Mémoire

Auteur : El Mehdi-Lamghari, Yassine
Promoteur(s) : Durkin, Keith; Baurain, Denis
Faculté : Faculté des Sciences
Diplôme : Master en bioinformatique et modélisation, à finalité approfondie
Année académique : 2024-2025
URI/URL : http://hdl.handle.net/2268.2/23819

Exploring Human Papillomavirus Genomic Diversity Across Geographic and Epidemiological Contexts

University of Liège
Faculty of Science
CHU de Liège

InBioS – Unit of Eukaryotic Phylogenomics
Department of Biomedical Sciences
GIGA – University of Liège

Presented by

El Mehdi-lamghari Yassine

For the Master in Bioinformatics and Modeling

Academic Year 2024–2025

Promotors: Prof. Keith Durkin & Prof. Denis Baurain

Table of Contents

Acknowledgments

I would like to express my sincere gratitude to Prof K. Durkin and Dr M. Artesi for their guidance, valuable advice, and continuous support throughout this thesis. Their expertise, availability, and constructive feedback were instrumental in the successful completion of my work.
I also thank all the members of the laboratory for their contributions to my project, whether through technical assistance or insightful suggestions.

A special and heartfelt acknowledgment goes to Prof D. Baurain, whose help, guidance, and encouragement significantly contributed to this work. Beyond his role as a professor, he generously shared his time, expertise, and methodological insights, allowing me to develop many skills while always fostering a positive and motivating atmosphere. I am also grateful to Prof F. Kerff for his assistance and valuable input on the 3D protein modeling aspect of this work.

Finally, I would like to thank my classmate W. Parla, whose support and collaboration throughout my academic journey were a great source of help and motivation.

1 Introduction

1.1 Human papillomavirus (HPV)

1.1.1 General background

Human papillomavirus (HPV) is a small, non-enveloped, circular double-stranded DNA virus measuring approximately 50–60 nm in diameter, and belongs to the Papillomaviridae family [1]. It is an epitheliotropic pathogen, capable of infecting the basal cells of cutaneous or mucosal epithelia [2]. Although HPV was discovered in the early 20th century, it was not until the 1980s that its causal role in precancerous lesions and cervical cancer was clearly demonstrated. This breakthrough was largely due to the pioneering work of German scientist Harald zur Hausen, who was awarded the Nobel Prize in Medicine in 2008 for establishing the link between HPV and cervical carcinogenesis. He notably showed that HPV16 and HPV18 are responsible for a significant proportion of cervical cancers, and that the E6 and E7 oncogenes play a central role in HPV-mediated oncogenesis [3].

To date, more than 200 distinct HPV genotypes have been identified. These are classified based on the sequence of the L1 gene, and grouped into five genera: Alpha, Beta, Gamma, Mu, and Nu [1,4]. Among them, we distinguish cutaneous types, commonly associated with benign skin warts, and mucosal types, which infect the anogenital and oropharyngeal regions. Mucosal types are further categorized into low-risk types (e.g., HPV6, HPV11) and high-risk oncogenic types (e.g., HPV16, HPV18, HPV31, HPV33, etc.), depending on their potential to induce high-grade intraepithelial lesions and cancers [5].

HPV typically enters the host through microabrasions in the epithelial barrier, allowing access to the basal stem cells of the skin or mucosa. Once inside the host cell, the early (E1–E7) and late (L1–L2) genes of its genome are expressed to produce viral proteins required for replication and assembly. Initially maintained as an episomal element within the nucleus, the viral DNA can, in some cases, integrate into the host genome. This integration tends to occur at chromosomal fragile sites, which are regions more susceptible to DNA strand breaks [6,7].

HPV is primarily transmitted through sexual contact, including vaginal, anal, and oral intercourse, but can also be spread by direct skin-to-skin or mucosal contact. Several risk factors have been associated with both the acquisition and persistence of HPV infection. These include early age at first sexual intercourse, a high number of sexual partners, smoking, prolonged use of oral contraceptives (over five years), and immunosuppression [8,9]. Additional environmental and behavioral cofactors, such as chewing betel nut and exposure to ultraviolet, may also contribute to HPV-related disease progression.

Given the wide genetic diversity of HPV types and their variable pathogenic potential, especially among mucosal types, a deeper understanding of their molecular, phylogenetic,

and structural characteristics remains essential. Such insights are key to improving diagnostics, vaccine development, and ultimately, strategies for the prevention of HPV-associated cancers, notably cervical cancer as well as other anogenital and oropharyngeal malignancies.

1.1.2 Epidemiology and Clinical Relevance

HPV is the most common sexually transmitted infection [10]. Although most infections are asymptomatic and spontaneously resolved by the immune system in more than 90% of individuals, persistent infections with high-risk HPV types can lead to serious health consequences. These include the development of genital warts and, more critically, several forms of cancer. Persistent infection with oncogenic types particularly HPV16 and HPV18 is the primary cause of cervical cancer and is also associated with cancers of the vulva, vagina, anus, penis, and oropharyngeal region (WHO, 2024). The five most common HPV types in HPV-positive women worldwide were HPV16, HPV18, HPV31, HPV58, and HPV52, together accounting for approximately 50% of all infections. Among these, HPV16 and HPV18 are particularly oncogenic and are responsible for about 70% of HPV-related disease burden worldwide [11].

Human papillomavirus (HPV) is the most prevalent sexually transmitted infection worldwide and a major public health concern. Cervical cancer alone accounts for over 90% of HPV-related cancers in women, and nearly all cases (around 99%) are caused by persistent infection with high-risk HPV types In 2022, cervical cancer alone remained the fourth most common cancer in women globally, with an estimated 660,000 new cases and 350,000 deaths, 85% of which occurred in low- and middle-income countries (WHO, 2024).

However, HPV is not only a concern in developing regions: in the United States alone, over 6.2 million people acquire an HPV infection each year. Among sexually active young individuals, about 66% contract genital HPV during the early years of sexual activity [12,13]. Importantly, the burden of HPV-related diseases is not equally distributed across the globe. The highest prevalence of cervical HPV infection is observed in sub-Saharan Africa (24%), followed by Latin America and the Caribbean (16%), Eastern Europe (14%), and South-East Asia (14%).

Several factors contribute to these disparities, including limited access to prophylactic HPV vaccination, cervical cancer screening, and timely treatment of pre-cancerous lesions. Vulnerable groups such as women living with HIV, men who have sex with men, immunocompromised individuals, and those co-infected with other sexually transmitted infections are at particularly high risk of persistent infection and progression to malignancy. Several studies have shown that HPV prevalence varies significantly with age. It tends to be highest among women under 35, followed by a decline in the 35-44 age group. Interestingly, a second increase in HPV prevalence is frequently observed among older women, particularly those aged 45 and above. This age-related pattern has been reported in most global regions, with the notable exception of Asia, where HPV prevalence continues to decrease with age (Figure 1) [14].

While high-income countries are witnessing a gradual decline in HPV-associated cancers due to widespread vaccination and screening programs, projections from the World Health Organization show a significant increase in HPV-related cervical cancer incidence and mortality in low- and middle-income countries if access to preventive strategies is not urgently expanded. The WHO graph (Figure 2) illustrates the sharp contrast in projected cervical cancer cases and deaths between high-income and low- and middle-income countries from 2022 to 2050, highlighting the urgent need for global equity in prevention.

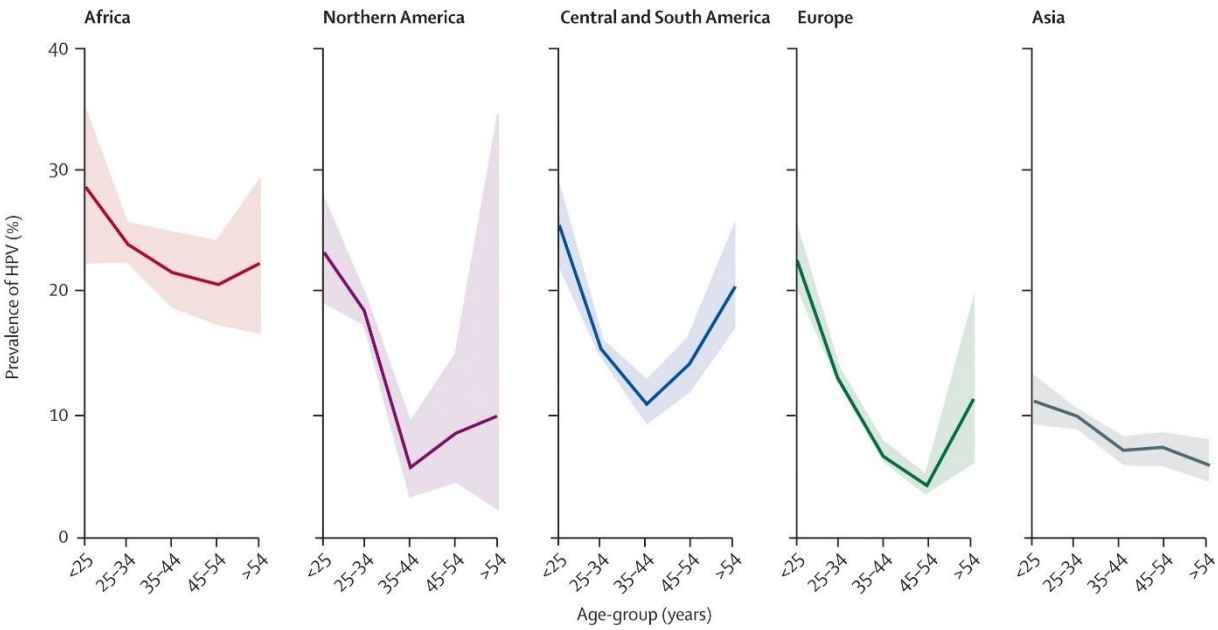

Figure 1: **Age-specific HPV prevalence among women with normal cervical cytology, stratified by world region.** Shaded areas represent the 95% confidence intervals. HPV prevalence is highest among women under 35 years of age, declines in the 35–44 group, and increases again in women aged 45 and older, except in Asia where it continues to decrease.

Source: https://doi.org/10.1016/S1473-3099(07)70158-5 [113]

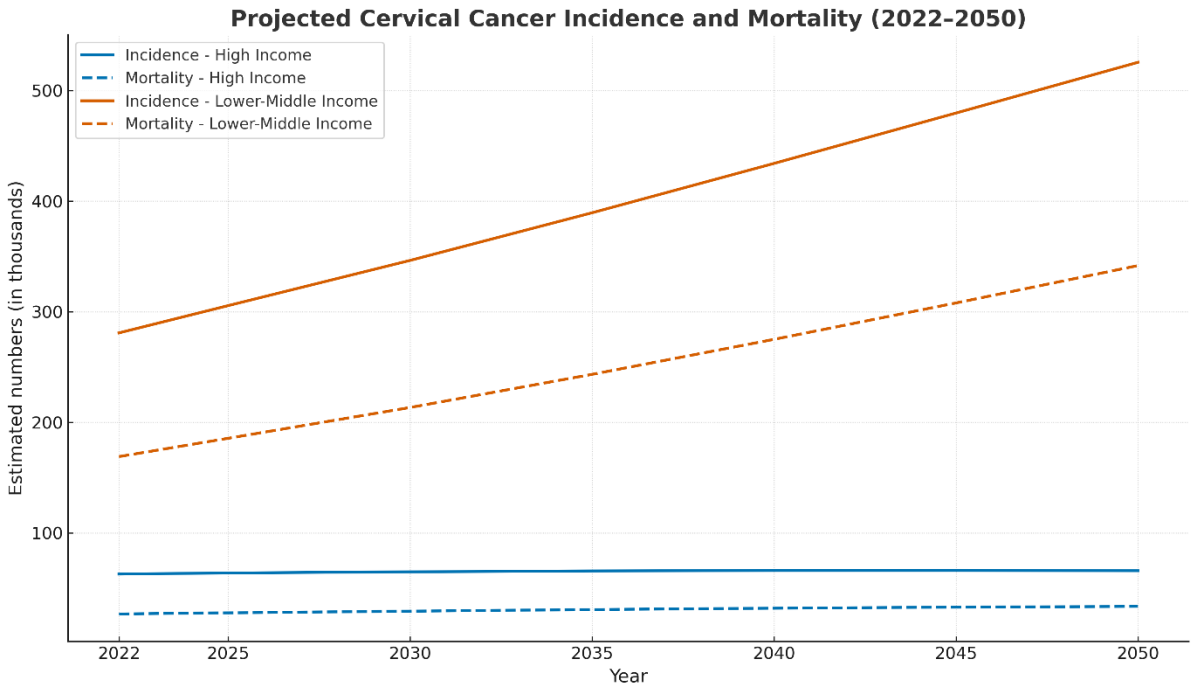

Figure 2. Projected cervical cancer incidence and mortality from 2022 to 2050 in high-income and lower-middle-income countries. This figure compares the estimated number of cervical cancer cases and deaths between high-income countries (blue) and lower-middle-income countries (orange), based on WHO projections. Solid lines represent projected incidence, while dashed lines indicate mortality. Although high-income countries show relatively stable trends over time, a significant increase in both incidence and mortality is expected in lower-middle-income countries.

Source: WHO Cancer Tomorrow – IARC Globocan 2022 (v1.1). Available at: https://gco.iarc.who.int/tomorrow

1.2 HPV lifecycle and Cancer Development

1.2.1 HPV Genome Organization and Viral proteins

Human papillomaviruses possess a genome of approximately 8,000 base pairs, typically maintained in infected cells as episomes. Despite the compact size of the genome, which contains eight to nine open reading frames (ORFs), the virus can produce a variety of proteins through the use of multiple promoters and complex splicing patterns [4]. Papillomaviruses rely on the host cell's replication machinery, which possesses high-fidelity proofreading activity, resulting in a relatively low mutation rate during viral genome replication [15]. The viral genome is functionally divided into three main regions: the early region (E), encoding regulatory and replication-related proteins (E1, E2, E4, E5, E6, and E7); the late region (L),

encoding structural proteins (L1 and L2); and a non-coding upstream regulatory region (URR), which controls transcription and replication [16].

Among the early proteins, E1 is a virus-specific helicase essential for the replication of viral DNA, while E2 assists E1 and plays a key role in genome partitioning and transcriptional regulation [16]. The oncoproteins E6 and E7, especially in high-risk HPV types, contribute to carcinogenesis by targeting tumor suppressor proteins p53 and pRb, respectively, disrupting normal cell cycle control and promoting genomic instability [17].

On the other hand, L1 and L2 form the capsid of the virus. L1, the major capsid protein, assembles into 72 pentamers (360 molecules total) arranged in an icosahedral structure with T=7 symmetry [4]. Each pentamer consists of five L1 molecules featuring a conserved β-jellyroll fold, and inter-capsomeric stabilization is mediated by the C-terminal tails of L1, which form disulfide bridges with adjacent pentamers [18,88]. L2, the minor capsid protein, is less abundant but essential for infection; it facilitates genome packaging, endosomal escape, and trafficking to the nucleus. During viral entry, the N-terminal domain of L2 is exposed and cleaved by host protease furin, a step critical for successful infection [19].

The surface loops of L1 are hypervariable and type-specific, limiting the cross-reactivity of neutralizing antibodies across HPV types, a factor that restricts the breadth of protection conferred by current prophylactic vaccines [20]. Despite some variations in genome organization among HPV types, all humanpapillomaviruses conserve the core genes necessary for replication and assembly, underscoring their essential roles in the viral life cycle.

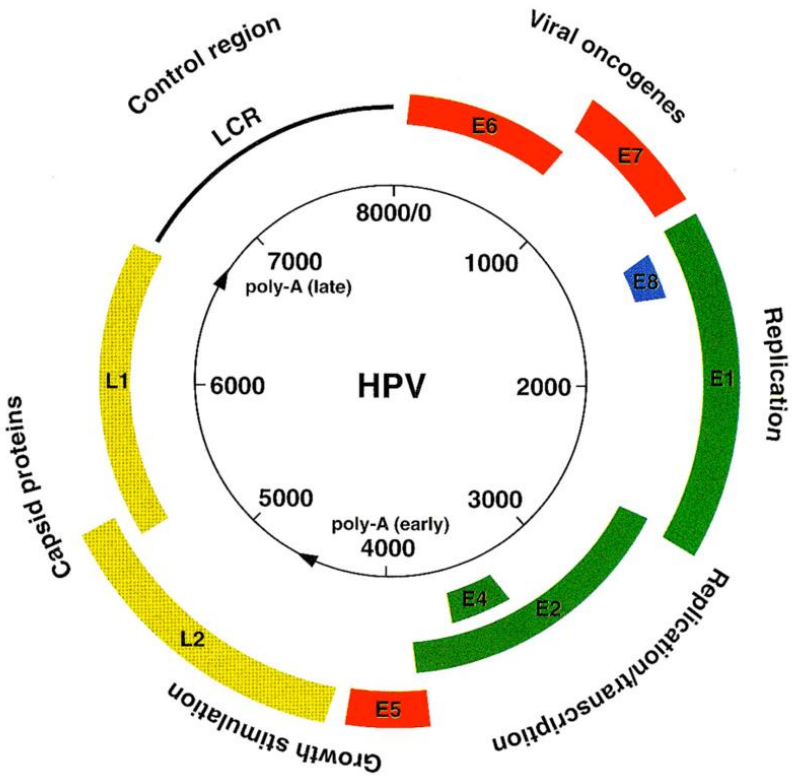

Figure 3: Schematic diagram of the genomic organization of human papillomavirus type 16 (HPV16). The circular double-stranded DNA genome comprises approximately 8,000 base pairs and includes eight open reading frames (ORFs), divided into early (E6, E7,E8, E1, E2, E4, E5) and late (L1, L2) regions. Early ORFs are involved in viral replication, transcriptional regulation, and host cell cycle modulation, while the late ORFs encode structural capsid proteins expressed in differentiated epithelial cells. The upstream regulatory region (URR), also known as the long control region (LCR), contains essential promoter and enhancer elements for transcriptional control.

Adapted from Prendiville, 2004 [114].

1.2.2 Viral Life Cycle of Human Papillomavirus

Upon entry into the host cell, the human papillomavirus (HPV) genome is transported to the nucleus in association with the viral L2 protein. Within the nucleus, the genome establishes itself as low-copy episomal DNA [21]. This initial phase involves rapid replication of the viral DNA, reaching up to 10–200 copies per cell, signifying the early amplification stage of infection. During this period, only the early promoter is transcriptionally active, leading to the expression of early viral proteins such as E1, E2, E6, and E7 (Figure 4) [22]. Infected basal keratinocytes replicate the viral genome in synchrony with cellular DNA during the S phase, and newly synthesized episomes are equally distributed between the two daughter cells. One of these cells remains in the basal layer to maintain proliferation, while the other

migrates upwards into the suprabasal layers and begins differentiation [23]. As the infected cells differentiate, HPV switches to a productive replication mode marked by enhanced E1 and E2 expression, resulting in the synthesis of thousands of genome copies. In the upper, terminally differentiated layers of the epithelium, activation of the late promoter triggers expression of the L1 and L2 capsid proteins, which facilitate the assembly of mature virions. These virions are subsequently shed along with exfoliating squamous cells at the epithelial surface, promoting transmission to new hosts [22]. Throughout the productive cycle in stratified epithelia, HPV genomes undergo three distinct replication modes [24,25]. The first phase, known as the initial amplification, involves rapid episomal replication to establish a stable copy number, relying heavily on E1 and E2 proteins, which recruit the host replication machinery via the upstream regulatory region (URR).

The second phase, or stable maintenance replication, ensures constant episomal copy number through bidirectional theta replication that is synchronized with host chromosomal replication [26]. This form of replication features circular DNA molecules with two replication forks originating at the viral origin (ORI) and proceeding in opposite directions again dependent on E1 and E2 activity.

The third phase, termed vegetative amplification, is characterized by a surge in viral genome production. Both bidirectional theta replication and recombination-dependent replication (RDR) have been identified during this stage [29]. Unlike theta replication, RDR is initiated at various regions across the genome and proceeds unidirectionally without relying on a defined origin of replication [29,31]. These replication intermediates are perceived as damaged DNA by the host cell, triggering a DNA damage response that the virus co-opts for its benefit, this replication occurs in the G2/M phase rather than the canonical S phase. Notably, individual expression of E1, E6, or E7 is sufficient to induce replication stress and activate these host responses [24,27].

While uninfected epithelial cells normally exit the cell cycle during differentiation, the viral oncoproteins E6 and E7 override this control by pushing differentiating cells back into a replication-permissive G2 environment [28]. They achieve this by interfering with tumor suppressor pathways: E7 promotes the degradation of pRb, leading to aberrant activation of E2F transcription factors and premature re-entry into the S phase of the cell cycle [29]. Normally, this unscheduled proliferation would typically activate p53 and trigger cell cycle arrest or apoptosis, but E6 circumvents this checkpoint by binding the cellular ubiquitin ligase E6AP to promote p53 degradation [30–32].

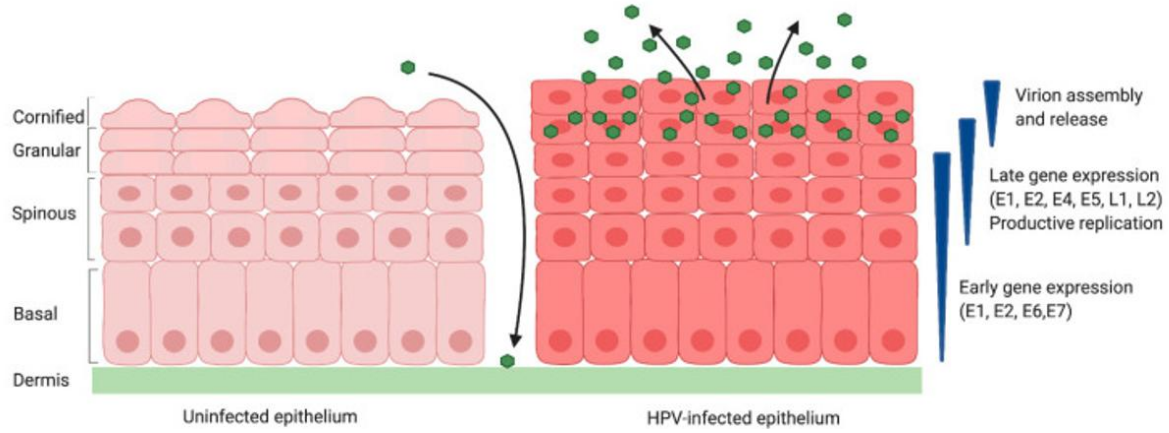

Figure 4: HPV life cycle in stratified epithelium. HPV infects basal keratinocytes through epithelial microwounds. Viral genomes are established as episomes and early genes (E1, E2, E6, E7) are expressed. As cells differentiate, viral replication shifts to a productive phase, leading to genome amplification, late gene expression (L1, L2), and virion assembly. E6 and E7 maintain cell cycle activity in differentiating cells to support replication.

Source: Mac, M.; Moody, C.A. Epigenetic Regulation of the Human Papillomavirus Life Cycle. Pathogens 2020, 9, 483 [115].

1.2.3 Molecular Mechanisms of HPV-Induced Carcinogenesis

Cervical Cancer

The progression from HPV infection to cancer involves a complex interplay between deregulated viral gene expression and genomic instability in host cells. These molecular alterations are frequently triggered by viral genome integration and epigenetic modifications, which disrupt both viral and host gene regulation [33].

In the early stages of infection, particularly in precancerous lesions harboring high-risk HPV (HR-HPV) types, the viral DNA generally remains episomal, existing independently of the host genome. However, as lesions progress, especially toward high-grade neoplasia and invasive cancer, viral genomes often integrate into the host's chromosomal DNA [23,34]. This integration frequently interrupts the E2 open reading frame, a regulatory gene that normally inhibits expression of the viral oncogenes E6 and E7. Loss of E2 leads to overexpression of E6 and E7, two key proteins involved in oncogenic transformation [23,33].

These molecular changes give rise to precancerous lesions referred to as cervical intraepithelial neoplasia (CIN), which are graded based on the extent of epithelial dysplasia [35]. CIN1, or low-grade lesions, typically affect the lower third of the cervical epithelium and often resolve spontaneously within two years. CIN2 involves up to two-thirds of the epithelial layer, while CIN3, considered a high-grade lesion, spans the full epithelial thickness and carries a substantial risk of progressing to invasive cervical cancer if left untreated [36]. This transformation may occur over several years or even decades. Carcinoma in situ (CIS)

represents the stage where malignant cells are still confined to the epithelium and have not yet crossed the basement membrane. Once this barrier is breached, the disease becomes invasive and may spread to adjacent tissues and distant organs [37].

Routine cervical screening, including Pap smears and HPV DNA testing, plays a crucial role in detecting such precancerous changes at early stages. This allows timely intervention and significantly reduces the risk of progression to cervical cancer.

Cutaneous Squamous cell Carcinoma (cSCC)

In addition to mucosal HR-HPVs, which lead to cervical cancer, beta-papillomaviruses (β-HPVs), such as HPV-5 and HPV-8, are associated with the development of cutaneous squamous cell carcinoma (cSCC), particularly in immunosuppressed individuals and those with chronic exposure to ultraviolet (UV) radiation [38,39]. Under normal conditions, UV-induced DNA damage in keratinocytes leads either to efficient DNA repair or to apoptosis if the damage is irreparable. However, β-HPV early proteins E6 and E7 impair these protective mechanisms by inhibiting apoptosis and interfering with nucleotide excision repair and p53-mediated cell cycle arrest. This promotes the survival of UV-damaged cells, increasing the risk of oncogenic mutations and malignant transformation (Figure 5) [39,40].

This sequence of events supports a “hit-and-run” model of carcinogenesis, where early β-HPV involvement initiates tumor development, but persistent viral gene expression is not necessarily required for tumor maintenance [41,42].

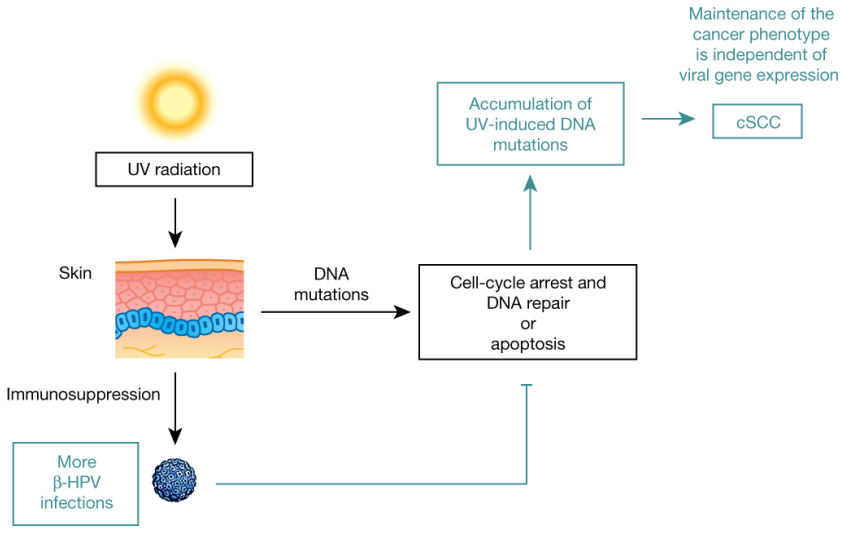

Figure 5: Proposed model of β-HPV contribution to UV-induced cutaneous squamous cell carcinoma (cSCC). Ultraviolet (UV) radiation induces DNA damage in epidermal keratinocytes, triggering either cell-cycle arrest and DNA repair or apoptosis when the damage is

irreparable. In immunosuppressed individuals, increased β-HPV infections further impair this protective response. The early viral proteins E6 and E7 interfere with apoptosis and DNA repair pathways, allowing DNA-damaged cells to survive and accumulate oncogenic mutations. This facilitates the early stages of carcinogenesis. Once transformation is initiated, the maintenance of the malignant phenotype becomes independent of viral gene expression, supporting a "hit-and-run" mechanism of β-HPV-mediated skin carcinogenesis.

Source: Lambert, P.F., Münger, K., Rösl, F. et al. Beta human papillomaviruses and skin cancer. Nature 588, E20–E21 (2020) [116].

1.3 HPV Classification

As previously described in the general background section, the more than 200 known types of HPV are grouped into five genera Alpha, Beta, Gamma, Mu, and Nu based on their L1 open reading frame (ORF). Within each genus, HPV types are further organized into species, designated by numbers (e.g., Alphapapillomavirus 9, abbreviated α-9), which cluster together closely related HPV types. The L1 gene is used for HPV classification because it is highly conserved and aligns well across different HPV types, allowing for a gene-based nomenclature. A difference of more than 10% in the L1 sequence defines distinct HPV types within each genus, as determined by global and pairwise sequence alignment [43].

Beta, Gamma, Mu, and Nu papillomaviruses are typically found on the skin and are primarily associated with benign lesions such as warts; they are generally less likely to cause cancer. In contrast, Alphapapillomaviruses primarily infect mucosal tissues and include types responsible for cervical cancer and other anogenital malignancies [4]. These are divided into two main categories based on their oncogenic potential: high-risk (HR) and low-risk (LR) types.

Based on the International Agency for Research on Cancer (IARC) classification, Alphapapillomaviruses are classified into four groups. Group 1 carcinogens, defined as having sufficient evidence of carcinogenicity in humans, include notably HPV16, HPV18, HPV31, HPV33, HPV35, HPV39, HPV45, HPV51, HPV52, HPV56, HPV58, and HPV59. Some studies also include HPV68 in this group. These are collectively referred to as high-risk HPVs, due to their well-established involvement in the pathogenesis of anogenital and oropharyngeal cancers, including those affecting the cervix, vulva, vagina, anus, penis, and oropharynx [45].

Additionally, types such as HPV26, HPV53, HPV66, HPV67, HPV68, HPV70, HPV73, and HPV82 have been classified as Group 2A (probably carcinogenic) or Group 2B (possibly carcinogenic), reflecting a lower, but still significant, oncogenic potential according to IARC evaluations. It is estimated that approximately 96% of cervical cancer cases are associated with the 13 HPV types classified as Group 1 and 2A (including HPV68) by the IARC [46]. Conversely, 12 HPV types (HPV6, HPV11, HPV40, HPV42, HPV43, HPV44, HPV54, HPV61, HPV70, HPV72, HPV81) are classified as low-risk, and are responsible for about 90% of

anogenital wart cases [48]. The classification of HPV types by oncogenic potential (Groups 1, 2A, 2B, and low-risk types) is summarized in Appendix (Table A1).

According to the ICTV Papillomavirus Study Group criteria, each HPV type may be further divided into lineages and sub-lineages, which differ by approximately 1–10% and 0.5–1% across the full genome, respectively [49]. These genetic variants can differ in geographic distribution as well as in their oncogenic potential.

Taking HPV16 as an example, studies have shown that HPV16 sub-lineages show distinct geographic patterns (Figure 6). Sub-lineage A1 is the most globally widespread, prevalent in Europe, the Americas, South Asia, and Oceania. A2, which is closely related to A1, is also found in Europe, North America, and Oceania. In contrast, A3 and A4 are mostly restricted to East Asia, where they account for approximately 70% of isolates. B and C sub-lineages are primarily found in Africa, representing nearly half the isolates in Sub-Saharan and North Africa, with B1 and C1 being the most common. D lineages, especially D3, are prevalent in South and Central America but are also found in other regions, with the exception of Oceania. D2 is largely confined to the Americas, while D4 appears to be specific to North Africa [50].

Moreover, HPV16 genetic variation has a strong impact on cervical cancer risk. Some studies have shown that D2 and D3, along with A3 and A4 sub-lineages, are associated with an increased risk of cervical cancer compared to the widespread A1 sub-lineage. For D2 and D3 in particular, this elevated risk is notable in adenocarcinoma (ADC) cases. In contrast, the B lineage has been associated with a lower risk of CIN3 [52].

Interestingly, the risk of developing precancerous lesions or cancer also varies depending on the match between a woman's ethnicity and the origin of the infecting variant. Women were found to be at higher risk of CIN3 when infected with an HPV16 variant that geographically aligned with their own ancestral background. For example, an Asian woman infected with an Asian HPV16 variant (such as A4) may have a higher risk of developing CIN3 than a non-Asian woman infected with the same variant [52].

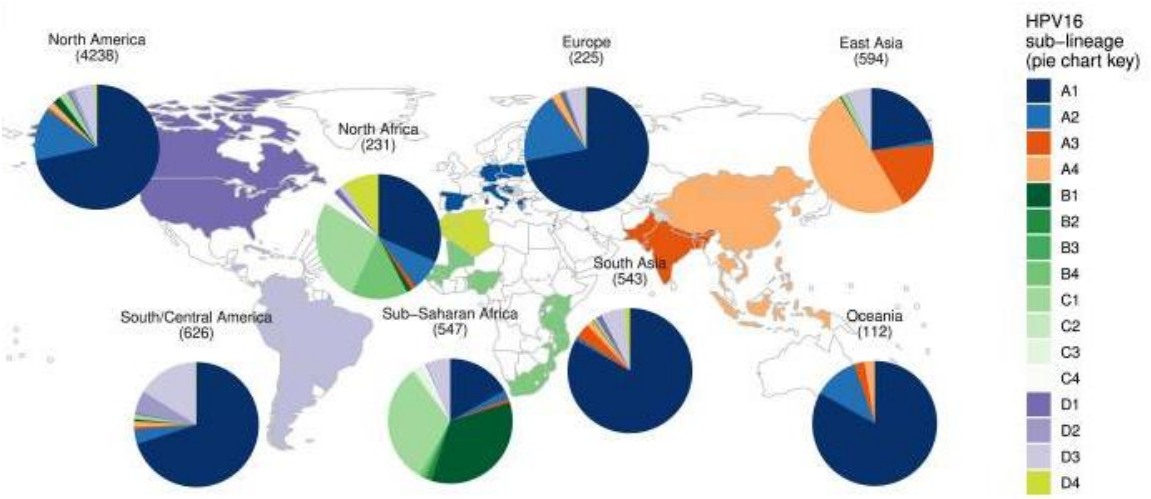

Figure 6: Geographic distribution of HPV16 sub-lineages across 7116 positive samples worldwide.

Data source: IARC. Map generated by IARC and the World Health Organization (WHO), 2018.

1.4 Vaccination and Prevention Strategies

Major HPV vaccination programs began around 2008 with the introduction of the bivalent vaccine Cervarix® (GSK), targeting HPV types 16 and 18, and the quadrivalent vaccine Gardasil® (MSD), covering HPV types 6, 11, 16, and 18. In 2015, MSD introduced the nonavalent vaccine, which is now the most widely used. It offers broader protection, covering HPV types 6, 11, 16, 18, 31, 33, 45, 52, and 58.

These vaccines are prophylactic, meaning they are preventive and therefore not intended to treat individuals already infected with HPV. Their preventive nature is demonstrated by their higher efficacy among girls vaccinated prior to sexual debut, which aligns with the fact that cervical cancer typically takes around ten years to develop following infection [53]. The vaccine has been shown to provide strong protection against cervical precancer in adolescent girls and young women. Moreover, within 5 to 8 years of vaccine implementation, a significant decline in anogenital wart diagnoses has been observed: an 88% reduction in girls under 20 years old and 86% in boys of the same age group in countries with high vaccination coverage [54,55].

