ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Identification of Human Papillomavirus Integration Sites in the Human Genome Using NGS-Based Bioinformatics Analysis

Hemavathi D[1], Sindhu S[2], Revathi Devi S[3], Jonesh Linso[4]

[1,2,4]Department of Data Science and Business Systems, SRM Institute of Science and Technology Kattankulathur, chennai, India

[3]School of Computing, Department of Genetic Engineering, SRM Institute of Science and Technology Kattankulathur, chennai, India

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150800011

Received: 15 August 2026; Accepted: 20 August 2026; Published: 02 September 2026

ABSTRACT

The majority of cases of cervical cancer worldwide are caused by high-risk genotypes such as HPV-16 and HPV-18, making the human papillomavirus (HPV) one of the most clinically significant viral pathogens. The physical integration of viral DNA into the host genome, which sets off a series of genomic disruptions, is one of the most important stages in the development of HPV-related cancer. Using publicly available next-generation sequencing (NGS) data, we developed and implemented a bioinformatics pipeline in this work to locate these integration sites. We began with raw sequencing reads from the NCBI Sequence Read Archive, performed quality filtering, aligned the cleaned reads to a merged human-HPV reference genome (GRCh38 plus HPV-16), and used variant calling to identify areas exhibiting viral insertion. Our findings provide unambiguous proof of genomic breakpoints associated with HPV and possible integration hotspots present in the human genome. Beyond the particular results, this work demonstrates that open-source computational tools by themselves can reconstruct significant biological signals from publicly available data, providing a useful and repeatable way to investigate how HPV modifies the genomes of the cells it infects.

Index Terms: Next-generation sequencing, bioinformatics pipeline, human papillomavirus, viral integration, and cervical cancer**.**

INTRODUCTION

Cervical cancer results in the death of hundreds of thousands of women annually, and in almost every instance, there is a single causative agent: Human Papillomavirus. It is a double-stranded DNA virus with over 100 different types of HPV. However, not all types of HPV have an equally high potential for causing damage. A very small number of types have an alarmingly high potential for causing malignant transformation: notably HPV-16 and HPV-18. These types of HPV greatly dominate the molecular landscape of cervical cancer. Their prevalence is of particular concern in low- and middle-income countries where screening for these infections is either nonexistent or underfunded, so that undetected infections occur and progression to disease is unimpeded. However, the more sinister aspect of the virus is the manner in which it incorporates the host's genome. While certain other viruses persist as an episome, high-risk HPV has a propensity to integrate its viral DNA directly into the host's chromosomes. When this integration does take place, the viral genome often disrupts its own controlling genes, E1 and E2. This disruption allows the unregulated expression of the oncogenes E6 and E7. This oncogene-mediated disruption targets two of the cell's most potent tumor suppressors: p53 and the retinoblastoma protein (pRb). This gives the cell the uncontrolled ability to replicate and accumulate further mutations over time [5], [9]. For a while, it was technically difficult to research this. You could find out if a person was infected by standard cytology and serology, but it told you nothing about whether or where it was integrated. That limitation has now been greatly alleviated by the use of next-generation sequencing, which generates a large enough number of reads and at a deep enough coverage to detect the chimeric fragments formed at the junctions of the virus and the host [2], [3].

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

The problem is now one of bioinformatics, and this is what this research tackles. We developed a step-by-step computational strategy to find HPV integration sites based on public NGS data. This strategy ranges from quality control and adapter removal to mapping against a combined human and HPV genome and finally to variant calling after mapping. We chose to use public domain tools for our strategy instead of proprietary platforms to ensure maximum flexibility and portability of our strategy for use by different research groups. This study is part of our understanding of how genetic events in viruses translate into molecular events relevant to cancer development and how public domain NGS data can be used to gain new insights into biology.

