



OPEN Scalable long-read nanopore HPV16 amplicon-based whole-genome sequencing

Maina K. Titus¹, David Giesbrecht¹, Cliff I. Oduor¹, Kapten Muthoka^{2,3}, Karamoko Niaré^{1,2}, Rebecca Crudale¹, Abebe A. Fola^{1,2}, Sin Ting Hui¹, Isaac E. Kim Jr.¹, Philip K. Tonui^{3,4}, Peter M. Itsura^{3,4}, Ronald Tonui³, Ann M. Moormann⁷, Patrick J. Loehrer^{5,6}, Darron R. Brown^{5,6}, Omenge E. Orang'o^{3,4,9}, Susan Cu-Uvin⁸, Aaron C. Ermel⁵, Rachel A. Katzenellenbogen⁵ & Jeffrey A. Bailey^{1,2}✉

Human papillomavirus 16 (HPV16) drives precursor cervical lesions that often progress to cervical cancer (CC). Variation within the HPV16 genome has been associated with CC risk. Here, we developed an affordable and portable amplicon-based long-read whole genome sequencing (WGS) approach using Oxford Nanopore Technologies to investigate HPV16 genetic diversity among women in sub-Saharan African countries. Applied to a control CaSki cell line and clinical samples ($n=12$), our method generated complete HPV16 genomes at high coverage (median read coverage: 5,899–15,279 $\times$). Benchmarking our HPV16 controls showed high accuracy for two variant calling pipelines (Clair3 and PEPPER-Margin DeepVariant). Phylogenetic analysis identified all four previously defined HPV16 lineages (A–D) and their high-risk sublineages. All lineages exhibited strong concordance across de novo assembly, reference-based phylogenetics, and unsupervised clustering. Our pipeline effectively captured the full extent of genomic variation, including putative lineage-informative SNPs. This method offers a robust amplicon-based WGS and analysis pipeline for HPV16, making it well-suited for integration into surveillance, diagnostics, and epidemiological efforts in low-resource areas.

Cervical cancer (CC) is a preventable and potentially curable disease; however, it remains the fourth most common cancer among women, with global estimates of 604,000 new cases and 342,000 deaths occurring annually¹. The vast majority of these deaths occur in countries with low Human Development Index (HDI), particularly in sub-Saharan Africa². Cervical cancer prevention primarily involves human papillomavirus (HPV) vaccination and screening for precursor lesions, followed by appropriate treatment. In 2020, the World Health Organization (WHO) aimed to eliminate CC by 2030, establishing the 90–70–90 targets: (1) vaccinating 90% of girls with the HPV vaccine by age 15, (2) screening 70% of women with a high-performance test at ages 35 and 45, and (3) ensuring that 90% of women with cervical pre-cancer or invasive CC have access to appropriate treatment³.

When implemented at scale, current screening methods effectively reduce the incidence and mortality of CC, with evidence suggesting HPV DNA detection alone is the most effective approach^{4–6}. In 2021, the WHO updated its guidelines for screening and treating cervical precancerous lesions, supported by two pivotal studies that reinforced the superiority of primary HPV nucleic acid detection as a screening strategy^{4,7}. Cervical intraepithelial neoplasia (CIN), a precursor to CC, is often driven by high-risk HPV (HR-HPV) infection. To date, more than 200 HPV genotypes have been identified and classified as either high-risk (HR) or low-risk (LR) based on their association with CC⁸. While WHO recognizes primary HPV testing as central to advancing CC screening^{5,6}, sequencing HR-HPV genotypes offers more profound insight into disease risk, with specific lineages and sub-lineages carrying different prognostic implications^{9–11}.

¹Department of Pathology and Laboratory Medicine, Warren Alpert Medical School, Brown University, Providence, RI 02912, USA. ²Center for Computational Molecular Biology, Brown University, Providence, RI, USA. ³Academic Model Providing Access to Healthcare (AMPATH) Cervical Cancer Screening Program, Eldoret, Kenya. ⁴Department of Reproductive Health and Gynecology, Moi University, Eldoret, Kenya. ⁵Indiana University School of Medicine, Indianapolis, IN, USA. ⁶Indiana University Melvin and Bren Simon Comprehensive Cancer Centre, Indianapolis, IN, USA. ⁷Division of Infectious Diseases and Immunology, Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, USA. ⁸Division of Infectious Diseases, Warren Alpert School of Medicine, Brown University, Providence, RI, USA. ⁹Department of Obstetrics and Gynecology, Medical College, Aga Khan University, Nairobi, Kenya. ✉email: jeffrey_bailey@brown.edu

Human papillomavirus type 16 (HPV16) is the most common HR-HPV type globally and is responsible for over 50% of worldwide CC cases¹². In sub-Saharan Africa, HPV16 remains the dominant HR-HPV among women, regardless of cervical lesion status^{13–17}. This viral type is further divided into four major evolutionary lineages: A (formerly European type), B (African type-1), C (African type-2), and D (Asian type), with specific variant sub-lineages, such as A4 and D2, associated with an increased risk of CC^{9,18–20}. While variants from Europe and Asia A lineage have been extensively characterized, HPV16 isolates from Africa remain understudied, and few have been sequenced, creating a significant knowledge gap regarding their lineage and sub-lineage, as well as their specific variations and associations with CC in sub-Saharan Africa.

Most HPV16 genomic studies in Africa have relied on targeted sequencing, focusing primarily on the oncogenic early gene E6, the major capsid late gene L1, and the non-coding long control region (LCR), leaving significant portions of the genome unexplored^{21–23}. Moreover, these approaches may fail to detect integrated HPV sequences. Whole-genome sequencing (WGS) provides a comprehensive landscape of HPV16 genetic diversity, which has been challenging to achieve with short-read sequencing platforms (e.g., Illumina) and is relatively expensive to implement in low-resource settings, such as sub-Saharan Africa. The recent development and maturation of long-read Oxford Nanopore Technologies (ONT) sequencing presents a promising alternative, offering portable, cost-effective, and real-time sequencing solutions through platforms such as the MinION and PromethION2 (P2). These technologies have been successfully deployed for pathogen genomic surveillance, including Tuberculosis, SARS-CoV-2, and Ebola virus, demonstrating the potential for improving genome assembly and variant detection through this long-read sequencing (LRS)^{24–26}. Given its affordability and adaptability, ONT sequencing holds promise for fully covering the genome in one or a few amplicons, thereby expanding HPV16 WGS in sub-Saharan Africa, where the burden of CC is highest.

Despite the increasing adoption of Nanopore LRS, no universally accepted variant caller exists, and it continues to be an area of active development. This differs from short-read sequencing, where the Genome Analysis Toolkit (GATK) and other variant callers are well-established as sensitive and specific^{27,28}. Several tools are available for ONT variant calling, including Medaka, Clairvoyante, Clair, and Nanocaller, which utilize pileup-based methods^{29,30}, while PEPPER employs a full-alignment approach²⁹. It maximizes precision and recall through full alignment. Clair3, the successor to Clair, merges pileup-based and full-alignment strategies to enhance efficiency and accuracy³⁰. These variant callers are typically benchmarked using well-characterized human genome datasets from the Genome in a Bottle (GIAB) consortium rather than viruses or other organisms³¹. Thus, it is essential to determine optimal parameters and validate which program has the best recall and precision for HPV. It requires proper controls that are already accurately sequenced and assembled (aka a “truth set”). We aimed to establish an HPV16 ‘truth set’ variants dataset to benchmark PEPPER and Clair3, two of the most widely used long-read variant callers.

We developed a scalable HPV16 WGS and analysis pipeline using ONT’s MinION and P2 platforms. By combining multiple primer sets for long-range PCR, including both tiling and full-length, we aimed to traverse the entire genome and potentially capture the circular form of the genome. We applied this approach to the HPV16-positive CaSki cell line, as well as clinical isolates from western Kenya and Rwanda. To benchmark variant calling accuracy, we established a high-confidence HPV16 “truth set” of variants and evaluated PEPPER and Clair3, both of which exhibited strong performance for whole-genome variant detection. Our method achieved complete genome recovery with high coverage, enabling accurate phylogenetic classification into lineages A, B, C, and D. This robust and adaptable HPV16 WGS workflow facilitates comprehensive genomic surveillance and diagnostics, particularly in resource-limited settings.

Results

WGS primer design to capture HPV16

To comprehensively analyze the genetic diversity of HPV16, we developed a long-range PCR viral enrichment strategy that leveraged three primer set designs for amplification. The primer designs were (1) a near full-length primer set generating amplicons up to 7.7 kb to capture intact or nearly full-length HPV16 DNA (complete genome measures ~8 kb), (2) a tiling-set of three primer pairs yielding overlapping amplicons of 2.1 kb, 3.9 kb, and 2.6 kb, (Fig. 1a) and a junction primer pair designed to capture the genome undergoing linearization and integration into the host DNA, as illustrated in (Fig. 1b). This last primer pair addresses HPV16’s episomal form and tandem integration into the host genome (extrachromosomal form)³² (Fig. 1b). We first tested our primers using DNA extracted from the HPV 16-positive, cervical cancer CaSki cell line and sequencing amplicons with the Oxford Nanopore MinION sequencer. Our method was sensitive, allowing us to amplify and sequence as few as five copies of HPV16 per reaction in 2 µL of CaSki DNA input (Supplementary Fig. 1a, b, and c). Our analysis workflow (Fig. 1c and Supplementary Fig. 5) enabled us to accurately align the amplicon sequences to the expected region of the HPV16 reference genome (GenBank: K02718.1) (Supplementary Fig. 2a and b). We pooled and sequenced all PCR products to ensure sufficient coverage for accurate variant calling and other downstream analyses. We achieved a deep (> 1000X) read depth across the HPV16 genome from the CaSki cell line DNA (Fig. 1d and Supplementary Fig. 1d).