However, evidence has emerged of a possible type-replacement phenomenon, where high-risk HPV types not targeted by the vaccines, such as HPV35 and HPV56, increase in prevalence, potentially replacing vaccine-covered types like HPV16 and 18. This trend has been observed in both vaccinated and unvaccinated women across different age groups and countries [56,57]. While ongoing surveillance is needed to understand the impact of type-replacement, it is well established that infections with HPV16, 18, or 45 are associated with

earlier development of invasive cervical cancer (ICC), emphasizing the need for continued screening and monitoring into older age [58].

It is therefore essential that long-term epidemiological surveillance extending over at least 30 years from now, be maintained to track vaccine-related changes in HPV type distribution and their impact on the development of cervical and other anogenital neoplasias. Additionally, continued screening and molecular epidemiological analysis are crucial to adapt prevention strategies accordingly.

At the global level, HPV vaccination coverage remains low. As of 2024–2025, it is estimated that 27% of eligible girls and women have received at least one dose, while only 20% are fully vaccinated [59]. Coverage is also highly uneven across countries: low- and middle-income countries (LMICs) face significant challenges in accessing both HPV vaccination and cervical cancer screening programs.

A recent study found significant viral flow and genotype migration from Africa to Europe, reinforcing the argument that comprehensive vaccination and screening programs must be sustained even in high-income countries, particularly those receiving migrants from under-vaccinated regions [60]. In Belgium, as of 2024, the first-dose vaccination coverage reached approximately 91% in Flanders, but only 36% in Wallonia. The real-world effectiveness of the vaccine has also been demonstrated: among women aged 16–22, a 43% coverage rate resulted in a 72% reduction in genital warts, with an estimated vaccine efficacy of 88% [61,61].

1.5 Molecular Detection and NGS-Based Analyses

Molecular detection methods, including polymerase chain reaction (PCR), whole genome amplification, and next-generation sequencing (NGS), have become essential tools for the characterization of human papillomavirus (HPV) infections. These techniques enable not only the detection of viral DNA but also the identification of specific genotypes, co-infections, and even genetic variants. In contrast to traditional cytological methods, molecular approaches offer higher sensitivity and specificity, especially in detecting low-copy viral genomes or multiple HPV types within a single sample [62]. The most current HPV detection methods that are commercially available are type-specific target amplification DNA PCR and signal amplification DNA ISH, which are approved for cervical samples [63]. The advent of NGS technologies has further enhanced HPV research, enabling high-throughput genotyping, strain diversity analysis, and better resolution of mixed infections [64].

1.5.1 Polymerase chain reaction

Polymerase Chain Reaction (PCR) is a cornerstone of molecular biology and remains a critical component of numerous diagnostic and research protocols [65]. In the context of HPV detection, PCR is used as a target amplification technique to amplify small quantities of viral DNA present in biological samples composed of mixed cell populations [66].

Depending on the primers used, PCR can be designed to detect a single HPV type or multiple types simultaneously.

Most commonly, primers are designed to target the L1 gene, which is well-conserved among HPV types, or alternatively the E6 and E7 oncogenes, which are more directly implicated in oncogenesis. Notably, PCR primers directed at E6 and E7 regions are often considered preferable in advanced disease stages, as the L1 and E1 regions may be deleted during the integration of viral DNA into the host genome, potentially leading to false negatives if these regions are used as targets [63,89].

In the present study, we employed a tiling multiplex PCR approach using multiple overlapping primer sets to ensure comprehensive amplification of the entire HPV genome. This strategy increases the likelihood of capturing both episomal and integrated forms of HPV DNA, allowing for whole-genome recovery prior to sequencing. Such approaches have been successfully applied in viral genomics, including for HPV, to enable full-genome resolution and strain-level analysis [67].

The Cobas 4800 HPV Test (Roche Diagnostics) is a widely used high-throughput clinical assay for HPV detection based on multiplex real-time PCR combined with nucleic acid hybridization. Approved by the FDA in 2011 and available in Europe since 2009, it allows the simultaneous detection of HPV16 and HPV18 individually, as well as a pooled result for 12 other high-risk HPV types (HPV31, HPV33, HPV35, HPV39, HPV45, HPV51, HPV52, HPV56, HPV58, HPV59, HPV66, HPV68), by targeting the L1 gene through four fluorescent probes [68]. The test is performed on the automated Cobas 4800 platform, which can process up to 280 cervical samples per day, making it suitable for large-scale screening programs. Despite its high sensitivity and specificity, the assay may yield false negatives in cases where the L1 region is deleted or disrupted, particularly in advanced lesions where the viral genome is integrated into the host DNA [69].

1.5.2 Multiple Displacement Amplification (MDA)

Multiple Displacement Amplification (MDA) is a powerful isothermal amplification method capable of generating large quantities of DNA from very small amounts of starting material, down to just a few copies. It relies on Phi29 DNA polymerase, whose strand displacement activity enables the synthesis of long fragments often exceeding 10 kb, with significantly higher fidelity than that of Taq polymerase [70,71].

In our study, MDA was particularly suited for HPV genome amplification because phi29 DNA polymerase preferentially amplifies circular DNA molecules, making it ideal for enriching the circular HPV genome. This feature represents one of the strengths of MDA in virology applications [72].

However, this preference for circular DNA is also a limitation, as phi29 will non-specifically amplify any circular DNA present in the sample, including mitochondrial DNA, plasmids, or other circular elements from host or contaminating organisms. This can result in unwanted

background amplification and may reduce the proportion of sequencing reads covering the HPV genome.

Several studies have shown that MDA can be effectively combined with long-read sequencing technologies, particularly Oxford Nanopore Technologies (ONT). For instance, Agyabeng Dadzie et al. [73] demonstrated that nearly complete viral genomes can be reconstructed from as little as 0.025 ng of input DNA when MDA is followed by ONT sequencing, provided that appropriate bioinformatic processing is applied to remove artificial concatemeric sequences generated during amplification.

This strategy is especially well-suited for whole-genome analysis of viruses such as HPV, as it enables the reconstruction of full-length genomes from samples containing low or degraded amounts of viral DNA. Moreover, unlike targeted PCR approaches, MDA does not require the use of specific primers for each HPV type, making it more inclusive. In our study, this approach allowed for the capture of both low-risk and high-risk HPV types, in contrast to the tiling PCR method, which was specifically designed to target only high-risk HPV genomes.

1.5.3 Sequencing technology

Early-generation sequencing technologies, such as Sanger sequencing, offered high accuracy but were time-consuming, low-throughput, and cost-prohibitive for large-scale genomic studies. The advent of next-generation sequencing (NGS) revolutionized genomics by enabling massively parallel sequencing, high sensitivity, and reduced cost per base. Platforms like Illumina have become widely adopted for their short-read accuracy and depth, though they often require extensive library preparation and do not easily resolve complex or repetitive regions [74].

In recent years, Oxford Nanopore Technologies (ONT) has emerged as a powerful alternative, providing long-read capabilities, real-time sequencing, and portable devices such as the MinION, GridION, and PromethION. Unlike short-read technologies, ONT directly senses nucleotides by measuring changes in electrical current as DNA or RNA molecules pass through biological nanopores. Each base induces a unique current shift, producing a raw signal that is decoded in real time [75,76].

One notable advantage of ONT is the “Read Until” feature, a real-time selective sequencing strategy. As a DNA strand begins to pass through the pore, the system compares its initial signal to a reference database (e.g., HPV genomes). If no match is detected, the sequencer reverses the voltage, ejecting the molecule and freeing the pore for another read [77]. This allows for targeted sequencing without prior enrichment, enhancing both efficiency and cost-effectiveness. In our context, this approach is particularly valuable. During upstream amplification steps, non-target DNA, including human genomic DNA, mitochondrial DNA, or bacterial DNA, can be unintentionally amplified through off-target primer binding (in PCR) or through non-specific amplification by MDA. Read Until enables us to minimize sequencing of irrelevant DNA, optimizing the flow cell capacity for the genomes of interest.

1.6 Current Insights into the Structure of HPV E6 and E7 Proteins

Understanding the structural basis of interactions between HPV's oncoproteins and key host tumor suppressors is critical to decipher their transforming mechanisms, including the disruption of cell cycle checkpoints, evasion of apoptosis, and induction of uncontrolled proliferation. Recent high-resolution structures highlight how E6 targets p53 and E7 targets pRb, thereby facilitating oncogenesis.

1.6.1 HPV E6–p53 Interaction

The oncogenic activity of HPV16 E6 relies on its ability to mediate the ubiquitin-dependent degradation of the tumor suppressor p53, in cooperation with the host E3 ubiquitin ligase E6AP (Figure 7). Structural studies, particularly the recent high-resolution cryo-EM structure of the full-length E6–E6AP–p53 complex (PDB: 8GCR), have provided detailed insight into this interaction. E6 contains two zinc-binding domains (residues ~10–43 and ~90–130), which form a hydrophobic interdomain pocket that specifically accommodates the LxxLL motif (LQELL; residues 403–407) of E6AP. This binding stabilizes E6 and induces a conformation that exposes a large interaction surface for the p53 DNA-binding domain (residues ~94–292). The resulting ternary complex, stabilized by hydrophobic contacts and hydrogen bonds over +-1200 $Å^2$ of buried surface area, effectively positions p53 for ubiquitination by E6AP. These structural insights explain the molecular mechanism by which HPV16 E6 disables p53 surveillance pathways, a critical event in HPV-mediated carcinogenesis [78,79].

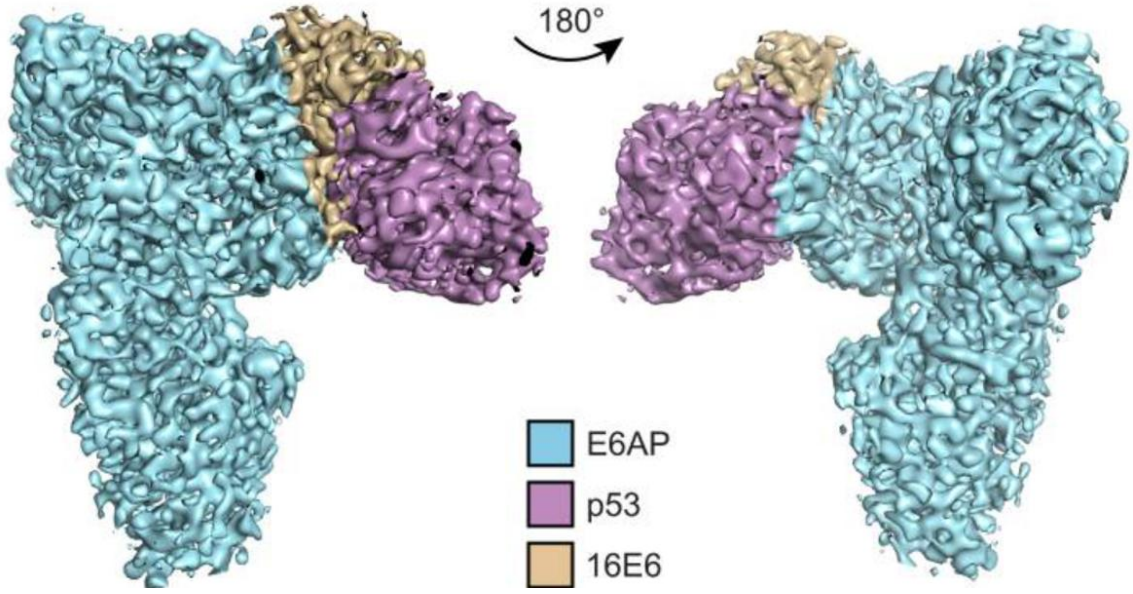

Figure 7: Cryo-EM density map of the HPV16 E6/E6AP/p53 ternary complex (PDB: 8GCR). The map is colored by component: HPV16 E6 is shown in orange, E6AP in blue, and

the p53 core domain in pink. The structure highlights the spatial arrangement of the three proteins and the extensive interface facilitating p53 degradation.

Source: Adapted from Structure of the p53 degradation complex from HPV16. Nature Communications, 15(1), 1842 [117].

1.6.2 HPV E7–pRb Interaction

The high-risk HPV16 E7 oncoprotein exerts its transformative effect primarily through binding to and inactivating the retinoblastoma tumor suppressor protein (pRb). This interaction is mediated by the highly conserved LxCxE motif located in the CR2 region (residues ~20–26) of E7, which specifically docks into the pocket B domain of pRb, displacing E2F transcription factors and thereby promoting uncontrolled cell cycle progression (Figure 8). Adjacent to this motif, the conserved serine residues (Ser31 and Ser32) serve as phosphorylation sites for casein kinase II (CKII), and this phosphorylation enhances E7's binding affinity for pRb [80].

Structural analysis of the CR3 domain, a zinc-binding dimer at physiological concentrations, revealed two conserved surface patches: one facilitating pRb binding, and the other targeting E2F, enabling E7 to disrupt the pRb–E2F complex effectively [81].

The zinc-binding CR3 domain (residues ~39–98) contributes not only to E7 dimer stability but also provides the structural architecture necessary for high-affinity interaction with pRb. These combined interactions subvert the pRb, E2F pathway and promote uncontrolled cell proliferation, an essential step in HPV-mediated carcinogenesis.

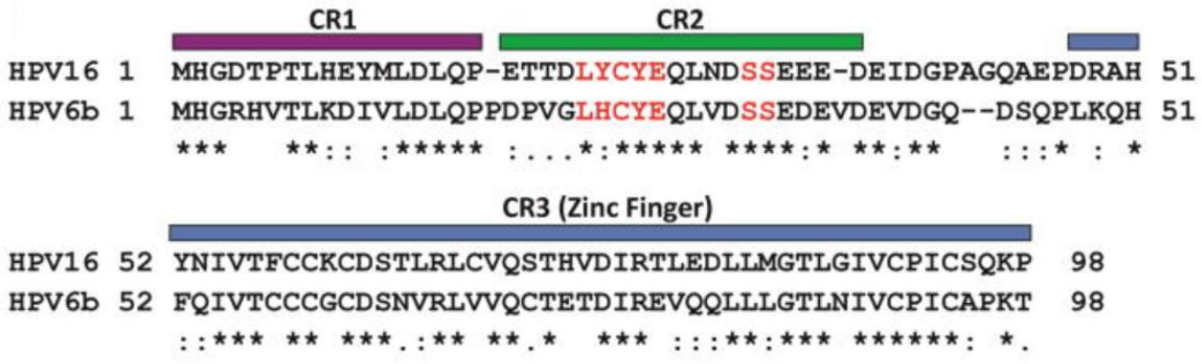

Figure 8: Sequence alignment of high-risk HPV16 E7 and low-risk HPV6b E7 proteins. Conserved domains CR1, CR2, and CR3 are indicated by colored horizontal bars**.** The LxCxE motif, responsible for pRb binding, as well as the conserved serine residues (phosphorylation targets of casein kinase II, CKII), are highlighted in red.

Source: Adapted from Jansma et al., J. Mol. Biol., 2014 [80].

2 Objectives

The present study aims to investigate the genetic diversity, lineage distribution, and molecular characteristics of human papillomavirus (HPV) types detected in two geographically and epidemiologically distinct cohorts: patients from the Democratic Republic of the Congo (DRC) and Belgium. By comparing these populations, the study seeks to analyse the spectrum of circulating HPV types in each region and to identify potential differences in type prevalence, sub-lineage distribution, and genomic variation, thereby providing insights into HPV epidemiology across contrasting settings.

To achieve this, two complementary whole-genome amplification strategies were employed: multiplex tiling PCR and multiple displacement amplification (MDA). The amplified viral genomes were subsequently sequenced using Oxford Nanopore Technologies (ONT), enabling the reconstruction of complete HPV genomes with long-read resolution and facilitating high-resolution genomic comparisons between isolates from both cohorts. This combined approach ensures both sensitive detection and comprehensive characterization of circulating HPV strains.

In addition to refining type distribution, this study also aims to explore the discovery of potentially novel HPV lineages, sub-lineages, or types, and to provide preliminary insights into the molecular determinants that may underlie the oncogenic potential of high-risk HPVs. To complement genomic analyses, structural bioinformatics approaches were applied, notably through the examination of AlphaFold-predicted 3D protein models, in order to investigate host–virus protein interactions and gain preliminary insights into the structural features driving the transforming capacity of HR-HPV oncoproteins.

Ultimately, the integration of molecular, structural, and epidemiological data is expected to enhance our understanding of HPV type circulation, genetic diversity, and their potential associations with disease risk in diverse populations. Such insights may inform targeted prevention strategies and support global public health efforts in HPV surveillance and control.

3 Materials and Methods

3.1 Lab work and genome assembly

3.1.1 Sample collection and HPV screening

Cervical swabs were collected in the CHU of Liege (Belgium) and Kinshasa (Democratic Republic of Congo) between 2022 and 2024. This study includes 191 women, 131 from DRC and 60 from Liege.

Samples were firstly tested using the Cobas 4800 HPV Test (Roche), which employs real-time PCR to detect the presence of HPV16, HPV18, a pool of 12 other high-risk HPV types, or to confirm the absence of HPV DNA [90].

Based on COBAS test results, samples were classified into three categories: (1) positive for HPV16 and/or HPV18, (2) positive for other high-risk HPV types, and (3) negative for HPV DNA. For the DRC cohort, all HPV-positive samples (categories 1 and 2) were retained for further analysis. In contrast, only samples that tested positive for HPV16 or HPV18 were selected from the Liège cohort. This selective approach was driven by the initial objective of studying HPV integration using the PCIP [82] method. However, the available PCIP protocol in our laboratory is currently validated only for HPV16 and HPV18, which limited the inclusion of other HPV types for integration analysis.

These two different selection strategies introduce a bias and do not allow for a direct comparison of HPV type distribution between the two populations. Due to this bias, the only meaningful comparison between the two populations could be the analyse of co-infections involving HPV16 and/or HPV18.

3.1.2 DNA extraction and Amplification

Genomic DNA was extracted from dry cell pellets obtained from cervical swab samples using the DNeasy® Blood & Tissue Kit (Qiagen). The concentration of extracted HPV DNA was then quantified using the Qubit™ dsDNA HS Assay Kit (Invitrogen™, Thermo Fisher Scientific Inc.).

Based on the initial DNA concentrations, dilutions were performed to normalize the DNA input to 100 ng per sample. A tiling multiplex PCR amplification was first carried out. When sufficient DNA was available, an additional MDA (Multiple Displacement Amplification) approach was also performed.

PCR Amplification

This method employs two pools of primers specifically designed to cover the entire genomes of the 14 high-risk HPV genotypes, allowing for full-genome amplification even in complex or mixed infections.

The PCR amplifications were carried out using a Veriti® 96-Well Thermal Cycler (Applied Biosystems™, Thermo Fisher Scientific, Waltham, MA, USA). Reference genome sequences for the 14 high-risk HPV types were retrieved from the PaVE database (Papillomavirus Episteme, https://pave.niaid.nih.gov/) [118].

To enable complete genome amplification through a multiplex tiling PCR strategy, primers were designed using Primal Scheme (https://primalscheme.com/). For this purpose, the reference sequences of all target HPV genotypes were concatenated into a single file, with each genome separated by an artificial stretch of 2,500 'N' bases to prevent overlap in primer design. Additionally, to account for the circular nature of HPV genomes and ensure seamless coverage at genome junctions, the final ~300 nucleotides of each genome were appended to the beginning (5'-end) of the corresponding sequence.

The tiling strategy was configured to produce amplicons of approximately 1,500 bp, resulting in the creation of two primer pools: one comprising 96 primers, the other 92 primers. Each pool was assembled by combining primers in equimolar ratios. All primers were synthesized by Integrated DNA Technologies (IDT, Leuven, Belgium).

After obtaining HPV PCR products, electrophoresis in a 1% agarose gel was conducted for the two pools of primers. This is a common technique for the separation and analysis of DNA, frequently employed to examine the size, purity, and concentration of PCR amplification products.

MDA Amplification

To amplify genomic DNA prior to sequencing, we applied Multiple Displacement Amplification (MDA) using the REPLI-g® Midi Kit (QIAGEN), a high-fidelity isothermal method for whole-genome amplification. This technique relies on the phi29 DNA polymerase, which possesses strong strand displacement activity and high proofreading accuracy, enabling uniform amplification of DNA fragments greater than 10 kb in length with minimal sequence bias. Importantly, MDA preferentially amplifies circular DNA molecules, such as papillomavirus genomes, whereas linear DNA templates are amplified less efficiently.

The protocol, illustrated in Figure 9, involves the following main steps:

1. Denaturation step: Purified genomic DNA is mixed with TE buffer and denaturation solution, followed by vortexing and a short incubation (3 minutes) at room temperature (15–25 °C), which helps to separate DNA strands.

2. Neutralization: A neutralization buffer is then added to stop the denaturation process.

3. Amplification: The reaction mixture is supplemented with the REPLI-g master mix, which contains the phi29 polymerase and random hexamer primers. The reaction proceeds isothermally:

o 8–16 hours at 30°C for the Midi protocol

o Followed by a 3-minute inactivation at 65°C

The resulting product is high molecular weight amplified DNA suitable for downstream applications such as long-read sequencing.

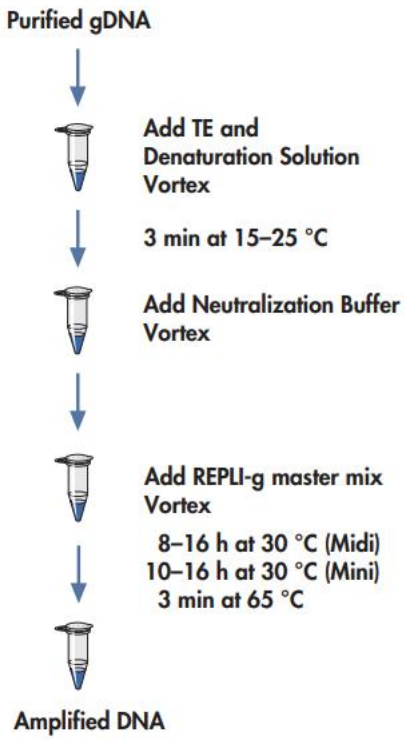

Figure 9: Workflow of the MDA reaction using the REPLI-g® Midi Kit (QIAGEN). After denaturation and neutralization, genomic DNA is amplified under isothermal conditions using phi29 polymerase.

Source: Adapted from REPLI-g® Mini/Midi Handbook, QIAGEN, July 2011 (https://www.qiagen.com).

The DNA products or fragments obtained from both amplification methods are subsequently purified using AMPure XP (AXP) magnetic beads (Beckman Coulter Inc., Brea, CA, USA). This purification process efficiently removes small non-specific fragments (<200 bp), residual primers, and other impurities generated during the amplification steps. As a result, it improves the concentration and quality of the DNA, which is essential for enhancing the efficiency of downstream library construction, increasing library yield, and reducing contamination or inhibitory substances that could interfere with sequencing.

Before use, the AXP beads should be equilibrated to room temperature for 30 minutes. The amplified DNA fragments are first pooled, then mixed with the beads to allow DNA to bind to

the magnetic surface. Impurities such as salts and unbound primers are then removed through washing steps. Finally, the purified DNA is eluted from the beads, yielding high-quality material suitable for subsequent workflows.

3.1.3 Library construction and Sequencing

Before sequencing, libraries were prepared using the Native Barcoding Kit 96 V14 (SQK-NBD114.96, Oxford Nanopore Technologies plc., Oxford, UK). This kit allows for the multiplexing of up to 96 samples by ligating a unique barcode to each DNA fragment, enabling the simultaneous sequencing of multiple samples while preserving the traceability of each individual HPV genome to its corresponding patient.

The DNA library preparation begins with an end-repair and A-tailing step, which generates DNA molecules with a 3' A overhang. This modification provides a suitable substrate for the ligation of barcoded adapters. Each sample receives a distinct barcode, and these barcodes are then used to identify reads during demultiplexing after sequencing.

Following barcoding, sequencing adapters coupled to a motor protein are ligated to the DNA ends. This protein is essential for nanopore sequencing, as it modulates the speed at which the DNA strand passes through the pore, ensuring optimal signal resolution. A clean-up step is then performed to remove unligated adapters, excess enzymes, residual primers, and other contaminants that may interfere with sequencing performance. Only properly ligated DNA fragments are retained for sequencing. Finally, the purified, barcoded DNA library is ready to be loaded onto the Oxford Nanopore flow cell for real-time sequencing.

Following library preparation, sequencing was performed using Oxford Nanopore Technologies (ONT).

With ONT sequencing there is the possibility of applying a real-time selective sequencing strategy known as Read Until. This approach increases efficiency by allowing the sequencer to reject DNA molecules that do not match a predefined list of target genomes. In this study, reference genomes of various HPV types were retrieved from the PaVE database and used to guide the selection process during sequencing.

In our context, this strategy helps avoid sequencing unwanted genomes that may be co-amplified during upstream steps. During PCR amplification, human DNA may be unintentionally amplified due to off-target primer binding, while circular genomes, such as mitochondrial or bacterial DNA, can also be preferentially amplified during MDA.

The sequencing run was initiated using the MinKNOW software (Oxford Nanopore Technologies), which controls the sequencing device, monitors performance in real time, and manages data acquisition. MinKNOW collects the raw electrical signal data generated as DNA molecules pass through the nanopores and performs real-time basecalling, converting these signals into readable nucleotide sequences (basecalled reads). These basecalled reads represent the raw sequencing data and serve as the foundation for downstream analyses.

3.1.4 Genome assembly

PCR Reads

Basecalling was performed using Guppy in super high-accuracy mode (https://nanoporetech.com/document/Guppy-protocol). This software, developed by Oxford Nanopore Technologies, translates raw electrical signals generated by the nanopore sequencer into nucleotide sequences. The super high-accuracy mode ensures very precise results but is computationally intensive and slower.

Adapter trimming was then conducted using Porechop (https://github.com/rrwick/Porechop), which removes sequencing adapters from both ends of each read. To further refine read quality, we used NanoFilt (https://github.com/wdecoster/nanofilt), applying the following filters:

- Removal of 40 bases from both the 5' and 3' ends of each read to eliminate possible primer sequences or low-quality regions.
- Retention of reads with a minimum quality score of 12 and a minimum length of 1000 bases.

In the second step, the cleaned reads were mapped to HPV reference genomes obtained from the PAVE database. Coverage across the genomes was then calculated using Mosdepth (https://github.com/brentp/mosdepth), generating depth matrices for each sample.

Next, consensus genome sequences were generated with Medaka (https://github.com/nanoporetech/medaka). To enhance consensus accuracy, reads were realigned to their newly generated consensus, and the process was repeated three times. This iterative correction improves the overall sequence quality and fidelity. Regions with a coverage lower than 15x were masked and replaced by "N" to indicate uncertainty.

MDA Reads

For MDA-derived reads, the assembly pipeline was essentially identical to that used for PCR amplicons, except that no primer trimming was required. After basecalling, adapter removal, and quality filtering, reads were mapped to their corresponding HPV reference genomes, and coverage was calculated. Consensus genomes were generated using Medaka, with multiple iterative rounds of read realignment to improve accuracy. Regions with coverage below 15× were masked with "N".

3.2 Quality control and Filtering

For global statistical analyses of HPV type distribution, all available sequences were included, regardless of completeness. However, for phylogenetic and classification

analyses, only the filtered full-length genomes were used to ensure consistency and avoid biases linked to incomplete data.

Among the 636 sequenced HPV genomes, we identified a total of 145 duplicates. Of these, 55 genomes were found to be identical across both amplification techniques (PCR and MDA). The remaining 90 genomes, although expected to be identical, exhibited minor discrepancies between the two amplification methods. These genomes likely correspond to the same biological sequence, but slight differences were observed between the versions generated by PCR and MDA. To investigate this, we manually inspected the alignments using Integrative Genomics Viewer (IGV) [83] to assess the nature of these discrepancies and identify potential sequencing or amplification artefacts.

After deduplication, we retained 491 unique sequences and subsequently applied a second filter to restrict the dataset to complete genomes.

We retained only complete HPV genomes in order to avoid bias in phylogenetic and classification analyses, which rely on full genome sequences. To achieve this, I used the script inst-qual-filter.pl from Bio-MUST-Core, a suite of tools and Perl modules developed by Prof D. Baurain to automate several sequence processing tasks required for phylogenomic analyses. The script was run as follows:

```
inst-qual-filter.pl final_samples --out=_filtered
```

Following this filtering step for complete genomes, we retained 357 unique, full-length HPV genomes for downstream analyses.

Each HPV genome sequence was processed using the PuMA software [84] , a dedicated papillomavirus genome annotation tool. PuMA (Papillomavirus Multi-genome Annotator) automatically generates GenBank-format files and provides a visual and tabular mapping of all predicted open reading frames (ORFs) across the viral genome. The tool compares each input sequence against reference genomes available in the PaVE database, enabling highly accurate annotation based on sequence homology, conserved protein domains, and gene architecture.

PuMA is written in Python and integrates tools such as BLAST+, MUSCLE, and MEME/FIMO to identify genes, regulatory elements, and conserved motifs [85–87]. In this study, it was used to standardize genome annotation across all samples, ensuring consistent ORF prediction and gene naming. The resulting GenBank files were subsequently used for downstream comparative and phylogenetic analyses

3.3 Design and Implementation of a Tool for HPV Genomic Classification

The main objective of this section was to develop a tool enabling rapid and accurate classification of our HPV sequences based on the official criteria from PAVE (https://pave.niaid.nih.gov/explore/variants/variant_nomenclature). The goal was to annotate each genome with its corresponding type, lineage, and sublineage.

3.3.1 HPVxHunter: Classification Software for HPV Genotyping

The tool is based, as previously described in the Introduction, on a global pairwise alignment using the VSEARCH software [91]. Input FASTA sequences are compared and aligned against one or several databases, depending on the selected option. The user may choose to use the well-annotated PaVE database, the GenBank database (which is not always properly annotated), or a custom database. This last option is particularly useful when a new lineage or sublineage has been identified but is not yet listed in the PaVE database, and one still wishes to continue analyzing newly sequenced genomes.

The source code and installation instructions are available at: https://github.com/ElmYassine29/HPVxHunter.

When the chosen reference is the PaVE database, the output includes an Excel file that provides the HPV type, lineage, and sublineage, based on the percentage of sequence similarity (in L1 ORF and complete genome) and in accordance with established classification and nomenclature rules (see Section 1.3: HPV classification). Additionally, the tool computes the length difference between sequences, which is useful to filter out false positives, as significant length variations can artificially lower the global similarity score and lead to incorrect identifications of new variants. High-risk HPV genomes are automatically highlighted in red (Figure 10).

Each analysis also generates three graphical plots:

- A bar plot of the top 20 most frequent HPV types
- A breakdown of lineages per type
- A breakdown of sublineages per lineage

When the reference database is not PaVE (e.g., GenBank or a user-provided custom database), the tool first identifies the best match for each input sequence based on global pairwise similarity. It reports the top hit along with the percentage of identity. Then, if possible, the matched sequence is further compared to the PaVE database in order to determine its most likely HPV type, lineage, and sublineage. This secondary mapping step ensures accurate classification even when the original database lacks full annotation. This classification does not result from a direct alignment of the input sequence to the PaVE database, but rather reflects the most likely classification of the best GenBank hit or custom database, based on a subsequent alignment against PaVE.

An optional feature, which is more computationally intensive, can be activated using the '–deep_analysis' flag. This function detects potential novel HPV types by analyzing the L1 open reading frame (ORF), in line with PaVE guidelines. Specifically, it compares the L1 region of sequences showing less than 95% overall similarity to known reference genomes. If the L1 region differs by more than 10% from known types, the tool flags the sequence as a potential new HPV type, with results summarized in a dedicated Excel report.

To extract the L1 ORF from each genome, the tool integrates the PuMA software, which automatically annotates all ORFs in papillomavirus genomes. This enables reliable extraction of the L1 region for downstream analysis.

Overall, this tool provides a rapid and scalable approach to analyze HPV types, lineages, and sublineages in sequencing datasets. It facilitates the detection of high-risk HPV infections and supports epidemiological studies by offering an automated and standardized framework for genome classification.

Sequence	Type	Lineage	Sub-lineage	Similarity	Length Difference	Results
HPV59_MDA_80189	HPV59	B	B-1	99,9	0	Same sublineage
HPV30_MDA_24360	HPV30	A	A-1	99,9	1	Same sublineage
HPV35_PCR_80107	HPV35	A	A-2	99,8	0	Same sublineage
HPV31_PCR_80087_MDA	HPV31	C	C-1	99,7	0	Same sublineage
HPV44_MDA_40395	HPV44	A	A-1	100	0	Same sublineage
HPV59_PCR_80648	HPV59	B	ND	99,5	0	Probably new sublineage
HPV16_PCR_35419	HPV16	A	A-1	99,9	0	Same sublineage
HPV18_PCR_40818_MDA	HPV18	A	A-5	99,7	0	Same sublineage
HPV35_PCR_80678	HPV35	A	A-2	99,7	0	Same sublineage
HPV56_MDA_80310	HPV56	A	A-2	99,8	0	Same sublineage
HPV6_MDA_80263	HPV6	B	B-5	99,9	0	Same sublineage
HPV90_MDA_80440	HPV90	A	A-1	99,7	0	Same sublineage
HPV35_MDA_80274	HPV35	A	A-2	99,8	0	Same sublineage
HPV16_MDA_80495_PCR	HPV16	C	C-1	99,8	0	Same sublineage
HPV68_PCR_80517	HPV68	D	D-1	99,9	0	Same sublineage
HPV42_MDA_39963	HPV42	A	A-1	99,6	16	Same sublineage
HPV35_PCR_80013	HPV35	A	A-2	99,7	1	Same sublineage
HPV42_MDA_25162	HPV42	A	ND	99,5	3	Probably new sublineage
HPV61_MDA_80131	HPV61	C	C-1	99,9	0	Same sublineage
HPV70_MDA_32089	HPV70	A	A-1	99,8	0	Same sublineage
HPV53_MDA_80114	HPV53	A	A-1	99,9	0	Same sublineage
HPV39_PCR_80181_MDA	HPV39	B	B-1	99,8	0	Same sublineage
HPV56_PCR_80019	HPV56	A	A-2	99,8	0	Same sublineage

Figure 10: Example of HPVxHunter output using the PAVE reference database, applied to our dataset of HPV genomes. For each sample (initially annotated only at the type level), the tool reports the refined classification, including the HPV type, lineage, and sub-lineage, along with the sequence similarity (%) to the closest reference genome and the length difference (in pb). The final column provides an interpretation of the classification result: whether the sequence corresponds to the *same sub-lineage*, *a probable new sub-lineage*, or in rare cases, *no significant match* (suggesting a potential novel type or the need for deeper analysis). This output exemplifies how HPVxHunter enables fine-grained and automated classification of HPV sequences.