MATERIALS AND METHOD

Data Source

In this research, all the data that was required for the sequencing was obtained from publicly accessible data repositories rather than creating the data de novo. The raw sequence data of the HPV-positive samples, in the form of FASTQ format, was retrieved from the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI), which stores sequence data submitted by research groups around the world.For the reference data required during the sequence alignment, two widely used sources were employed: the well-assembled human genome sequence, known as GRCh38, and the complete genome sequence of the most common high-risk type of HPV, known as HPV-16, with the accession number NC001526.4. The reason for the selection of this particular type of HPV is that it is the most common cause of cancer cases among the high-risk types of the virus.

Data Preprocessing and Quality Control

However, before we can do any kind of meaningful analysis, the raw sequence data needs to be cleaned. This is because the sequencing run is not perfect; reads may have adapter contamination from the library preparation step, bases may have been sequenced with low confidence, and the reads may be too short to be reliably aligned. This is noise that can further corrupt the eventual results. We used the tool "fastp" for this step. We chose this tool for its speed and for the fact that it can perform several cleaning steps in a single step. In one step, "fastp" automatically detected and removed adapter contamination from the reads, removed reads and bases with low confidence, computed Q20 and Q30 as a measure of the quality of the reads, and removed reads that were too short for reliable alignment. "Fastp" also produced HTML and JSON reports for each sample, which we verified to confirm whether the quality had indeed improved before we proceeded further. This step is often easy to overlook but is an important one; cleaning the data may look good on paper but does not always translate to cleaner data.

Reference Genome Preparation

The overall bioinformatics workflow adopted in this study is illustrated in Fig. 1.

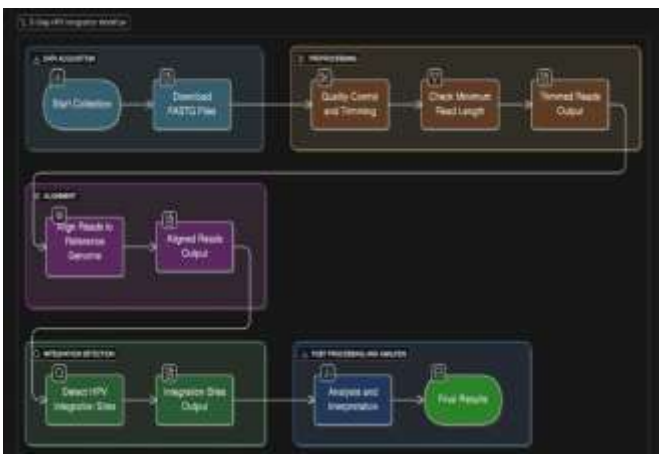


Fig. 1. Overview of the NGS-based bioinformatics pipeline used to identify HPV integration sites in the human genome.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

One of the important choices that was made in this process was the way to deal with the two different genomes simultaneously. Instead of running two different alignment tasks, one for the human genome and the other for the HPV genome, the two genomes were joined into one file in FASTA format. This way, every read, regardless of whether it is from the host, the virus, or the junction between the two, has the opportunity to compete for the best alignment. Once the genome was prepared, it was indexed using the Burrows-Wheeler Aligner (BWA). Indexing is a one-time computation that rearranges the genome into a form that allows subsequent searches to be extremely fast.

Sequence Alignment

With our reference prepared and our reads cleaned, we used the bwa mem algorithm for alignment. This particular algorithm has become a standard for short read data, as it is well-suited for the types of complex alignment that are most relevant for our detection of viral integration. This is particularly evident in that this algorithm supports split read alignment, which means that a read that spans a junction between virus and host will have each half of the read align separately. This is, perhaps, the clearest indicator that a particular location has integrated viral sequence [4]. Alignment yielded a SAM file, which contains a record for every read, where they align, how well they align, and any discrepancies they may have with the reference. This is where all subsequent work was based.

Post-Alignment Processing

SAM files are human-readable but aren't particularly practical for large-scale analysis because they're cumbersome, inefficient for querying, and aren't indexed. We then used samtools to convert the files to BAM format, which compresses the same data in a binary form that's much more efficient to store and query. We then sorted each BAM file by genome coordinate and created an index on each one. These are not particularly glamorous steps, but they're absolutely essential ones. If a pipeline fails or slows down at this step, it's going to fail or slow down at every step that follows.