Benchmarking PEPPER and Clair3 for HPV16 WGS variant calling

To assess the accuracy of long-read variant calling, we benchmarked HPV16 controls comparing PEPPER and Clair3 variant calling against GATK4 (Methods). We evaluated the performance of the variant callers using precision, recall (also known as sensitivity), and F1-score (the harmonic mean of precision and recall). Among the 29 consensus variants identified by all three callers, Clair3 and GATK4 jointly detected two additional variants that PEPPER missed (Fig. 2a). Meanwhile, PEPPER and GATK4 identified one variant that Clair3 missed. Additionally, PEPPER identified one variant that was not detected by either Clair3 or GATK4 (Fig. 2a).

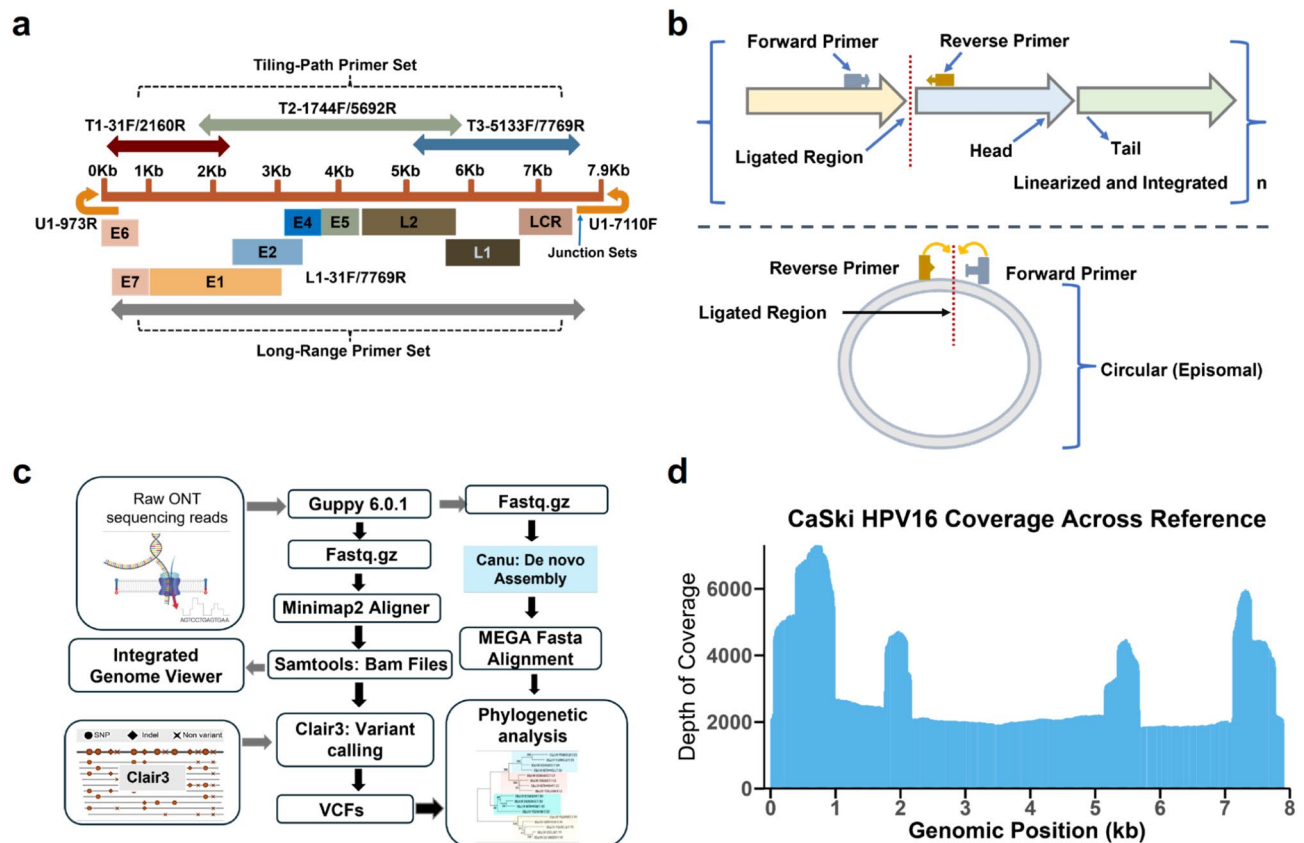


Fig. 1. A trio primer set for HPV16 whole-genome sequencing enrichment. **a** Schematic of the primer design for viral pre-sequencing enrichment. On top of the panel, the tiling-path primers are set: T1 (dark red), T2 (light green), and T3 (blue), along with the U1 (yellow) primer set, which is essential for amplifying the ligated region at the genome's terminus. At the bottom of the panel is the near full-length primer set (gray). **b** Representation of HPV16 in its integrated extrachromosomal form as a multi-tandem repeat (upper) and episomal form (lower). **c** HPV16 long-read analysis pipeline. ONT reads were base-called with Guppy, aligned using Minimap2, and processed with SAMtools. Variants were called with Clair3. De Novo Assembly was executed using Canu, followed by sequence alignment and phylogenetic analysis in MEGA. **d** Representative depth-of-coverage plot showing sequencing read depth across the CaSki HPV16 genome relative to the reference for combined amplicons.

Overall, Clair3 achieved 100% precision, 96.9% recall, and a 98.4% F1-score, whereas PEPPER attained 96.8% precision, 93.8% recall, and a 95.2% F1-score (Fig. 2b; Table 1).

When assessing SNPs, Clair3 achieved 96.7% precision and 100% recall, while PEPPER attained 100% precision and 96.6% recall, resulting in equivalent F1 scores of 98.3% (Fig. 2c; Table 1). For indel detection, Clair3 exhibited 100% precision but only 33.3% recall, resulting in an F1 score of 50% due to its conservative calling approach. In contrast, PEPPER achieved 66.7% precision, 66.7% recall, and a 66.7% F1 score (Fig. 2d; Table 1). These findings are consistent with known challenges of Oxford Nanopore Technologies (ONT) sequencing, particularly with R9.4.1 flow cells and Guppy basecalling models, where small indel errors frequently occur, particularly in homopolymer regions³³.

HPV16 WGS of clinical samples as proof of concept

Next, we assessed the performance of our approach using 12 HPV16-positive isolate samples derived from cervical exfoliated cells, which had previously been genotyped with the Roche Linear Array. These clinical specimens were from women ($n = 12$; mean age, 30.8 years; range, 24–45 years) enrolled in cervical cancer-related studies conducted in western Kenya ($n = 10$) and Rwanda ($n = 2$), and were archived for future research^{14,34}. All samples yielded a good depth of coverage (median read coverage ranging from 5899 to 15,279) (Fig. 3; Supplementary Fig. 3), and 100% complete genome coverage was observed in clinical samples. A localized coverage drop in sample PMTRHP10 near the 3 kb region revealed a 540 bp deletion spanning positions 2851–3390, as confirmed by IGV visualization (Supplementary Fig. 4). This structural variant represents a bona fide biological event rather than a methodological artifact. The deletion disrupts the E2 gene, a critical negative regulator of the viral oncogenes E6 and E7, whose loss is known to result in oncogene overexpression and drive cellular transformation and cervical carcinogenesis. Given the use of overlapping primer pairs, we observed peaks in the sequencing data, reflecting the overlapping regions of the amplicons across all samples (Fig. 3). This pattern consistently appeared