3.3.2 Classification and Phylogenetic Analyses

We then applied our in-house HPV analysis tool to the complete genome sequences, using the *deep_analysis* option. This mode specifically targets sequences that do not match any known reference genome, in order to assess whether they may represent novel HPV types.

The tool performs accurate classification of HPV genomes down to the sublineage level, based on full-genome similarity with a curated reference database. When a sequence cannot be matched to any known type, the *deep_analysis* option triggers a more in-depth comparison to evaluate its degree of divergence from existing HPV references.

To classify the collected HPV sequences, we used the HPVxHunter tool, referencing both the Papillomavirus Episteme (PaVE) database for established type and lineage definitions, and GenBank to identify sequences that may not yet be characterized in PAVE but share high similarity with our data. This dual-reference strategy allowed us to detect not only known HPV types and variants, but also sequences potentially representing uncharacterized or novel lineages. The resulting classifications were exported in two detailed spreadsheets summarizing type, lineage, sublineage, and closest matches. We also activated the option 'deep_analysis' to detect potential novel types.

For HPV types where candidate new lineages or sublineages were identified, we constructed phylogenetic trees using MAFFT v7.453 for multiple sequence alignment and IQ-TREE MPI multicore version 1.6.12 for maximum-likelihood phylogenetic inference, with 1,000 ultrafast bootstrap replicates to assess node support [92,93]. To validate the existence of potential new groups and to integrate our findings within the broader genomic context, additional sequences were incorporated into the phylogenetic trees. These sequences were selected as the best GenBank hits corresponding to our own sequences.

In addition, we used a short script (see below) to compute patristic distances within the trees using the ape package in R providing a complementary and quantitative criterion to verify clustering with known sublineages [98] providing:

```
#Example with the HPV226 and the sample 80440

#Load the ape package for phylogenetic analysis
library(ape)

#Read the phylogenetic tree generated from a multiple sequence alignment
tree <- read.tree("HPV226_aligned.fasta.treefile")

#Define the sample of interest
sample <- "HPV226_MDA_80440"

#Compute the cophenetic distance matrix from the tree
#This gives the pairwise distances between all tips in the tree
dist_matrix <- round(cophenetic(tree) * 100, 2)  # distances converted to per
centage and rounded
```

```r
#Extract the distances between the target sample and all other sequences
dist_vector <- dist_matrix[sample, ]

#Convert these distances to a matrix of strings with percentage symbols
dist_vector_pct <- as.matrix(paste0(dist_vector, "%"))

#Set the appropriate row and column names for clarity
rownames(dist_vector_pct) <- names(dist_vector)
colnames(dist_vector_pct) <- sample

#Display the percentage distance matrix for the sample
dist_vector_pct

#Sort the distances to identify the closest sequence (i.e., most similar)
sort(dist_vector_pct[, sample])
```

In cases where novel HPV types were detected, we additionally constructed individual phylogenetic trees for each major open reading frame (ORF). We then compared the phylogenetic placement of the candidate sequences across these ORF-specific trees. This approach allowed us to rule out the possibility that these sequences resulted from inter-type recombination, rather than representing genuine novel HPV types.

3.3.3 Comprehensive Analysis of Human Papillomavirus Genomes from the EMBL-EBI Database

To complement our classification analyses, we applied HPVxHunter to a comprehensive dataset of publicly available HPV genomes retrieved from the EMBL-EBI nucleotide sequence database. The sequences were obtained using the following search query:

```
description="human papillomavirus" AND base_count>7000
```

This query was designed to collect complete papillomavirus genomes (minimum length of 7,000 base pairs) by targeting sequences explicitly annotated as human papillomaviruses. As of May 27, 2025, this search yielded a total of 10,729 HPV genomes, which served as a reference background dataset for large-scale comparison.

We also performed the equivalent search on GenBank, which provided a comparable number of genomes with similar characteristics, confirming the consistency of publicly available datasets across major repositories.

In addition to taxonomic classification, this dataset was intended to support a spatial and geographic epidemiological analysis by leveraging available metadata such as country of origin and collection date. The objective of this step was to explore the global diversity and distribution of HPV types and lineages. However, the feasibility and reliability of such

analyses depended on the completeness and quality of the associated metadata, which are further discussed in the results section.

3.4 Structural Analysis of E6/E7 Viral Proteins

To investigate the molecular interactions between the HPV oncoproteins E6 and E7 and their respective cellular targets, p53 and pRB, we performed a structural bioinformatics analysis. This approach aimed to explore potential differences in the structural behavior of these interactions between high-risk (hrHPV) and low-risk (lrHPV) types, which may contribute to their distinct oncogenic potential.

We based our modeling on experimentally resolved structures available in the Protein Data Bank: the E6–E6AP–p53 complex (PDB: 8GCR) and the E7–pRB complex (PDB: 4YOZ). These structures served as templates to analyze the corresponding interactions across our full set of HPV genomes

To achieve this, we retrieved the E6 and E7 protein sequences from all complete HPV genomes in our dataset. In order to avoid redundant modeling of identical sequences present across multiple samples, we applied CD-HIT clustering with a 100% sequence identity threshold to both E6 and E7 datasets [94]. This ensured that each unique protein sequence was modeled only once. We then used these unique sequences to model their interactions with their respective cellular targets using homology-based and structure-guided approaches, with protein structure prediction performed using the AlphaFold server [95].

3.4.1 E6-E6AP-p53 interaction

Based on the available experimental structure of the E6–E6AP–p53 complex (PDB: 8GCR), we substituted the original E6 sequence in the model with each of our unique E6 sequences, in order to reconstruct a full E6–E6AP–p53 complex specific to each viral variant.

To quantitatively assess the structural features of these complexes, we used PRODIGY, a computational software tool that estimates binding affinities between two proteins based on the physicochemical properties of their interaction interfaces [96]. For each HPV type, we analyzed the protein–protein contacts between E6 and E6AP, as well as between E6 and p53.

This allowed us to extract key descriptors of the interaction surfaces, including contact areas and predicted binding energies. We then performed statistical analyses to assess whether certain structural features could discriminate between high-risk and low-risk HPV types. For each feature, we first computed descriptive statistics (mean, standard deviation, and median) within each group. We then tested for significant differences between groups using the Wilcoxon rank-sum test, and adjusted the resulting p-values for multiple testing using the Benjamini–Hochberg procedure.

3.4.2 E7-pRB interaction

For the structural analysis of the E7–pRB interaction, we used the crystal structure of the E7 peptide bound to the pRB pocket domain (PDB: 4YOZ) as a reference. This model includes only the minimal E7 peptide fragment (residues 20 to 35), which encompasses the conserved LxCxE motif responsible for binding to the retinoblastoma protein family (pRB/p107).

Accordingly, for each of our E7 sequences, we extracted only the residues corresponding to positions 20 to 35, and substituted them into the original E7 peptide of the reference model. This allowed us to generate E7–pRB complexes tailored to each HPV variant.

To ensure that this peptide region reliably represented the interaction interface, we also analyzed a full-length E7–pRB complex model manually using PyMOL, to visually inspect the interacting residues and confirm their relevance (see Supplementary Figure S1) [97]. This step helped us validate the structural assumptions based on the truncated peptide model.

As with the E6 complex, we used the PRODIGY software to evaluate the binding affinity and interaction interface between the modeled E7 peptides and the pRB pocket domain. These structural features were then used to explore potential differences in binding properties between high-risk and low-risk HPV types.

3.5 Ethical Approval

The study made use of leftover Pap smears and received approval from the Medical Ethics Committee of the University Hospital of Liège (Ref. 2019/139; Ref. 2024/495) as well as the Ethics Committee of the School of Public Health, University of Kinshasa (Ref. ESP/CE/65/2025).

4 Results

4.1 HPV Genomic Dataset Overview

To provide a comprehensive view of the diversity of HPV types detected in our study population, we first performed a global analysis on all sequenced viral genomes, including both PCR- and MDA-amplified samples, and regardless of genome completeness. This inclusive dataset allowed us to capture the broadest possible picture of the HPV landscape across the two cohorts prior to any sequence deduplication or quality filtering.

In this initial step, HPV types were assigned to each genome based on BLAST analyses against reference databases. These type annotations were subsequently appended to the sample names to facilitate downstream analyses and visualizations. At this stage, the classification was limited to the type level, without yet considering lineage or sublineage resolution.

These patterns are illustrated in the following plots showing the distribution of high-risk and low-risk HPV types detected in the DRC cohort using both amplification methods (Figures 11 and 12). For comparison, similar plots from the Liège cohort are provided in the Appendix (Figures S2 and S3), but due to selection bias, they are less suitable for direct comparison with other datasets.

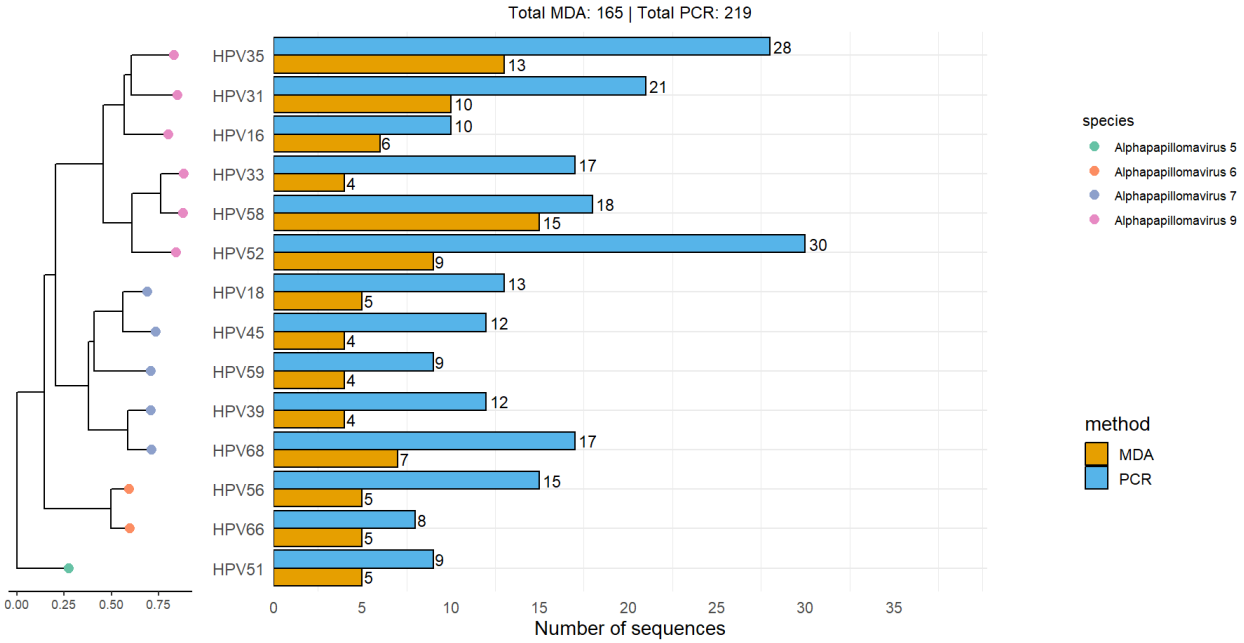

Figure 11. Distribution and phylogeny of high-risk HPV types identified in DRC. This figure combines two elements: (1) a bar plot showing the distribution of high-risk HPV (hrHPV) types detected in samples from the DRC, and (2) a phylogenetic tree of the detected hrHPV types. Sequences were aligned using MAFFT and the tree was constructed with IQ-TREE. In the tree, HPV types are color-coded according to their species classification, as indicated in the legend. Together, these visualizations provide an overview of both the prevalence and evolutionary relationships of circulating hrHPV types in the region.

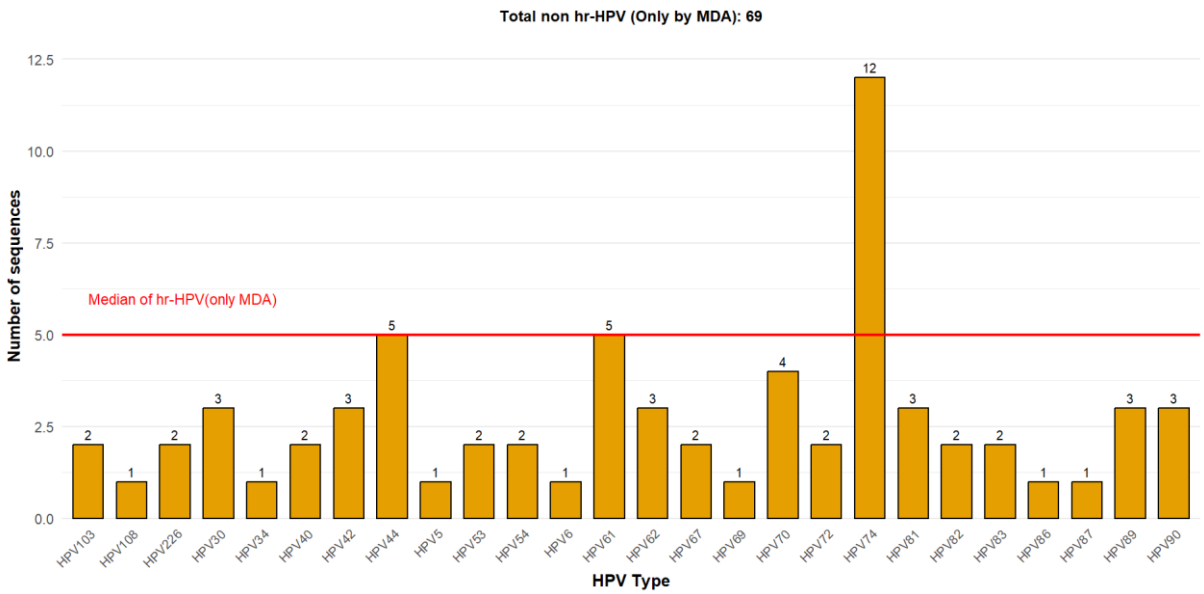

Figure 12 : Distribution of non hr-HPV types (only captured by MDA) in RDC. The red line indicates the median number of high-risk HPV (hrHPV) types detected by MDA, providing a reference for comparison. This visualization highlights the relative efficiency of MDA in capturing hrHPV versus non-hrHPV types within the same dataset.

When comparing the two amplification methods, a clear distinction emerges. PCR tends to yield a higher number of sequences per HPV type, highlighting its efficiency and accuracy in amplifying targeted viral genomes. However, MDA offers broader coverage, enabling the detection of a greater diversity of HPV types, particularly low-risk types, thanks to its primer-independent amplification strategy. These complementary profiles illustrate how PCR excels in depth, while MDA provides breadth, especially valuable for capturing HPV types not already known or targeted by PCR primers

Moreover, analysis of type distribution reveals that the most prevalent high-risk HPV types in the DRC cohort are HPV52, HPV35, HPV31, and HPV68. Among the low-risk types, HPV74 stands out as the most frequently detected. In the Liège cohort, as expected due to the inclusion criteria, HPV16 and HPV18 are the dominant high-risk types, followed by HPV31. Regarding low-risk HPV, type 42 is the most commonly observed (Figures S2 and S3).

A deduplication step was applied to the HPV genomic dataset to retain only one unique genome per HPV type and sample, in cases where amplification had been performed using both MDA and PCR. As a result, the initial dataset of 636 sequences was reduced to 491 unique genomes. Among the 145 cases where both methods had detected the same HPV type in the same individual, 90 genome pairs were found to be strictly identical and were directly merged. The remaining 55 genome pairs displayed near-identical sequences but with minor differences that required further inspection.

Manual review of these 55 cases revealed that discrepancies were typically attributable to technical issues already identified during the sequencing process. Most commonly, these included partial genome recovery resulting from the failure of one of the two primer pools, sequencing artifacts in homopolymeric regions (leading to indels), or insufficient coverage in one of the methods (especially MDA), resulting in stretches of unresolved nucleotides ("N"). Based on this evaluation, the most complete and high-quality sequence was retained for each case, ensuring the reliability of downstream analyses. Examples of read alignments visualized in IGV are provided in supplementary files (Figures S4 and S5), and a detailed table reporting all issues for each PCR/MDA sample pair is also available in the supplementary files (Table A2).

This curated dataset of 491 non-redundant HPV genomes was specifically used to analyze co-infection patterns within individual samples, ensuring that each HPV type was only counted once per sample.

In the Liège cohort, where samples were selected based on the presence of HPV16 or HPV18, we explored the co-infection landscape involving HPV16. The most frequently co-detected types with HPV16 were HPV31 and HPV42 (Figure 13), both of which are also classified as high-risk and low-risk types, respectively. This analysis highlights potential patterns of viral coexistence within the same host. Due to the low number of HPV16-positive cases in the DRC cohort, no equivalent analysis was performed for that population.

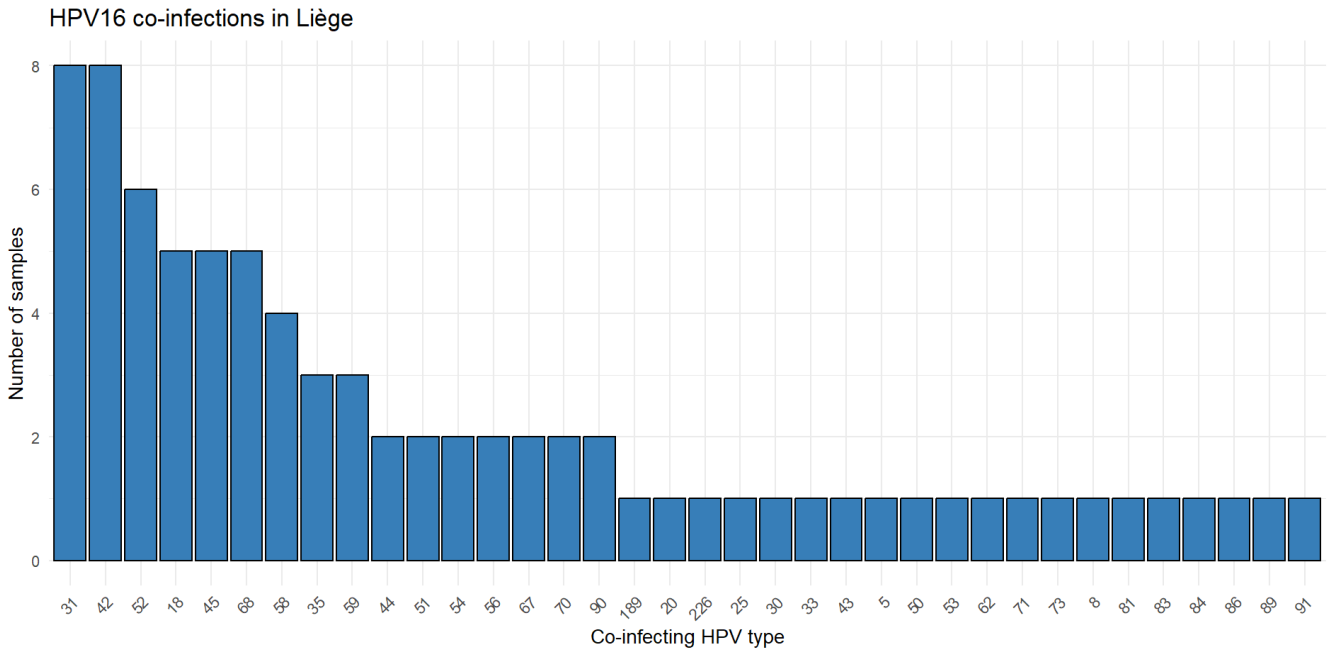

Figure 13. Barplot showing the distribution of HPV types co-infecting with HPV16 in the Liège cohort. The x-axis indicates the HPV types detected alongside HPV16, and the y-axis represents the number of co-infection cases identified.

To further characterize the diversity of HPV infections at the individual level, we analyzed the number of distinct HPV types detected per sample in both cohorts. Figure 14 presents the distribution observed in the DRC cohort, where most individuals (42.1%) carried a single HPV type. Co-infections were nevertheless common, with 23.8% of individuals harboring two types and 14.3% three types. One sample displayed up to ten different HPV types

The corresponding analysis for the Liège cohort is provided in supplementary files (Figure S6), as the selection bias in that dataset precludes direct comparison. We also identified one individual carrying a notably high number of HPV types (17).

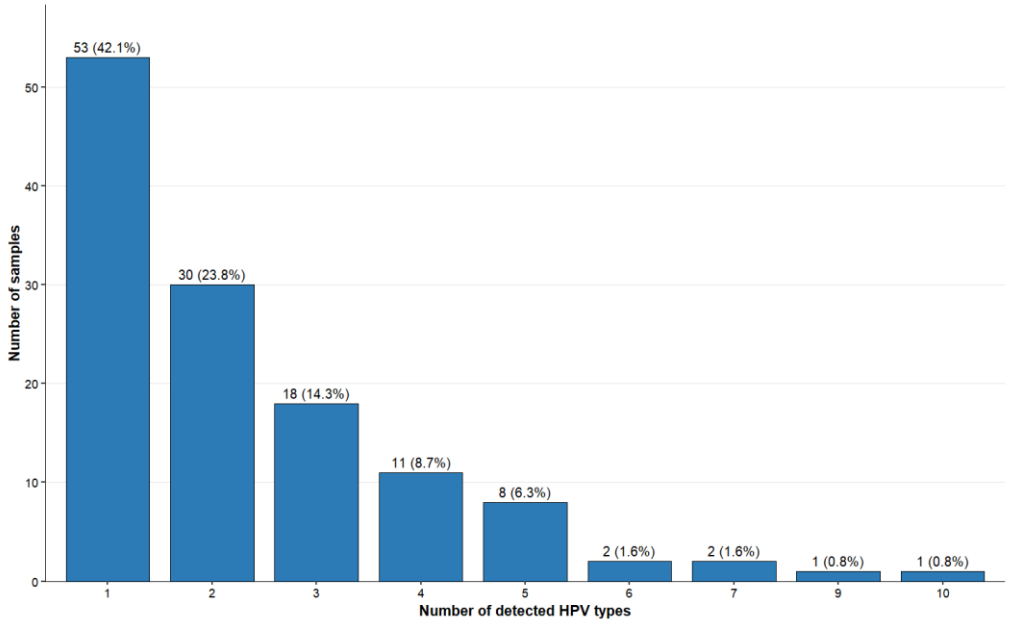

Figure 14: : Distribution of the number of HPV types per sample in Kinshasa,DRC. The x-axis represents the number of distinct HPV types detected in a sample, while the y-axis indicates how many samples fall into each category. This highlights the frequency of co-infections across the DRC data.

4.2 Classification Results Based on Full HPV Genomes

4.2.1 Classification of Sequences in Our Dataset

Following the classification workflow described previously, we applied a second filtering step to retain only complete HPV genome sequences with no ambiguous bases. This filtering resulted in a final dataset of 357 complete sequences, which served as the basis for downstream lineage, sublineage, and variant classification using the HPVxHunter tool. This step allowed for accurate typing and the identification of potential novel lineages among the collected samples.

We first examined the output table generated by the deep_analysis option of our tool (see Table 1). This table lists the names of sequences flagged as potentially novel HPV types, along with a second column indicating whether each sequence was ultimately confirmed as a true novel type or not, based on the L1 ORF (see section 1.3 *Classification*), which is the official criterion used for HPV type designation.

Among these candidates, 17 sequences were initially flagged as potentially new types due to their low similarity to any known HPV reference in the PaVE database. After verification by our pipeline, including additional phylogenetic placement, 8 sequences were confirmed as novel HPV types.

Potential New Type Sequence	Final Assignment
25162-196l	is a new HPV type
25162-175l	is a new HPV type
80131-223l	is a new HPV type
27492-195l	is a new HPV type
36510-226l	is a new HPV type
80189-103l	is a new HPV type
V7-222l	is a new HPV type
V8-168l	is a new HPV type
HPV44_MDA_23311	is NOT a new HPV type
HPV44_MDA_80159	is NOT a new HPV type
HPV44_MDA_80317	is NOT a new HPV type
HPV74_MDA_80087	is NOT a new HPV type
HPV74_MDA_80126	is NOT a new HPV type
HPV74_MDA_80228	is NOT a new HPV type
HPV86_MDA_35343	is NOT a new HPV type
HPV74_MDA_80503	is NOT a new HPV type
HPV86_MDA_80018	is NOT a new HPV type

Table 1: Results of the analysis performed with the '-deep_analysis' option enabled, aiming to identify novel HPV types based on PAVE classification criteria.

To further support the classification of these sequences as true novel HPV types, and to rule out the possibility that they are recombinant forms of existing types, we constructed phylogenetic trees based on five conserved viral ORFs: E6, E7, L1, L2, and E1. For each putative novel type, we examined its phylogenetic neighborhood across all gene trees. When a sequence consistently clustered with the same known HPV clade across all ORF-based trees, this stable topological pattern supported its identity as a distinct type rather than a recombinant artifact.

For all analyzed ORFs (E6, E7, L1, and L2), the eight newly identified HPV types showed identical phylogenetic neighborhood patterns. In contrast, the E1 phylogeny displayed a different clustering for some of these types, which may suggest a potential recombination event in the E1 region. However, this should be interpreted with caution: E1 is the largest ORF in the HPV genome (~2 kb), providing a greater number of informative nucleotide sites and thus higher phylogenetic resolution. This increased resolution may influence the observed neighborhood of certain samples without necessarily indicating genuine recombination (see Supplementary Table A3).

Among the eight newly identified HPV types, two appear to belong to the *Betapapillomavirus* genus, while the remaining six likely belong to the *Gammapapillomavirus* genus (see Supplementary Figure S7). This classification is supported by phylogenetic analysis, which shows that the novel types cluster with known HPV types from the Beta and Gamma genera, respectively.

Clones derived from these 8 samples were submitted to the International HPV Reference Center, where they were confirmed as novel HPV types. This external validation further supports the reliability and effectiveness of our tool.

After the identification of potential novel HPV types, we focused on the classification of all sequences using two reference databases: PaVE and GenBank. The output files from these comparisons provided key information including percent identity, assigned type, lineage, and sublineage (see an excerpt of the Excel outputs in Appendix Figures S8 and S9).

Initially, we observed a few false positives in sublineage assignment, primarily due to variations in sequence length between the query and the reference sequence, a factor that can bias similarity-based matching. To address this, we added a new column in our results sheet reporting the sequence length difference, so that classifications can be interpreted with caution when this difference is large, as such cases are often associated with false novel findings.

To provide an overview of lineage and sublineage diversity, we show the plot generated by our tool, displaying the prevalence of the twelve most frequent HPV types in samples from the DRC and Liège, broken down by lineage and sublineage (Figure 15,16). The full table including all identified types is provided in the Supplementary Material (Appendix, Table A4). Sequences that did not match any known reference are labeled 'ND' (Not Determined) in the table. These may represent novel lineage or sublineage candidates, especially when the best match to a known reference showed <99.5% similarity.

Additional plots generated by our HPV tool, providing complementary visualizations of the dataset, are available in the Supplementary Material (Supplementary Figures S10-13).

All HPV types identified belong to the *Alphapapillomavirus* genus, with the exception of HPV5, HPV8, and HPV25, which are members of the *Betapapillomavirus* genus. The detection of Beta types in cervical swabs is unusual, given their preferential tropism for cutaneous rather than mucosal epithelia [5].

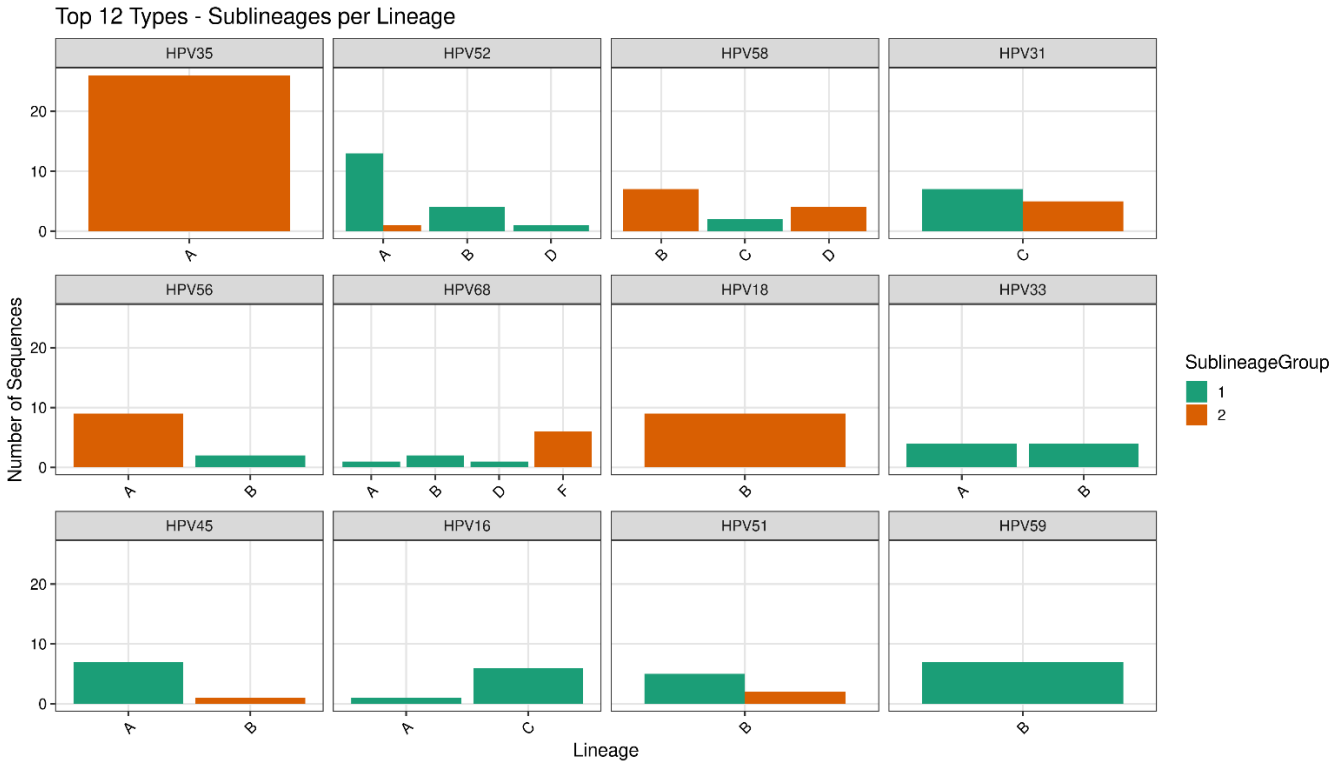

Figure 15. Distribution of sublineages within lineages for the twelve most prevalent HPV types in samples from the DRC. The x-axis represents the lineages for each HPV type, while the y-axis shows the number of sequences. Bars are colored according to sublineage group. Facets are ordered from left to right and top to bottom according to the total number of sequences per HPV type.

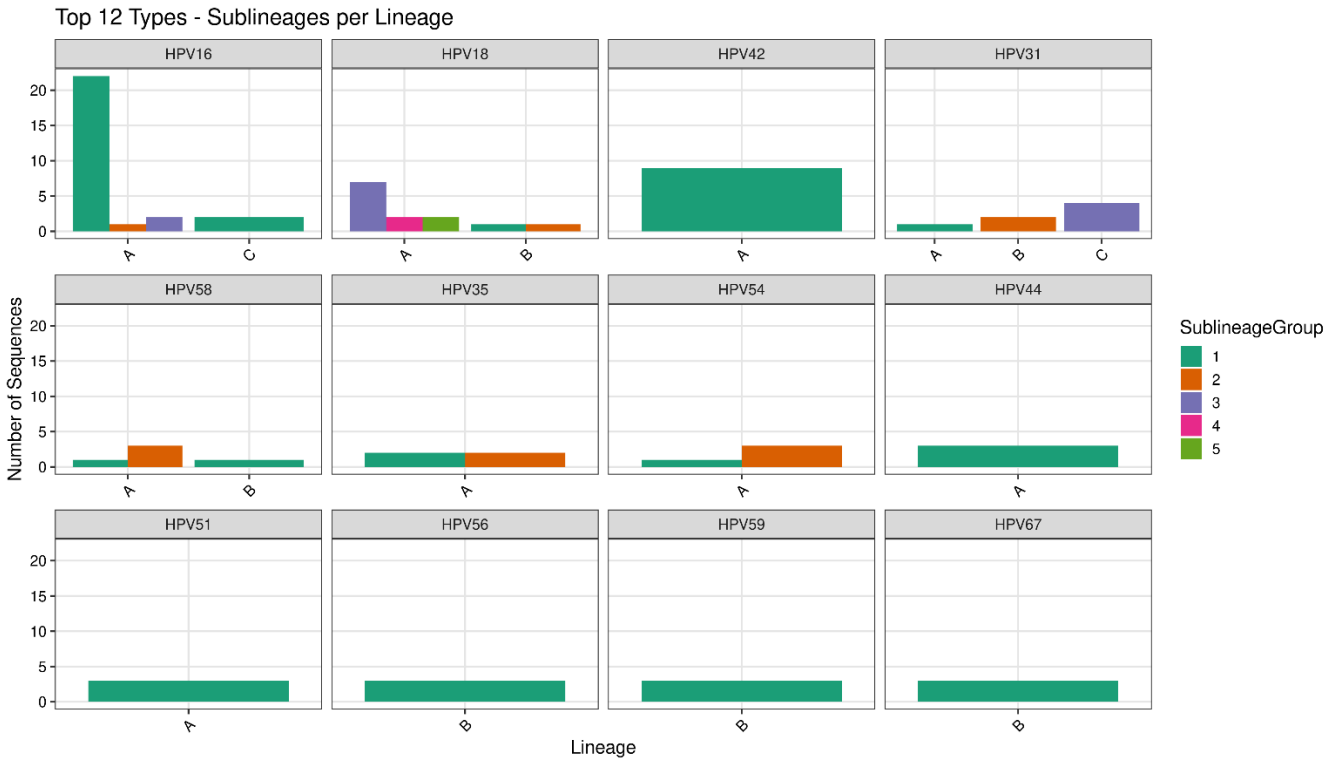

Figure 16. Distribution of sublineages within lineages for the twelve most prevalent HPV types in samples from Liege, Belgium. The x-axis represents the lineages for each HPV type, while the y-axis shows the number of sequences. Bars are colored according to sublineage group. Facets are ordered from left to right and top to bottom according to the total number of sequences per HPV type.

In the Democratic Republic of Congo, the five most prevalent HPV types are HPV35, HPV52, HPV58, HPV31 and HPV56. The major sublineages associated with these types are HPV35-A2, HPV52-A1, HPV31-C1, HPV58-B2, and HPV56-A2. HPV68-F2 is the dominant sublineage among HPV68 sequences in this population.
Notably, all sequences identified as HPV74 could not be assigned to any known lineage or sublineage, suggesting they may represent novel variants.Focusing on the two high-risk types most commonly targeted by vaccination, HPV18 and HPV16, we found relatively low frequencies in this cohort. Their dominant sublineages were HPV18-B2 and HPV16-C1, respectively.