Variant Calling and Integration Evidence Analysis

To determine where these positions are, the bcftools package was utilized to perform a variant calling analysis. The mpileup tool was used to count the number of bases observed at any given site, and the subsequent call tool was used to compare this to the bases that are expected by the genome reference sequence. Sites where the observed bases are different from the expected bases, particularly where the bases are observed in the HPV genome, are likely to be the result of viral integration into the genome. In addition to the automated analysis, the data was also examined manually by using a genome browser to examine the alignment data. Examining the data in this way allowed us to get a better overall feel for the data, particularly where the data was irregular, as the automated analysis might not have picked up on this. This is particularly important in integration analysis, as this type of data is necessarily complex.

RESULTS AND DISCUSSION

Quality Assessment and Preprocessing

The raw reads were in a condition expected of sequencing data retrieved from public repositories, i.e., mostly of good quality but with a significant proportion of adapter contamination and poor quality bases, which had to be cleaned. The quality of the output was clear from running it through fastp, as it showed a significant improvement in the quality metrics. The Q20 and Q30 qualities increased, showing that the reads we were left with were of much higher confidence. The short, unreliable reads were removed, and we were left with FASTQ files of a quality we were confident about moving into alignment.

Sequence Alignment to the Combined Reference Genome

The process of aligning these cleaned reads against our new human and HPV reference was successful. The great majority of these reads were found to map to human chromosomes. This is perfectly expected, given

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

how much human DNA there is compared to HPV in an HPV-positive tissue sample. However, aside from this large majority of reads, there was a very distinct minority of reads that specifically mapped to the HPV-16 genome. This was good evidence that HPV genetic material was indeed present in our dataset and not simply an artifact. The reason why using our combined human and HPV genome was so powerful was that it allowed us to see everything in one space. We didn't have to try to infer integration by combining two different alignments. We could see our reads mapping to both human and HPV genome sequences in one run. This was particularly important for seeing our chimeric reads. These reads overlap with both human and HPV genome sequences. They represent our most direct evidence of integration.

Post-Alignment Analysis

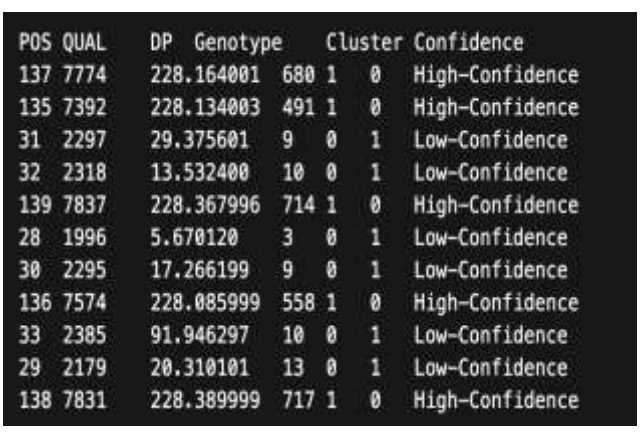
This sorting and indexing of the BAM files really opened up the files for a more detailed inspection. Upon inspection of the coverage distribution over the genomic coordinates of the HPV genome, it was noted that this was not a uniform distribution. In addition, there were regions of pronounced dips, spikes, or discontinuities. These irregular coverage profiles are not noise. These are patterns of coverage that have been consistently associated with integrated viral DNA [9], as the physical disruption of the viral genome at the point of integration causes such coverage irregularities, which are not seen following a standard episomal infection. This was a promising sign of a meaningful alignment.

Variant Detection and Integration-Related Signals

Variant calls provided another tier of evidence. The VCF files indicated a variety of genomic locations where variations from the expected sequence occurred, and by cross-checking these with locations of HPV read accumulations, a cohesive picture began to emerge. Variants near locations of HPV read accumulations are unlikely random occurrences; they are, in fact, more likely due to the genomic disruption caused by viral insertion. Needless to say, we are cautious not to overstate the case here. Variant calls near viral read accumulations are suggestive of integration, but they are not, in themselves, conclusive evidence for integration. To be certain of the exact location of a virus-host junction, long-read sequencing or experimental validation of these regions, for example, with inverse PCR or junction-specific amplifications, must be used. What we can be certain of here, however, is that the aggregate evidence of coverage irregularities, viral read mappings, and variant calls all converge in a similar direction, which does increase the degree of confidence in the result.