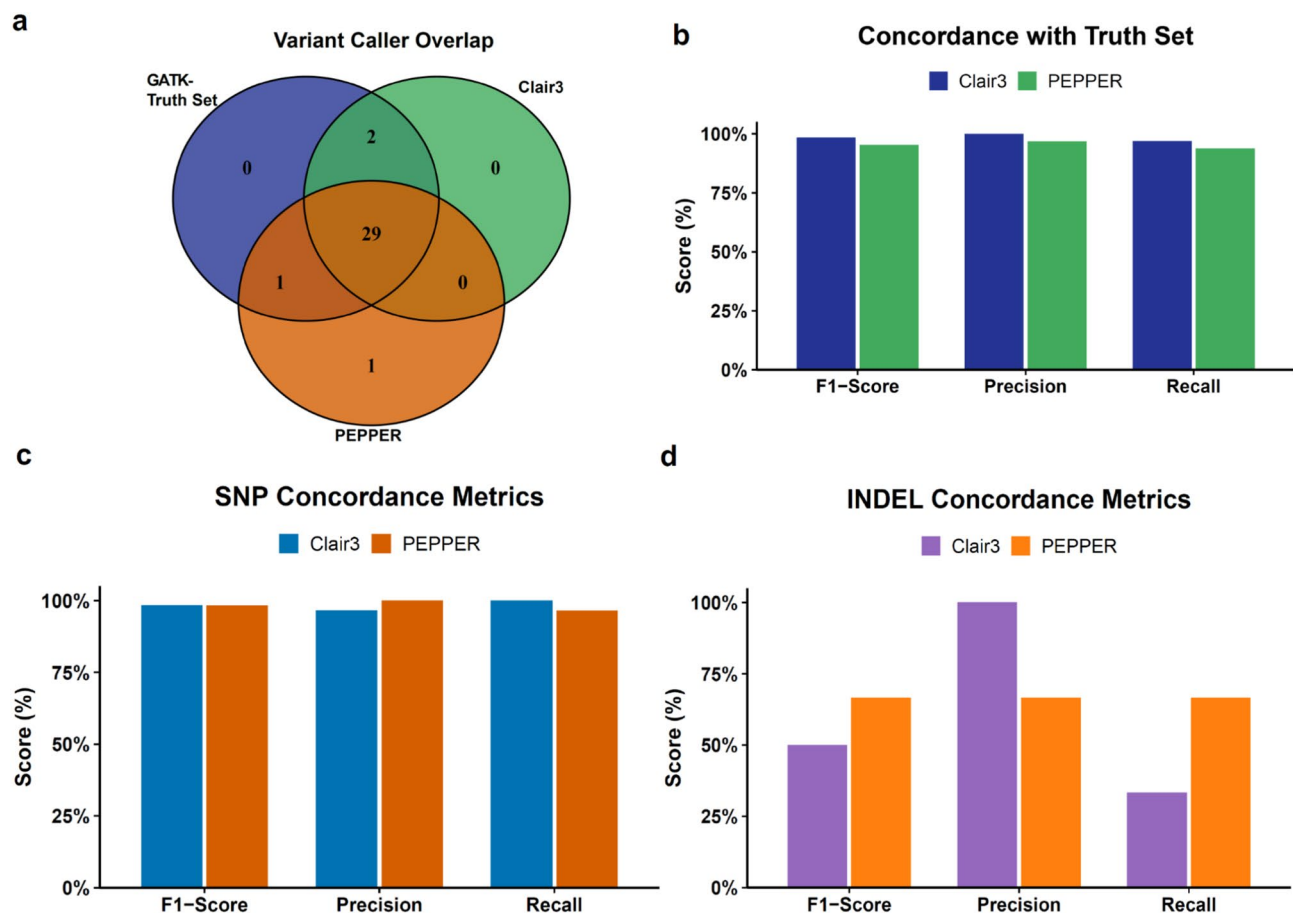


Fig. 2. Performance comparison of Clair3 and PEPPER against GATK. **a** A Venn diagram shows the overlap among the variant callers: PEPPER (brown), Clair3 (green), and the Truth-set GATK. **b** Overall concordance performance of Clair3 (blue) and PEPPER (green) relative to the Truth-set GATK, with F1-Score, Precision, and Recall on the x-axis and percentages on the y-axis. **c** SNP concordance metrics for Clair3 (blue) and PEPPER (brown), displaying F1-Score, Precision, and Recall on the x-axis and percentages on the y-axis. **d** INDEL concordance metrics for Clair3 (purple) and PEPPER (yellow), featuring F1-Score, Precision, and Recall on the x-axis and percentages on the y-axis.

Tool	Variant type	Precision	Recall	F1-Score
Clair3	Overall	1.000	0.969	0.984
PEPPER	Overall	0.968	0.938	0.952
Clair3	SNP	0.967	1.000	0.983
PEPPER	SNP	1.000	0.966	0.983
Clair3	INDEL	1.000	0.333	0.500
PEPPER	INDEL	0.667	0.667	0.667

Table 1. Performance comparison of Clair3 and PEPPER Margin DeepVariant (PEPPER) with GATK4. The table summarizes the performance metrics of Clair3 and PEPPER in relation to the GATK truth set. It presents overall performance alongside specific metrics for SNPs and INDELs, providing precision, recall, and F1-score values for each variant type.

in the HPV16 coverage bar plots, regardless of lineages (Fig. 3), demonstrating robust and uniform coverage across samples.

Robust recovery and phylogenetic analysis of HPV16 genomes from clinical samples

Next, we evaluated the compatibility of sequencing data from clinical samples ($n = 12$) generated by our WGS approach for downstream analyses, including de novo assembly and reference-based variant calling. Reference genomes for HPV16 variant lineages and sublineages were retrieved from GenBank and merged with de novo–

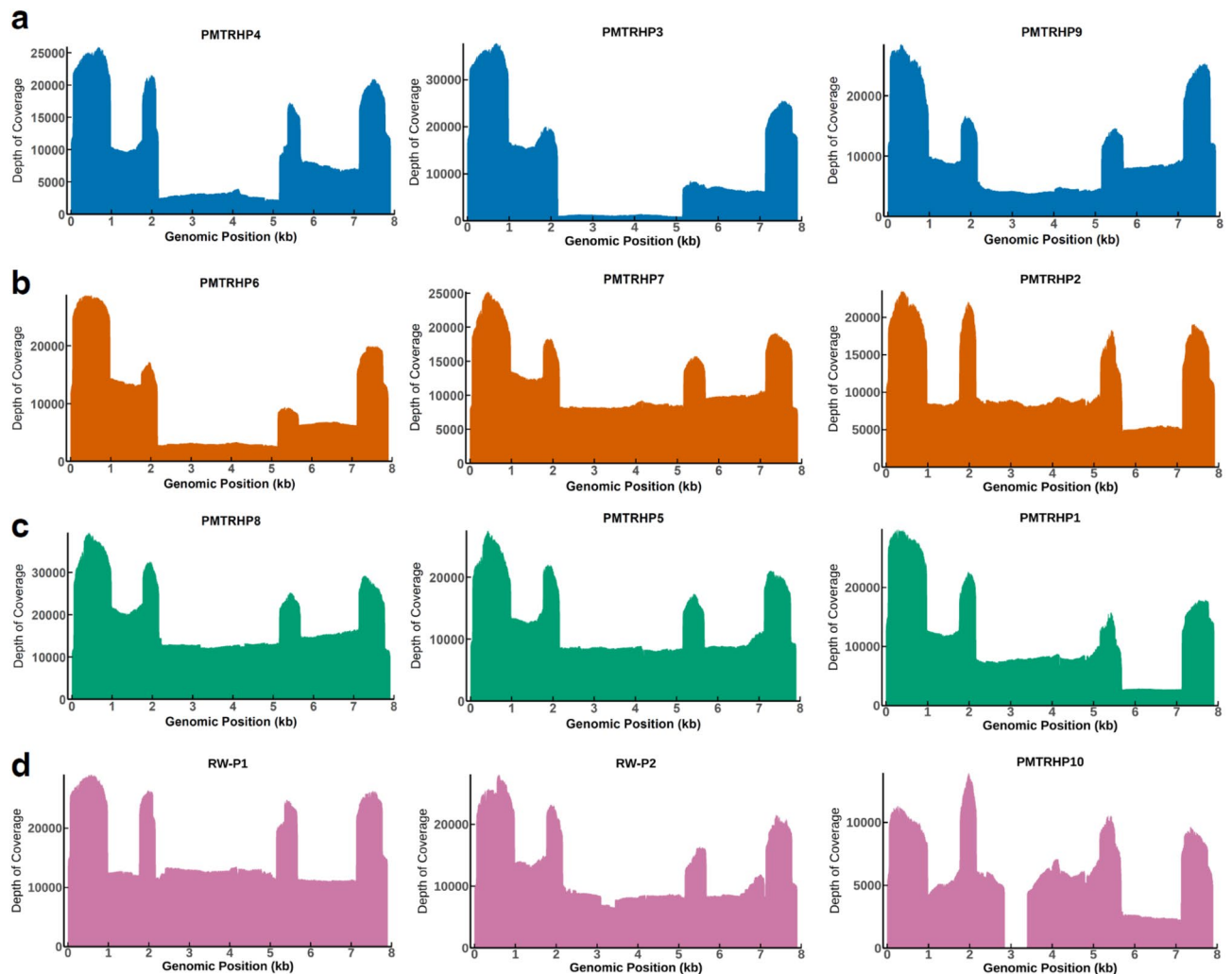


Fig. 3. Depth of coverage analysis of HPV16 genomes in clinical cervical isolates. Read depth and coverage profiles for 12 HPV16-positive clinical samples, categorized by lineage, were generated from bedgraph files. **a** Lineage A (blue), **b** Lineage B (vermilion), **c** Lineage C (green), and **d** Lineage D (magenta), with each panel representing three clinical samples per lineage. The x-axis indicates the genomic position (kb) of HPV16, and the y-axis shows the sequencing depth. PMTRHP3 showed $>1,000\times$ coverage, while PMTRHP10 exhibited a drop near 3 kb, consistent with a 540 bp deletion (positions 2851–3390). High-intensity peaks reflect overlapping amplicons, indicating strong and consistent coverage across samples.

assembled sequences generated using the canu package. Molecular Evolutionary Genetic Analysis (MEGA) software was used for pairwise and multiple sequence alignments, followed by the construction of a maximum likelihood phylogenetic tree.

Our method successfully recovered complete HPV16 genomes from all clinical samples (Fig. 4a). Phylogenetic analysis revealed four distinct lineages: A, B, C, and D (Fig. 4a). Within lineage A, two samples were classified as sublineage A1 and one as A2. Samples from lineage B aligned with sublineage B4, Lineage C with sublineage C1, and Lineage D samples with sublineage D3. These results demonstrate that our WGS method reliably recovers high-quality complete HPV16 genomes, enabling precise phylogenetic classification.

In parallel, reference-based alignment and variant calling were performed on sorted BAM files using Clair3, with variant call format (VCF) files subsequently merged via BCFtools. A custom bash script and R workflow were used to process VCFs for phylogenetic tree construction and Uniform Manifold Approximation and Projection (UMAP) clustering (Fig. 4b, c). As expected, the reference-based phylogeny reproduced the four lineage clusters (A, B, C, and D) (Fig. 4b) observed in the de novo analysis (Fig. 4a). UMAP clustering further confirmed these four distinct groups (Fig. 4c).

The concordance across de novo assembly, reference-based phylogeny, and unsupervised clustering underscores the robustness of our method for recovering complete HPV16 genomes. This method enabled precise lineage classification and downstream evolutionary analyses, affirming the extensive genetic heterogeneity of HPV16 as observed in previous studies^{10,35,36}.