In the Liège cohort, HPV16 and HPV18 were by far the most prevalent types, with frequencies of 22.1% and 9.7% respectively. This overrepresentation is expected given the selection bias favoring samples positive for these high-risk types.
HPV16 sequences were overwhelmingly assigned to lineage A, particularly sublineage A1 (90%), with minor proportions of A2 and A3. A small proportion of sequences belonged to

lineage C (C1). Similarly, HPV18 sequences mainly clustered in lineage A, with a predominance of sublineages A3 to A5, while lineage B was also represented equally by B1 and B2 sublineages.
Excluding HPV16 and HPV18, the most frequent types were HPV42, HPV31, HPV58, HPV67, and HPV35.
For HPV31, lineage C (sublineage C3) was the most prevalent, followed by lineages B2 and A1. HPV58 showed a predominance of lineage A (notably A2 and A1), while lineage B (B1) was less frequent. HPV67 was dominated by lineage B, mostly B1, and included a proportion (25%) of lineage B sequences that could not be confidently assigned to a known sublineage (B-ND), suggesting possible novel variation.
Finally, HPV35 appeared in equal proportions of sublineages A1 and A2, all belonging to lineage A. These lineage and sublineage distributions provided a first insight into HPV diversity in the two cohorts and highlighted several unclassified sequences that warranted further phylogenetic investigation.

4.2.2 Phylogenetic Validation of Novel Lineage/Sublineage Candidates

To validate the identification of potential novel lineages or sublineages, phylogenetic trees were constructed for all HPV types containing at least one candidate labeled as 'ND' in the lineage or sublineage column. Reference sequences from the PAVE database and GenBank were incorporated to provide context and facilitate comparison with previously characterized HPV types. This phylogenetic approach allowed us to support or revise the initial classification based on sequence similarity.

We chose to display three representative phylogenetic trees in the main text, focusing on HPV types for which the novel lineage or sublineage candidates appeared the most distinct or informative. These selections illustrate key cases where phylogenetic analysis supported the initial classification based on sequence similarity.

Specifically, we included:
– HPV45, a high-risk type, for which a potential novel sublineage was identified;
– HPV74, which showed evidence of novel lineages with sequences from both study cohorts (RDC and Liège);
– HPV66, a possibly oncogenic type, also exhibiting candidate novel variants.

These examples allow us to illustrate diverse phylogenetic patterns across clinically relevant HPV types (Figure 17-19).

The remaining trees, covering all other types with at least one 'ND'-labeled sequence, are available in the Supplementary Material (Figures S14-S28).

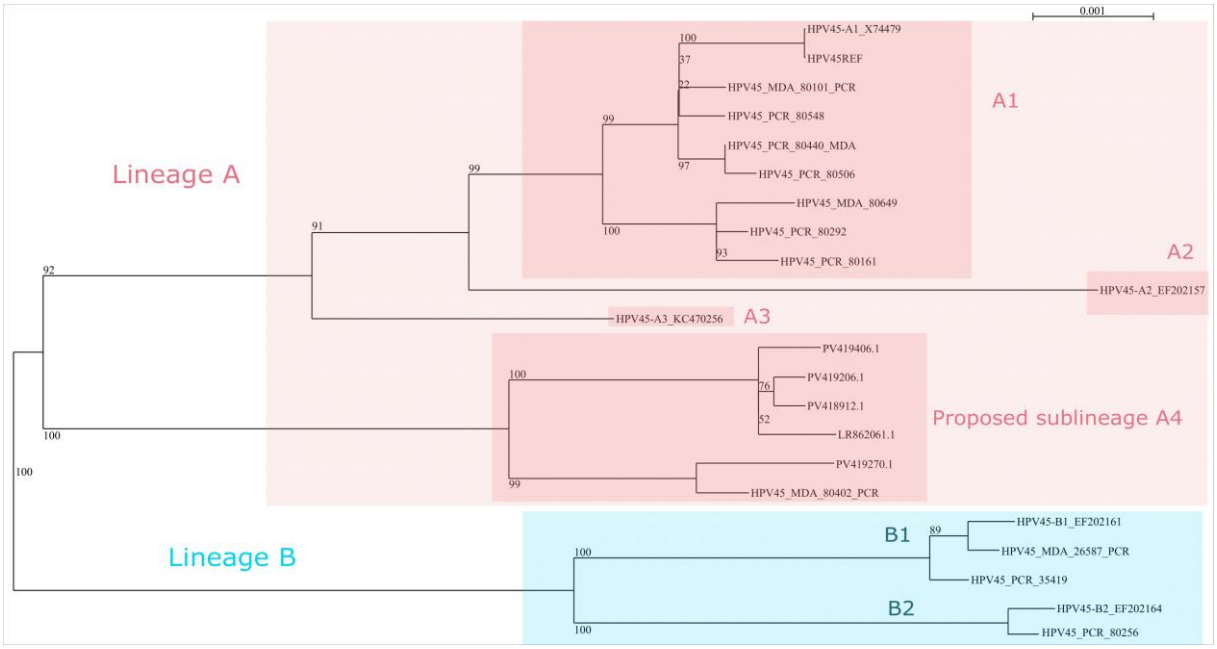

Figure 17. Maximum likelihood phylogenetic tree of complete HPV45 genomes, including study samples with reference sequences from PaVE and GenBank, highlighting the proposed novel sublineage A4. The tree includes all reference sequences from PAVE (lineages A and B shown in red and blue, respectively), our study sequences, and the five closest GenBank matches to *HPV45_MDA_80402_PCR*. All other HPV45 sequences from our dataset clustered with known sublineages (A1–A3, B1–B2). In contrast, *HPV45_MDA_80402_PCR* and its GenBank matches formed a distinct cluster, separated by ≥0.9% from HPV45-A3, and showing ≥99.5% identity within the group. This supports the proposal of a new sublineage, provisionally labeled A4.

In addition, analysis of amino acid substitutions in the E6 oncoprotein revealed a mutation specific to the proposed sublineage A4: a Cysteine-to-Tyrosine (C→Y) substitution at position 53. This mutation was not observed in any of the known sublineages and may represent a distinctive molecular signature of this new variant. Since E6 plays a key role in oncogenesis, such a substitution could potentially influence the protein's functional properties, although further experimental validation would be needed to assess its impact.

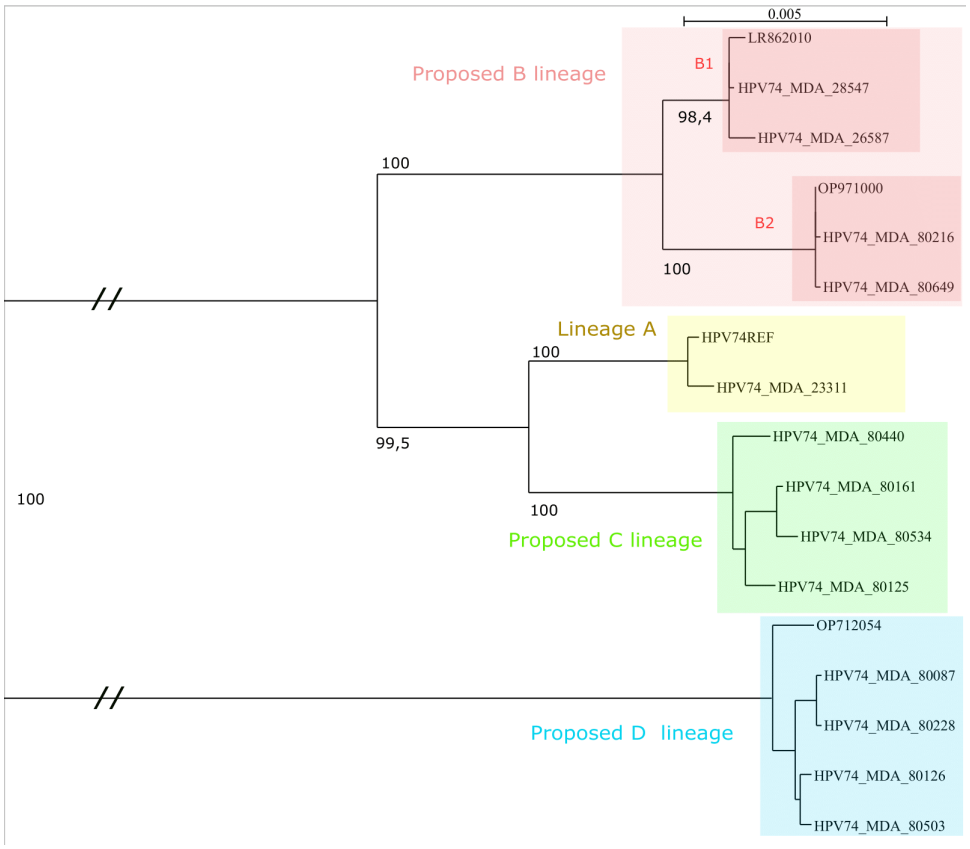

Figure 18. Maximum likelihood phylogenetic tree of complete HPV74 genomes, including study samples with reference sequences from PaVE and GenBank. Sample 23311 clusters with the reference sequence and belongs to the previously described lineage A1 (yellow). A probable lineage B (red) includes two distinct subgroups: 28547/26587 and 80216/80649, each clustering with their closest GenBank match. These subgroups show a 0.6% nucleotide divergence and are collectively at least 1.7% divergent from other lineages. A newly proposed lineage C (green) is formed by samples 80440, 80161, 80534, and 80125, which differ from each other by no more than 0.3%, and show no GenBank match above 99% similarity, supporting their classification as a novel lineage. Finally, lineage D (blue) comprises samples 80087, 80228, 80126, and 80503, with intra-lineage identity above 99.9% and more than 1% divergence from other clusters. This group also includes its closest GenBank reference and shows the highest inter-lineage divergence (≥5%).

The geographic origin of the samples further supports the proposed sublineage structure. For instance, the proposed *B1* sublineage is composed exclusively of samples from Belgium,

and its closest GenBank match (LR862010) originates from Luxembourg, which suggests a potential European sublineage. In contrast, the *B2* group includes only samples from the Democratic Republic of Congo (DRC), with its closest reference genome (OP971000) coming from South Africa. This geographical consistency strengthens the hypothesis of region-specific sublineages. A similar pattern is observed for the proposed *C* and *D* lineages, which are exclusively formed by samples from DRC. Notably, the best GenBank match for lineage *D* (OP712054) is from Togo, further supporting a potential African origin.

While these observations suggest a possible geographical clustering of lineages, they should be interpreted cautiously due to the limited sampling in public databases and potential biases in geographic representation.

The analysis of amino acid mutations in the E6 and E7 proteins further supports the proposed lineage structure. In the E6 protein, the proposed D lineage harbors the highest number of specific mutations (n=5), which is consistent with its deep divergence in the phylogenetic tree. The C lineage presents two unique amino acid changes, while B1 and A show only one specific mutation in this region.

In the E7 protein, D again stands out with five specific mutations, reinforcing its distinctiveness. Notably, B1 carries a two-amino-acid insertion (L–E) at positions 17–18, absent in B2, along with an additional mutation, which together differentiate B1 from B2.

These lineage-specific amino acid changes highlight molecular divergence that may have accumulated in geographically or evolutionarily distinct viral.

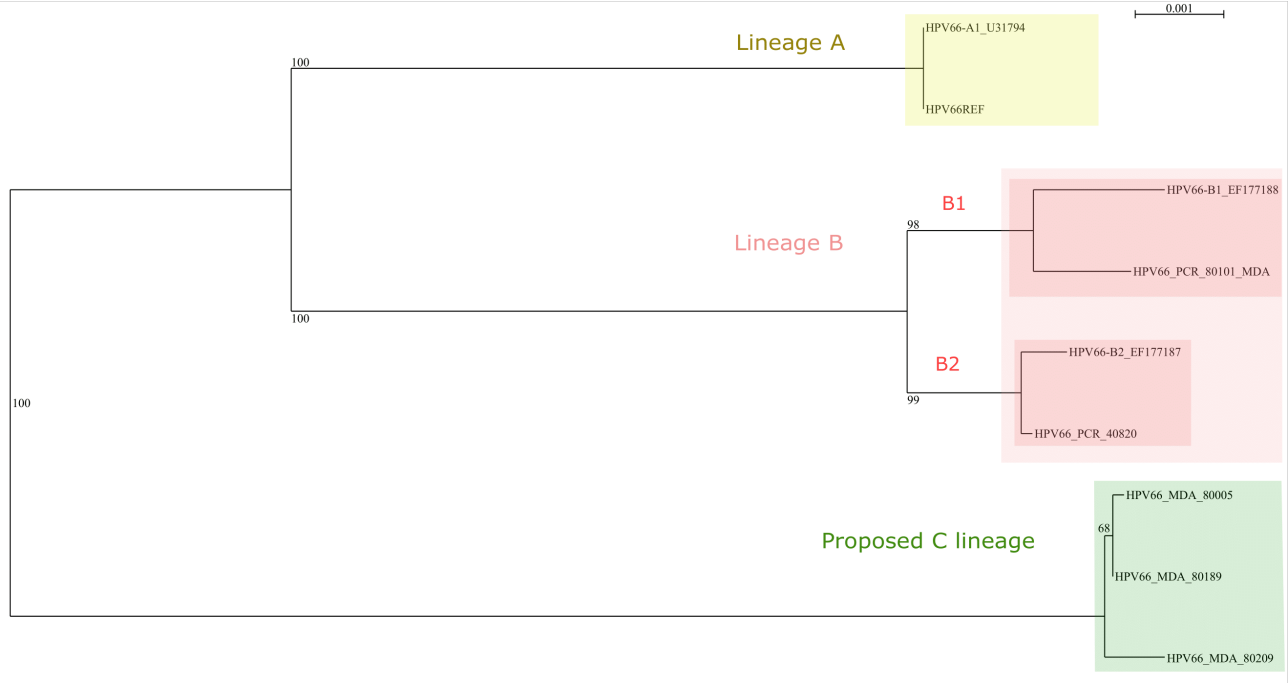

Figure 19. Maximum likelihood phylogenetic tree of complete HPV66 genomes, including study samples with reference sequences from PaVE. The tree was constructed based on full genome sequences of HPV66. Groups are defined by a minimum divergence of 1% between them and a maximum intra-group divergence of 0.2%. The reference sequence from lineage A is shown in yellow. Two samples from this study (80101 and 40820) cluster closely with the B1 and B2 sublineages, respectively, and are highlighted in red. In contrast, three other samples (80005, 80189, and 80209) form a distinct and well-supported clade, highlighted in green, which is sufficiently divergent to be proposed as a novel lineage C.

In addition to its phylogenetic divergence, the proposed lineage C displays a distinct molecular signature, with three specific amino acid changes in the E6 protein and two in E7). While these E7 mutations (L70F and D75N) lie outside the well-characterized LXCXE motif responsible for binding to the retinoblastoma (pRB) family proteins, located approximately at positions 20–26, they may still contribute to subtle functional differences through effects on protein stability, folding, or secondary interactions. Further experimental validation will be necessary to determine the precise functional consequences of these changes in both E6 and E7

4.2.3 Classification Results from EMBL HPV Reference Database

To further assess the robustness of our classification tool and to gain a broader overview of HPV genomic diversity, we applied our pipeline to a large dataset of publicly available HPV genome sequences from the European Nucleotide Archive (ENA). While this large-scale

analysis allowed us to evaluate the performance of our method beyond our original cohorts (RDC and Liège), it also provided an opportunity to explore the current state of HPV genomic representation in public databases.

Out of the 10,729 genomes retrieved from the EMBL/ENA database, only 280 sequences (+-2.6%) could not be assigned to a known HPV type by our tool. This low proportion of unclassified sequences suggests robust performance of the pipeline, despite the variable quality of publicly deposited data.

Interestingly, over 80% of the dataset is composed of only seven HPV types, highlighting a strong imbalance in the representation of HPV genomic diversity. The most frequent types are: HPV16 (4059 sequences), HPV31 (2161), HPV35 (904), HPV45 (802), HPV6 (306), untyped (ND, 280), and HPV58 (188). This skewed composition reflects database submission biases rather than the true prevalence of HPV types worldwide.

Moreover, metadata completeness is strikingly limited. Only 77% of the sequences included a country of origin, and a mere 22% had a collection date annotated. The geographic representation is also highly skewed, with the United States alone contributing 5559 sequences, more than half of the entire dataset (Figure 20). Finally, only 1404 sequences contain both country and collection date information (see Supplementary Figure S29).

This lack of comprehensive spatial and temporal metadata seriously hinders the feasibility of robust, large-scale epidemiological analyses, especially those that aim to track transmission dynamics, local lineage expansion, or the emergence of new variants. Similar approaches have been extensively applied to SARS-CoV-2, where integrating genomic, spatial, and temporal data enabled the reconstruction of transmission chains and the tracking of local lineage expansion. These analyses have been crucial for monitoring the emergence and spread of new variants, demonstrating the importance of comprehensive metadata for effective epidemiological surveillance [119,120].

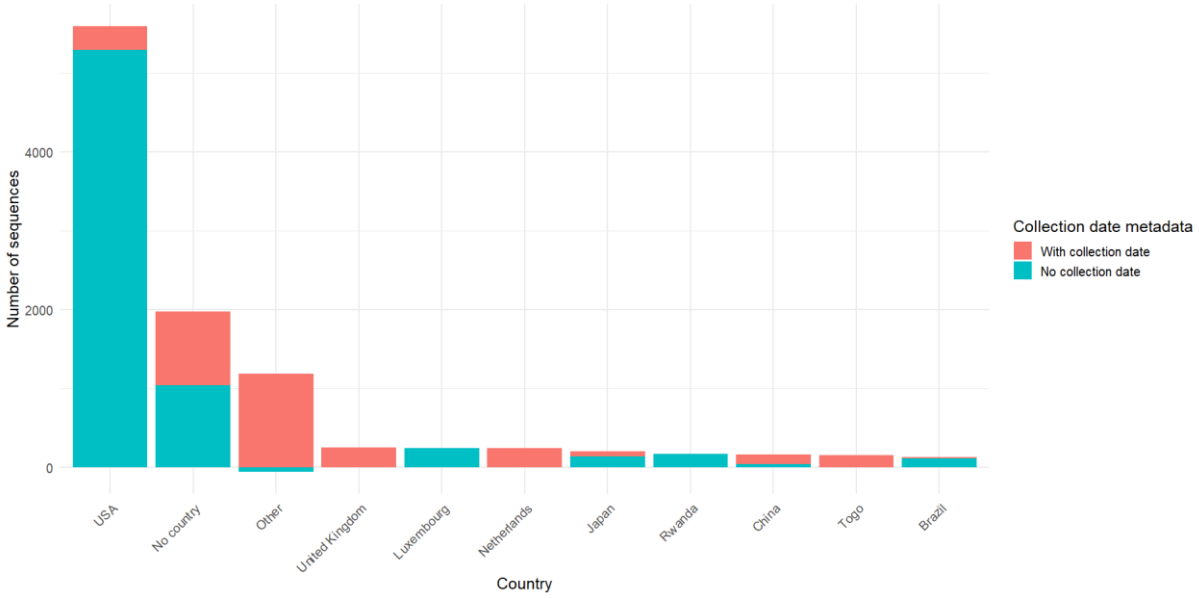

Figure 20. Geographic distribution of HPV genome sequences deposited in the EMBL/ENA database, highlighting the imbalance across countries. The bar heights represent the number of sequences per country, and the bars are split by availability of collection date metadata (red: with collection date; turquoise: without). The United States alone contributes over half of all sequences, while many entries lack complete metadata

4.3 Comparative Structural Analysis of Oncoprotein–Host Interactions in High- and Low-Risk HPV Types

To investigate differences in host–viral interactions between low- and high-risk HPV types, E6 and E7 complexes with their cellular targets were modeled using AlphaFold-Multimer, and interaction features were extracted with PRODIGY for statistical comparison.

4.3.1 Structural Modeling and Interaction Features of E6–E6AP–p53 Complexes

E6-E6AP

To provide an evolutionary context for the structural investigation of the E6 oncoprotein, we first constructed a maximum-likelihood phylogenetic tree based on the E6 amino acid sequences from both high-risk (HR) and low-risk (LR) HPV types (Figure 21). As shown in the tree, all HR types (highlighted in red) form a well-supported monophyletic clade, suggesting a shared evolutionary origin.

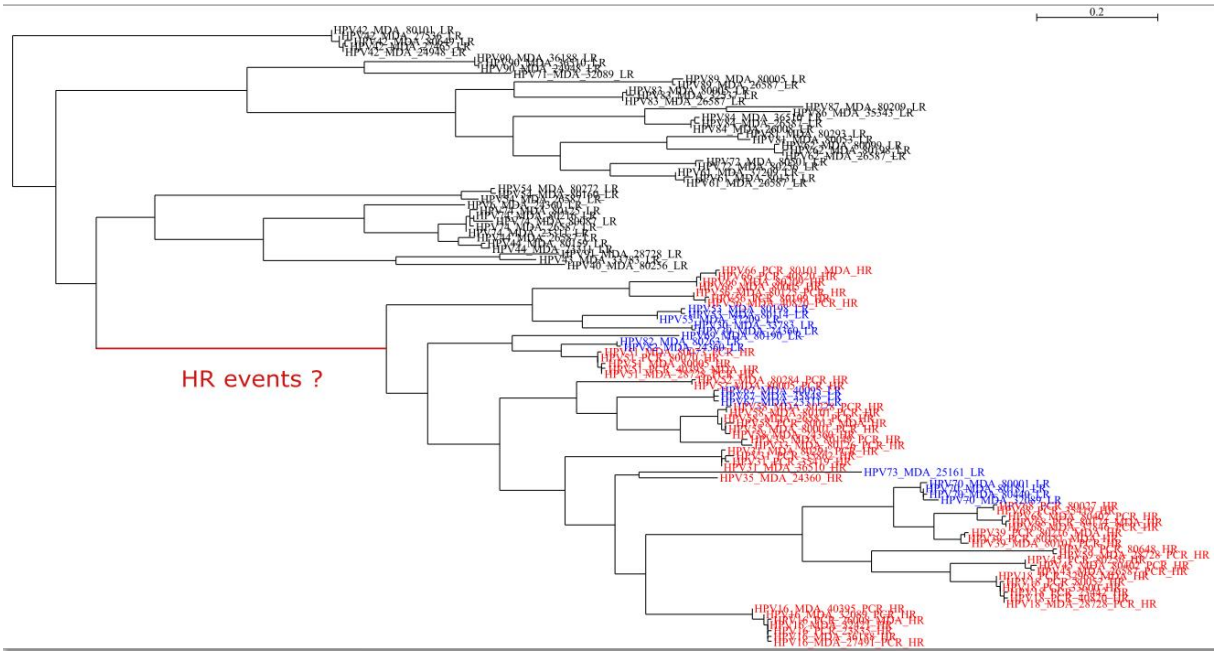

Figure 21: Maximum-likelihood phylogenetic tree of E6 amino acid sequences from HPV types. High-risk (HR) types are shown in red, low-risk (LR) types in black, and LR types that cluster within the HR subtree are highlighted in blue. A potential ancestral branch leading to HR divergence is indicated in red, raising the hypothesis of key mutations associated with oncogenicity. The unexpected clustering of some LR types within the HR clade suggests possible reverse evolution.

Interestingly, a subset of LR types (in blue) also clusters within this HR-dominated subtree, which may suggest potential functional or evolutionary similarities. These LR sequences appear to diverge from a branch inferred to be associated with HR-specific mutational events (marked in red), raising the possibility that they could represent reversal mutations, for example, types that may have lost oncogenic potential over time.

This phylogenetic topology, with its nested clustering of HR and certain LR types, points to evolutionary transitions between risk groups and supports the hypothesis that sequence-level changes in E6 could underlie the differential interactions with host proteins such as E6AP and p53.

Given the phylogenetic clustering patterns observed, especially the interspersing of certain LR types within a primarily HR clade, we hypothesized that specific sequence variations in E6 may influence its biochemical behavior and oncogenic potential. In particular, these variations might alter the interaction capabilities of E6 with its host partners, notably E6AP and p53, two critical components of HPV-mediated oncogenesis.

To test this hypothesis, we first examined whether primary sequence patterns could account for the phylogenetic clustering. A multiple sequence alignment of E6 proteins from both risk groups was performed, but no clear, consistent motifs emerged that could reliably

discriminate HR from LR types. This lack of obvious sequence-level signatures led us to consider that functional differences might instead arise from subtler structural or interfacial variations that are not apparent in the primary sequence.

We therefore turned to structural modeling as a complementary approach. Using AlphaFold-Multimer, we generated models of E6 in complex with its known cellular targets, E6AP and p53. The predicted structures displayed reasonably good folding confidence, with ptm (predicted TM-score) values ranging from 0.53 to 0.66, indicating that the individual protein folds were reliably modeled. We also examined the iptm scores, which estimate interface prediction accuracy: HR types exhibited a higher mean iptm (0.437) compared to LR types (0.3938), suggesting more stable or extensive predicted interactions in the HR group.

To explore these differences further, we used the PRODIGY tool to derive structural and energetic descriptors for the E6–E6AP complexes, including predicted binding affinities, detailed interfacial composition, and the nature of intermolecular contacts.

We first compared structural descriptors between HR types (red) and the HR-clade LR types (blue) using the Wilcoxon rank-sum test. Three features: charged apolar (number of residues at the interface surface exhibiting both charged and apolar characteristics), binding affinity, and dissociation constant, displayed nominal statistical significance (raw $p < 0.05$) but did not remain significant after Benjamini–Hochberg (BH) correction for multiple testing, suggesting a modest trend toward structural differentiation but overall similarity between these two groups.

We then asked whether the HR-clade LR types differed from the phylogenetically distant LR types (black). Here, two features, percent_charged_nis and percent_apolar_nis, remained significant even after BH correction. These metrics represent the proportion of charged or apolar residues located in the non-interaction surface (NIS) of the complex, i.e., the solvent-exposed protein surface outside the binding interface. This indicates that the HR-clade LR types possess distinct surface property profiles compared with LR types outside the HR branch.

Next, we compared all HR types against all LR types (blue + black). Seven features passed BH correction, with the strongest signal ($p = 5 \times 10^{-6}$) indicating robust structural differentiation between canonical risk groups. However, to specifically test the hypothesis that HR-clade LR types resemble HR types more than other LR types, we compared the combined group of HR and HR-clade LR types (red + blue) against the distant LR types (black). This yielded the same significant features after BH correction, with the most significant at $p = 5.9 \times 10^{-8}$, stronger than in the HR vs. all-LR comparison, further supporting the notion that HR-clade LR types are structurally more similar to HR types than to other LR types.

From an evolutionary perspective, these findings suggest a scenario in which an ancestral LR lineage acquired mutations leading to the HR phenotype (the “HR events” branch), followed by partial reversion in certain lineages (the HR-clade LR types), attenuating but not entirely eliminating oncogenic potential. Notably, HR-clade LR types (“53”, “30”, “69”, “82”,

"67", "73", "70") are classified by the IARC as probably oncogenic, consistent with their intermediate position in both phylogenetic and structural space.

All statistical test values for the four comparisons are provided in Supplementary Table A5.

Interestingly, low-risk HPV E6 proteins show a higher proportion of apolar residues on their non-interacting surface compared to high-risk types (Figure 22). Conversely, charged residues are more abundant in the non-interacting surface of high-risk E6 proteins (adjusted $p = 6.55 \times 10^{-5}$). Although these regions are not part of the binding interface, their composition may influence surface properties such as solubility, flexibility, or transient interactions. This pattern is consistent with, but does not by itself demonstrate, the idea that high-risk E6 variants can form more stable and functionally relevant complexes with E6AP, possibly contributing to their oncogenic potential.

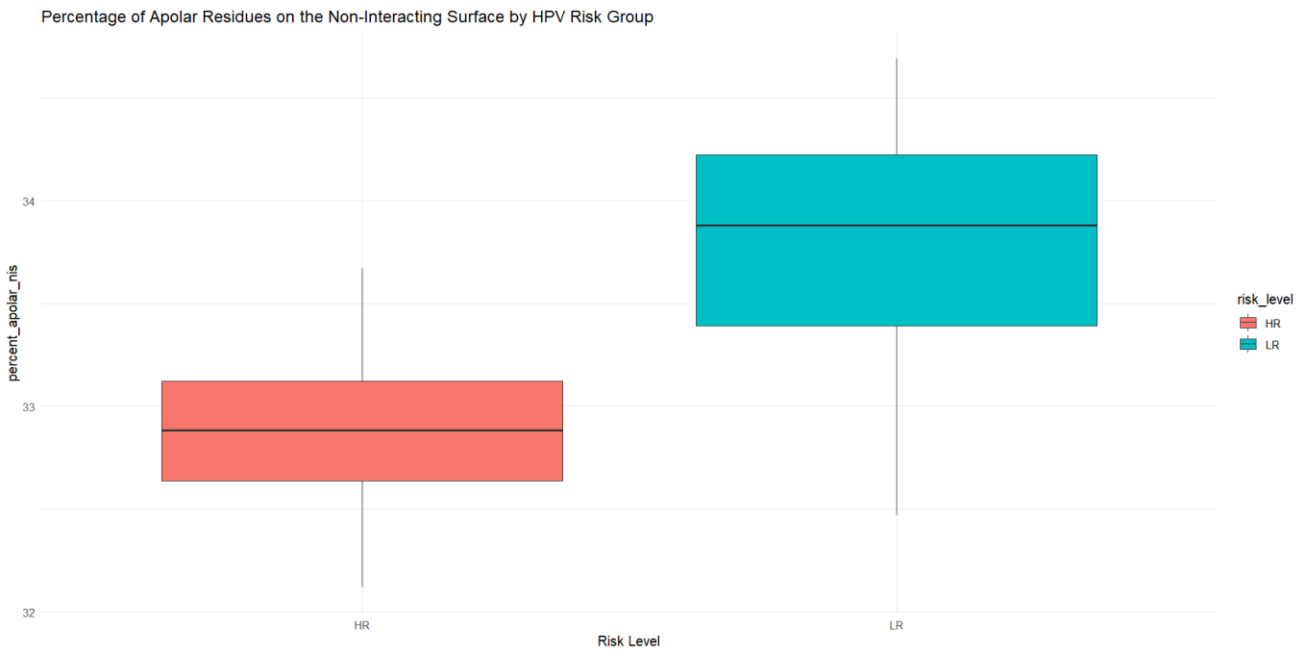

Figure 22. Distribution of the percentage of apolar residues on the non-interacting surface of E6 proteins, stratified by HPV risk group. Low-risk (blue) types show a significantly higher proportion of apolar residues on non-interacting surfaces compared to high-risk (red) types (adjusted p value = 5.04×10^{-6}, Wilcoxon test with Benjamini–Hochberg correction). This feature may contribute to differences in surface properties affecting protein–protein interaction potential.

E6-p53

Similar trends were observed for the E6–p53 interaction as for E6–E6AP (data not shown). However, one additional finding emerged: for certain HPV types, Prodigy software detected no contacts between E6 and p53. Most of these HPV types belong to the low-risk (LR) group, indicating weaker interactions and further supporting the lower oncogenic potential of LR-HPVs

4.3.2 Structural Modeling and Interaction Features of E7–pRB Complexes

Phylogenetic reconstruction of E7 protein sequences revealed two distinct monophyletic clusters enriched in high-risk (HR) types (Figure 23). These clusters share no immediate common ancestor within the HPV phylogeny, suggesting that the HR phenotype may have arisen independently at least twice during papillomavirus evolution. Such a pattern is compatible with convergent evolution, where distinct lineages acquire similar molecular properties conferring an enhanced ability to disrupt pRB function.

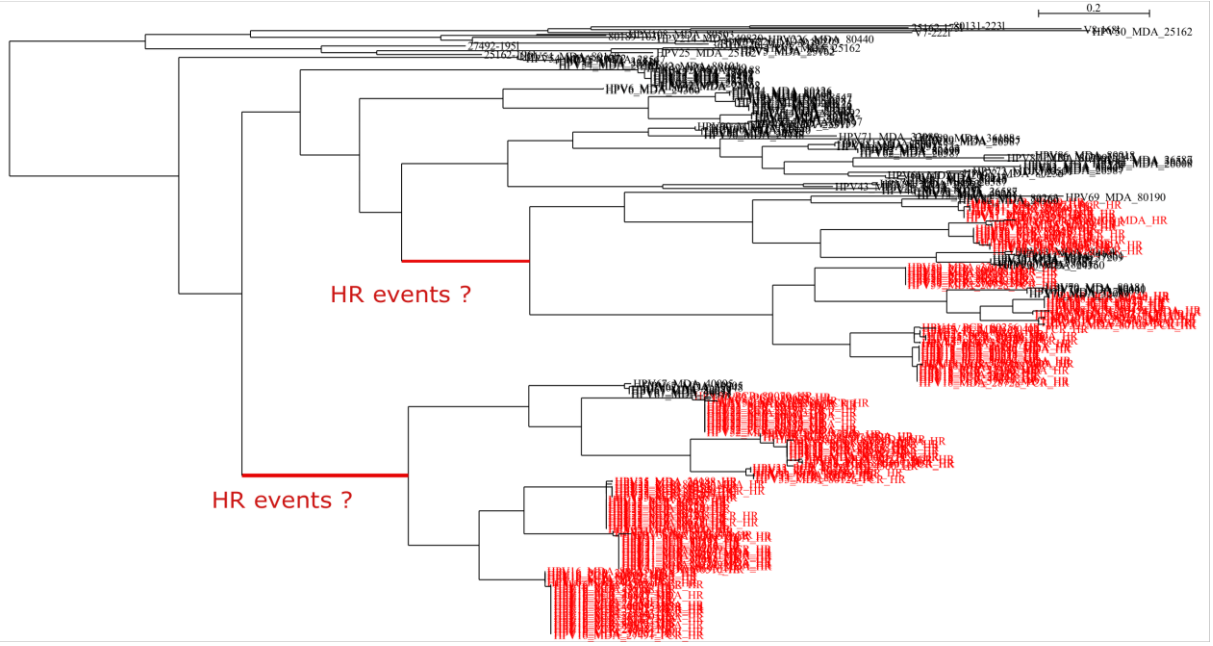

Figure 23: : Maximum-likelihood phylogenetic tree of E7 amino acid sequences from HPV types. High-risk (HR) variants are shown in red, and low-risk (LR) variants in black. Two major clades enriched in HR variants are observed, suggesting the occurrence of distinct evolutionary events that independently converged toward structural profiles associated with high oncogenic risk.

To explore whether these two HR lineages display comparable structural features in their E7–pRB interfaces, we applied our structural analysis pipeline, combining AlphaFold-Multimer

for complex prediction and PRODIGY for interface characterization. Pairwise comparisons between the “upper” and “lower” HR clusters identified several statistically significant differences after Benjamini–Hochberg correction (Wilcoxon rank-sum test). Most notably, intermolecular contacts, charged–apolar interactions, and the proportion of apolar and charged residues in non-interacting surfaces showed highly significant differences (adjusted $p < 0.01$), suggesting that these two HR lineages may have evolved distinct physicochemical strategies to achieve similar oncogenic outcomes.