Summary of Findings

Thus, the entire pipeline fulfilled each of the objectives. The quality filtering step genuinely improved the read set. The alignment against the combined reference successfully incorporated both host and viral reads. Inspection after alignment confirmed the expected patterns for coverage. And the variants step provided additional corroborating genomic information. While none of this information by itself would constitute evidence for integration at a specific point, collectively it forms a cohesive and biologically plausible argument for integration events being present in this data.



POS	QUAL	DP	Genotype		Cluster	Confidence
137	7774	228.164001	680	1	0	High-Confidence
135	7392	228.134003	491	1	0	High-Confidence
31	2297	29.375601	9	0	1	Low-Confidence
32	2318	13.532400	10	0	1	Low-Confidence
139	7837	228.367996	714	1	0	High-Confidence
28	1996	5.670120	3	0	1	Low-Confidence
30	2295	17.266199	9	0	1	Low-Confidence
136	7574	228.085999	558	1	0	High-Confidence
33	2385	91.946297	10	0	1	Low-Confidence
29	2179	20.310101	13	0	1	Low-Confidence
138	7831	228.389999	717	1	0	High-Confidence

Fig. 2. Example output of the variant analysis stage showing genomic positions with their associated

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

quality score (QUAL), read depth (DP), genotype classification, cluster assignment, and resulting confidence level (High-Confidence or Low-Confidence)

DISCUSSION

The overall significance of this work is also partially contained within the results, as is the significance of the methodology used. From a biological point of view, the integration of HPV within the host genome is not a singular process with a singular outcome, but rather varies by location, process, and outcome from tumor to tumor. Identifying where, within the genome, this process preferentially occurs, and which genes or regulatory regions are thereby affected, is an active area of cancer genomics research [7], [8].

Our method, even without providing a singular location for this process, identifies the types of signatures that make this detectable computationally.

From a methodological perspective, it is also worth noting that this whole effort was done using only open-source tools and publicly available data. This is important because access to expensive software or new sequencing runs can be a significant limitation for many research groups, particularly in resource-limited settings. A pipeline like ours, using fastp, bwa, samtools, and bcftools, can be performed on a standard research cluster and can be easily adapted for different HPV types or host organisms without additional cost. This is valuable for research.

There are honest limitations to acknowledge. The dataset we worked with is small, and a more expansive study covering a variety of HPV genotypes and sample types would provide a more comprehensive picture of integration patterns. Short-read sequencing is a powerful tool, but it does have a fundamental limitwith regard to the exact nature of complex integration events — a read only 150 base pairs long can only do so much to explore an integration junction that may stretch over thousands of base pairs. Future work using long-read technologies such as Oxford Nanopore and PacBio, or even integration-detection tools such as SearcHPV [4] and DisV-HPV16 [3], would greatly improve the resolution of what we have started to explore.

CONCLUSIONS

This study was designed to prove that a well-thought-out bioinformatics pipeline, constructed from readily available and freely accessible open-source tools, can significantly identify the genomic footprint of HPV integration, even from publicly available sequencing data sets. In our opinion, the results of this study prove this hypothesis well. From raw sequence reads, filtered alignments, and variant calls, the entire pipeline contributed to a result that makes sense in light of what is known about the behavior of integrated HPV in the human genome.

Of course, we are not claiming that we have mapped integration site locations at the base pair level, and we have been explicit about that in our paper. What we have done, however, is show a workflow that can be used as a starting point for other researchers wishing to explore the complex interplay between HPV and the host genome without the need for specialized software or newly generated data sets. As the cost of long-read sequencing continues to decrease and tools for integration site analysis continue to improve, this type of workflow will become even more powerful, and for now, it represents a useful starting point for research in the field of computational cancer genomics.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the Department of Computer Science and Engineering, SRM University, Chennai, India.