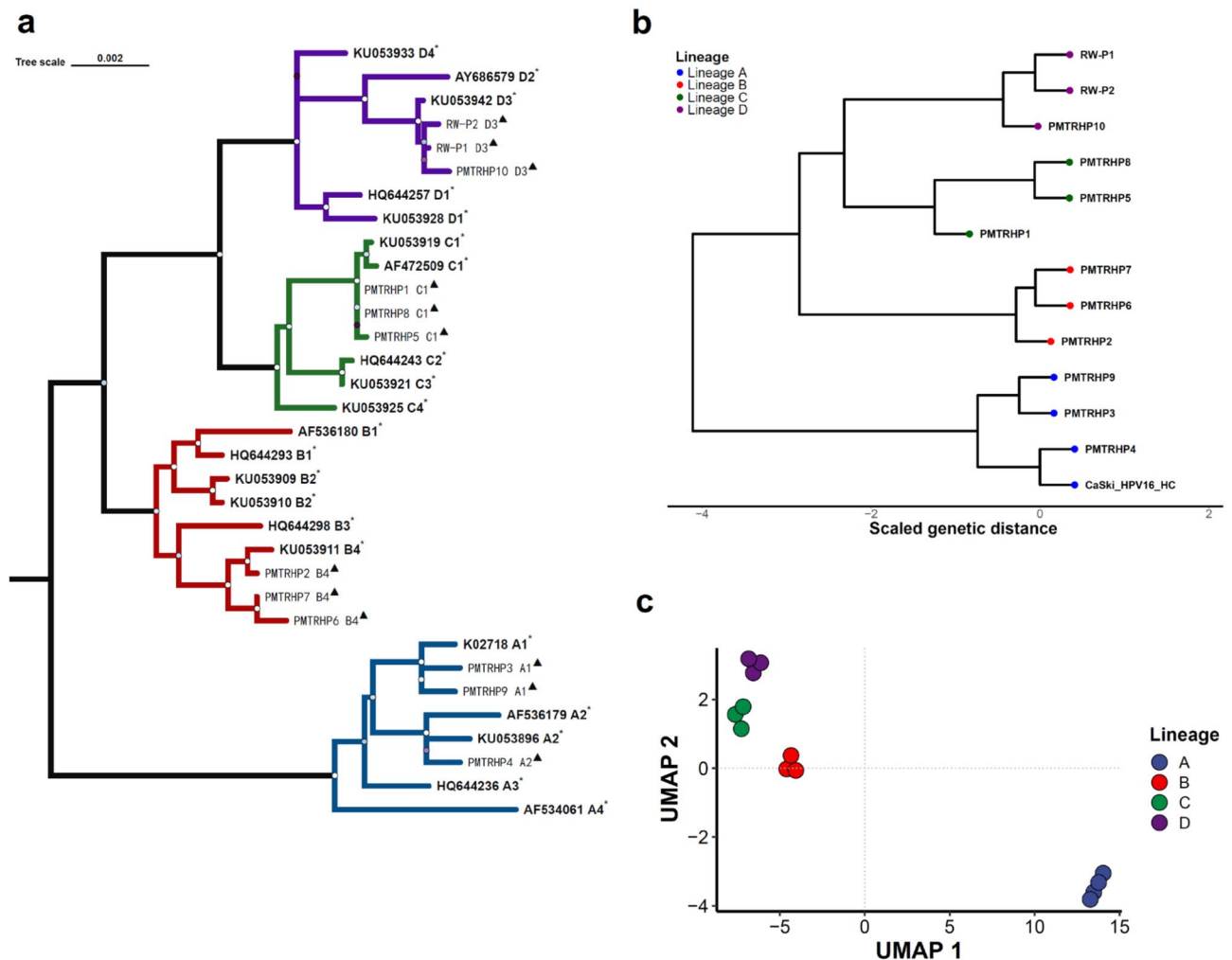


Fig. 4. Phylogenetic tree and UMAP clustering of HPV16 genomes. **a** A maximum likelihood phylogenetic tree constructed from reference HPV16 genomes (GenBank; marked with an asterisk) and assembled clinical isolates from cervical swabs (black triangles). Branch colors indicate the four major HPV16 lineages. The scale bar represents genetic distance. Node support is visualized with gradient circles, where brighter shades indicate higher bootstrap values. **b** A phylogenetic tree generated from variant call format (VCF) files, with tip colors corresponding to HPV16 lineages as in panel a. **c** A UMAP plot based on processed VCF data, illustrating unsupervised clustering of HPV16 genomes. In all panels, HPV16 lineages are color-coded as follows: A (blue), B (red), C (green), and D (purple).

Mutational profiles of clinical HPV16 isolates

To further evaluate the suitability of ONT WGS and our optimized bioinformatics pipeline for HPV16 genomic analysis, we analyzed viral genomes from our clinical samples using the PaVe HPV16 reference (accession number K02718). Across all samples, we identified a total of 253 variants, the vast majority of which were SNPs (244/253, 96.4%), while the remainder was insertions and deletions (indels; 9/253, 3.6%), including two insertions and seven deletions (Fig. 5a). This corresponded to an average of one variant per 31 nucleotides. When classified by predicted functional impact, 52.8% of the variants were synonymous and non-coding SNPs, and 47.2% were missense mutations, resulting in a missense-to-silent ratio of 0.89 (Fig. 5b). The absence of stop codons supports the accuracy of the sequencing, as these clinical isolates are expected to be intact. Excluding the highly variable upstream regulatory region (URR), variants were broadly distributed across the remaining coding regions (Fig. 5c), with the following gene-level counts: E1 (n = 44), E2 (n = 37), E5α (n = 12), E6 (n = 14), E6 (n = 7), E7 (n = 4), E8^E2 (n = 26), L1 (n = 28), and L2 (n = 56). Regarding normalized variant frequency by gene, E5 exhibited the highest number of variants, followed by L2 and E2. Meanwhile, E7 was the least variable gene, consistent with previous studies^{37,38} (Fig. 5c). Focusing specifically on missense mutations, E2 harbored the highest number (n = 24), followed by E1 and L2 (n = 17). In contrast, E7 contained only a single missense SNP (Fig. 5d).

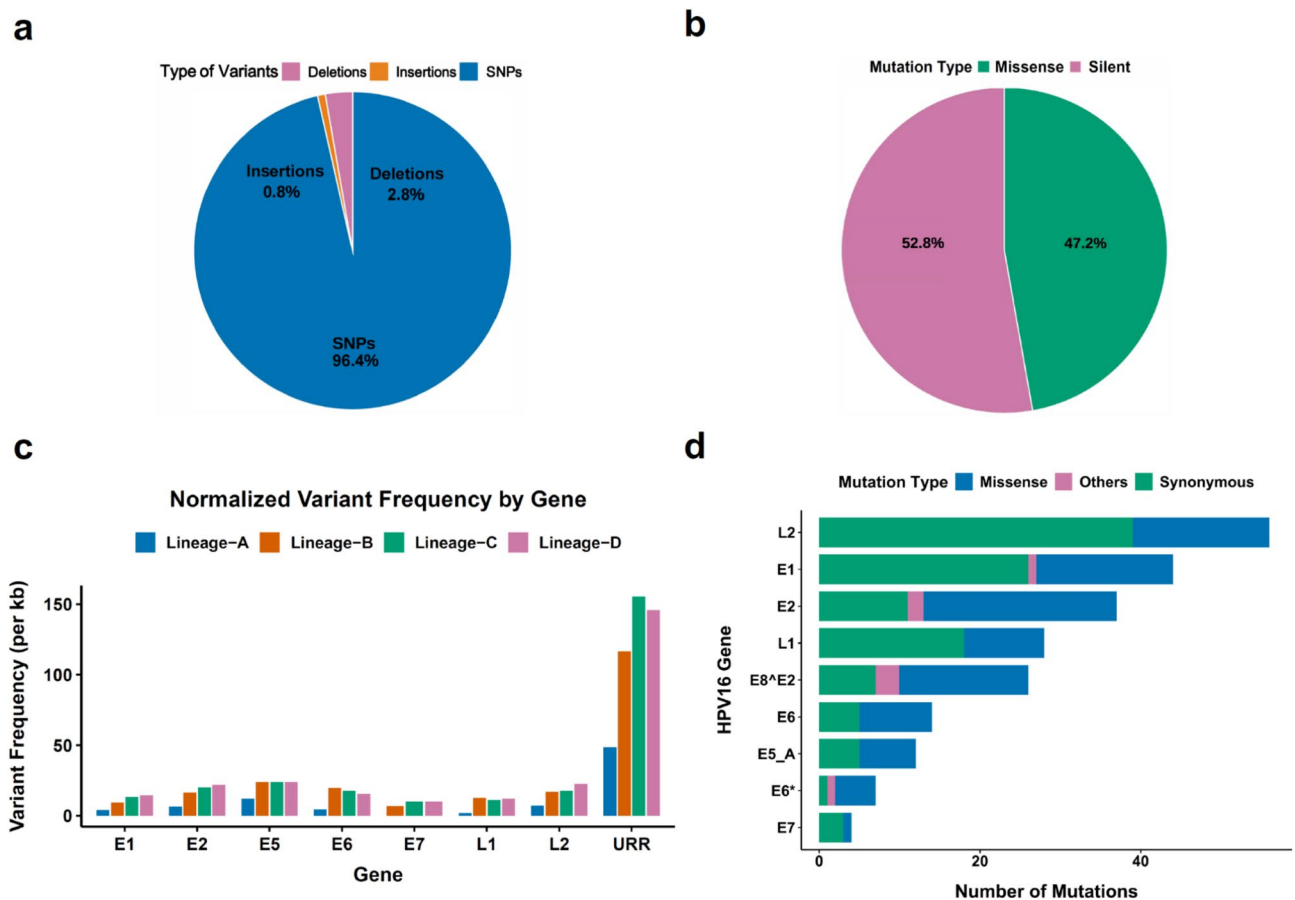


Fig. 5. Analysis of fixed mutations in HPV16 clinical isolates. **a** Distribution of variant types: single-nucleotide polymorphisms (SNPs; blue), insertions (vermilion), and deletions (magenta). **b** Ratio of missense (green) to silent (magenta) mutations. **c** Normalized variant frequency by gene across HPV16 lineages, color-coded as A (blue), B (vermilion), C (green), and D (magenta). **d** Mutation counts across HPV16 genes: synonymous (green), missense (blue), and other protein-altering mutations (magenta), including disruptive in-frame, frameshift, and stop-gained variants.