When comparing all HR types collectively against low-risk (LR) types, only charged–polar and apolar–apolar interfacial contacts remained significantly different (adjusted $p < 0.05$), whereas other descriptors, including total intermolecular contacts and predicted binding affinity, did not reach significance. This contrast implies that, although HR types share some broad interface trends compared to LR types, the two independent HR clusters also retain lineage-specific molecular signatures, potentially reflecting different evolutionary trajectories towards high-risk phenotypes.

The full statistical results for both comparisons (upper vs. lower HR clusters, and all HR vs. LR types) are summarized in supplementary files , Table A6.

5 Discussion

In the Congolese cohort, the most prevalent types were HPV35 (12.26%), HPV52 (9.91%), HPV31 (8.02%), HPV58 (7.08%), HPV56 (5.19%), HPV68 (5.19%), HPV18 (4.25%), HPV45 (4.25%), and HPV33 (3.77%). This distribution is consistent with previous reports from Congo, where HPV35, HPV52, HPV31, and HPV58 are frequently detected among high-risk infections [99,100]. Notably, several of the most common genotypes in our study, including HPV35, HPV52, HPV58, and HPV68, are not covered by the currently available bivalent, quadrivalent, or even nonavalent vaccines. This highlights a critical gap in vaccine protection for the Congolese population, emphasizing the importance of continued surveillance and potential future adaptation of vaccine formulations to better match the local epidemiological profile.

For HPV31, some patterns observed in our dataset are consistent with previously reported trends in the literature. In particular, the predominance of the C1 sub-lineage in the DRC is in line with the distribution typically reported for African populations [101]. For HPV58, we observed a dominance of the B2 sub-lineage in our DRC cohort, whereas European populations mainly carry A/B1 lineages. This is consistent with recent reports showing that B2, C, and D lineages are primarily found in African populations, while A/B1 predominates in Europe [102].

In our cohort, HPV42 was rare in the DRC compared to Belgium, suggesting a possible lower prevalence in this region of Africa. Moreover, according to a study conducted in Italy, HPV42 was the most frequent low-risk type, which supports our assumption about geographic variability in its distribution [103].

In the case of HPV16, the DRC cohort was predominantly represented by the C1 sub-lineage, whereas the Belgian cohort mainly harbored the A1 sub-lineage, consistent with previous reports on global HPV16 distribution [104]. For HPV18, only the B2 sub-lineage was observed in the DRC cohort, while the Belgian cohort was largely composed of A3–A5 sub-lineages, also aligning with published data [105].

Some of our results were surprising, notably the high number of HPV types detected per sample, as well as the presence of Beta-papillomaviruses in cervical swabs. In the first case, the high HPV multiplicity could be explained by factors such as immunodeficiency or increased exposure, for example in the context of sex work—conditions known to be associated with a higher HPV burden [106]. In the second case, the detection of Beta-papillomaviruses likely reflects incidental or transient presence, as these viruses are commonly found on the skin and may be transferred to the cervical swab during sampling. Indeed, as reported in the literature, "detection of β-HPV types in the cervix tends to occur as random and transient episodes not explained via the sexual-transmission correlates that characterize infections by α-HPVs" [107].

Our tool performed exceptionally well, rapidly classifying HPV sequences down to the sub-lineage level. This capability represents a significant advantage for both epidemiology and clinical practice, as it allows timely monitoring of infections and early identification of high-risk HPV types and more aggressive sub-lineages. Furthermore, the tool performed reliably on publicly available databases such as EMBL, demonstrating its potential for large-scale epidemiological surveillance. However, the metadata associated with these public sequences were often incomplete. Such limitations in both diversity and metadata completeness, particularly the lack of consistent spatial and temporal information, severely restrict the ability to conduct robust epidemiological analyses. In particular, methods like BEAST [108,109], which require both sampling dates and geographic origins for time-calibrated phylogeographic modeling, cannot be meaningfully applied to most of these public sequences. These findings emphasize the need for more diverse, better-curated, and consistently annotated public HPV genomic databases, especially if such resources are to be used for surveillance or global diversity assessments. Similar efforts have proven invaluable during the COVID-19 pandemic [120], where initiatives such as the COVID-19 Genomics UK Consortium (COG-UK) played a key role in sequencing SARS-CoV-2 genomes and tracking the emergence and spread of viral variants.

Despite its large global health impact and the significant mortality it causes each year, HPV remains comparatively under-studied (Figure 24). This research gap may be partly explained by the fact that HPV-related mortality is highest in low- and middle-income countries, where research infrastructure is often limited [110]. In addition, HPV-related deaths disproportionately affect women, which may contribute to lower prioritization compared to infectious diseases with broader or more immediate perceived impact across populations [111,112].

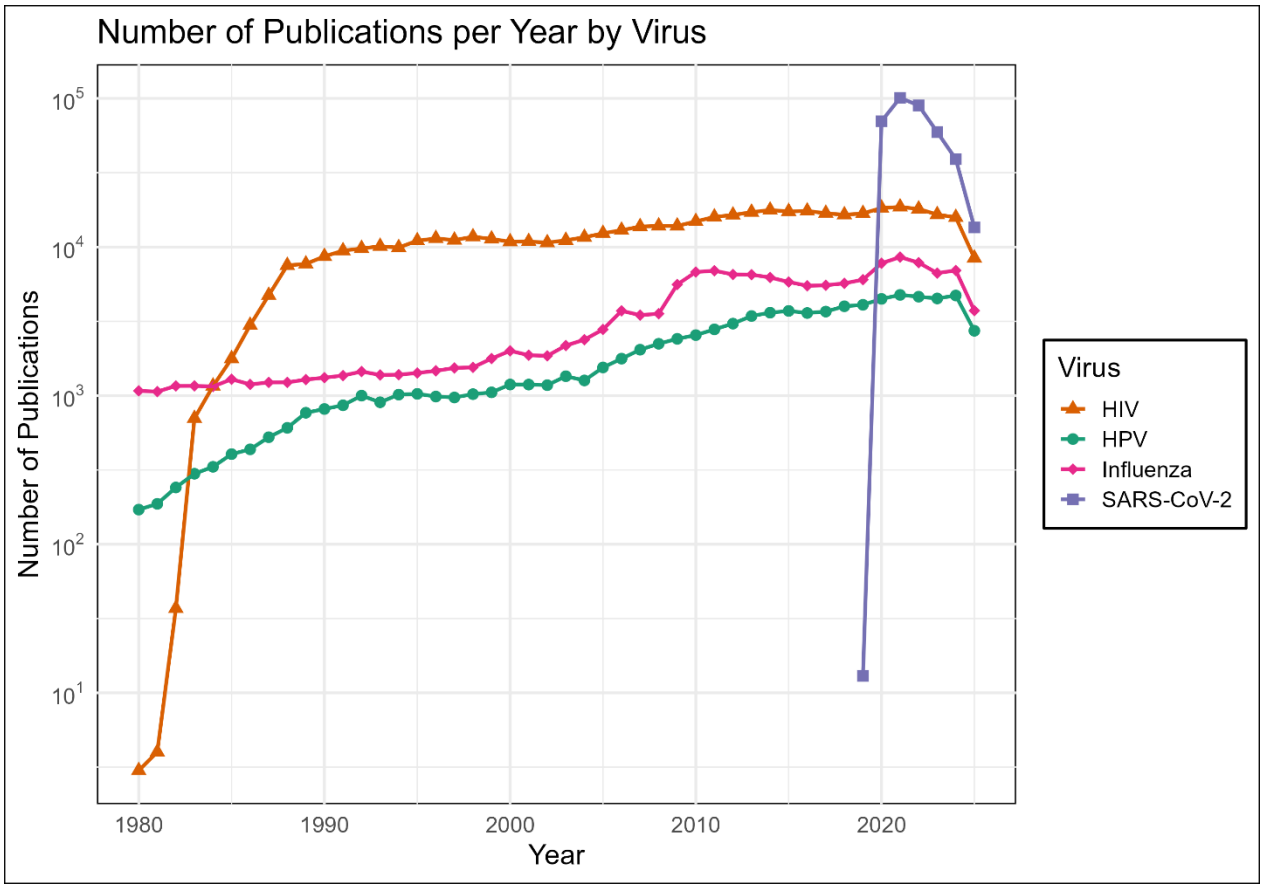

Figure 24. Annual number of scientific publications for HIV, HPV, Influenza, and SARS-CoV-2 (1980–2024). Despite its substantial health burden, HPV remains comparatively underrepresented in research output, as illustrated by the number of articles per year retrieved from PubMed for these viruses.

Our structural and phylogenetic analyses of the E6 oncoprotein suggest the existence of three functional–evolutionary "states" within papillomaviruses: canonical high-risk (HR) types, canonical low-risk (LR) types, and a third category of LR types embedded within the HR clade that retain intermediate structural and biochemical characteristics. These "HR-clade LR" types show interface and surface residue properties more similar to HR types than to distant LR types, despite their current classification as low-risk. This intermediate profile is consistent with an evolutionary scenario in which an ancestral LR lineage acquired mutations conferring HR-like oncogenic potential, followed by partial reversion events that attenuated, but did not entirely abolish, these capabilities. The positioning of such types within phylogenetic space, together with their structural similarity to HR E6, aligns with IARC's designation of some as "probably oncogenic" and supports a more nuanced, continuum-like view of HPV oncogenic risk, potentially conceptualized as a triangular relationship rather than a simple binary.

In the case of E7–pRB interactions, phylogenetic reconstruction revealed two monophyletic HR clusters lacking a recent common ancestor, indicating convergent evolution towards high-risk phenotypes. Structural comparisons confirmed that these clusters exhibit distinct interface chemistries, suggesting that different evolutionary pathways can yield comparable oncogenic capabilities. This convergence underscores the functional constraints shaping E7–pRB interactions and highlights the potential for multiple structural “solutions” to achieving host protein disruption.

We implemented a preliminary machine learning approach using a Random Forest model trained on E7–pRB interaction features to evaluate whether novel HPV types could be classified according to their oncogenic risk. For the eight newly characterized types, the model consistently predicted a low-risk profile, consistent with their clustering within Gamma and Beta papillomaviruses. Some low-risk types located within high-risk clades were predicted as high-risk, indicating sensitivity to subtle structural differences (data not shown).

Although limited by the small dataset, this approach demonstrates the potential to classify HPV types based on structural descriptors alone. Combining these models with complementary data such as epidemiological records, host–immune interaction profiles, transcriptomic or proteomic signatures, or in vitro functional assays could improve predictive accuracy and provide an early-warning framework for emerging high-risk variants. Larger and better-annotated structural and genomic datasets will be needed to fully realize this potential. This exploratory analysis highlights a promising direction for future work.

6 Conclusion and perspectives

This work demonstrates the effectiveness of our tool in rapidly and accurately classifying HPV sequences down to the sub-lineage level, enabling both timely epidemiological monitoring and clinically relevant risk assessment. Coupled with PCR, this approach offers a low-cost, fast, and scalable solution for large-scale surveillance, applicable even in low-resource healthcare settings. In the Democratic Republic of the Congo, identification of the most prevalent HPV types revealed that they are not covered by current vaccines, highlighting a critical gap in preventive strategies. Furthermore, the ability to detect novel variants, lineages, and sub-lineages provides an essential framework for tracking viral evolution and monitoring potential shifts toward increased oncogenicity. These capabilities position our tool as a valuable asset for integrated HPV control efforts, from local public health initiatives to global surveillance programs.

References

Some computational scripts were developed with assistance from ChatGPT (OpenAI, March 2025 – August 2025), and English language refinement was supported by DeepL (DeepL SE) and QuillBot (QuillBot Inc.).

1. Oyouni AAA. Human papillomavirus in cancer: Infection, disease transmission, and progress in vaccines. Journal of Infection and Public Health [Internet]. 2023 Apr [cited 2025 Jul 14];16(4):626–31. Available from: https://www.sciencedirect.com/science/article/pii/S1876034123000564

2. Humans IWG on the E of CR to. Human Papillomavirus (HPV) Infection. In: Human Papillomaviruses [Internet]. International Agency for Research on Cancer; 2007 [cited 2025 Jul 13]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK321770/

3. Lowy DR. Harald zur Hausen (1936 to 2023): Discoverer of human papillomavirus infection as the main cause of cervical cancer. Proceedings of the National Academy of Sciences of the United States of America [Internet]. [cited 2025 Jul 7];121(11):e2400517121. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10945753/

4. Doorbar J, Egawa N, Griffin H, Kranjec C, Murakami I. Human papillomavirus molecular biology and disease association. Reviews in Medical Virology. 2015 Mar;25 Suppl 1(Suppl Suppl 1):2–23.

5. Gheit T. Mucosal and Cutaneous Human Papillomavirus Infections and Cancer Biology. Frontiers in Oncology [Internet]. 2019 May [cited 2025 Jul 9];9:355. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6517478/

6. Araldi RP, Sant'Ana TA, Módolo DG, Melo TC de, Spadacci-Morena DD, Cassia Stocco R de, et al. The human papillomavirus (HPV)-related cancer biology: An overview. Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie. 2018 Oct;106:1537–56.

7. Thorland EC, Myers SL, Persing DH, Sarkar G, McGovern RM, Gostout BS, et al. Human papillomavirus type 16 integrations in cervical tumors frequently occur in common fragile sites. Cancer Research. 2000 Nov;60(21):5916–21.

8. Luria L, Cardoza-Favarato G. Human Papillomavirus. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 [cited 2025 Jul 14]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK448132/

9. Cervical Cancer Causes, Risk Factors, and Prevention - NCI [Internet]. 2022 [cited 2025 Jul 14]. Available from: https://www.cancer.gov/types/cervical/causes-risk-prevention

10. Dunne EF, Park IU. HPV and HPV-Associated Diseases. Infectious Disease Clinics of North America [Internet]. 2013 Dec [cited 2025 Jul 14];27(4):765–78. Available from: https://www.sciencedirect.com/science/article/pii/S089155201300072X

11. Pimple S, Mishra G. Cancer cervix: Epidemiology and disease burden. CytoJournal [Internet]. 2022 Mar [cited 2025 Jul 14];19:21. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9063649/

12. Stokley S, Jeyarajah J, Yankey D, Cano M, Gee J, Roark J, et al. Human Papillomavirus Vaccination Coverage Among Adolescents, 2007–2013, and Postlicensure Vaccine Safety Monitoring, 2006–2014 — United States. Morbidity and Mortality Weekly Report [Internet]. 2014 Jul [cited 2025 Jul 15];63(29):620–4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5779422/

13. Forhan SE, Gottlieb SL, Sternberg MR, Xu F, Datta SD, McQuillan GM, et al. Prevalence of Sexually Transmitted Infections Among Female Adolescents Aged 14 to 19 in the United States. Pediatrics [Internet]. 2009 Dec [cited 2025 Jul 15];124(6):1505–12. Available from: https://doi.org/10.1542/peds.2009-0674

14. Sanjosé S de, Diaz M, Castellsagué X, Clifford G, Bruni L, Muñoz N, et al. Worldwide prevalence and genotype distribution of cervical human papillomavirus DNA in women with normal cytology: A meta-analysis. The Lancet Infectious Diseases [Internet]. 2007 Jul [cited 2025 Jul 14];7(7):453–9. Available from: https://www.sciencedirect.com/science/article/pii/S1473309907701585

15. Burk RD, Harari A, Chen Z. Human papillomavirus genome variants. Virology [Internet]. 2013 Oct [cited 2025 Feb 27];445(1):232–43. Available from: https://www.sciencedirect.com/science/article/pii/S0042682213004388

16. Han F, Guo X, Jiang M, Xia N, Gu Y, Li S. Structural biology of the human papillomavirus. Structure [Internet]. 2024 Nov [cited 2025 Jul 15];32(11):1877–92. Available from: https://www.sciencedirect.com/science/article/pii/S0969212624003800

17. Schiller JT, Lowy DR. Understanding and learning from the success of prophylactic human papillomavirus vaccines. Nature reviews Microbiology [Internet]. 2012 Oct [cited 2025 Jul 15];10(10):681–92. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6309166/

18. Chen XS, Garcea RL, Goldberg I, Casini G, Harrison SC. Structure of small virus-like particles assembled from the L1 protein of human papillomavirus 16. Molecular Cell. 2000 Mar;5(3):557–67.

19. Buck CB, Day PM, Trus BL. The Papillomavirus Major Capsid Protein L1. Virology [Internet]. 2013 Oct [cited 2025 Jul 15];445(0):169–74. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3783536/

20. Schiller JT, Müller M. Next generation prophylactic human papillomavirus vaccines. The Lancet Oncology. 2015 May;16(5):e217–225.

21. DiGiuseppe S, Bienkowska-Haba M, Guion LG, Sapp M. Cruising the cellular highways: How human papillomavirus travels from the surface to the nucleus. Virus Research [Internet]. 2017 Mar [cited 2025 Jul 15];231:1–9. Available from: https://www.sciencedirect.com/science/article/pii/S0168170216306967

22. McKinney CC, Hussmann KL, McBride AA. The Role of the DNA Damage Response throughout the Papillomavirus Life Cycle. Viruses [Internet]. 2015 May [cited 2025 Jul 15];7(5):2450–69. Available from: https://www.mdpi.com/1999-4915/7/5/2450

23. Moody CA, Laimins LA. Human papillomavirus oncoproteins: Pathways to transformation. Nature Reviews Cancer [Internet]. 2010 Aug [cited 2025 Jul 15];10(8):550–60. Available from: https://www.nature.com/articles/nrc2886

24. Orav M, Geimanen J, Sepp EM, Henno L, Ustav E, Ustav M. Initial amplification of the HPV18 genome proceeds via two distinct replication mechanisms. Scientific Reports [Internet]. 2015 Nov [cited 2025 Jul 15];5(1):15952. Available from: https://www.nature.com/articles/srep15952

25. Moody CA, Laimins LA. Human Papillomaviruses Activate the ATM DNA Damage Pathway for Viral Genome Amplification upon Differentiation. PLOS Pathogens [Internet]. 2009 Oct [cited 2025 Jul 15];5(10):e1000605. Available from: https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1000605

26. Flores ER, Lambert PF. Evidence for a switch in the mode of human papillomavirus type 16 DNA replication during the viral life cycle. Journal of Virology [Internet]. 1997 Oct [cited 2025 Jul 15];71(10):7167–79. Available from: https://journals.asm.org/doi/10.1128/jvi.71.10.7167-7179.1997

27. Human Papillomaviruses Preferentially Recruit DNA Repair Factors to Viral Genomes for Rapid Repair and Amplification mBio [Internet]. [cited 2025 Jul 15]. Available from: https://journals.asm.org/doi/10.1128/mbio.00064-18

28. Moody C. Mechanisms by which HPV Induces a Replication Competent Environment in Differentiating Keratinocytes. Viruses [Internet]. 2017 Sep [cited 2025 Jul 15];9(9):261. Available from: https://www.mdpi.com/1999-4915/9/9/261

29. HPV31 E7 facilitates replication by activating E2F2 transcription through its interaction with HDACs The EMBO Journal [Internet]. [cited 2025 Jul 15]. Available from: https://www.embopress.org/doi/full/10.1038/sj.emboj.7600651

30. Scheffner M, Werness BA, Huibregtse JM, Levine AJ, Howley PM. The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. Cell [Internet]. 1990 Dec [cited 2025 Jul 15];63(6):1129–36. Available from: https://www.sciencedirect.com/science/article/pii/0092867490904098

31. Scheffner M, Huibregtse JM, Vierstra RD, Howley PM. The HPV-16 E6 and E6-AP complex functions as a ubiquitin-protein ligase in the ubiquitination of p53. Cell [Internet]. 1993 Nov [cited 2025 Jul 15];75(3):495–505. Available from: https://www.sciencedirect.com/science/article/pii/0092867493903843

32. Howie HL, Katzenellenbogen RA, Galloway DA. Papillomavirus E6 proteins. Virology [Internet]. 2009 Feb [cited 2025 Jul 15];384(2):324–34. Available from: https://www.sciencedirect.com/science/article/pii/S0042682208007253

33. Senapati R, Senapati NN, Dwibedi B. Molecular mechanisms of HPV mediated neoplastic progression. Infectious Agents and Cancer [Internet]. 2016 Nov [cited 2025 Jul 15];11(1):59. Available from: https://doi.org/10.1186/s13027-016-0107-4

34. Dürst M, Kleinheinz A, Hotz M, Gissmann L. The Physical State of Human Papillomavirus Type 16 DNA in Benign and Malignant Genital Tumours. Journal of General Virology [Internet]. 1985 [cited 2025 Jul 15];66(7):1515–22. Available from: https://www.microbiologyresearch.org/content/journal/jgv/10.1099/0022-1317-66-7-1515

35. Burk RD. Human papillomavirus and the risk of cervical cancer. Hospital Practice (1995). 1999 Nov;34(12):103–111; quiz 112.

36. Loopik DL, Bentley HA, Eijgenraam MN, IntHout J, Bekkers RLM, Bentley JR. The Natural History of Cervical Intraepithelial Neoplasia Grades 1, 2, and 3: A Systematic Review and Meta-analysis. Journal of Lower Genital Tract Disease. 2021 Jul;25(3):221–31.

37. Jastreboff AM, Cymet T. Role of the human papilloma virus in the development of cervical intraepithelial neoplasia and malignancy. Postgraduate Medical Journal. 2002 Apr;78(918):225–8.

38. Quint KD, Genders RE, Koning MNC de, Borgogna C, Gariglio M, Bouwes Bavinck JN, et al. Human Beta-papillomavirus infection and keratinocyte carcinomas. The Journal of Pathology. 2015 Jan;235(2):342–54.

39. Tampa M, Mitran CI, Mitran MI, Nicolae I, Dumitru A, Matei C, et al. The Role of Beta HPV Types and HPV-Associated Inflammatory Processes in Cutaneous Squamous Cell Carcinoma. Journal of Immunology Research [Internet]. 2020 Apr [cited 2025 Jul 15];2020:5701639. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7165336/

40. Lambert PF, Münger K, Rösl F, Hasche D, Tommasino M. Beta human papillomaviruses and skin cancer. Nature [Internet]. 2020 Dec [cited 2025 Jun 11];588(7838):E20–1. Available from: https://www.nature.com/articles/s41586-020-3023-0

41. Smola S. Human Papillomaviruses and Skin Cancer. Advances in Experimental Medicine and Biology. 2020;1268:195–209.

42. Sichero L, El-Zein M, Nunes EM, Ferreira S, Franco EL, Villa LL, et al. Cervical Infection with Cutaneous Beta and Mucosal Alpha Papillomaviruses. Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology. 2017 Aug;26(8):1312–20.

43. Bernard HU, Burk RD, Chen Z, Doorslaer K van, Hausen H zur, Villiers EM de. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. Virology [Internet]. 2010 May [cited 2025 Jul 16];401(1):70–9. Available from: https://www.sciencedirect.com/science/article/pii/S0042682210001005

44. Wang R, Pan W, Jin L, Huang W, Li Y, Wu D, et al. Human papillomavirus vaccine against cervical cancer: Opportunity and challenge. Cancer Letters [Internet]. 2020 Feb [cited 2025 Jul 16];471:88–102. Available from: https://www.sciencedirect.com/science/article/pii/S0304383519306044

45. Muñoz N, Bosch FX, Sanjosé S de, Herrero R, Castellsagué X, Shah KV, et al. Epidemiologic Classification of Human Papillomavirus Types Associated with Cervical Cancer. New England Journal of Medicine [Internet]. 2003 Feb [cited 2025 Jul 16];348(6):518–27. Available from: https://www.nejm.org/doi/full/10.1056/NEJMoa021641

46. Arbyn M, Tommasino M, Depuydt C, Dillner J. Are 20 human papillomavirus types causing cervical cancer? The Journal of Pathology [Internet]. 2014 [cited 2025 Jul 16];234(4):431–5. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/path.4424

47. Hawkins MG, Winder DM, Ball SL, Vaughan K, Sonnex C, Stanley MA, et al. Detection of specific HPV subtypes responsible for the pathogenesis of condylomata acuminata. Virology Journal [Internet]. 2013 May [cited 2025 Jul 16];10(1):137. Available from: https://doi.org/10.1186/1743-422X-10-137

48. Human Papillomavirus Type 33 - an overview ScienceDirect Topics [Internet]. [cited 2025 Jul 16]. Available from: https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/human-papillomavirus-type-33/

49. Chen Z, Schiffman M, Herrero R, DeSalle R, Anastos K, Segondy M, et al. Evolution and Taxonomic Classification of Human Papillomavirus 16 (HPV16)-Related Variant Genomes: HPV31, HPV33, HPV35, HPV52, HPV58 and HPV67. PLoS ONE [Internet]. 2011 May [cited 2025 Jul 16];6(5):e20183. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3103539/

50. Clifford GM, Tenet V, Georges D, Alemany L, Pavón MA, Chen Z, et al. Human papillomavirus 16 sub-lineage dispersal and cervical cancer risk worldwide: Whole viral genome sequences from 7116 HPV16-positive women. Papillomavirus Research [Internet]. 2019 Feb [cited 2025 Jul 16];7:67–74. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6374642/

51. Clifford GM, Tenet V, Georges D, Alemany L, Pavón MA, Chen Z, et al. Human papillomavirus 16 sub-lineage dispersal and cervical cancer risk worldwide: Whole viral genome sequences from 7116 HPV16-positive women. Papillomavirus Research [Internet]. 2019 Feb [cited 2025 Jun 29];7:67–74. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6374642/

52. Mirabello L, Yeager M, Cullen M, Boland JF, Chen Z, Wentzensen N, et al. HPV16 Sublineage Associations With Histology-Specific Cancer Risk Using HPV Whole-Genome Sequences in 3200 Women. Journal of the National Cancer Institute. 2016 Sep;108(9):djw100.

53. Hampson IN, Oliver AW. Update on Effects of the Prophylactic HPV Vaccines on HPV Type Prevalence and Cervical Pathology. Viruses [Internet]. 2024 Aug [cited 2025 Jul 19];16(8):1245. Available from: https://www.mdpi.com/1999-4915/16/8/1245

54. Arbyn M, Xu L, Simoens C, Martin-Hirsch PP. Prophylactic vaccination against human papillomaviruses to prevent cervical cancer and its precursors. The Cochrane Database of Systematic Reviews. 2018 May;5(5):CD009069.

55. Drolet M, Bénard É, Pérez N, Brisson M. Population-level impact and herd effects following the introduction of human papillomavirus vaccination programmes: Updated systematic review and meta-analysis. Lancet (London, England) [Internet]. 2019 Aug [cited 2025 Jul 19];394(10197):497–509. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7316527/

56. Hampson IN. Effects of the Prophylactic HPV Vaccines on HPV Type Prevalence and Cervical Pathology. Viruses. 2022 Apr;14(4):757.

57. Du J, Ährlund-Richter A, Näsman A, Dalianis T. Human papilloma virus (HPV) prevalence upon HPV vaccination in Swedish youth: A review based on our findings 2008-2018, and perspectives on cancer prevention. Archives of Gynecology and Obstetrics. 2021 Feb;303(2):329–35.

58. Powell N, Cuschieri K, Cubie H, Hibbitts S, Rosillon D, De Souza SC, et al. Cervical cancers associated with human papillomavirus types 16, 18 and 45 are diagnosed in younger women than cancers associated with other types: A cross-sectional observational study in Wales and Scotland (UK). Journal of Clinical Virology [Internet]. 2013 Nov [cited 2025 Jul 19];58(3):571–4. Available from: https://www.sciencedirect.com/science/article/pii/S1386653213003442

59. Analysis highlights very low level of HPV vaccine uptake globally CIDRAP [Internet]. 2025 [cited 2025 Jul 19]. Available from: https://www.cidrap.umn.edu/human-papillomavirus-hpv/analysis-highlights-very-low-level-hpv-vaccine-uptake-globally

60. Allali M, El Fermi R, Errafii K, Abdelaziz W, Al Idrissi N, Fichtali K, et al. HPV genotypes in Africa: Comprehensive analysis of genetic diversity and evolutionary dynamics. Archives of Virology. 2025 Apr;170(6):116.

61. Huyghe E, Abrams S, Bogers JP, Verhoeven V, Benoy I. Evolution of human papilloma virus prevalence in a highly vaccinated region in Belgium: A retrospective cohort study in Flemish women (2010-2019). European journal of cancer prevention: the official journal of the European Cancer Prevention Organisation (ECP). 2023 Jan;32(1):48–56.

62. Williams J, Kostiuk M, Biron VL. Molecular Detection Methods in HPV-Related Cancers. Frontiers in Oncology [Internet]. 2022 Apr [cited 2025 Jul 20];12:864820. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9092940/

63. Pagliusi SR, Garland SM. International Standard Reagents for HPV Detection. Disease Markers [Internet]. 2007 [cited 2025 Jul 20];23(4):591826. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1155/2007/591826

64. Ronco G, Dillner J, Elfström KM, Tunesi S, Snijders PJF, Arbyn M, et al. Efficacy of HPV-based screening for prevention of invasive cervical cancer: Follow-up of four European randomised controlled trials. Lancet (London, England). 2014 Feb;383(9916):524–32.

65. Waters DLE, Shapter FM. The polymerase chain reaction (PCR): General methods. Methods in Molecular Biology (Clifton, NJ). 2014;1099:65–75.

66. Westra WH. Detection of Human Papillomavirus in Clinical Samples. Otolaryngologic Clinics of North America [Internet]. 2012 Aug [cited 2025 Jul 20];45(4):765–77. Available from: https://www.sciencedirect.com/science/article/pii/S0030666512000370

67. Malet I, Draa I, Leducq V, Vuong F, Bonnafous P, Marcelin AG, et al. An amplicon-based approach for full-genome characterization of HPV16. Microbiology Spectrum. 2025 Jul;13(7):e0307324.

68. Cui M, Chan N, Liu M, Thai K, Malaczynska J, Singh I, et al. Clinical Performance of Roche Cobas 4800 HPV Test. Journal of Clinical Microbiology [Internet]. 2014 Jun [cited 2025 Jul 20];52(6):2210–1. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4042746/

69. Tjalma WaA, Depuydt CE. Cervical cancer screening: Which HPV test should be used–L1 or E6/E7? European Journal of Obstetrics, Gynecology, and Reproductive Biology. 2013 Sep;170(1):45–6.

70. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. Proceedings of the National Academy of Sciences of the United States of America. 2002 Apr;99(8):5261–6.

71. Mai M, Hoyer JD, McClure RF. Use of multiple displacement amplification to amplify genomic DNA before sequencing of the α and β haemoglobin genes. Journal of Clinical Pathology [Internet]. 2004 Jun [cited 2025 Jul 20];57(6):637–40. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1770312/

72. Johne R, Müller H, Rector A, Ranst M van, Stevens H. Rolling-circle amplification of viral DNA genomes using phi29 polymerase. Trends in Microbiology. 2009 May;17(5):205–11.

73. Agyabeng-Dadzie F, Beaudry MS, Deyanov A, Slanis H, Duong MQ, Turner R, et al. Evaluating the Benefits and Limits of Multiple Displacement Amplification With Whole-Genome Oxford Nanopore Sequencing. Molecular Ecology Resources [Internet]. 2025 [cited 2025 Jul 20];25(6):e14094. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.14094

74. Mardis ER. Next-generation sequencing platforms. Annual Review of Analytical Chemistry (Palo Alto, Calif). 2013;6:287–303.

75. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. Genome Biology. 2016 Nov;17(1):239.

76. Rang FJ, Kloosterman WP, Ridder J de. From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. Genome Biology. 2018 Jul;19(1):90.

77. Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. Nature Biotechnology. 2021 Apr;39(4):442–50.

78. Wang JCK, Baddock HT, Mafi A, Foe IT, Bratkowski M, Lin TY, et al. Structure of the p53 degradation complex from HPV16. Nature Communications. 2024 Feb;15(1):1842.

79. Martinez-Zapien D, Ruiz FX, Poirson J, Mitschler A, Ramirez J, Forster A, et al. Structure of the E6/E6AP/p53 complex required for HPV-mediated degradation of p53. Nature. 2016 Jan;529(7587):541–5.

80. Jansma AL, Martinez-Yamout MA, Liao R, Sun P, Dyson HJ, Wright PE. The high-risk HPV16 E7 oncoprotein mediates interaction between the transcriptional coactivator CBP and the retinoblastoma protein pRb. Journal of molecular biology [Internet]. 2014 Dec [cited 2025 Jul 21];426(24):4030–48. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4258470/

81. Liu X, Clements A, Zhao K, Marmorstein R. Structure of the human Papillomavirus E7 oncoprotein and its mechanism for inactivation of the retinoblastoma tumor suppressor. The Journal of Biological Chemistry. 2006 Jan;281(1):578–86.

82. Artesi M, Hahaut V, Cole B, Lambrechts L, Ashrafi F, Marçais A, et al. PCIP-seq: Simultaneous sequencing of integrated viral genomes and their insertion sites with long reads. Genome Biology [Internet]. 2021 Apr [cited 2025 Jul 6];22:97. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8025556/

83. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature Biotechnology [Internet]. 2011 Jan [cited 2025 Jul 27];29(1):24–6. Available from: https://www.nature.com/articles/nbt.1754

84. Pace J, Youens-Clark K, Freeman C, Hurwitz B, Van Doorslaer K. PuMA: A papillomavirus genome annotation tool. Virus Evolution [Internet]. 2020 Jul [cited 2025 Jul 23];6(2):veaa068. Available from: https://doi.org/10.1093/ve/veaa068

85. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. BMC bioinformatics. 2009 Dec;10:421.

86. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32(5):1792–7.

87. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., & Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, *37*(Web Server), W202–W208. https://doi.org/10.1093/nar/gkp335

88. Bishop B, Dasgupta J, Klein M, Garcea RL, Christensen ND, Zhao R, et al. Crystal structures of four types of human papillomavirus l1 capsid proteins: Understanding the specificity of neutralizing monoclonal antibodies. J Biol Chem. 2007 Oct 26;282(43):31803–11.

89. Tjalma WaA, Depuydt CE. Cervical cancer screening: Which HPV test should be used–l1 or e6/e7? Eur J Obstet Gynecol Reprod Biol. 2013 Sep;170(1):45–6.

90. Castle PE, Stoler MH, Wright TC, Sharma A, Wright TL, Behrens CM. Performance of carcinogenic human papillomavirus (HPV) testing and HPV16 or HPV18 genotyping for cervical cancer screening of women aged 25 years and older: A subanalysis of the ATHENA study. Lancet Oncol. 2011 Sep;12(9):880–90.

91. VSEARCH: A versatile open source tool for metagenomics [PeerJ] [Internet]. [cited 2025 Jul 23]. Available from: https://peerj.com/articles/2584/

92. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution [Internet]. 2013 Apr 1 [cited 2025 Jul 27];30(4):772–80. Available from: https://doi.org/10.1093/molbev/mst010

93. Nguyen LT, Schmidt HA, Haeseler A von, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution [Internet]. 2015 Jan 1 [cited 2025 Jul 27];32(1):268–74. Available from: https://doi.org/10.1093/molbev/msu300

94. Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics [Internet]. 2006 Jul 1 [cited 2025 Jul 27];22(13):1658–9. Available from: https://doi.org/10.1093/bioinformatics/btl158

95. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature [Internet]. 2021 Aug [cited 2025 Jul 27];596(7873):583–9. Available from: https://www.nature.com/articles/s41586-021-03819-2

96. Xue LC, Rodrigues JP, Kastritis PL, Bonvin AM, Vangone A. PRODIGY: A web server for predicting the binding affinity of protein–protein complexes. Bioinformatics [Internet]. 2016 Dec 1 [cited 2025 Jul 27];32(23):3676–8. Available from: https://doi.org/10.1093/bioinformatics/btw514

97. Lilkova, e., et al. (2015) the PyMOL molecular graphics system, version 2.0 schrodinger, LLC. - references - scientific research publishing [Internet]. [cited 2025 Jul 27]. Available from: https://www.scirp.org/reference/referencespapers?referenceid=2403147

98. Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in r language. Bioinformatics. 2004 Jan 22;20(2):289–90.

