human papillomavirus integration with better patient outcomes in oropharyngeal squamous cell carcinoma. *Head Neck*. 2021;43:544-557. doi:10.1002/hed.26501
- 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*. 2019;40:1220-1228. doi:10.1093/carcin/bgz094
- McBride AA, Warburton A. The role of integration in oncogenic progression of HPV-associated cancers. *PLoS Pathog*. 2017;13:e1006211. doi:10.1371/journal.ppat.1006211
- Bodelon C, Untereiner ME, Machiela MJ, Vinokurova S, Wentzensen N. Genomic characterization of viral integration sites in HPV-related cancers. *Int J Cancer*. 2016;139:2001-2011. doi:10.1002/ijc.30243
- Akagi K, Li J, Broutian TR, et al. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res*. 2014;24:185-199. doi:10.1101/gr.164806.113
- Pinatti LM, Walline HM, Carey TE. Human papillomavirus genome integration and head and neck cancer. *J Dent Res*. 2018;97:691-700. doi:10.1177/0022034517744213
- 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*. 2001;92:9-17.
- 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*. 1999;59:6132-6136.
- 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:e64465. doi:10.1371/journal.pone.0064465
- Wang Q, Jia P, Zhao Z. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. *Genome Med*. 2015;7:2. doi:10.1186/s13073-015-0126-6
- 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*. 2013;29:266-267. doi:10.1093/bioinformatics/bts665
- 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
- 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*. 2016;24:490-495. doi:10.1097/PAL.0000000000000215
- Morel A, Neuzillet C, Wack M, et al. Mechanistic signatures of human papillomavirus insertions in anal squamous cell carcinomas. *Cancers (Basel)*. 2019;11:1846. doi:10.3390/cancers11121846
- 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.* 2019;9:1504. doi:10.1038/s41598-018-37871-2
19. 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.* 2017;39:840-852. doi:10.1002/hed.24729
  20. Shuman AG, Gornick MC, Brummel C, et al. Patient and provider perspectives regarding enrollment in head and neck cancer research. *Otolaryngol Head Neck Surg.* 2020;162:73-78.
  21. Heft Neal ME, Bhargale AD, Birkeland AC, et al. Prognostic significance of oxidation pathway mutations in recurrent laryngeal squamous cell carcinoma. *Cancers (Basel).* 2020;12:3081. doi:10.3390/cancers12113081
  22. Papillomavirus Episteme. National Institute of Allergy and Infectious Diseases. Updated January 13, 2020. Accessed May 1, 2020. <https://pave.niaid.nih.gov/>
  23. Van Doorslaer K, Li Z, Xirasagar S, et al. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Res.* 2017;45:D499-D506. doi:10.1093/nar/gkw879
  24. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25:1754-1760. doi:10.1093/bioinformatics/btp324
  25. Picard toolkit. Broad Institute GitHub Repository. Accessed April 15, 2020. <https://github.com/broadinstitute/picard>
  26. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20:1297-1303. doi:10.1101/gr.107524.110
  27. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd merger. *Bioinformatics.* 2014;30:614-620. doi:10.1093/bioinformatics/btt593
  28. Huang X, Madan A. CAP3: a DNA sequence assembly program. *Genome Res.* 1999;9:868-877. doi:10.1101/gr.9.9.868
  29. Team RC. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2019.
  30. Van Rossum G, Drake FL. Python 3 Reference Manual: Python Documentation Manual Part 2. CreateSpace Independent Publishing Platform; 2009.
  31. Liu J, Pan S, Hsieh MH, et al. Targeting Wnt-driven cancer through the inhibition of porcupine by LGK974. *Proc Natl Acad Sci U S A.* 2013;110:20224-20229.
  32. 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.* 2018;9:30419-30433. doi:10.18632/oncotarget.25754
  33. Myers JE, Guidry JT, Scott ML, et al. Detecting episomal or integrated human papillomavirus 16 DNA using an exonuclease V-qPCR-based assay. *Virology.* 2019;537:149-156. doi:10.1016/j.virol.2019.08.021
  34. 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.* 2015;136:E207-E218. doi:10.1002/ijc.29112
  35. Poirson J, Biquand E, Straub M-L, et al. Mapping the interactome of HPV E6 and E7 oncoproteins with the ubiquitin-proteasome system. *FEBS J.* 2017;284:3171-3201.
  36. Thompson DA, Zacny V, Belinsky GS, et al. The HPV E7 oncoprotein inhibits tumor necrosis factor alpha-mediated apoptosis in normal human fibroblasts. *Oncogene.* 2001;20:3629-3640.
  37. 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.* 2015;47:158-163. doi:10.1038/ng.3178
  38. Ferber MJ, Thorland EC, Brink AA, et al. Preferential integration of human papillomavirus type 18 near the c-myc locus in cervical carcinoma. *Oncogene.* 2003;22:7233-7242. doi:10.1038/sj.onc.1207006
  39. Schmitz M, Driesch C, Jansen L, Runnebaum IB, Durst M. Non-random integration of the HPV genome in cervical cancer. *PLoS One.* 2012;7:e39632. doi:10.1371/journal.pone.0039632
  40. 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.* 2016;14:941-952. doi:10.1158/1541-7786.MCR-16-0105
  41. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature.* 2015;517:576-582. doi:10.1038/nature14129
  42. Groves IJ, Coleman N. Human papillomavirus genome integration in squamous carcinogenesis: what have next-generation sequencing studies taught us? *J Pathol.* 2018;245:9-18. doi:10.1002/path.5058
  43. Pannunzio NR, Li S, Watanabe G, Lieber MR. Non-homologous end joining often uses microhomology: implications for alternative end joining. *DNA Repair (Amst).* 2014;17:74-80. doi:10.1016/j.dnarep.2014.02.006
  44. 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.* 2020;117:8476-8485. doi:10.1073/pnas.1921791117