Discussion

Here, we present a robust amplicon-based WGS approach for HPV16, employing a trio-primer enrichment strategy optimized for ONT. This method achieves full genome coverage, including previously challenging regions such as E5, E2, and URR²⁰, and performs well with low-abundance samples. Junction-specific primers facilitate the recovery of the complete genome in both circular and tandem forms. The method leverages ONT platforms, including PromethION 2 Solo (P2) and MinION, which are well-suited for deployment in resource-limited settings. A previous study employed the MinION for HPV-ICB2 WGS³⁹ but required rolling circle amplification (RCA) before long-range PCR, a limitation for integrated HPV types such as HPV16 in CaSki cells, where circular templates are absent. The ONT Rapid Barcoding Kit enables cost-effective multiplexing of up to 96 samples. Our workflow begins with the collection of cervical swabs, followed by DNA extraction, PCR-based viral enrichment, and library preparation using this kit. Sequencing is performed on a single flow cell, with the overall cost from extraction to sequencing estimated at approximately USD \$25–\$30 per sample, depending on the specific Lab settings. Costs may be further reduced by repurposing the flow cell for additional sequencing runs targeting taxonomically unrelated organisms, thereby minimizing the risk of barcode cross-reactivity. Additional scalability could be achieved through the incorporation of unique molecular identifiers (UMIs), as demonstrated by Karst et al.,⁴⁰ which may enhance high-throughput sequencing, an area historically dominated by Illumina platforms that offer high base-call accuracy (~0.1% error) but are limited to short (~500 bp) amplicons⁴¹. This scalable and cost-effective strategy has strong potential for advancing HPV16 genomic surveillance.

Applied to 12 cervical clinical isolates previously genotyped as HPV16-positive using the Roche Linear Array, our method achieved complete genome coverage across all samples. Median read depths ranged from 5,899× to 15,279×, with minimum depths exceeding 500× even in challenging regions such as E2, E4, E5, and L2. The high and uniform sequencing coverage achieved across all samples enabled accurate de novo assembly and high-confidence variant calling, facilitating the recovery of complete, high-quality HPV16 genomes from clinical samples. Read-based variant calling, wherein individual sequencing reads are aligned directly to

a reference genome, offers high sensitivity for detecting single-nucleotide polymorphisms (SNPs) and small insertions or deletions (indels), including low-frequency variants. In contrast, assembly-based approaches align consensus contigs to a reference genome, thereby improving the detection of larger structural variants and complex genomic rearrangements. However, these methods may overlook minor or low-frequency variants due to consensus smoothing or errors introduced during the assembly process. Both reference-based and de novo analyses consistently identified four major phylogenetic clusters corresponding to the established HPV16 lineages: A (European), B (African 1), C (African 2), and D (Asian)^{9,10,35,36}. These lineage assignments were further confirmed by unsupervised clustering using UMAP, which preserves local and global data structures, revealing a clear separation of the four groups. The precise classification of samples into their respective sub-lineages aligns with previous phylogenetic studies^{9,10,35,36}, validating the accuracy and strength of our approach. The concordance across de novo assembly, reference-guided phylogeny, and dimensionality reduction-based clustering underscores the robustness and resolution of our method and analysis pipeline.

Importantly, our approach significantly enhances accurate lineage assignment, enabling downstream evolutionary and epidemiological analyses that capture the extensive genetic heterogeneity of HPV16. The ability to resolve lineage diversity with high confidence is critical for molecular epidemiology, public health surveillance, and studies on lineage-specific pathogenicity. Furthermore, our pipeline's capability to generate phylogenies and UMAP or t-SNE clusters directly from processed VCF files offers an alternative and scalable framework for HPV16 genotyping, eliminating the need for full-genome assemblies. Additionally, our method effectively captured the full extent of genomic variation, with one variant occurring approximately every 31 nucleotides, predominantly as SNPs. Coding changes were evenly split between synonymous and missense mutations (ratio = 0.89). Variants were broadly distributed, showing enrichment in E5, L2, and E2, while E7 remained highly conserved, consistent with prior reports^{37,38}. Collectively, these findings establish our WGS workflow as a reliable and high-resolution platform for HPV16 genomic analysis, with broad utility in clinical diagnostics, viral evolution studies, and public health surveillance.

PEPPER and Clair3 variant calling demonstrated strong overall performance in HPV16 whole-genome variant analysis. Both tools achieved comparable F1-scores and demonstrated high precision and recall for SNP detection, underscoring their suitability for analyzing single-nucleotide variants in viral genomes. However, indel calling remained a challenge, with reduced performance across both callers. This limitation reflects well-documented constraints of ONT data, particularly from R9.4.1 flow cells and Guppy basecalling, where systematic errors are prevalent in homopolymeric and repetitive regions^{31,33}. These platform-specific biases underscore the need for improved indel resolution and benchmarking strategies tailored to ONT-based viral genome analysis. Furthermore, our study integrated ONT data with publicly available Illumina data from CaSki cell lines; however, minor differences in cell line sources may have slightly affected indel concordance. Future comparative analyses should include a wider panel of controls, such as additional HPV16-positive cell lines (e.g., SiHa), and well-characterized clinical specimens sequenced across multiple platforms using the same DNA material to improve benchmarking accuracy. As ONT technology continues to evolve, gains in basecalling accuracy and chemistry are expected to enhance variant detection. Both callers will likely benefit, improving their utility for high-resolution HPV16 genomic studies. Importantly, the tools employ distinct computational strategies. PEPPER uses full alignment, while Clair3 combines pileup-based and full alignment approaches, offering different trade-offs between accuracy and computational efficiency. These considerations are critical when scaling analyses across large cohorts. Overall, this study provides a methodological proof of concept that lays the groundwork for deeper investigations into HPV16 genomic diversity. While the current findings offer valuable insights, they also underscore the need for validation across larger and more diverse sample sets to ensure robustness and generalizability. Such efforts will be critical for translating these approaches into meaningful clinical and epidemiological applications.

This study has limitations. First, we did not quantify HPV16 viral load using qPCR to correlate with sequencing depth. Instead, we relied on Qubit-based estimates using known CaSki cell copy numbers, which approximate a detection limit of five copies per 2 μ L. Second, although the study was designed as a proof-of-concept and currently utilizes three singleplex reactions, only minimal advantages are anticipated from transitioning to a two-tube multiplex format. Future research should prioritize primer optimization to enable single-tube multiplexing, which would substantially improve scalability and reduce per-sample costs in high-throughput settings. Third, INDEL detection remained suboptimal, likely due to known limitations of Oxford Nanopore Technologies (ONT) R9.4.1 flow cells and Guppy basecalling in homopolymeric regions. However, recent improvements, including Kit 14 chemistry, R10.4.1 flow cells, and enhanced basecalling models, are expected to improve variant detection accuracy. Finally, we did not analyze lineage or sublineage diversity and their correlation with clinical risk in women from sub-Saharan Africa due to the small sample size ($n = 12$). A future study with a larger cohort from sub-Saharan Africa is planned to examine the relationships between lineage, sublineage, and the progression of cervical disease.

In conclusion, we have developed a scalable and deployable WGS workflow for both linear integrated and episomal HPV16 using ONT long-read technology, coupled with an optimized analysis pipeline. We generated a benchmark set of HPV16 'truth set' variants and demonstrated that PEPPER and Clair3 variant callers perform reliably for SNP detection, supporting their potential utility in HPV16 variant analysis. While further validation in larger, diverse cohorts is warranted for broader clinical adoption, the strong concordance observed across de novo assembly, reference-based phylogenetics, and unsupervised clustering highlights the robustness of our approach for reconstructing complete HPV16 genomes. A forthcoming study will leverage this strategy to investigate the epidemiological implications of HPV16 in sub-Saharan Africa. Notably, the capacity of our WGS framework to resolve fine-scale variation across the HPV16 genome enables robust analyses of coding diversity and its potential functional implications. This work contributes to global efforts in HPV surveillance and CC elimination, particularly in high-burden, resource-limited settings.