99. Mutombo AB, Benoy I, Tozin R, Bogers J, Van geertruyden JP, Jacquemyn Y. Prevalence and distribution of human papillomavirus genotypes among women in kinshasa, the democratic republic of the congo. JGO [Internet]. 2019 Jul [cited 2025 Aug 13];(5):1–9. Available from: https://ascopubs.org/doi/10.1200/JGO.19.00110

100. Tsimba Lemba PC, Boumba LMA, Péré H, Nganga PC, Veyer D, Puech J, et al. Human papillomavirus genotype distribution by cytological status and associated risk factors in the general population of congolese women living in urban and rural areas: Implications for cervical cancer prevention. Infectious Diseases Now [Internet]. 2023 Oct 1 [cited 2025 Aug 13];53(8):104762. Available from: https://www.sciencedirect.com/science/article/pii/S2666991923001240

101. Pinheiro M, Harari A, Schiffman M, Clifford GM, Chen Z, Yeager M, et al. Phylogenomic analysis of human papillomavirus type 31 and cervical carcinogenesis: A study of 2093 viral genomes. Viruses [Internet]. 2021 Sep 28 [cited 2025 Aug 13];13(10):1948. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8540939/

102. Shabanpour M, Jalali-Alhosseini P, Shoja Z, Ghafoori-Ghahdarijani F, Taherkhani S, Jalilvand S. Lineage and sublineage analysis of human papillomavirus type 58 in iranian women. Virol J [Internet]. 2024 Oct 3 [cited 2025 Aug 13];21:244. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11451209/

103. Serretiello E, Corrado F, Santella B, Chianese A, Iervolino D, Coppola A, et al. Prevalence and distribution of high- and low- risk HPV genotypes in women living in the metropolitan area of naples: A recent update. Asian Pac J Cancer Prev. 2023 Feb 1;24(2):435–41.

104. Clifford GM, Tenet V, Georges D, Alemany L, Pavón MA, Chen Z, et al. Human papillomavirus 16 sub-lineage dispersal and cervical cancer risk worldwide: Whole viral genome sequences from 7116 HPV16-positive women. Papillomavirus Res [Internet]. 2019

Feb 6 [cited 2025 Jun 29];7:67–74. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6374642/

105. Chen AA, Gheit T, Franceschi S, Tommasino M, Clifford GM. Human papillomavirus 18 genetic variation and cervical cancer risk worldwide. J Virol [Internet]. 2015 Aug 12 [cited 2025 Aug 13];89(20):10680–7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4580183/

106. Farahmand M, Moghoofei M, Dorost A, Abbasi S, Monavari SH, Kiani SJ, et al. Prevalence and genotype distribution of genital human papillomavirus infection in female sex workers in the world: A systematic review and meta-analysis. BMC Public Health. 2020 Sep 25;20(1):1455.

107. Sichero L, El-Zein M, Nunes EM, Ferreira S, Franco EL, Villa LL, et al. Cervical infection with cutaneous beta and mucosal alpha papillomaviruses. Cancer Epidemiol Biomarkers Prev. 2017 Aug;26(8):1312–20.

108. Hoffmann K, Bouckaert R, Greenhill SJ, Kühnert D. Bayesian phylogenetic analysis of linguistic data using BEAST. Journal of Language Evolution [Internet]. 2021 Jul 1 [cited 2025 Jun 29];6(2):119–35. Available from: https://doi.org/10.1093/jole/lzab005

109. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol [Internet]. 2018 Jun 8 [cited 2025 Jun 29];4(1):vey016. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6007674/

110. Yegros-Yegros A, Klippe W van de, Abad-Garcia MF, Rafols I. Exploring why global health needs are unmet by research efforts: The potential influences of geography, industry and publication incentives. Health Research Policy and Systems [Internet]. 2020 May 15 [cited 2025 Aug 14];18(1):47. Available from: https://doi.org/10.1186/s12961-020-00560-6

111. Mazure CM, Jones DP. Twenty years and still counting: Including women as participants and studying sex and gender in biomedical research. BMC Women's Health [Internet]. 2015 Oct 26 [cited 2025 Jun 27];15(1):94. Available from: https://doi.org/10.1186/s12905-015-0251-9

112. Mirin AA. Gender disparity in the funding of diseases by the u.s. National institutes of health. J Womens Health (Larchmt). 2021 Jul;30(7):956–63.

113. De Sanjosé, S., Diaz, M., Castellsagué, X., Clifford, G., Bruni, L., Muñoz, N., & Bosch, F. X. (2007). Worldwide prevalence and genotype distribution of cervical human papillomavirus DNA in women with normal cytology: a meta-analysis. *The Lancet Infectious Diseases*, *7*(7), 453–459. https://doi.org/10.1016/s1473-3099(07)70158-5

114. *The health Professional's HPV handbook | Human Papillomavirus and CERV*. (2006, January 13). Taylor & Francis.

https://www.taylorfrancis.com/books/edit/10.1201/9780367806170/health-professional-hpv-handbook-walter-prendiville

115: Mac, M., & Moody, C. A. (2020). Epigenetic regulation of the human papillomavirus life cycle. *Pathogens*, *9*(6), 483. https://doi.org/10.3390/pathogens9060483

116: Lambert, P. F., Münger, K., Rösl, F., Hasche, D., & Tommasino, M. (2020). Beta human papillomaviruses and skin cancer. *Nature*, *588*(7838), E20–E21. https://doi.org/10.1038/s41586-020-3023-0

117: Wang, J. C. K., Baddock, H. T., Mafi, A., Foe, I. T., Bratkowski, M., Lin, T., Jensvold, Z. D., López, M. P., Stokoe, D., Eaton, D., Hao, Q., & Nile, A. H. (2024). Structure of the p53 degradation complex from HPV16. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-024-45920-w

118: Dommer, J., Van Dooslaer, K., Afrasiabi, C., Browne, K., Ezeji, S., Kim, L., Dolan, M., & McBride, A. A. (2024). PaVE 2.0: behind the scenes of the Papillomavirus Episteme. *Journal of Molecular Biology*, 168925. https://doi.org/10.1016/j.jmb.2024.168925

119: Moreno, G. K., Braun, K. M., Riemersma, K. K., Martin, M. A., Halfmann, P. J., Crooks, C. M., Prall, T., Baker, D., Baczenas, J. J., Heffron, A. S., Ramuta, M., Khubbar, M., Weiler, A. M., Accola, M. A., Rehrauer, W. M., O'Connor, S. L., Safdar, N., Pepperell, C. S., Dasu, T., . . . Friedrich, T. C. (2020). Revealing fine-scale spatiotemporal differences in SARS-CoV-2 introduction and spread. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19346-z

120: Ardisson, J. S., Sagrillo, M. V. M., Athaydes, B. R., Vargas, A. M. C., Torezani, R., Ribeiro-Rodrigues, R., Spano, L. C., Paneto, G. G., Delatorre, E., Von Zeidler, S. V., & Filho, T. F. B. (2025). Comparative spatial–temporal analysis of SARS-CoV-2 lineages B.1.1.33 and BQ.1.1 Omicron variant across pandemic phases. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-95140-5

Appendix

IARC Group	HPV Types
Group 1 (Carcinogenic) – High-Risk	HPV 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59
Group 2A (Probably Carcinogenic)	HPV 68
Group 2B (Possibly Carcinogenic)	HPV 26, 30, 34, 53, 66, 67, 69, 70, 73, 82, 85, 97
Low-Risk (Benign Lesions)	HPV 6, 11, 40, 42, 43, 44, 54, 61, 70, 72, 81, CP6108

Table A1. Classification of human papillomavirus (HPV) types according to the International Agency for Research on Cancer (IARC). HPV types are categorized based on their oncogenic potential as follows: Group 1 includes high-risk types that are carcinogenic to humans; Group 2A includes types that are probably carcinogenic; Group 2B includes types that are possibly carcinogenic; and Low-Risk (LR) types are generally associated with benign lesions such as anogenital warts.

Source: Adapted from IARC Monographs Volume 100B (2012), Zhao, Hu, & Qiao, 2020.

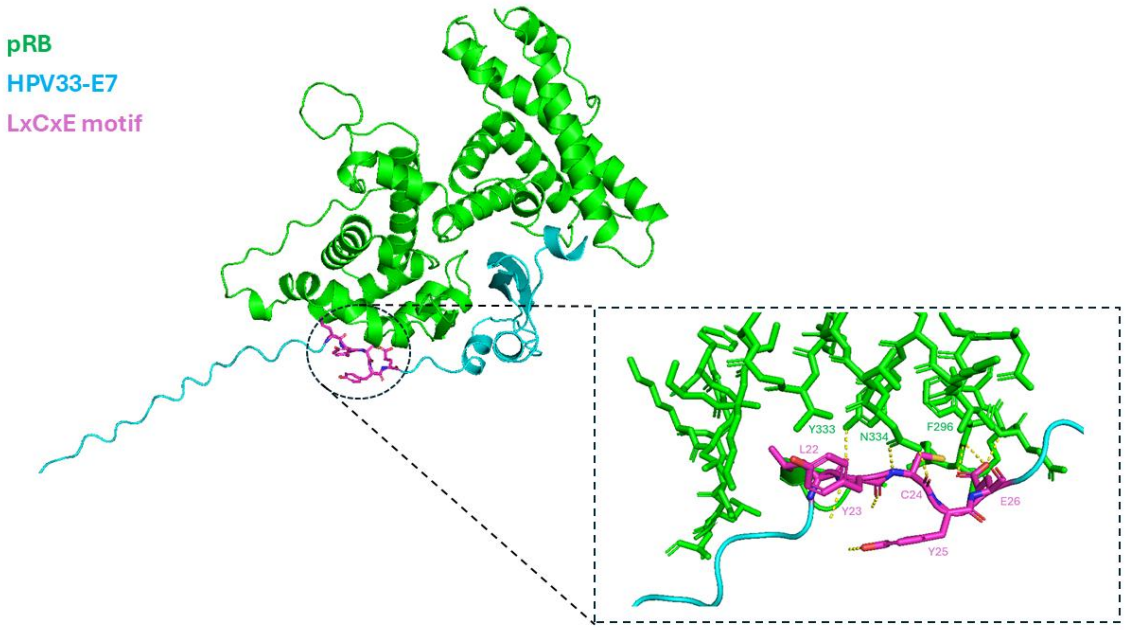

Figure S1. Model of the interaction between HPV33 E7 protein and the B-pocket domain of the retinoblastoma protein (pRB). The structure of the complex was predicted using the AlphaFold-Multimer server. The pRB protein is shown in green, the HPV33 E7 protein in cyan, and the full LxCxE motif (residues 22–26: LYCYE) is highlighted in magenta. A close-up view illustrates predicted polar contacts identified using PyMOL (cutoff: 3.6 Å): L22 interacts with Y333, C24 with N334, and E26 with F296 on pRB. These interactions suggest a conserved anchoring of the LxCxE motif into the B-pocket, consistent with previously described mechanisms in high-risk HPV types.

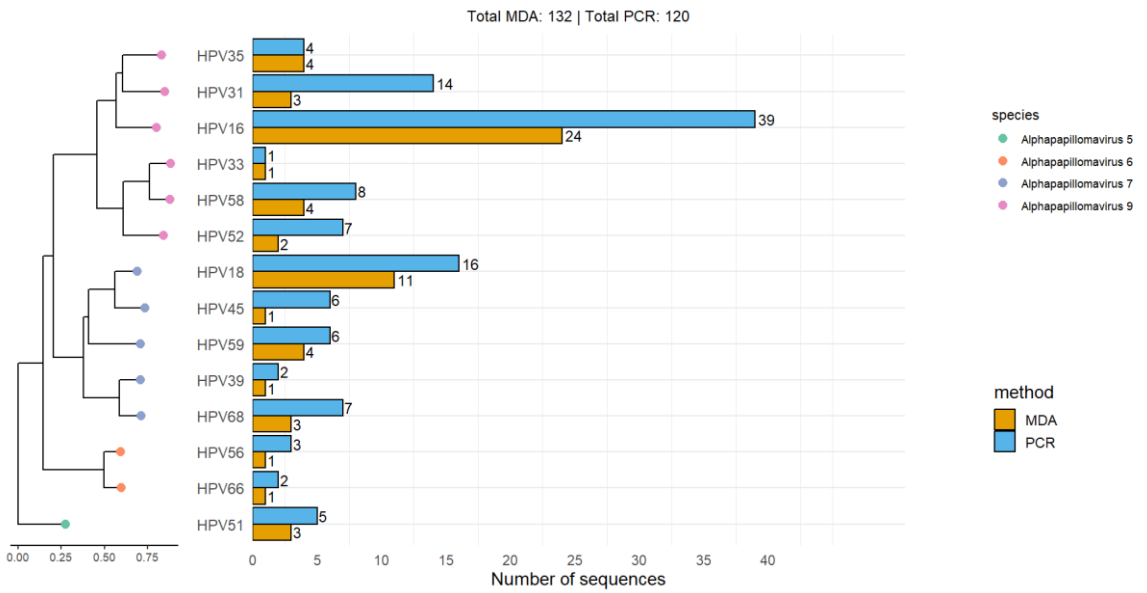

Figure S2. Distribution and phylogeny of high-risk HPV types identified in Liège. This figure combines two elements: (1) a bar plot showing the distribution of high-risk HPV types detected in Liège samples, and (2) a phylogenetic tree of the associated with hrHPV types. In the tree, HPV types are color-coded according to their species classification, as indicated in the legend. Together, these visualizations illustrate both the prevalence and evolutionary relationships of the circulating hrHPV types in the region.

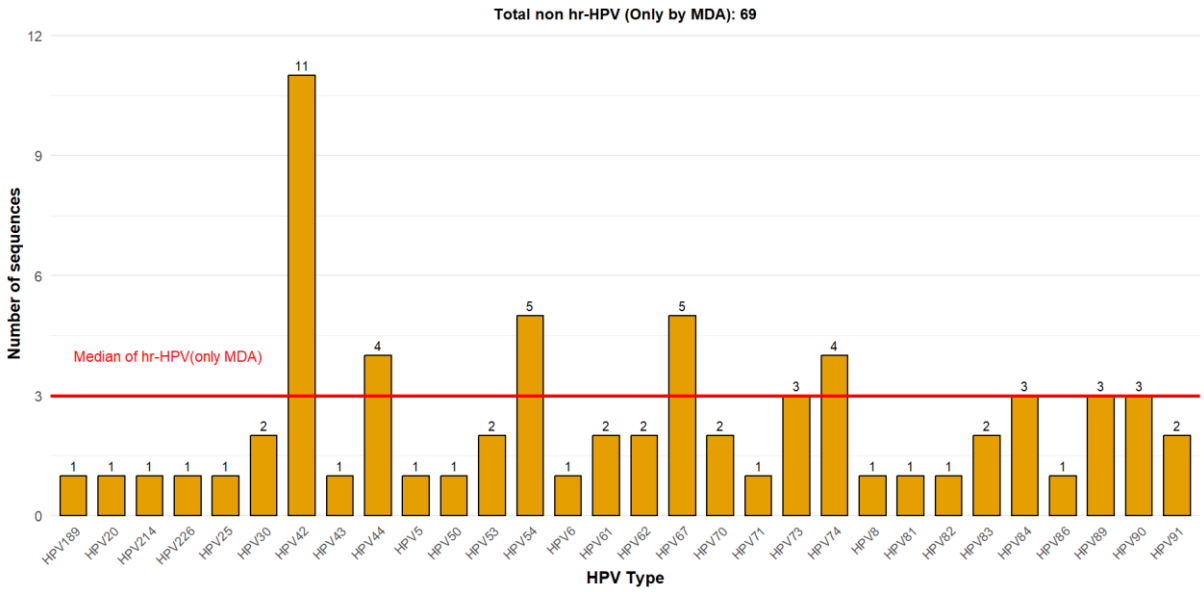

Figure S3 : Distribution of non hr-HPV types (only captured by MDA) in Liege. The red line shows the median of hr-HPV counts captured by MDA.

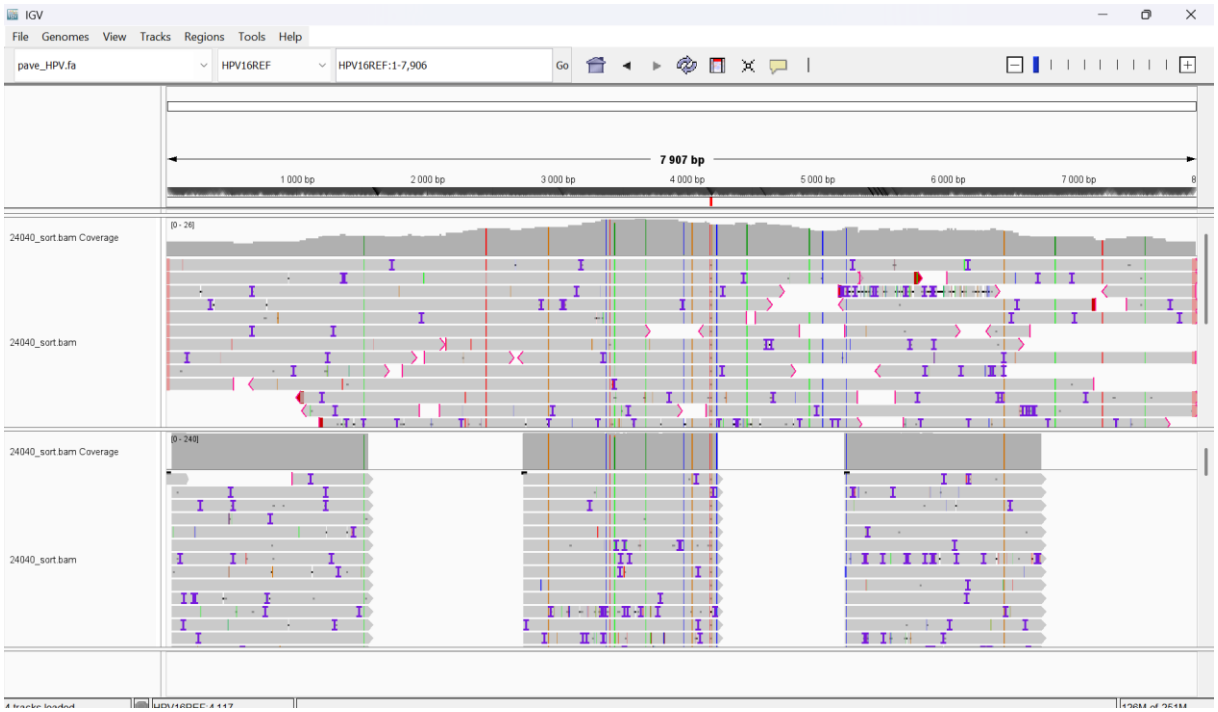

Figure S4: Example where one of the two PCR pools was missed, resulting in gaps in the read distribution where coverage is absent.

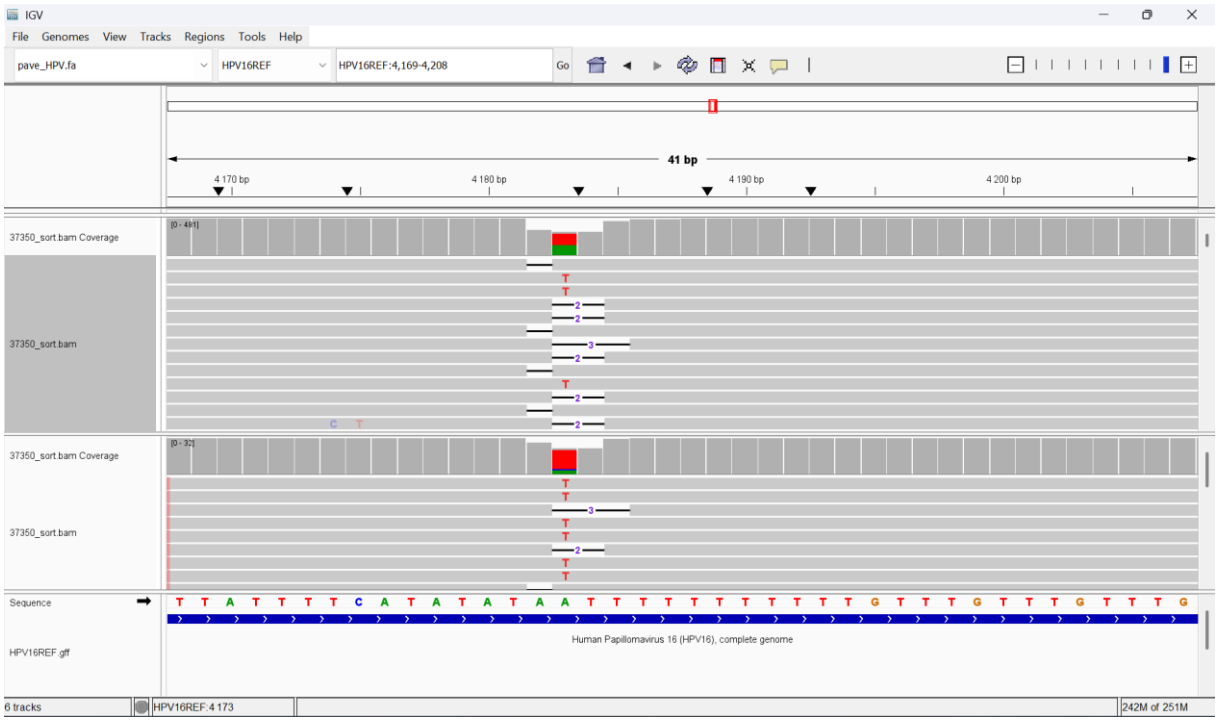

Figure S5: Example of an indel in an intergenic region, frequently occurring in homopolymeric stretches.

Samples	Observed issue
HPV16_24040	one PCR pool failed and low coverage MDA
HPV16_24948	different base at position 4184 but higher coverage in PCR
HPV16_32421	indel position 4195
HPV16_33203	indel position 4183
HPV16_34267	N in MDA
HPV16_35419	N in MDA
HPV16_37350	indel position 4195
HPV16_38062	indel
HPV16_39963	indel position 4195
HPV16_40071	N in MDA
HPV16_40083	indel position 4195
HPV16_80009	N in MDA
HPV16_80272	different base at position 5899 (MDA 10x more coverage)
HPV18_23311	N in MDA
HPV18_23442	N in MDA
HPV18_36188	one PCR pool failed
HPV18_80099	N in MDA
HPV18_80317	N in MDA
HPV31_32089	N in MDA
HPV31_36510	one PCR pool failed
HPV33_80211	indel
HPV33_80639	indel
HPV35_24360	one PCR pool failed
HPV35_24948	indel position 14-15
HPV35_26587	indel position 14-15
HPV35_80087	indel position 14-15
HPV35_80120	indel position 14-15
HPV35_80140	indel position 14-15
HPV35_80159	N in MDA
HPV35_80263	N in MDA
HPV35_80453	indel position 14-15
HPV35_80509	indel position 14-15
HPV35_80648	indel position 14-15
HPV51_80005	indel
HPV51_80070	indel
HPV51_80189	one PCR pool failed
HPV52_26777	N in MDA

HPV58_24360	one PCR pool failed
HPV58_80140	one PCR pool failed
HPV58_80235	N in MDA
HPV58_80287	N in PCR
HPV58_80292	N in MDA
HPV58_80648	N in MDA
HPV59_80189	one PCR pool failed
HPV66_80005	one PCR pool failed
HPV66_80109	one PCR pool failed
HPV66_80189	one PCR pool failed
HPV66_80209	one PCR pool failed
HPV66_40820	N in MDA
HPV68_24360	one PCR pool failed
HPV68_35419	N in MDA
HPV68_80027	N in MDA
HPV68_80189	2 different bases but higher coverage in MDA
HPV68_80230	N in MDA
HPV68_80313	N in MDA

Table A2. Summary of discrepancies between HPV genome sequences obtained using multiple displacement amplification (MDA) and PCR-based amplification, sorted by HPV type. For each sample, the observed issue is reported, including insertions/deletions (indels), nucleotide mismatches at specific positions, ambiguous nucleotides ("N") in one amplification method, failure of a PCR pool, or differences in coverage depth between methods.

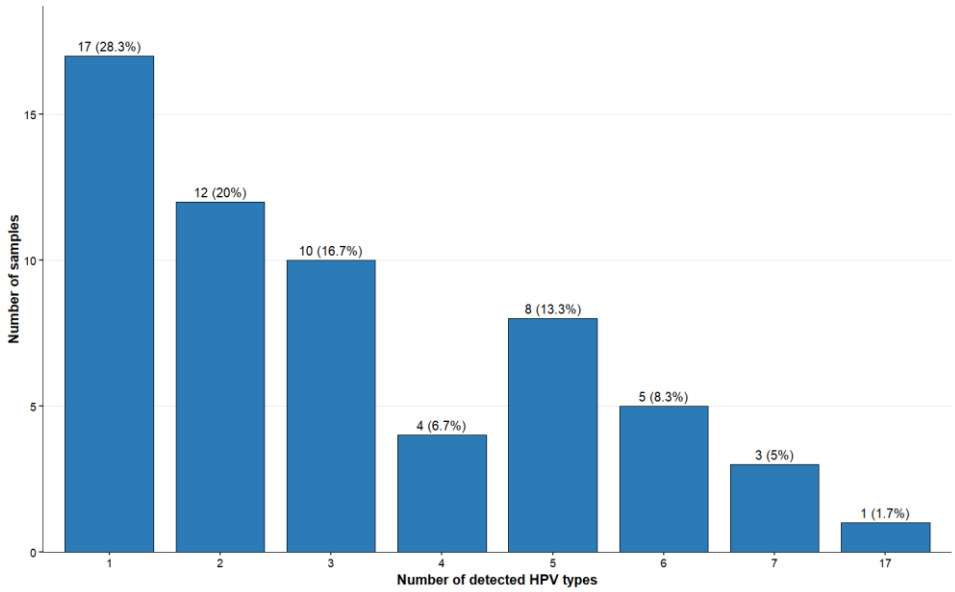

Figure S6: Distribution of the number of HPV types per sample in Liege, Belgium. The x-axis represents the number of distinct HPV types detected in a sample, while the y-axis indicates how many samples fall into each category. This highlights the frequency of co-infections across the Liege data.

SAMPLES/PROT	E6	E7	L1	L2	E1
25162-196l	118,195,124,152,196,195				25,8,5,143,152,124
25162-175l	175,223,172,156				175,201,197,166
80131-223l	223,172				
27492-195l	118,195,124,152,196				25,8,5,143,152,124
36510-226l	/	226,101, 214,108,103			
80189-103l	/	103,108,214			
V7-222l	222,162,166,161				138,203,123
V8-168l	168,112,164,147				157,131,187,176,119

Table A3. Closest phylogenetic neighbors for each newly identified HPV type across five individual ORF-based trees constructed from amino acid sequences (E6, E7, L1, L2, E1). Values indicate the HPV types showing the closest clustering with the study-derived sample in each ORF-specific phylogeny. A slash ("/") denotes the absence of the corresponding ORF in the assembled genome.

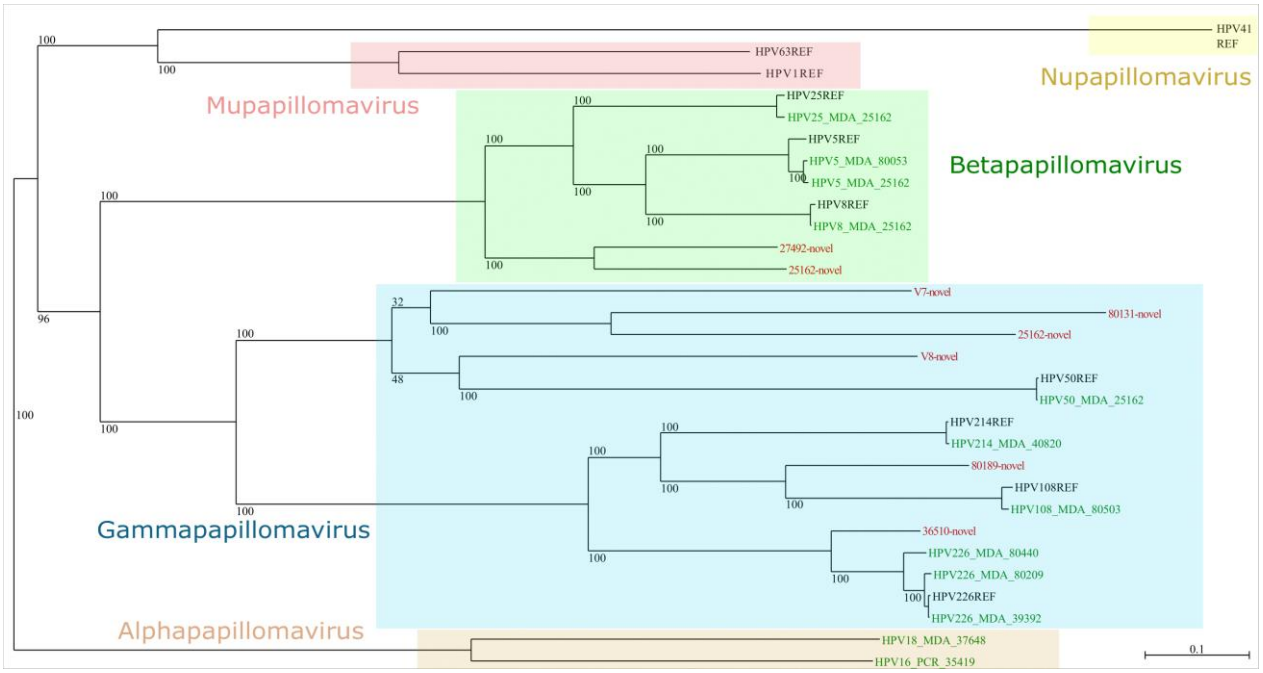

Figure S7: Maximum-likelihood phylogenetic tree of papillomaviruses, rooted on the *Alphapapillomavirus* clade. Novel HPV types identified in this study are highlighted in red, other study-derived sequences in green, and reference sequences from PaVE in black. The tree illustrates the separation of the five genera represented here (*Betapapillomavirus*, *Gammapapillomavirus*, *Mupapillomavirus*, and *Nupapillomavirus*, with *Alphapapillomavirus* used as the outgroup). Notably, the novel types cluster within the Beta and Gamma genera, strongly suggesting that they represent additional members of these lineages.

Sequence	Match	Type	Lineage	Sub-lineage	Similarity	Length Diff	Results	PAVE match
HPV70_MDA_32089	KC470285.1	Same type	Same lineage	Same sublineage	99,9	0	Perfect or near-perfect match	HPV70-A-A1
HPV70_MDA_33783	KC470285.1	Same type	Same lineage	Same sublineage	99,9	0	Perfect or near-perfect match	HPV70-A-A1
HPV70_MDA_80001	KC470291.1	Same type	Same lineage	Same sublineage	99,9	0	Perfect or near-perfect match	HPV70-B-ND
HPV70_MDA_80181	OP711991.1	Same type	Same lineage	Same sublineage	99,9	0	Perfect or near-perfect match	HPV70-B-B1
HPV70_MDA_80440	OP711991.1	Same type	Same lineage	Same sublineage	99,9	0	Perfect or near-perfect match	HPV70-B-B1
HPV71_MDA_32089	MT783419.1	Same type	Same lineage	Same sublineage	99,9	0	Perfect or near-perfect match	HPV71-A-A1
HPV72_MDA_80256	OP712107.1	Same type	Same lineage	Same sublineage	99,7	0	Perfect or near-perfect match	HPV72-A-ND
HPV72_MDA_80501	OP712107.1	Same type	Different lineage	Different sublineage	98,5	5	Minor divergent - review recommended	HPV72-A-ND
HPV73_MDA_25161	LR861989.1	Same type	Same lineage	Same sublineage	100	0	Perfect or near-perfect match	HPV73-A-A2
HPV73_MDA_26587	KF436827.1	Same type	Same lineage	Same sublineage	99,9	0	Perfect or near-perfect match	HPV73-A-A2
HPV73_MDA_40058	KF436828.1	Same type	Same lineage	Same sublineage	100	0	Perfect or near-perfect match	HPV73-A-A2

Figure S8. Classification output using GenBank as the reference database (complementary view 1). In this example, sequences assigned to HPV70 show high similarity to GenBank entries, and one of them maps to an undefined sublineage (HPV70-B-ND), suggesting a potential novel variant. For HPV71 and HPV73, all sequences show high similarity to existing GenBank entries and are consistently classified via PAVE. For HPV72,

the situation is more complex: while both sequences find GenBank matches, one aligns well (>99.5%), while the other falls below this threshold (~98.5%), indicating a possible lineage mismatch. Importantly, neither of the corresponding GenBank references map to a defined sublineage in PAVE, suggesting that these sequences may represent uncharacterized or novel sublineages within HPV72.

Sequence	Type	Lineage	Sub-lineage	Similarity	Length Difference	Results
HPV70_MDA_32089	HPV70	A	A-1	99,8	0	Same sublineage
HPV70_MDA_33783	HPV70	A	A-1	99,8	0	Same sublineage
HPV70_MDA_80001	HPV70	B	B-1	99,6	0	Same sublineage
HPV70_MDA_80181	HPV70	B	B-1	99,7	0	Same sublineage
HPV70_MDA_80440	HPV70	B	B-1	99,7	0	Same sublineage
HPV71_MDA_32089	HPV71	A	A-1	99,7	2	Same sublineage
HPV72_MDA_80256	HPV72	A	ND	99,5	1	Probably new sublineage
HPV72_MDA_80501	HPV72	ND	ND	98,5	6	Probably new lineage - Minor divergent
HPV73_MDA_25161	HPV73	A	A-2	99,9	0	Same sublineage
HPV73_MDA_26587	HPV73	A	A-2	99,8	0	Same sublineage
HPV73_MDA_40058	HPV73	A	A-2	99,9	0	Same sublineage

Figure S9: Classification output using PAVE as the reference database (complementary view 2). This table presents the classification results of selected sequences using PAVE as the reference database. All sequences from HPV70, HPV71, and HPV73 could be clearly classified within existing lineages and sublineages. However, two sequences assigned to HPV72 could not be fully resolved: one corresponds to a putative new *lineage*, and the other to a potential *new sublineage* not currently defined in PAVE. For the latter (HPV72_MDA_80501), despite the absence of sublineage assignment in PAVE, the GenBank comparison reveals a strong match to another sequence classified similarly, supporting its biological relevance and indicating that it likely represents a real, yet uncharacterized, sublineage rather than an artefactual variant. This example highlights the need to use both PAVE and GenBank in combination for robust classification, particularly in the detection of novel diversity.