REFERENCES

1. H. C. P. F. de Oliveira et al., “NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome,” British Journal of Cancer, vol. 112, no. 12, pp. 1882–1888, Jun.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

2015.
2. T. E. Olshen et al., “HPV Integration Site Mapping: A Rapid Method of Viral Integration Site (VIS) Analysis and Visualization Using Automated Workflows in CLC Microbial Genomics,” Current Protocols, vol. 2, no. 8, e702, Aug. 2022.
3. C. Vega et al., “DisV-HPV16, versatile and powerful software to detect HPV in RNA sequencing data,” Bioinformatics, vol. 35, no. 20,
4. pp. 4226–4228, Oct. 2019.
5. S. A. de Munnik et al., “SearcHPV: a novel approach to identify and assemble human papillomavirus integration sites,” Bioinformatics, vol. 37, no. 12, pp. 1715–1722, Jun. 2021.
6. L. Chen et al., “Comprehensive comparative analysis of methods and software for identifying viral integrations,” Briefings in Bioinformatics, vol. 20, no. 6, pp. 2088–2102, Nov. 2019.
7. K. Singh et al., “Next-generation sequencing of cervical DNA detects human papillomavirus-presence predicts quality of life in cervical in- traepithelial neoplasia,” Gynecologic Oncology, vol. 126, no. 2, pp. 317– 323, Aug. 2012.
8. J. Li et al., “HPVPool-Seq: a genotype-guided pooling strategy for cost- effective HPV integration detection,” Microbiology Spectrum, vol. 13, no. 8, e01399-25, Aug. 2025.
9. M. A. Smith et al., “Precise identification of viral-host integration events in HPV-positive cancers using WGS and RNA-seq,” Patterns, vol. 6, no. 3, 100289, Mar. 2025.
10. M. Scarpitta et al., “Characterization of HPV integration, viral gene expression and novel fusion transcripts in cervical cancers,” Genomics, vol. 111, no. 4, pp. 805–814, Aug. 2019.
11. R. Patel et al., “A Novel NGS-Based Algorithm for Precise HPV Genotyping and Co-Infection Detection,” Current Protocols in Bioin- formatics, vol. 82, no. 1, e70231, Oct. 2025.
12. Akagi et al., “Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability,” Genome Research, vol. 24, no. 2, pp. 185–199, Feb. 2014.
13. S. Zhao et al., “HPV integration landscape in cervical carcinoma identifies potential genomic drivers,” Nature Communications, vol. 7, 13556, Nov. 2016.
14. J. Hu et al., “Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism,” Nature Genetics, vol. 47, no. 2, pp. 158–163, Feb. 2015.
15. Ojesina et al., “Landscape of genomic alterations in cervical carcinomas,” Nature, vol. 506, no. 7488, pp. 371–375, Feb. 2014.
16. X. Tang et al., “ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution,” Bioinformatics, vol. 29, no. 5, pp. 649–651, Mar. 2013.
17. P. Chen et al., “VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data,” PLoS ONE, vol. 8, no. 5, e64465, May 2013.
18. J. Wang et al., “Detection of viral integration sites in host genomes using next generation sequencing,” Current Genomics, vol. 14, no. 8, pp. 545–556, Dec. 2013.
19. Y. Li et al., “ViFi: accurate detection of viral integration and fusion transcripts from RNA-seq data,” Bioinformatics, vol. 34, no. 19, pp. 3283–3289, Oct. 2018.
20. H. Xu et al., “Identification of viral integration sites using whole genome sequencing and targeted sequencing approaches,” BMC Genomics, vol. 14, Suppl. 8, S4, Dec. 2013.
21. S. Gao et al., “HPV integration drives carcinogenesis in cervical cancer by disrupting host genome stability,” Journal of Virology, vol. 91, no. 8, e01768-16, Apr. 2017