Abbreviation	Name	Primer sequence (5' → 3')	Amplicon size (bps)	Start	Stop
T1	HPV16-31FO	CGTAACCGAAATCGGTTGAACC	2151	31	52
	HPV16-2160RE	CCAATCACCTCCATCATCTACC		2181	2160
T2	HPV16-1744FO	TGTCTAACTATTATGTGTGTCTCCAATG	3950	1745	1773
	HPV16-5692RE	GATACTGGGACAGGAGGCAAGTAGACAGT		5694	5666
T3	HPV16-5133FO	TGGCATTAGGTACAGTAGAATTGG	2658	5133	5156
	HPV16-7769RE	ACTAGGGTGACATTAGTTGGC		7790	7769
U1	HPV16-7110FO	CCTCATCTACCTCTACAACCTGC	1795	7110	7131
	HPV16-973RE	CCTGTATCACTGTCATTTTCGTTCTC		998	973
L1	HPV16-31FO	CGTAACCGAAATCGGTTGAACC	7760	31	52
	HPV16-7769RE	ACTAGGGTGACATTAGTTGGC		7790	7769

Table 2. Primer pairs used in this study for complete whole-genome sequencing of HPV16. Primer names, abbreviations, and corresponding sequences are presented alongside their nucleotide start and stop positions. All coordinates provided relative to the HPV16 reference genome (GenBank accession: K02718.1).

Methods

Study design, population, and ethical approvals

Ten archived cervical swab specimens were obtained from women enrolled at cervical cancer screening centers under the Academic Model Providing Access to Healthcare (AMPATH) in western Kenya, and two specimens were obtained from the Einstein-Rwanda Research and Capacity Building Program (ER-RCBP) at Research for Development (RD Rwanda), Rwanda. These samples were collected as part of ongoing studies approved by the US National Cancer Institute (NCI) and the National Institutes of Health (NIH). Written informed consent was obtained from all participants. Ethical approvals were obtained from the Institutional Research and Ethics Committee (IREC) of Moi Teaching and Referral Hospital (MTRH) and Moi University, Eldoret, Kenya (reference number IREC/371/2022), as well as from the Institutional Review Board (IRB) of Brown University, Providence, Rhode Island, USA (IRB-ID: 2,023,003,553). Additionally, this study utilized the HPV16-positive CaSki and HPV16-negative C33A cervical carcinoma cell lines, which were provided by the laboratory of Dr. Rachel A. Katzenellenbogen at Indiana University School of Medicine in Indianapolis, IN, USA.

Human papillomavirus 16 detection and genotyping

Roche linear array

DNA extraction from archived cervical swab specimens was performed using the Qiagen Mini-Elute kit, following the manufacturer's instructions. Subsequently, typing was conducted using the Roche Linear Array® (LA) (Roche, Branchburg, NJ, USA) to detect 37 HPV types, including 12 high-risk, eight probable high-risk, and 17 low-risk HPV genotypes. Genotyping with LA was performed as previously described⁴², amplifying 450-bp fragments from the L1 region of the HPV virus by PCR and using a reverse line blot system for simultaneous detection of up to 37 HPV genotypes (i.e., 6, 11, 16, 18, 26, 31, 33, 35, 39, 40, 42, 45, 51, 52, 53, 54, 55, 56, 58, 59, 61, 62, 64, 66, 67, 68, 69, 70, 71, 72, 73, 81, 82, 83, 84, IS39, and CP6108). The LA assay, widely used in epidemiological studies, is regarded as the reference standard for HPV genotyping.

DNA purification for long-range PCR and downstream next-generation sequencing

The HPV16-positive cervical swab isolates were selected for Nanopore HPV16 whole genome sequencing. The DNA for the downstream HPV16 WGS was purified from the CaSki cell line and cervical exfoliate cells. Cervical exfoliated cells were first eluted in 1 ml of 1 × phosphate-buffered saline (PBS). Subsequently, DNA extraction was performed using 450 µL of the elute with the Monarch® Genomic DNA Purification Kit (New England Biolabs, Ipswich, MA, USA), following the manufacturer's instructions. This DNA purification kit yields an excellent amount of input material for applications such as long-range PCR and downstream next-generation sequencing (NGS) library preparation. The DNA extracted from the HPV-positive cervical carcinoma cell lines CaSki and C33A served as positive and negative controls, respectively. Purified DNA concentration was quantified using the Qubit dsDNA HS Assay kit (Thermo Fisher Scientific).

Trio-primer sets for pre-sequencing PCR viral enrichment.

We designed primers for viral enrichment based on the HPV16 whole genome reference sequence (accession number K02718) from the Papillomavirus Episteme (PaVE) database⁴³ (Table 2). The Integrated DNA Technology (IDT) and the NCBI Primer-BLAST software were used to design three primer sets (Fig. 1a and b): Primer set 1, a near full-length PCR primer set abbreviated **L1**, which generated long amplicons spanning up to 7.7 kb of the HPV16 genome. This primer set amplifies any intact or long HPV16 DNA sequences in a sample. Primer set 2- tiling-path primer sets abbreviated **T₁**, **T₂** (adopted from Yusuke Hirose et al., 2019)⁴⁴, and **T₃** give overlapping amplicons of approximately 2.1 kb, 3.9 kb, and 2.6 kb in length, respectively. This set of primers helps to enrich fragmented HPV16 DNA in a sample. Primer set 3, the junction primer set (abbreviated **U1**), amplifies the split region of HPV16 when it is linearized and integrated into the host. To assess the sensitivity of our primers, we performed serial dilutions of DNA extracted from the CaSki cell line, which is known to contain an integrated HPV16 genome.

Polymerase chain reaction (PCR) for HPV16 components and conditions

DNA samples were amplified using specific primers and a high-fidelity PCR kit (New England Biolabs) on an Eppendorf Vapo Protect MasterCycler Pro system. Each 50 µl PCR reaction typically included 2 µl sample template DNA, 10 µM forward primers (2.5 µl), 10 µM reverse primers (2.5 µl), 10 mM dNTPs (1 µl), 5X Q5 Reaction Buffer (10 µl), (Nuclease-Free Water (31.5 µl) and Q5 High-Fidelity DNA Polymerase (0.5 µl). PCR amplification was performed for 40 cycles in the thermocyclers under the following conditions. All reactions were performed with human genomic C33A negative control DNA and Caski HPV16 positive control. The cycling conditions for each primer set were optimized as follows: The **U1** primer set underwent an initial denaturation at 98 °C for 90 s, followed by 40 cycles of denaturation at 98 °C for 10 s, annealing at 65 °C for 20 s, extension at 72 °C for 60 s, and a final extension at 72 °C for 2 min. Similarly, the **T1** primer set had an initial denaturation at 98 °C for 90 s, followed by 40 cycles of denaturation at 98 °C for 10 s, annealing at 66 °C for 20 s, extension at 72 °C for 60 s, with a final extension at 72 °C for 2 min. For the **T2** primer set, the cycling conditions were as follows: an initial denaturation at 98 °C for 90 s, followed by 40 cycles of denaturation at 98 °C for 10 s, annealing at 65 °C for 20 s, extension at 72 °C for 5 min, with a final extension at 72 °C for 2 min. The **T3** primer set underwent an initial denaturation at 98 °C for 90 s, followed by 40 cycles of denaturation at 98 °C for 10 s, annealing at 64 °C for 20 s, extension at 72 °C for 75 s, and a final extension at 72 °C for 2 min. Finally, the **L1** primer set underwent an initial denaturation at 98 °C for 90 s, followed by 40 cycles of denaturation at 98 °C for 10 s, annealing at 65 °C for 20 s, extension at 72 °C for 5 min, with a final extension at 72 °C for 2 min.

Library preparation and nanopore WGS of HPV16

The PCR amplicons were amplified separately for each primer pair. They were then pooled for each sample, cleaned using in-house PCR cleanup magnetic beads, and prepared for ONT sequencing using the Oxford Nanopore Rapid Barcoding kit (SQK-RBK110.96, Oxford Nanopore). Briefly, 9 µl of template DNA (ranging from 50 to 400 ng) was mixed with 1.4 µl of Rapid Barcodes (RB01-96, one for each sample), and the mixture was incubated in a thermocycler at 30 °C for 2 min, followed by 80 °C for 2 min, and then cooled to 4 °C. The barcoded samples were pooled, and SPRI magnetic beads were added for bead clean-up in an equal volume. Beads were washed twice in 500 µl of 80% ethanol, and DNA was eluted in a 45 µl elution buffer. After pelleting the beads on a magnet, 45 µl of eluate containing the DNA library was collected, and the concentration of the library was measured using a Qubit fluorometer. Subsequently, 11 µl of the library was combined with 1.4 µl of Rapid Adapter F (RAP-F), mixed gently by flicking the tube, and then spun down. The mixture was incubated at room temperature for 5 min. The flow cell primer was prepared by mixing 30 µl Flush Tether (FLT) and 1170 µl Flush Buffer (FB) for loading onto the flow cell. For sequencing, 37.5 µl Sequencing Buffer (SQB), 25.5 µl Loading Beads (LB), and 12 µl of the DNA library were combined, and 75 µl of the library was loaded onto the Nanopore R9.4.1 flow cell for sequencing using a MinION sequencer.

Base calling and demultiplexing of ONT sequences

Raw data (FAST5 files generated) from the Nanopore sequences were base-called and demultiplexed using Guppy 6.0.1 to FASTQ using the command for base calling with the R9.4.1 l: `guppy_basecaller -x "cuda:0" -c dna_r9.4.1_450bps_sup.cfg -input_path ./fast5 -save_path guppy6_sup_out/ -compress_fastq -enable_trim_barcode_kits SQK-RBK110-96`.

De Novo assembly of HPV16 genomes

The HPV16 genome assembly was performed using Canu version 2.2⁴⁵, which specializes in assembling long-read sequences. HPV16 fastq files for each sample were concatenated and used in genome assembly. For the de novo assembling process, we set Canu parameters as follows: `./canu readSamplingCoverage=100 -p sampleID -d sampleID minReadLength=500 minOverlapLength=250 genomeSize=8 k -nanopore-raw sample.fastq.gz`. This process involved three phases: correction, trimming, and assembly, which improved base accuracy, removed low-quality regions, and generated contigs, consensus sequences, and alternate paths graphs. Finally, the assembled HPV16 genomes underwent manual correction and polishing using BioEdit v7.7.1.0 software.