Type	Lineage	Sub-lineage	X	Type	Lineage	Sub-lineage
HPV35 (12,26%)	A (100%)	A2 (100%)	X	HPV16 (22,07%)	A (93,75%)	A1 (90%)
HPV52 (9,91%)	A (71,43%)	A1 (93,33%)	X			A2 (3,33%)
		A2 (6,67%)	X			A3 (6,67%)
	B (19,05%)	B1 (100%)	X		C (6,25%)	C1 (100%)
	D (4,76%)	D1 (100%)	X	HPV18 (9,66%)	A (85,71%)	A3 (66,67%)
	E (4,76%)	E1 (100%)	X			A4 (16,67%)
HPV31 (8,02%)	C (100%)	C1 (70,59%)	X			A5 (16,67%)

		C2 (29,41%)	X		B (14,29%)	B1 (50%)
HPV58 (7,08%)	B (60%)	B2 (100%)	X			B2 (50%)
	C (13,33%)	C1 (100%)	X	HPV42 (7,59%)	A (100%)	A1 (100%)
	D (26,67%)	D2 (100%)	X	HPV31 (4,83%)	A (14,29%)	A1 (100%)
HPV56 (5,19%)	A (81,82%)	A2 (100%)	X		B (28,57%)	B2 (100%)
	B (18,18%)	B1 (100%)	X		C (57,14%)	C3 (100%)
HPV68 (5,19%)	A (9,09%)	A1 (100%)	X	HPV58 (3,45%)	A (80%)	A1 (25%)
	B (18,18%)	B1 (100%)	X			A2 (75%)
	D (9,09%)	D1 (100%)	X		B (20%)	B1 (100%)
	F (63,64%)	F2 (100%)	X	HPV67 (3,45%)	A (20%)	A1 (100%)
HPV74 (4,72%)	ND (100%)	ND (100%)	X		B (80%)	B1 (75%)
HPV18 (4,25%)	B (100%)	B2 (100%)	X			B-ND (25%)
HPV45 (4,25%)	A (88,89%)	A1 (87,5%)	X	HPV35 (2,76%)	A (100%)	A1 (50%)
		A-ND (12,5%)	X			A2 (50%)
	B (11,11%)	B2 (100%)	X	HPV44 (2,76%)	A (75%)	A1 (100%)
HPV33 (3,77%)	A (50%)	A1 (100%)	X		ND (25%)	ND (100%)
	B (50%)	B1 (100%)	X	HPV51 (2,76%)	A (100%)	A1 (100%)
HPV59 (3,77%)	B (100%)	B1 (100%)	X	HPV54 (2,76%)	A (100%)	A1 (25%)
HPV16 (3,3%)	A (14,29%)	A1 (100%)	X			A2 (75%)
	C (85,71%)	C1 (100%)	X	HPV59 (2,76%)	B (100%)	B1 (100%)
HPV51 (3,3%)	B (100%)	B1 (71,43%)	X	HPV56 (2,07%)	B (100%)	B1 (100%)
		B2 (28,57%)	X	HPV73 (2,07%)	A (100%)	A2 (100%)
HPV39 (1,89%)	A (50%)	A2 (100%)	X	HPV74 (2,07%)	A (33,33%)	A1 (100%)
	B (50%)	B1 (100%)	X		ND (66,67%)	ND (100%)
HPV44 (1,89%)	A (50%)	A1 (100%)	X	HPV84 (2,07%)	A (66,67%)	A1 (100%)
	ND (50%)	ND (100%)	X		ND (33,33%)	ND (100%)
HPV66 (1,89%)	B (25%)	B1 (100%)	X	HPV90 (2,07%)	A (100%)	A1 (33,33%)
	ND (75%)	ND (100%)	X			A-ND (66,67%)
HPV30 (1,42%)	A (33,33%)	A5 (100%)	X	HPV30 (1,38%)	A (50%)	A1 (100%)
	B (66,67%)	B1 (100%)	X		B (50%)	B1 (100%)
HPV61 (1,42%)	B (66,67%)	B1 (100%)	X	HPV39 (1,38%)	A (100%)	A1 (50%)
	C (33,33%)	C1 (100%)	X			A2 (50%)
HPV70 (1,42%)	B (100%)	B1 (100%)	X	HPV45 (1,38%)	B (100%)	B1 (100%)
HPV81 (1,42%)	A (100%)	A1 (33,33%)	X	HPV52 (1,38%)	A (100%)	A1 (100%)
		A-ND (66,67%)	X	HPV53 (1,38%)	D (100%)	D1 (100%)

HPV90 (1,42%)	A (100%)	A1 (33,33%)	X	HPV61 (1,38%)	A (50%)	A1 (100%)
		A-ND (66,67%)	X		B (50%)	B-ND (100%)
HPV226 (0,94%)	ND (100%)	ND (100%)	X	HPV62 (1,38%)	A (50%)	A-ND (100%)
HPV40 (0,94%)	A (100%)	A-ND (100%)	X		ND (50%)	ND (100%)
HPV42 (0,94%)	ND (100%)	ND (100%)	X	HPV68 (1,38%)	A (50%)	A1 (100%)
HPV53 (0,94%)	A (50%)	A1 (100%)	X		D (50%)	D1 (100%)
	ND (50%)	ND (100%)	X	HPV70 (1,38%)	A (100%)	A1 (100%)
HPV54 (0,94%)	B (50%)	B1 (100%)	X	HPV83 (1,38%)	ND (100%)	ND (100%)
	C (50%)	C1 (100%)	X	HPV89 (1,38%)	A (100%)	A-ND (100%)
HPV62 (0,94%)	A (50%)	A-ND (100%)	X	HPV91 (1,38%)	A (100%)	A-ND (100%)
	ND (50%)	ND (100%)	X	HPV214 (0,69%)	A (100%)	A1 (100%)
HPV67 (0,94%)	B (100%)	B1 (50%)	X	HPV226 (0,69%)	A (100%)	A1 (100%)
		B-ND (50%)	X	HPV25 (0,69%)	ND (100%)	ND (100%)
HPV72 (0,94%)	A (50%)	A-ND (100%)	X	HPV43 (0,69%)	A (100%)	A1 (100%)
	ND (50%)	ND (100%)	X	HPV5 (0,69%)	ND (100%)	ND (100%)
HPV83 (0,94%)	ND (100%)	ND (100%)	X	HPV50 (0,69%)	A (100%)	A1 (100%)
HPV108 (0,47%)	ND (100%)	ND (100%)	X	HPV6 (0,69%)	B (100%)	B3 (100%)
HPV5 (0,47%)	ND (100%)	ND (100%)	X	HPV66 (0,69%)	B (100%)	B2 (100%)
HPV6 (0,47%)	B (100%)	B5 (100%)	X	HPV71 (0,69%)	A (100%)	A1 (100%)
HPV69 (0,47%)	A (100%)	A2 (100%)	X	HPV8 (0,69%)	A (100%)	A-ND (100%)
HPV82 (0,47%)	B (100%)	B1 (100%)	X	HPV82 (0,69%)	B (100%)	B1 (100%)
HPV86 (0,47%)	ND (100%)	ND (100%)	X	HPV86 (0,69%)	ND (100%)	ND (100%)
HPV87 (0,47%)	A (100%)	A1 (100%)	X			
HPV89 (0,47%)	A (100%)	A1 (100%)	X			

Table A4. Distribution of HPV types, lineages, and sublineages in RDC and Liège. For each HPV type, the table reports the percentage of sequences assigned to each lineage and sublineage in the two study populations. The central 'X' column separates RDC (left) and Liège (right). Cases labeled 'ND' could not be assigned to any known reference in the PAVE database. These ND sequences showed <99.5% similarity to known references and are considered candidates for novel lineage or sublineage status.

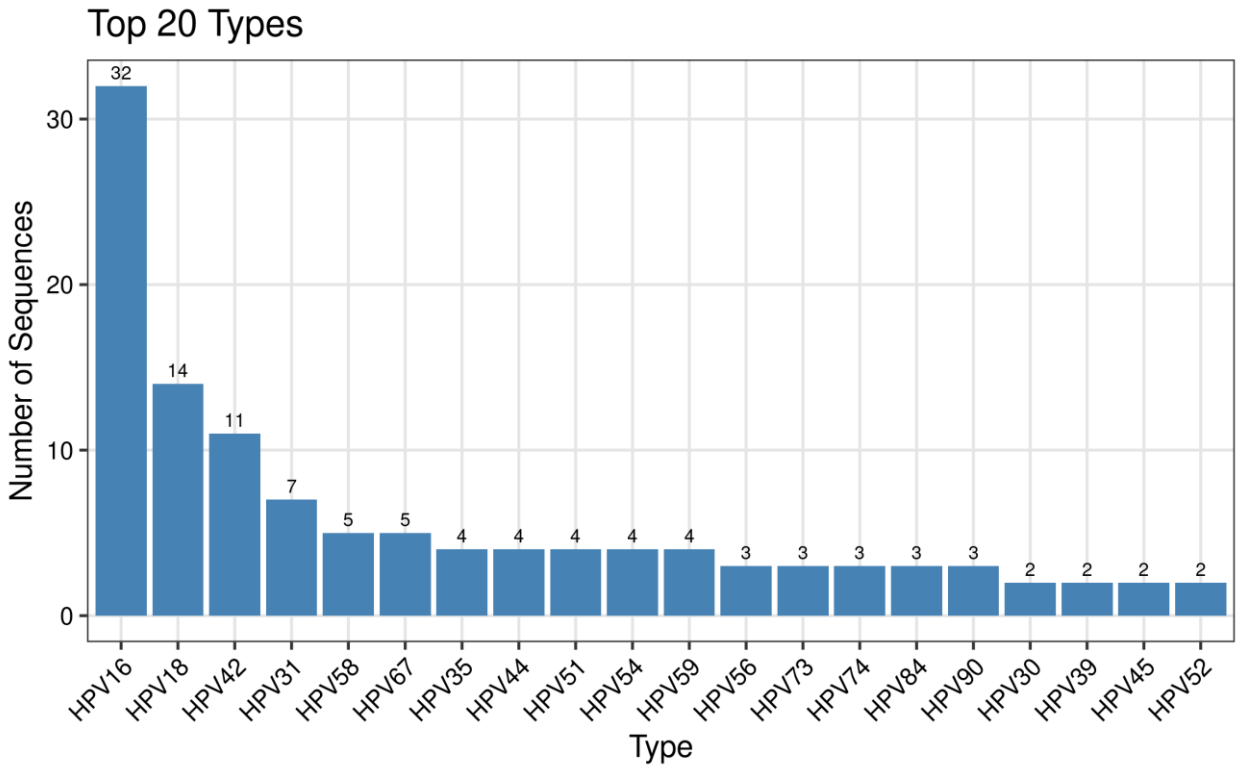

Figure S10: Distribution of the top 20 HPV types in the Liège, Belgium cohort. Data were processed and visualized using our in-house HPV analysis tool.

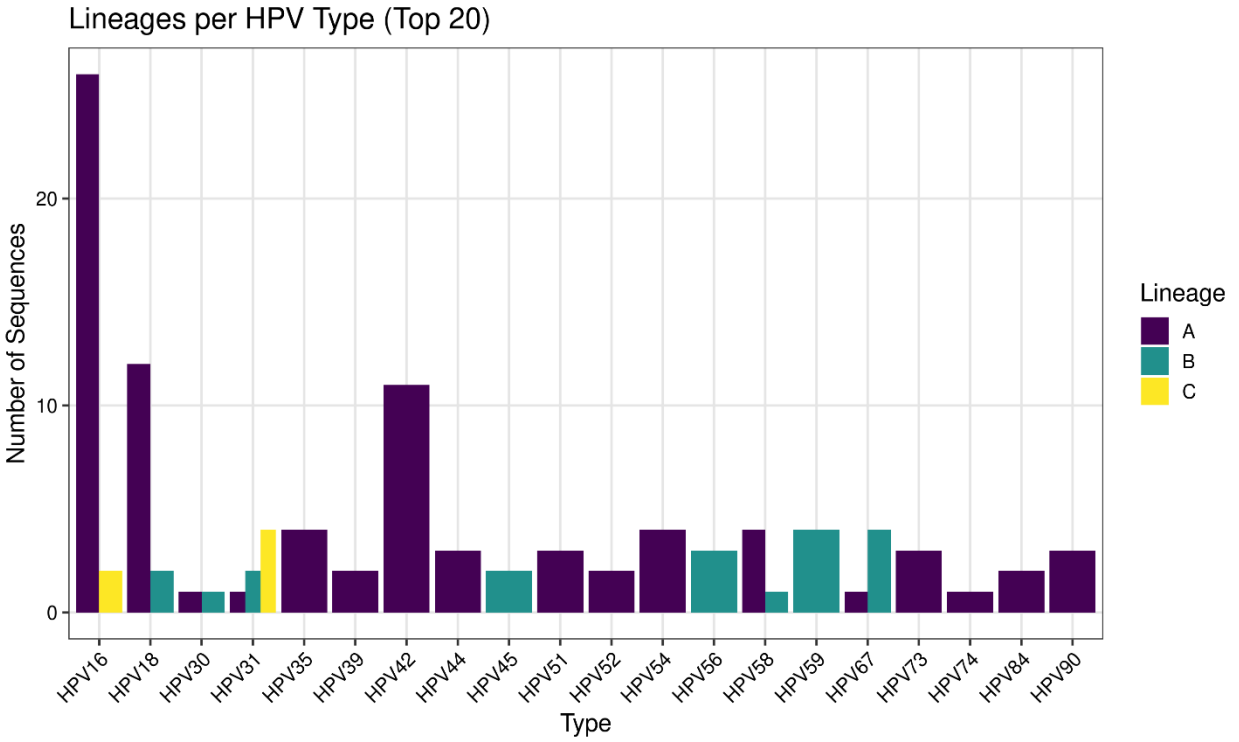

Figure S11: Distribution of lineages for the top 20 HPV types identified in the Liège, Belgium cohort. Lineage classification and visualization were performed using our in-house HPV analysis tool

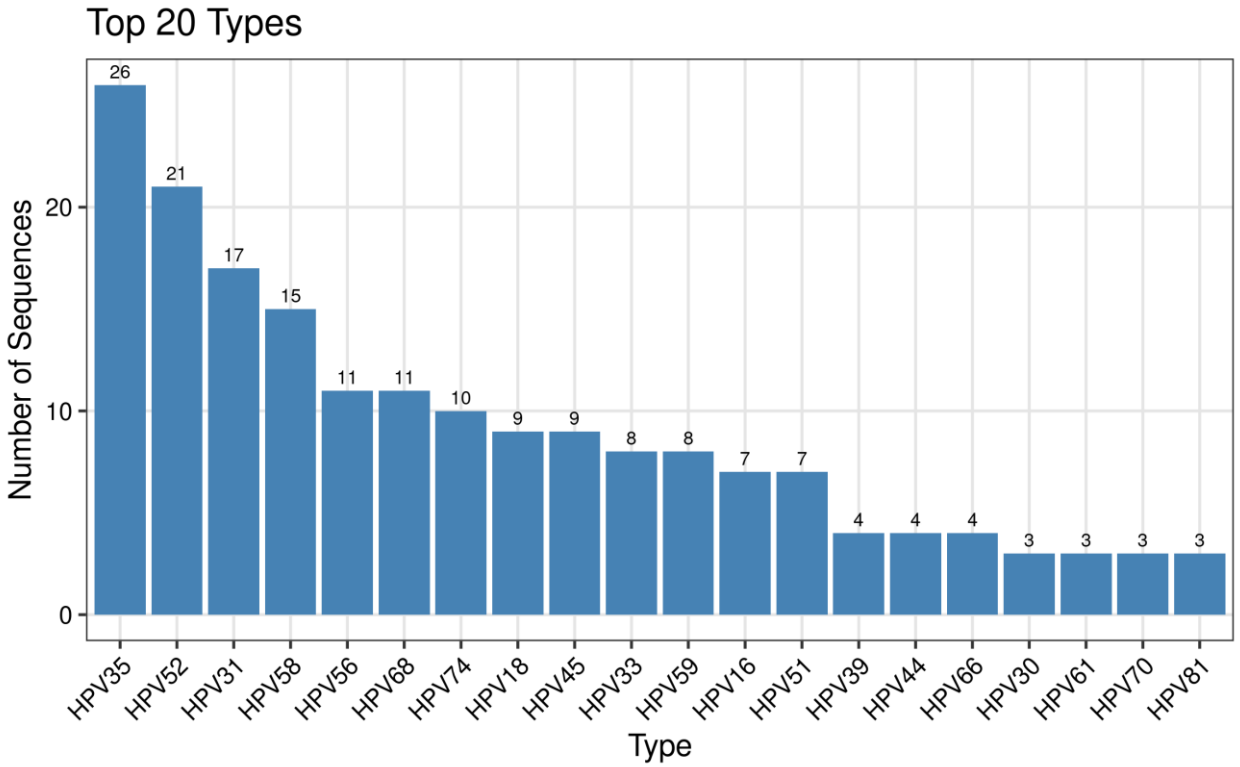

Figure S12: Distribution of the top 20 HPV types in the Kinshasa, DRC cohort. Data were processed and visualized using our in-house HPV analysis tool.

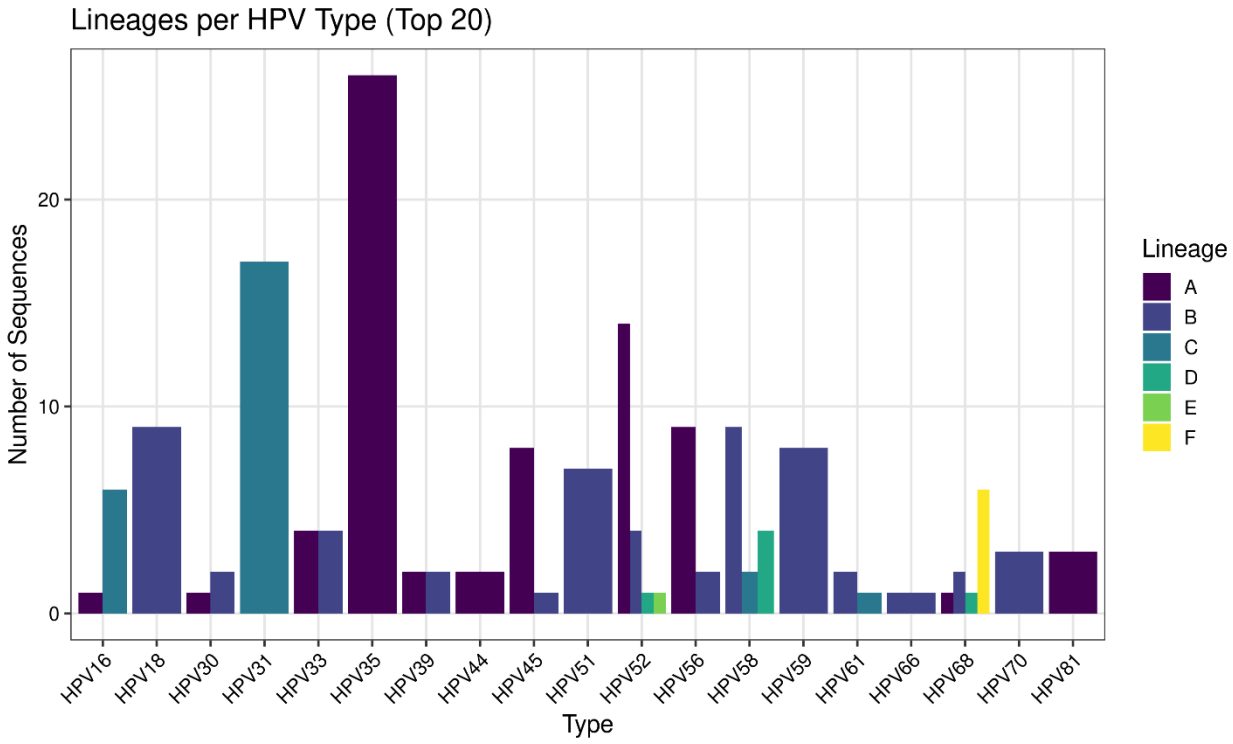

Figure S13: Distribution of lineages for the top 20 HPV types identified in the Kinshasa, DRC cohort. Lineage classification and visualization were performed using our in-house HPV analysis tool

Phylogenetic Trees

HPV44

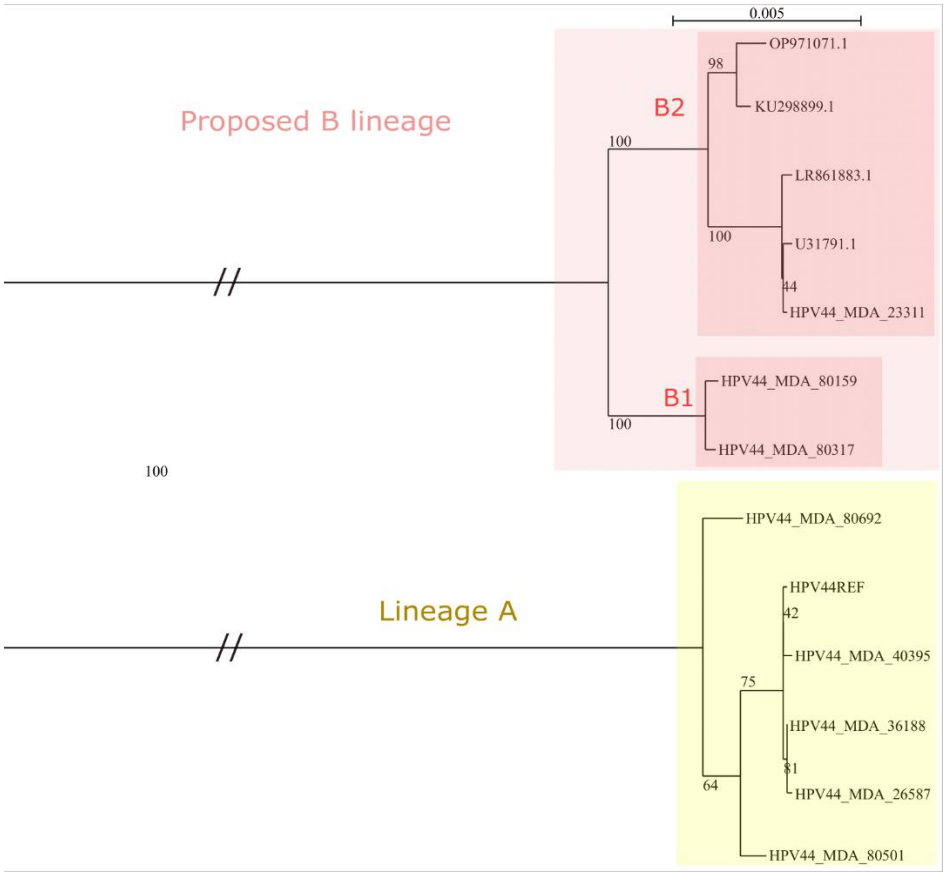

Figure S14: Maximum likelihood phylogenetic tree of HPV44 sequences, including study samples and reference genomes. The tree reveals two major clusters. The first cluster (yellow) corresponds to lineage A, which includes five of our samples alongside the HPV44 reference sequence from PAVE. The second cluster (red) represents a proposed lineage B, further subdivided into two sublineages. Sublineage B1 comprises sample *23311* and its four closest GenBank matches, with sequence identities ranging from 99.6% to 99.9% and are different by 0,7% from B2 and 7% from lineage A. Sublineage B2 includes samples *80159* and *80317*, which share 99.9% identity with each other but are ~0.7% divergent from B1 and ~7% divergent from lineage A. These two samples have no close GenBank matches (all below 99.5% identity), supporting the definition of a distinct sublineage. Scale bar indicates nucleotide substitutions per site. Bootstrap values are shown for major nodes.

Mutation analysis of the E6 gene reveals a specific amino acid change in sublineage B1 (R35K), and two unique mutations in sublineage B2 (N81D and L101F). For the E7 protein, a single mutation (Y5H) was identified exclusively in sublineage B1. These molecular distinctions further support the establishment of a novel lineage B and its subdivision into B1 and B2.

HPV226

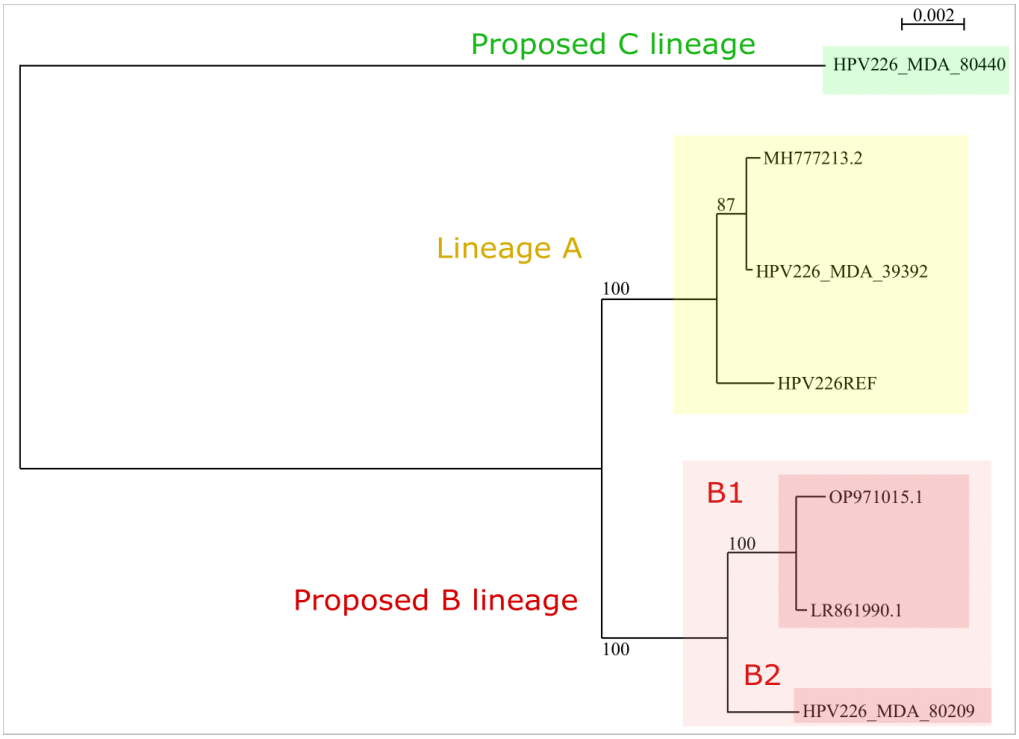

Figure S15. Maximum likelihood phylogenetic tree of HPV226 complete genome sequences, including study samples, PAVE reference genomes, and GenBank entries. The yellow-highlighted cluster represents Lineage A, formed by the reference genome (from PAVE), our sample 39392, and a GenBank match (MH777213.2), all showing less than 0.5% divergence. The green-highlighted sequence 80440 is the most divergent, showing >4% nucleotide difference from all other clusters, supporting its classification as a proposed Lineage C. The red-highlighted clade represents a proposed Lineage B, which is ~1% divergent from Lineage A. Within this lineage, two sub-lineages can be distinguished: B1, composed of GenBank sequences (OP971015.1, LR861990.1), and B2, represented by sample 80209, which differs from B1 by approximately 0.5%.

The GenBank sequences forming sub-lineage B1 were the closest matches to sample 80209 (B2).

HPV42

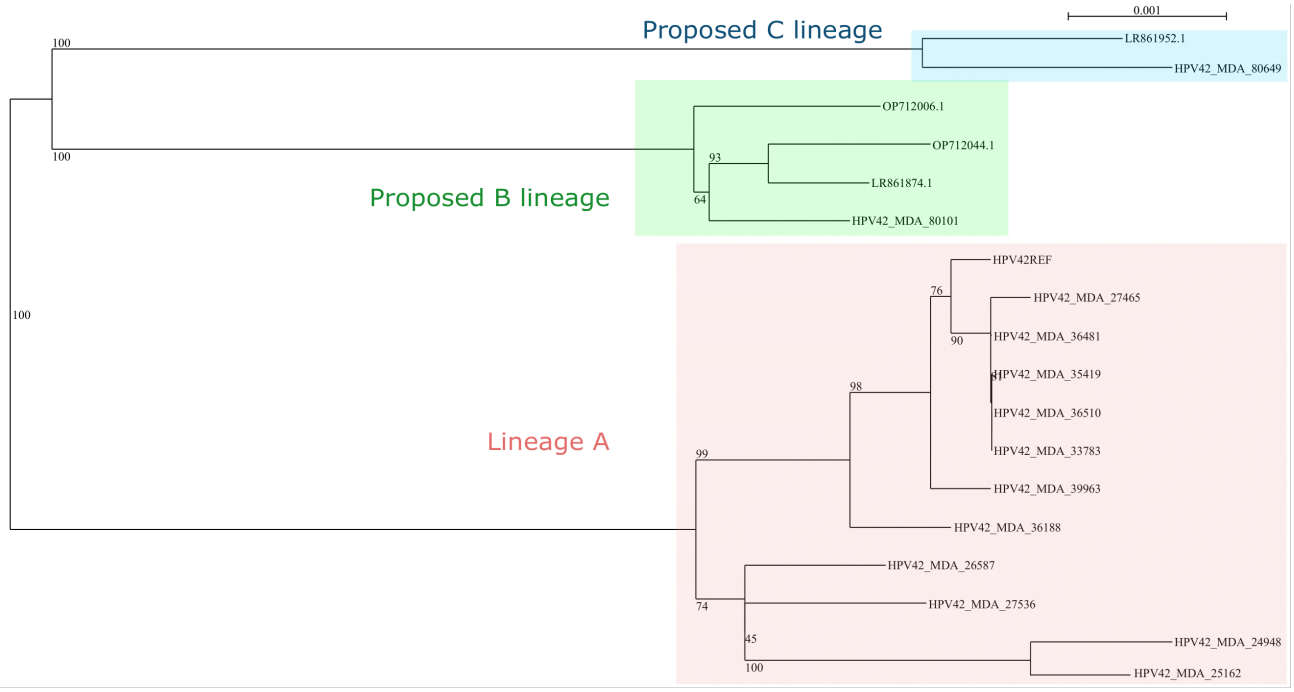

Figure S16: Maximum likelihood phylogenetic tree based on complete HPV42 genome sequences, including study-derived samples, PAVE reference genomes, and GenBank sequences. Eleven study samples cluster with the HPV42 reference genome from PAVE to form lineage A (pink). Sample 80101 and its closest GenBank matches form a proposed lineage B (green), differing by a maximum of 0.4% among themselves and separated by more than 1% from both lineage A and the proposed lineage C. Lineage C is composed of sample 80649 and its best GenBank match (99.7% similarity), and differs by at least 1% from the other two lineages, supporting its classification as a distinct lineage.

HPV53

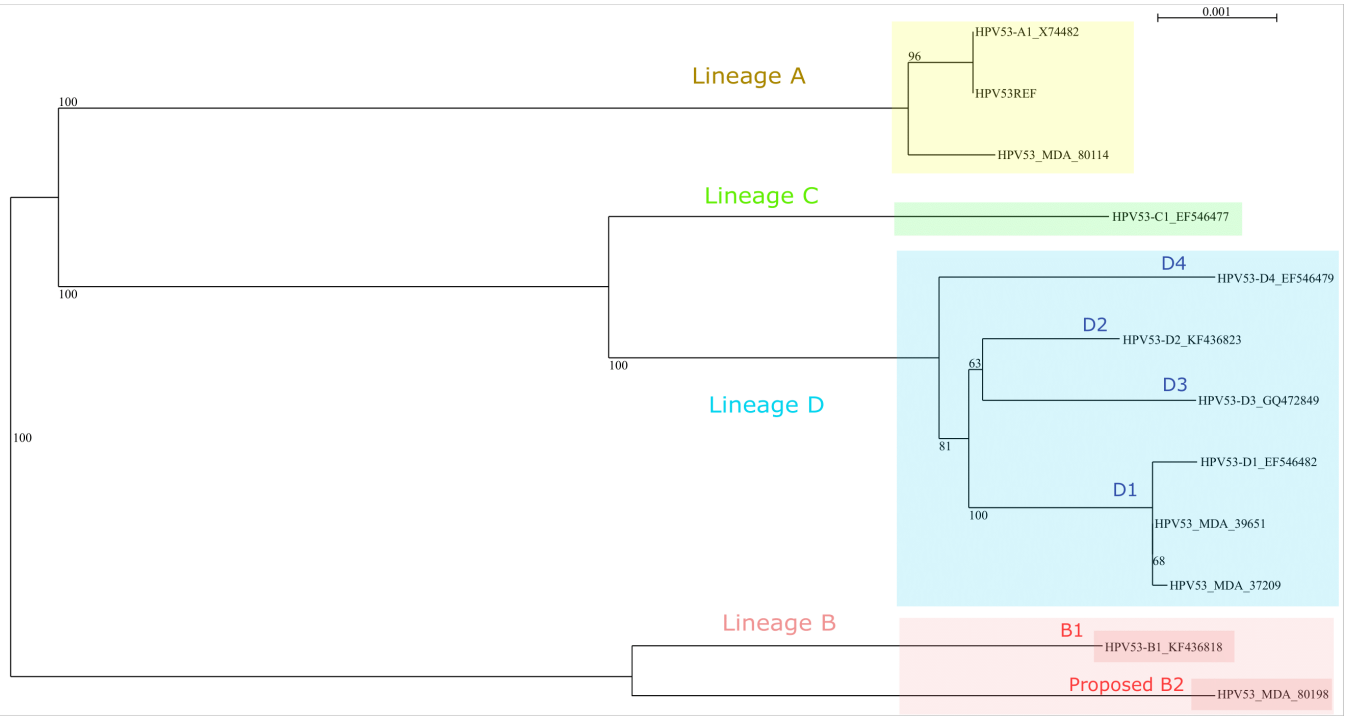

Figure S17. Maximum likelihood phylogenetic tree based on complete HPV53 genome sequences, including study-derived samples and PAVE reference genomes. Among the genomes analyzed, only HPV53_MDA_80198 exhibited notable divergence from both PAVE and GenBank sequences. Its closest match was the reference variant HPV53_B1, with a nucleotide difference of 0.9%, supporting the proposal of a new sublineage, B2. No closer match was identified in GenBank, with the best hit showing only 99% (red) similarity.