HPV16 reference-based variant calling pipeline

To analyze HPV16 sequences, we aligned long-read sequences to the HPV16 reference genome using Minimap2 with the command `minimap2 -ax map-ont -a ./HPV16REF.fa Sample.fastq.gz > Sample.sam`. The resulting SAM file was converted to a BAM file for efficient storage using `samtools view -bS Sample.sam > Sample.bam`. The BAM file was then sorted with `samtools sort -O BAM Sample.bam -o Sample_sort.bam`, and an index was created using `samtools index Sample_sort.bam`. This pipeline effectively transforms raw sequencing data into a sorted and indexed BAM file, ready for variant calling and further analysis.

Variant calling with PEPPER-Margin DeepVariant

To perform high-confidence variant calling from Oxford Nanopore Technologies (ONT) long-read data, we employed the PEPPER pipeline (version 0.8) using a Singularity container (`pepper_deepvariant_r0.8.sif`)²⁹. Reads were aligned to the HPV16 reference genome (GenBank: K02718.1), and variant calling was conducted using the `call_variant` subcommand with the `-ont_r9_guppy5_sup` model, optimized for super-accurate basecalled ONT data. The input BAM file (sorted BAM file *CaSki_HPV16h_sort.bam*) and reference genome HPV16REF.fa were processed using the specified pipeline parameters. Stringent filtering criteria were applied to improve precision and reduce false positives, including `-pepper_snp_q_cutoff 10`, `-pepper_indel_q_cutoff 30`, `-pepper_min_indel_baseq 20`, `-pepper_candidate_support_threshold 60`, and `-pepper_indel_candidate_frequency_threshold 0.85`. Additional thresholds were set for alignment quality and allele frequency: `-pepper_insert_p_value 0.26`, `-pepper_delete_p_value 0.26`, `-pepper_min_coverage_threshold 50`, `-pepper_min_mapq 44`, `-indel_min_af`

0.15, and `-pepper_snp_p_value` 0.000. Haplotype tagging was enabled via `-use_pepper_hp`, and phased BAM output was disabled using `-skip_final_phased_bam`. Variant calls were generated in compressed VCF format for downstream analysis.

A final filtering step was applied using `bcftools` to further enhance variant confidence, retaining only variants with a variant allele frequency (VAF) of 0.27 or higher. This threshold was chosen to minimize false positives while preserving true positive calls relevant to HPV16 genome analysis. The filtering was executed using the command `bcftools view -i 'FMT/VAF[*] >= 0.27' -Oz -o Filtered_PMDV.vcf.gz Input_PMDV.vcf.gz`, resulting in a refined VCF file used for downstream analyses.

Clair3 variant calling

Variant calling was performed using Clair3 (v1.0)³⁰, a tool optimized for long-read sequencing data. Aligned CaSki HPV16 reads (CaSki_HPV16h_sort.bam) were mapped to the HPV16 reference genome (GenBank: K02718.1). Clair3 was executed in a Singularity container (clair3_latest.sif), ensuring consistent deployment across computational platforms. The variant calling command included a model trained for Oxford Nanopore Technologies (ONT) data (*r941_prom_sup_g5014*), with 32 CPU threads allocated for parallel processing. The analysis was executed with parameters set to include all chromosomes (`-include_all_ctgs`), allow for long indels (`-enable_long_indel`), and a minimum SNP allele frequency of 7.5% (`-snp_min_af` 0.075). A minimum coverage threshold of 2 (`-min_coverage` 2) was established, and variant calling was enabled at both the head and tail of sequences (`-enable_variant_calling_at_sequence_head_and_tail`). The chunk size was set to 8000 bases (`-chunk_size` 8000) for more efficient processing. All other parameters were maintained at their default settings, including those for haploid sensitivity, removal of intermediate directories, and use of the specified model path (`-model_path`).

With Clair3, variant calls were generated from both the full-alignment and pileup models. These two outputs were merged and filtered using a custom script to obtain a high-confidence set of variants. The merging process prioritized variants from the pileup VCF, including full-alignment variants only if they were not present in the pileup output. After merging, variants were filtered to retain only those that passed standard quality control (FILTER=PASS), had a genotype quality (GQ) score of at least 5, and did not contain duplicates. This strategy facilitated the integration of high-quality variant calls while minimizing noise, ensuring a reliable call set for HPV16 genome analysis.

Establishment of an HPV16 truth set of variants

To establish proper control of HPV16 variants, also referred to as the 'truth set', we reprocessed publicly available short-read Illumina and in-house long-read Oxford ONT sequencing data derived from the CaSki cell line. The Illumina FASTQ files (European Nucleotide Archive accession: ERP111061) were initially aligned to both the human reference genome (GRCh38/hg38) and the HPV16 reference genome (GenBank: K02718.1) using HISAT2 (version 2.1.0) as described in the original study⁴⁶. We independently realigned the same short-read data to the HPV16 reference genome using BWA-MEM2 (v2.2.1) for our analysis. Long-read ONT data generated in-house were processed using the same pipeline. Following alignment, we used SAMtools for sorting and indexing, GATK's MarkDuplicates tool to mark PCR duplicates, and GATK's AddOrReplaceReadGroups tool to assign read group information. Variant calling was performed using GATK HaplotypeCaller. The resulting VCF files from both sequencing platforms were merged and filtered to exclude low-confidence sites, defined as genomic regions that are consistently difficult to sequence with high accuracy and coverage. In our study, this primarily included homopolymeric stretches, which are particularly prone to errors in Nanopore sequencing. Both the BAM file and corresponding VCF file were visually inspected using the Integrative Genomics Viewer (IGV) and manually curated against previously reported Sanger-derived HPV16 CaSki variants⁴⁷, which served as the reference baseline. Variants not present in the baseline or lacking sufficient read depth, strand balance, or allele frequency (below 30%) were excluded. This variant set served as a ground truth for benchmarking and concordance assessments of Clair3 and PEPPER.

Statistical analysis

The variant calls were analyzed using R (version 4.3.1; The R Project for Statistical Computing). Data were reported as means or medians for discrete and continuous variables, with counts expressed as percentages of the total. To ensure reliable sequencing across the low-coverage-prone region spanning positions 2,168 to 5,165, a minimum depth of 50× is typically recommended. In our dataset, the median read depth within this region was 7,026, indicating that our data exceeded the recommended coverage by approximately 140-fold and could be downsampled accordingly without compromising variant detection. Genomic benchmarking of variant callers is crucial for evaluating the concordance between ground-truth variants and called variants, as well as for identifying true positives (TPs), false negatives (FNs), and false positives (FPs). SNPs and indels were compared against our in-house "truth set" using a custom R script, classifying each variant as true positive (TP), false positive (FP), or false negative (FN). The script also computed key performance metrics, including precision (positive predictive value [PPV]), recall (sensitivity), and F1-score (harmonic mean of precision and recall), using the following formulas:

$$\text{Precision} = \text{True Positives} / (\text{True Positives} + \text{False Positives}) \quad (1)$$

$$\text{Recall} = \text{True Positives} / (\text{True Positives} + \text{False Negatives}) \quad (2)$$

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (3)$$

Phylogenetic analysis

Polished 12 clinical patient samples' FASTA sequences, generated with Canu, and HPV16 variant lineages and sub-lineages reference sequences were manually concatenated and subjected to Molecular Evolutionary Genetics Analysis version 11 (MEGA 11) software for pairwise and multiple sequence alignment. We analyzed the resulting Newick tree in the Chiplot Version 2.6.1 online tool.

Data availability

The datasets generated and/or analysed during the current study are available in the Sequence Read Archive (SRA) accession number SAMN50620040 to SAMN50620052 (<https://submit.ncbi.nlm.nih.gov/subs/>). The associated BioProject can be found under accession number PRJNA130603.

Received: 29 May 2025; Accepted: 3 September 2025

Published online: 07 October 2025

References

- Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* vol. 71 209–249 Preprint at <https://doi.org/10.3322/caac.21660> (2021).
- Sengayi-Muchengeti, M. et al. Cervical cancer survival in sub-Saharan Africa by age, stage at diagnosis and Human Development Index: A population-based registry study. *Int. J. Cancer* **147**, 3037–3048 (2020).
- World Health Organization. *Global Strategy to Accelerate the Elimination of Cervical Cancer as a Public Health Problem*.
- Simms, K. T. et al. Benefits, harms and cost-effectiveness of cervical screening, triage and treatment strategies for women in the general population. *Nat. Med.* **29**, 3050–3058 (2023).
- Recommendations and Good Practice Statements on Screening and Treatment to Prevent Cervical Cancer. (World Health Organization, 2021).
- Kelly, H. et al. Diagnostic accuracy of cervical cancer screening strategies for high-grade cervical intraepithelial neoplasia (CIN2+/CIN3+) among women living with HIV: a systematic review and meta-analysis. *EClinicalMedicine* **53**, 101645 (2022).
- Hall, M. T. et al. Benefits and harms of cervical screening, triage and treatment strategies in women living with HIV. *Nat. Med.* **29**, 3059–3066 (2023).
- Chen, Z. et al. Classification and evolution of human papillomavirus genome variants: Alpha-5 (HPV26, 51, 69, 82), Alpha-6 (HPV30, 53, 56, 66), Alpha-11 (HPV34, 73), Alpha-13 (HPV54) and Alpha-3 (HPV61). *Virology* **516**, 86–101 (2018).
- Mirabello, L. et al. HPV16 Sublineage Associations With Histology-Specific Cancer Risk Using HPV Whole-Genome Sequences in 3200 Women. *JNCI: Journal of the National Cancer Institute* **108**, (2016).
- Clifford, G. M. et al. Human papillomavirus 16 sub-lineage dispersal and cervical cancer risk worldwide: whole viral genome sequences from 7116 HPV16-positive women. *Papillomavirus Res.* **7**, 67–74 (2019).
- Galati, L., Di Bonito, P., Marinaro, M., Chiantore, M. V. & Gheit, T. HPV16 phylogenetic variants in anogenital and head and neck cancers: state of the art and perspectives. *Viruses* **16**, 904 (2024).
- Chen, W. et al. The variable clinicopathological categories and role of human papillomavirus in cervical adenocarcinoma: a hospital based nation-wide multi-center retrospective study across China. *Int. J. Cancer* **139**, 2687–2697 (2016).
- De Vuyst, H. et al. The Burden of human papillomavirus infections and related diseases in Sub-Saharan Africa. *Vaccine* **31**, F32–F46 (2013).
- Ermel, A. et al. A cross-sectional analysis of factors associated with detection of oncogenic human papillomavirus in human immunodeficiency virus-infected and uninfected Kenyan women. *BMC Infect. Dis.* **19**, 352 (2019).
- Ermel, A. et al. Invasive cervical cancers in the United States, Botswana and Kenya: HPV type distribution and health policy implications. *Infect. Agent. Cancer* **11**, 56 (2016).
- Gagliardi, A. et al. Analysis of Ugandan cervical carcinomas identifies human papillomavirus clade-specific epigenome and transcriptome landscapes. *Nat. Genet.* **52**, 800–810 (2020).
- Arbyn, M. et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. *Lancet Glob Health* **8**, e191–e203 (2020).
- Xi, L. F. et al. Lineages of oncogenic human papillomavirus types other than type 16 and 18 and risk for cervical intraepithelial neoplasia. *J. Natl. Cancer Inst.* **106**, 270 (2014).
- Burk, R. D., Harari, A. & Chen, Z. Human papillomavirus genome variants. *Virology* **445**, 232–243 (2013).
- Cullen, M. et al. Deep sequencing of HPV16 genomes: a new high-throughput tool for exploring the carcinogenicity and natural history of HPV16 infection. *Papillomavirus Res.* **1**, 3–11 (2015).
- Maueia, C. et al. Identification of HPV16 lineages in South African and Mozambican women with normal and abnormal cervical cytology. *Viruses* **16**, 1314 (2024).
- Iglesias, P. et al. Characterization of human Papillomavirus 16 from Kinshasa (Democratic Republic of the Congo)-implications for pathogenicity and vaccine effectiveness. *Microorganisms* **10**, 2492 (2022).
- Tawe, L. et al. Genetic diversity in L1 ORF of human papillomavirus in women with cervical cancer with and without human immunodeficiency virus in Botswana and Kenya. *BMC Infect. Dis.* **22**, 95 (2022).
- Tegally, H. et al. The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance. <https://doi.org/10.1101/2022.04.17.22273906>. (2022)
- Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).
- Yu, S. et al. Nanopore sequencing for precise detection of Mycobacterium tuberculosis and drug resistance: a retrospective multicenter study in China. *J. Clin. Microbiol.* e0181324 (2025).
- Van Der Auwera, G. A. & Connor, O. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. (Sebastopol: O'Reilly Media, 2020).
- Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* (2017) <https://doi.org/10.1101/201178>.
- Shafin, K. et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* **18**, 1322–1332 (2021).
- Luo, R. et al. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Research Square* <https://doi.org/10.21203/rs.3.rs-1881940/v1> (2022).
- Wagner, J. et al. Benchmarking challenging small variants with linked and long reads. *Cell Genom.* **2**, 100128 (2022).
- McBride, A. A. & Warburton, A. The role of integration in oncogenic progression of HPV-associated cancers. *PLoS Pathog.* **13**, e1006211 (2017).
- Kolmogorov, M. et al. Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nat. Methods* **20**, 1483–1492 (2023).

34. Murenzi, G. et al. Type-specific persistence, clearance and incidence of high-risk HPV among screen-positive Rwandan women living with HIV. *Infect. Agent. Cancer* **16**, 16 (2021).
35. Chen, Z. et al. Niche adaptation and viral transmission of human papillomaviruses from archaic hominins to modern humans. *PLoS Pathog.* **14**, e1007352 (2018).
36. Pimenoff, V. N., de Oliveira, C. M. & Bravo, I. G. Transmission between Archaic and modern human ancestors during the evolution of the oncogenic human papillomavirus 16. *Mol. Biol. Evol.* **34**, 4–19 (2016).
37. Liu, Y., Pan, Y., Gao, W., Ke, Y. & Lu, Z. Whole-genome analysis of human Papillomavirus types 16, 18, and 58 isolated from cervical precancer and cancer samples in Chinese women. *Sci. Rep.* **7**, (2017).
38. Arroyo-Mühr, L. S. et al. Human papillomavirus type 16 genomic variation in women with subsequent in situ or invasive cervical cancer: prospective population-based study. *Br. J. Cancer* **119**, 1163–1168 (2018).
39. Brancaccio, R. N. et al. MinION nanopore sequencing and assembly of a complete human papillomavirus genome. *J. Virol. Methods* **294**, 114180 (2021).
40. Karst, S. M. et al. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat. Methods* **18**, 165–169 (2021).
41. Johnson, J. S. et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 5029 (2019).
42. Brown, D. R. et al. A longitudinal study of genital human papillomavirus infection in a cohort of closely followed adolescent women. *J. Infect. Dis.* **191**, 182–192 (2005).
43. Van Doorslaer, K. et al. The papillomavirus episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Res.* **45**, D499–D506 (2017).
44. Hirose, Y. et al. Whole-genome analysis of human Papillomavirus type 16 prevalent in Japanese women with or without cervical lesions. *Viruses* **11**, 350 (2019).
45. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
46. Lagström, S. et al. TaME-seq: An efficient sequencing approach for characterisation of HPV genomic variability and chromosomal integration. *Sci. Rep.* **9**, 524 (2019).
47. Meissner, J. D. Nucleotide sequences and further characterization of human papillomavirus DNA present in the CaSki, SiHa and HeLa cervical carcinoma cell lines. *J. Gen. Virol.* **80**(Pt 7), 1725–1733 (1999).

Acknowledgements

The authors are grateful to the nurses and study participants at the Cervical Cancer Screening Program at Moi Teaching and Referral Hospital in Eldoret, Kenya, for their hard work, dedication, and generosity in supporting this research. We thank our collaborators in Rwanda, including Gad Murenzi and his team, for providing the two clinical samples used in this study.

Preprint server

The manuscript was deposited as a preprint to medRxiv and is available at <https://doi.org/10.1101/2025.05.22.25328149>. Licensing information states that the authors of this manuscript are copyright holders and have granted medRxiv a license to display the preprint under a CC-BY-NC-ND 4.0 International license.

Author contributions

Conceptualization and design of the study were led by M.K.T., D.G., C.I.O., A.C.E, R.A.K., and J.A.B. Optimization of the protocol M.K.T., D.G., C.I.O., R.M.C., A.A.F, S.T.H., K.N., and J.A.B. Data acquisition M.K.T., S.T.H., I.E.K., R.M.C., S.C., K.M., A.A.F, D.G., A.C.E, C.I.O., R.A.K., and J.A.B. Led the parent studies investigations P.K.T., P.M.I, P.J. L., A.M.M., R.T., D.R.B., O.E.O., S.C., A.C.E, R.A.K and J.A.B. Field work K.M., P.K.T., P.M.I, and O.E.O. Data analysis and interpretation M.K.T., D.G., C.I.O., R.M.C., A.A.F, S.T.H., K.N., R.A.K., and J.A.B. Writing original draft M.K.T., D.G., C.I.O., R.A.K., and J.A.B. Drafting and editing the article M.K.T., S.T.H., I.E.K., R.M.C., K.M., A.A.F, D.G., A.C.E, K.N., C.I.O., P.K.T., P.M.I, P.J. L., A.M.M., R.T., D.R.B., O.E.O., S.C., R.A.K and J.A.B. Funding Acquisition R.T., D.R.B., O.E.O., S.C., A.C.E, R.A.K, J.A.B., O.E.O., and P.J. L. Supervised R.A.K., and J.A.B. All authors contributed to the writing and editing of the manuscript and approved the final version before submission.

Funding

This study was supported by funding from the National Institutes of Health (NIH) (D43TW011317, United States of America; S. Cu-Uvin and O. Omengo) and the National Cancer Institute (NCI) (5U54CA190151-02 and 5U54CA254518-03, United States of America; Rachel A. Katzenellenbogen, D. Brown, O. Omengo, and P. Loehrer) and Providence/Boston CFAR (P30AI042853, United States of America; S. Cu-Uvin).

Declarations

Competing interests

The authors declare no competing interests.

Ethical statement

All participants provided written informed consent. Ethical approvals were obtained from the Institutional Research and Ethics Committee (IREC) of Moi Teaching and Referral Hospital (MTRH) and Moi University, Eldoret, Kenya (reference number IREC/371/2022) and from the Institutional Review Board (IRB) of Brown University, Providence, Rhode Island, USA (IRB-ID: 2,023,003,553). All study procedures were performed in accordance with relevant guidelines and regulations outlined by the Ethics Review Boards indicated above.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-18664-w>

[0.1038/s41598-025-18664-w](https://doi.org/10.1038/s41598-025-18664-w).

Correspondence and requests for materials should be addressed to J.A.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025