HPV62

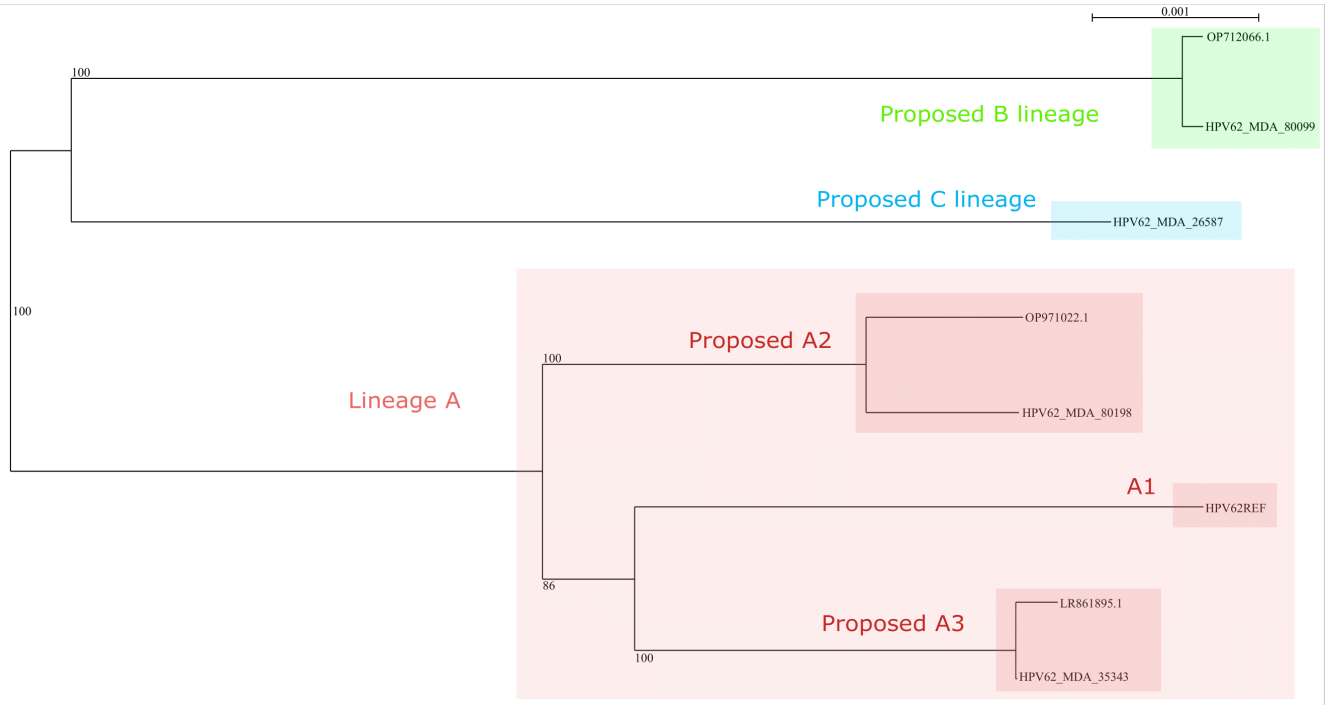

Figure S18. Maximum likelihood phylogenetic tree based on complete HPV62 genome sequences, including study-derived samples and selected GenBank references. Four HPV62 genomes were identified in our dataset. HPV62_MDA_80099 clusters with its best GenBank match (OP712066.1) and differs by ≥1.2% from other sequences, supporting a new lineage, B (green). HPV62_MDA_26587, with no close GenBank match and ≥1.1% divergence from all others, is proposed as lineage C (blue). HPV62_MDA_80198 clusters with its closest GenBank match (99.7% similarity), and both differ by 0.7% from the reference sequence (A1), supporting their classification as a new sublineage, A2. HPV62_MDA_35343 shows 99.9% similarity with its best GenBank match, but differs by 0.6% from the reference (A1) and 0.5% from the proposed sublineage A2, supporting the designation of a distinct sublineage, A3.

By analyzing lineage-specific mutations in the E6 and L1 genes, we identified amino acid changes that further support the proposed classification.

In the E6 protein, one specific substitution (I85L) was observed in sublineage A2, while three mutations (L67C, H80Y, and F83L) were unique to the proposed lineage B.

In the L1 protein, three amino acid changes (T279P, E329D, and A351T) were specific to sublineage A3, two (I138V and I146V) to lineage B, and three others (K190T, S434T, and A497T) were found in the proposed lineage C, highlighting distinct molecular profiles across lineages.

HPV72

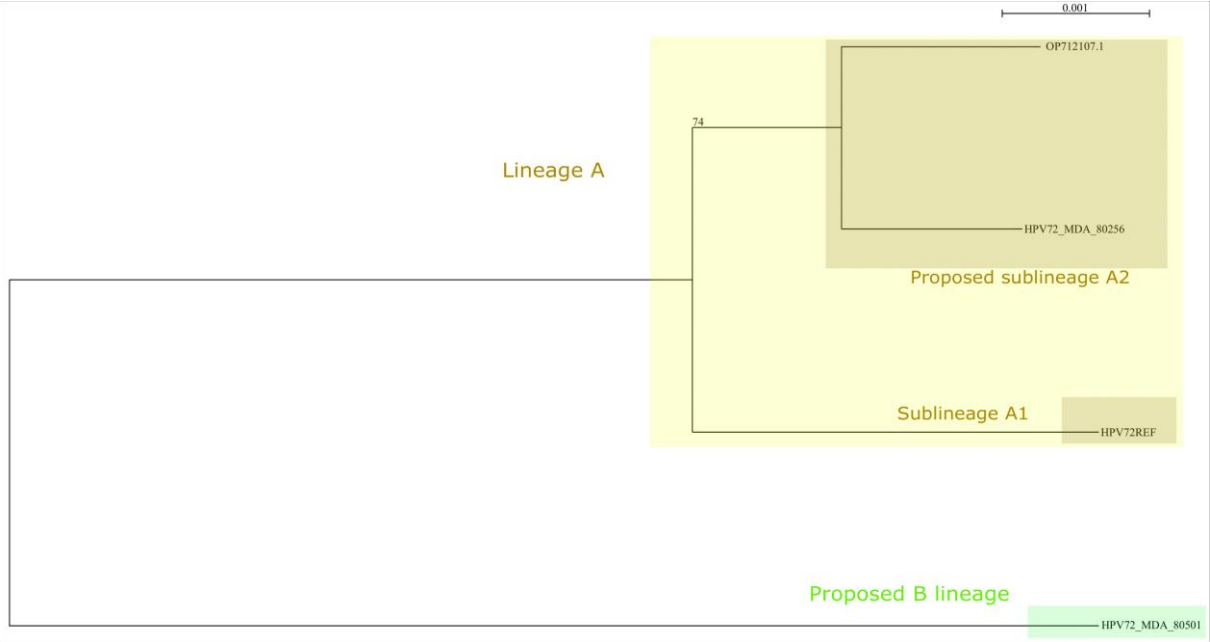

Figure S19. Maximum likelihood phylogenetic tree based on complete HPV72 genome sequences, including study-derived samples and selected GenBank references. Two HPV72 genomes were identified in our dataset. Sample HPV72_MDA_80256 differs by 0.5% from the reference sequence and clusters with its best GenBank match (99.7% similarity), forming a proposed sublineage A2 (yellow). Sample HPV72_MDA_80501 differs by at least 1% from lineage A and has no close GenBank match (>99.5% similarity), supporting its classification as a new lineage B (green).

The lineage B is further reinforced by the presence of mutation in the E6 gene (A69V) for the and four specific mutations in the E7 gene (L29I, Y67W, D78H, A83V), which are not observed in other lineages. In contrast, the two samples proposed as sub-lineage A2 share two specific E7 mutations (S32A and G58R), distinguishing them from the reference and supporting their classification as a new sub-lineage within HPV72.

HPV83

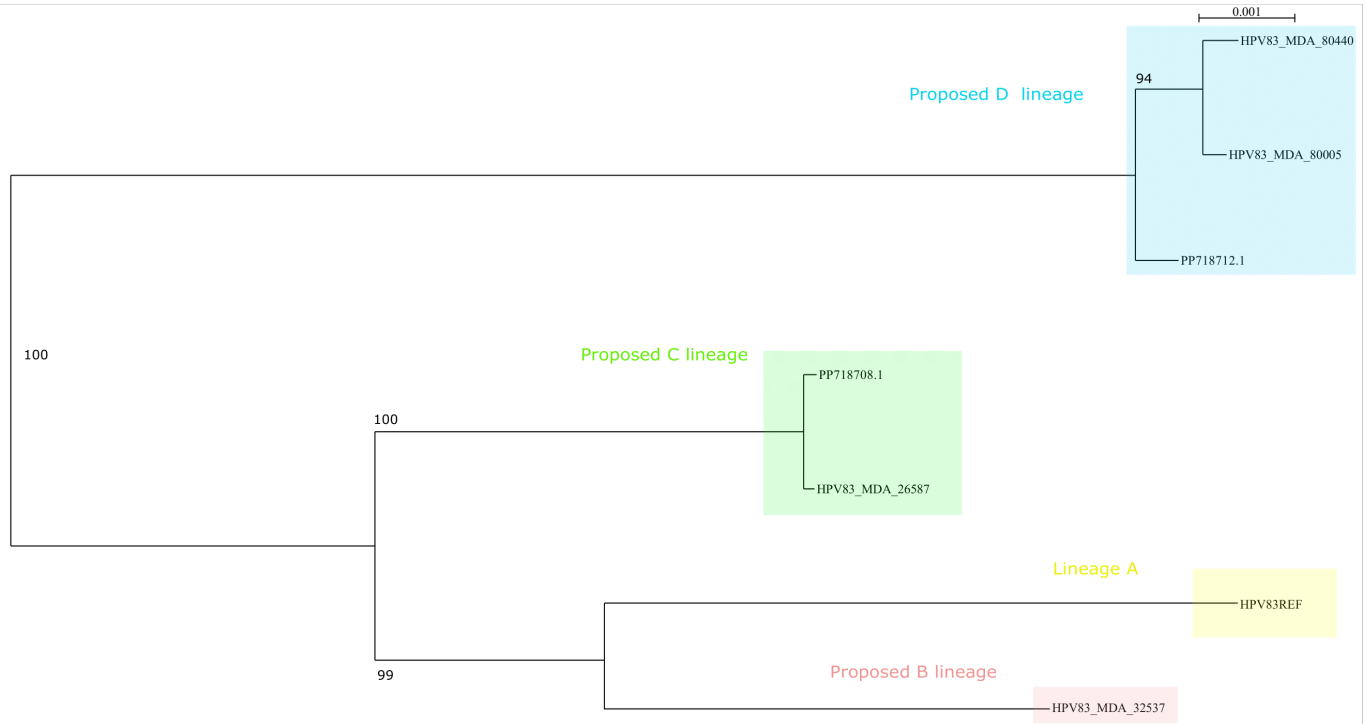

Figure S20. Maximum likelihood phylogenetic tree based on complete HPV83 genome sequences, including study-derived samples and selected GenBank references. The proposed lineage D (blue) comprises two study samples and their closest GenBank match, all sharing >99.5% similarity and differing by at least 1% from other clusters, supporting its classification as a distinct lineage. The proposed lineage C (green) includes sample HPV83_MDA_26587 and its best GenBank match, also differing by ≥1% from other clusters. The final sample, HPV83_MDA_32537, shows >1% divergence from all others and has no close GenBank match (>99% similarity), supporting its assignment to a new lineage B (pink).

HPV108

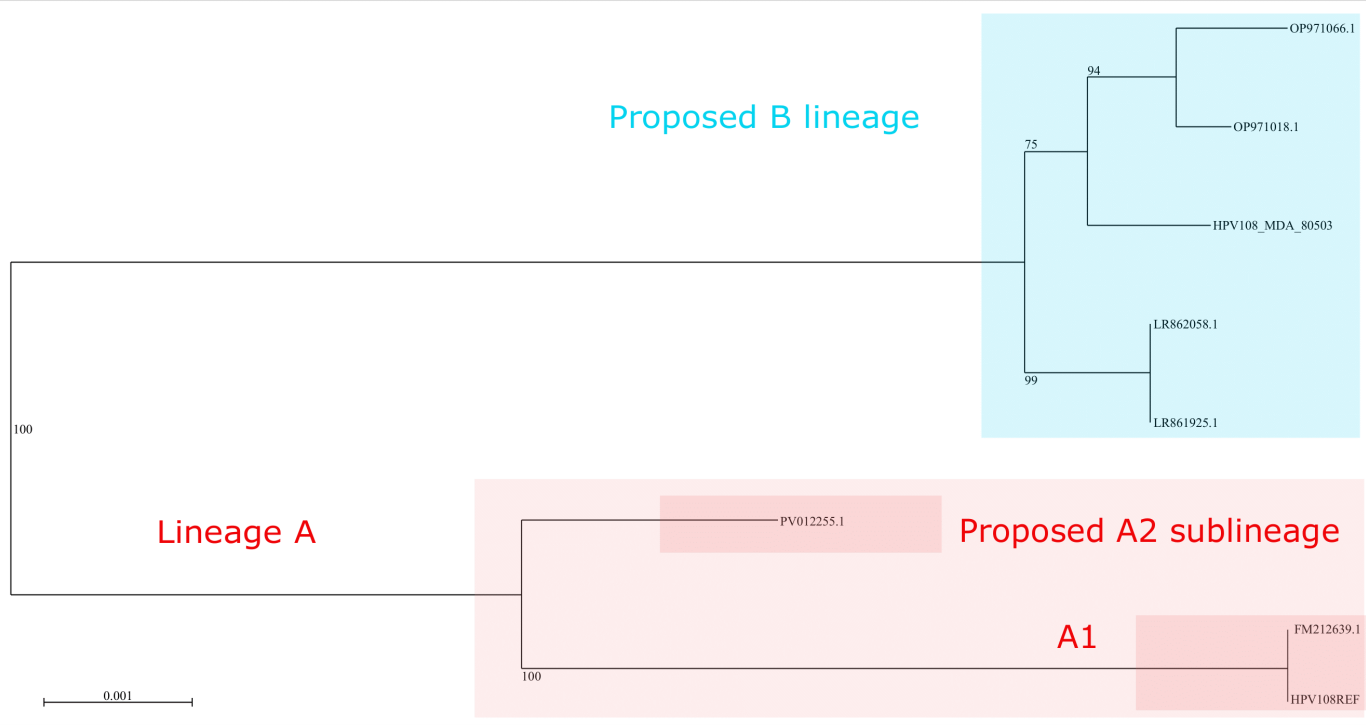

Figure S21. Maximum likelihood phylogenetic tree based on complete HPV108 genome sequences, including study-derived samples and selected GenBank references. We included the five closest GenBank matches to our HPV108 sample (HPV108_MDA_80503). Four of them cluster tightly with our sample (≥99.5% similarity), forming the proposed lineage B (blue), which differs from lineage A by over 1%, consistent with PAVE criteria. The fifth closest GenBank sequence is more distant, showing 1.6% divergence from our sample and 0.7% from the reference, and is proposed to represent a new A2 sublineage (red).

HPV5

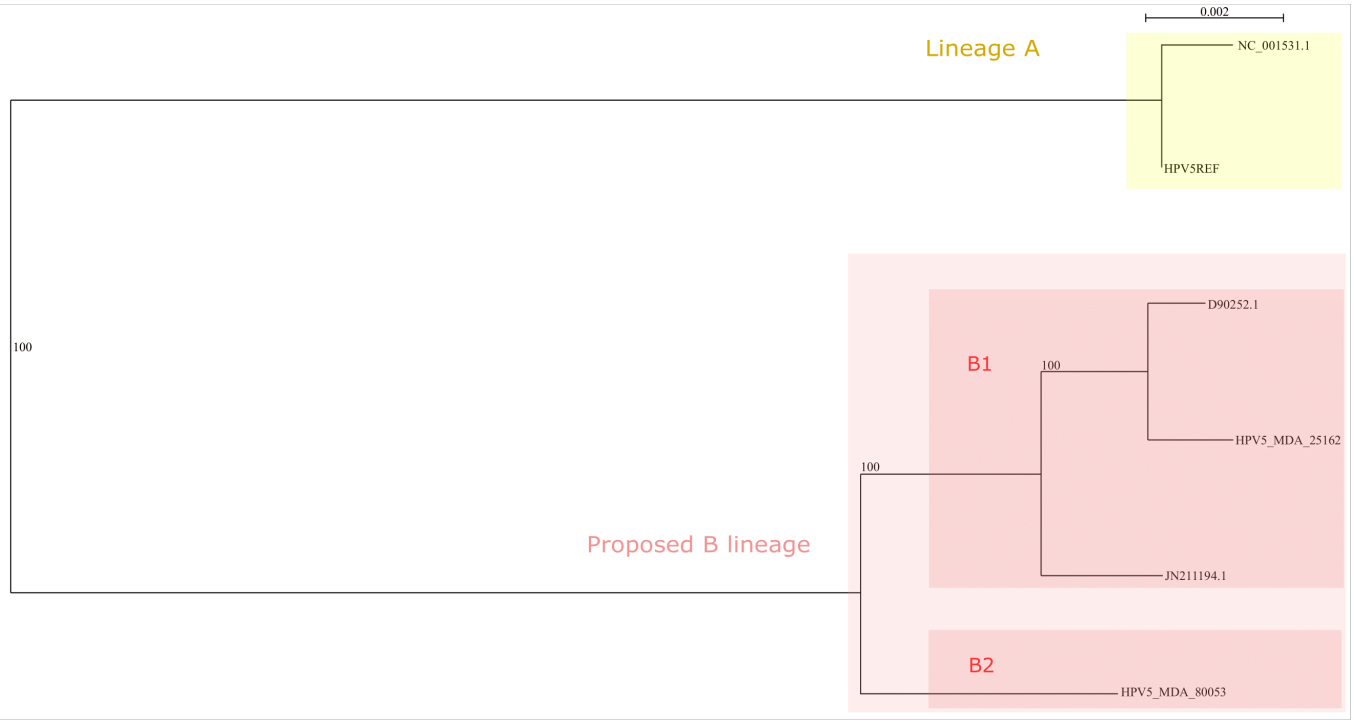

Figure S22. Maximum likelihood phylogenetic tree based on complete HPV5 genome sequences, including study-derived samples and selected GenBank references. Our dataset includes two HPV5 samples. HPV5_MDA_25162 clusters with its closest GenBank sequences, D90252 (99.7% similarity) and JN211194 (99.6%),forming a well-supported group. The second sample, HPV5_MDA_80053, does not have any GenBank match above 99.5% identity and shows approximately 0.9% divergence from the first group and more than 3% divergence from the reference genome. The reference itself differs by at least 3.2% from all other sequences. Based on these pairwise distances and phylogenetic structure, we propose a new lineage, B, which can be subdivided into B1 (including HPV5_MDA_25162 and its GenBank matches) and B2 (HPV5_MDA_80053). The only GenBank sequence closely related to the reference genome, NC_001531, was also included for context.

HPV86

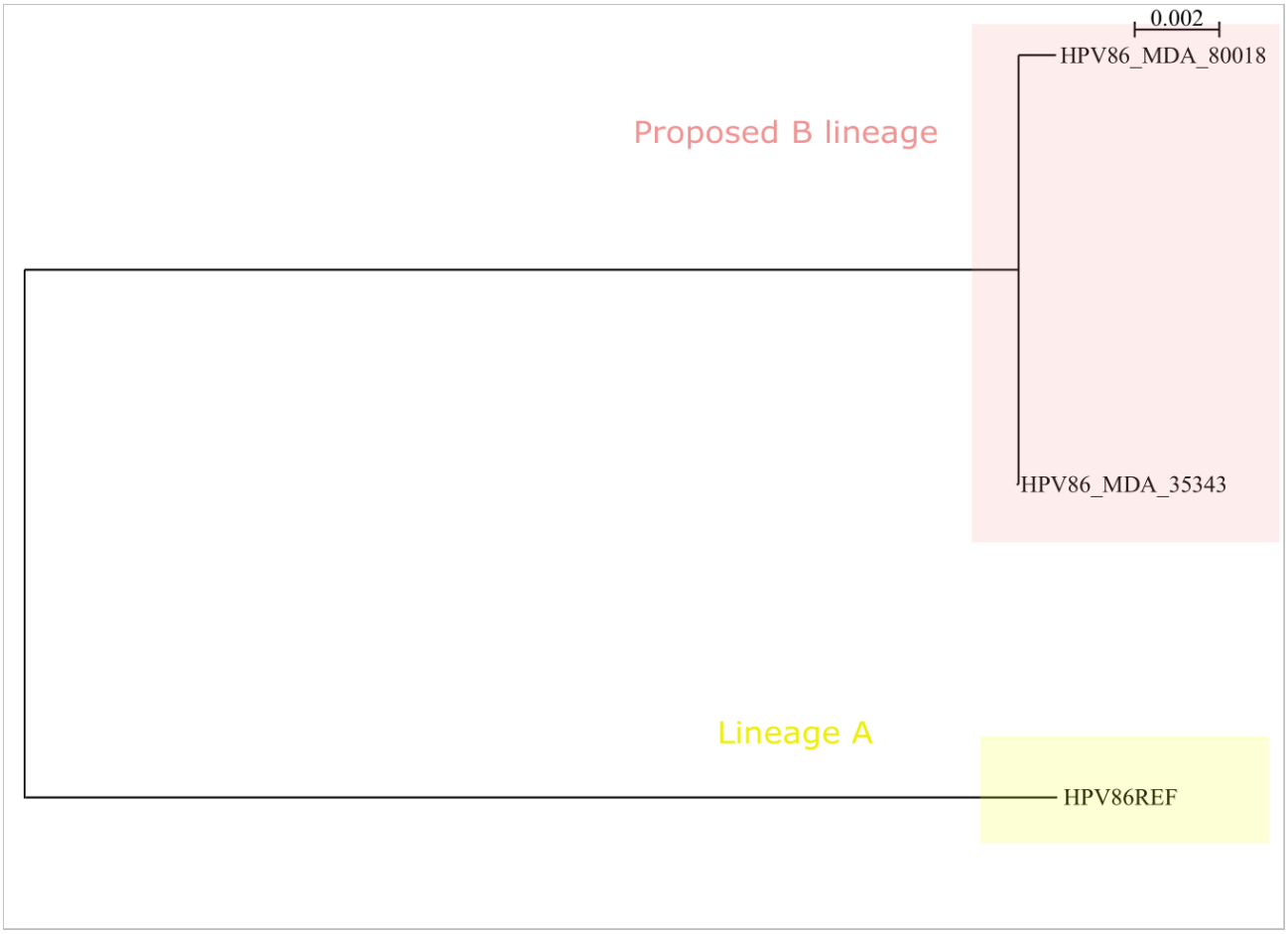

Figure S23. Maximum likelihood phylogenetic tree based on complete HPV86 genome sequences, including study-derived samples and PAVE reference. The two study samples cluster closely together (99.9% nucleotide identity) and are genetically distant from the reference genome (yellow) by approximately 4.6%. This level of divergence supports the proposal of a distinct lineage, designated as lineage B (pink).

HPV67

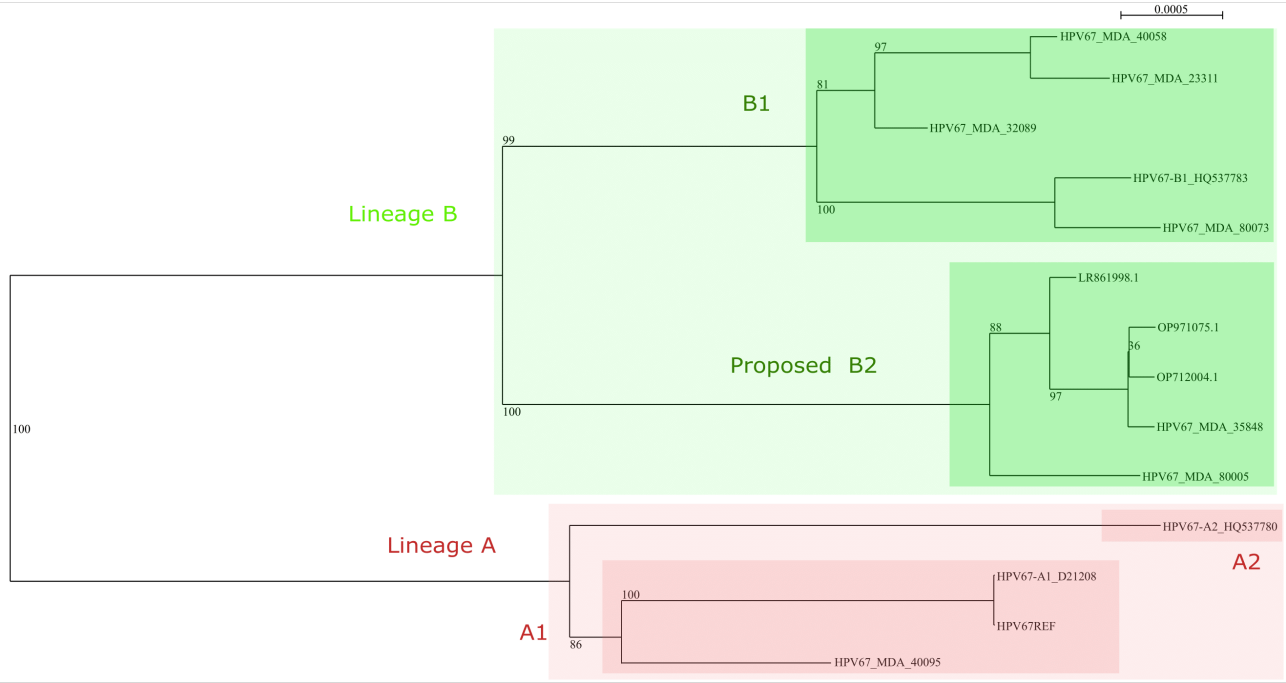

Figure S24. Maximum likelihood phylogenetic tree based on complete HPV67 genome sequences, including study-derived samples and selected GenBank references. One study sample clusters with the reference lineage A (pink) with 99.9% similarity, while several others group within reference lineage B (green), sharing at least 99.6% identity. Two samples, 80005 and 35848, do not match any known reference but cluster with their three closest GenBank matches. Together, these five sequences, sharing ≥99.8% similarity and form a distinct clade that differs from sublineage B1 by 0.5–0.6%, supporting the definition of a novel sublineage B2.

HPV90

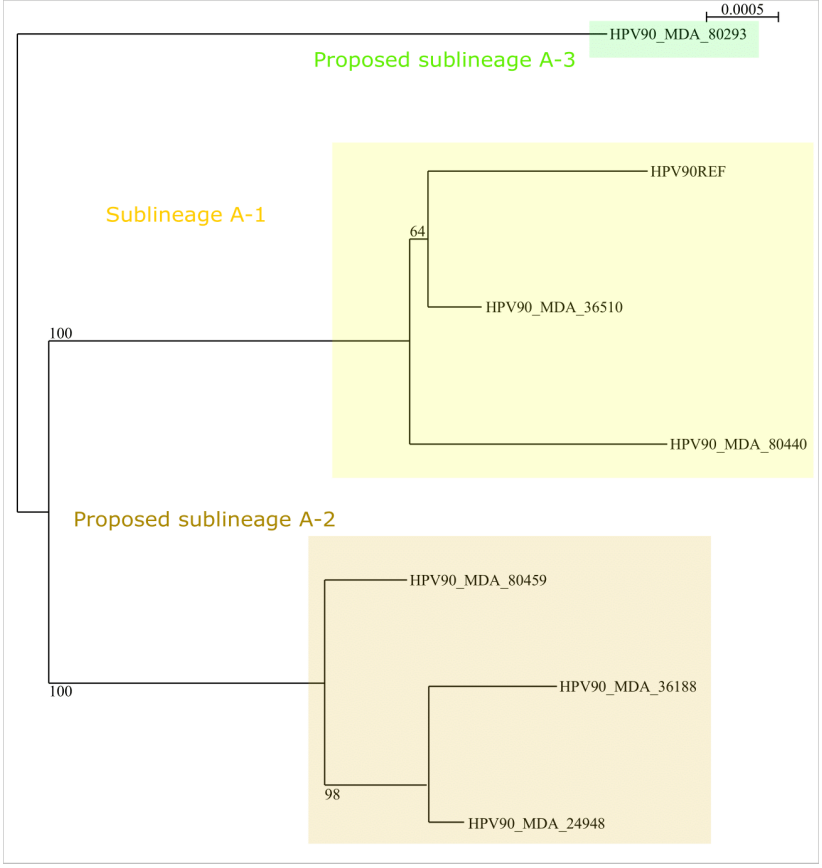

Figure S25. Maximum likelihood phylogenetic tree based on complete HPV90 genome sequences, including study-derived samples and PAVE references. The central cluster, corresponding to the known A1 sublineage, comprises the reference sequence together with

HPV90_MDA_36510 and HPV90_MDA_80440, which differ by no more than 0.3% from each other and by 0.5–0.7% from other groups, confirming their grouping within A1. The proposed A2 sublineage includes HPV90_MDA_80459, HPV90_MDA_36188, and HPV90_MDA_24948, with a maximum intra-group divergence of 0.3% and a divergence of 0.7% from the other sublineages. The proposed A3 sublineage is represented by a single sample, HPV90_MDA_80293, showing a minimum divergence of 0.6% from all other groups and lacking any close GenBank match (>99.5% identity).

HPV81

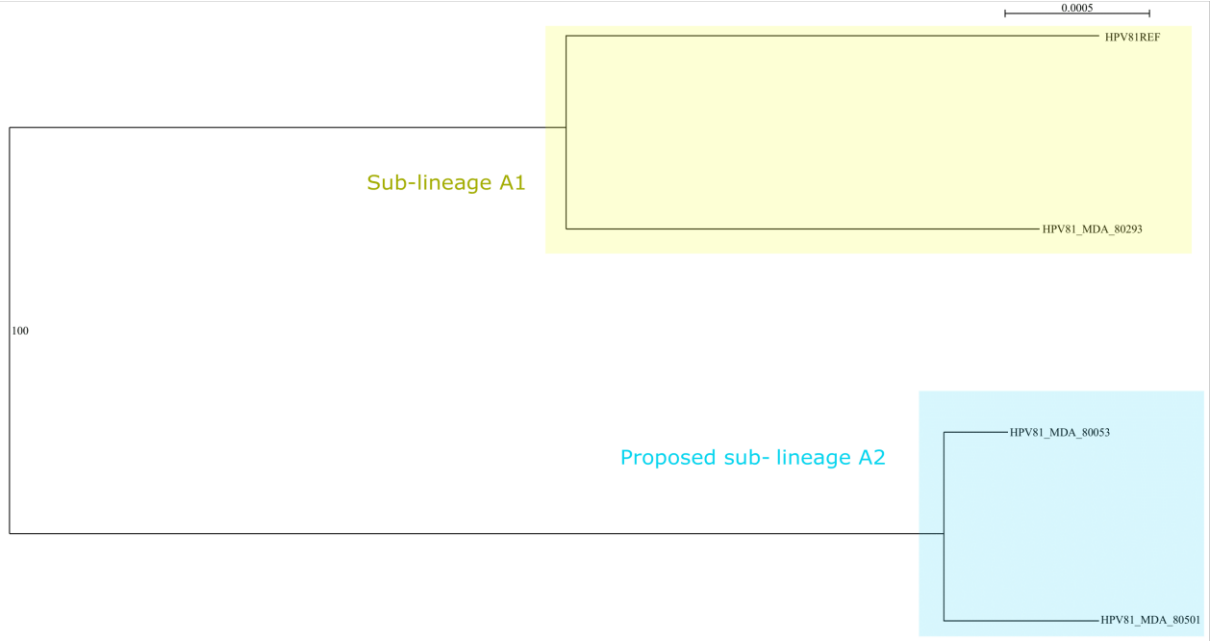

Figure S26. Maximum likelihood phylogenetic tree based on complete HPV81 genome sequences, including study-derived samples and PAVE references. One of the three study samples clusters well with the reference A1 (yellow). The two other genomes differ by at least 0.6% from this group and appear to form a distinct cluster, which we propose as a new sub-lineage, A2 (blue).

HPV84

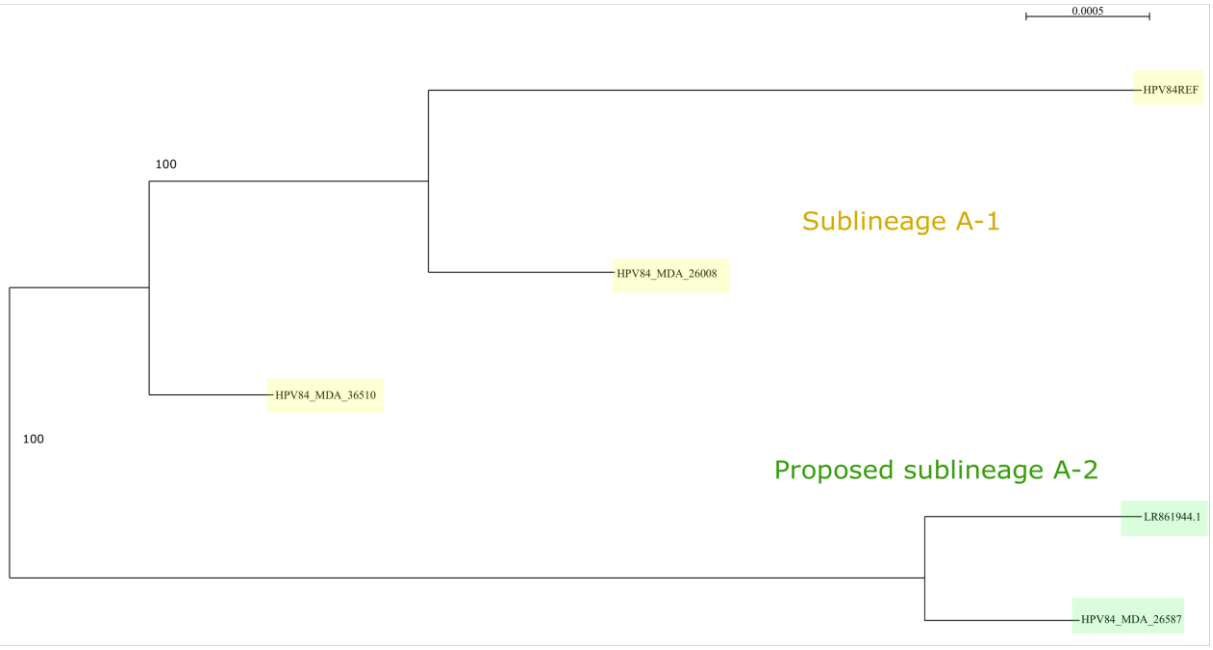

Figure S27. Maximum likelihood phylogenetic tree based on complete HPV84 genome sequences, including study-derived samples and selected GenBank references. Sample 26008 is 0.4% divergent from the reference sequence and 0.2% from sample 36510. Together with the reference, these sequences form the A1 cluster (yellow). The remaining sample, 26587, clusters with its closest GenBank match and is at least 0.5% divergent from all other sequences, supporting the proposal of a new sublineage, A2 (green).

HPV25

No phylogenetic tree could be generated because only two sequences were available (one from this study and the reference). The study-derived sample has no close match in GenBank and shows a 1.4% divergence from the reference, supporting the identification of a new lineage.

HPV8

No phylogenetic tree could be generated because only two sequences were available (one from this study and the reference). The study-derived sample has no close match in GenBank and shows a 0,6 % divergence from the reference, supporting the identification of a new sub-lineage.

HPV89

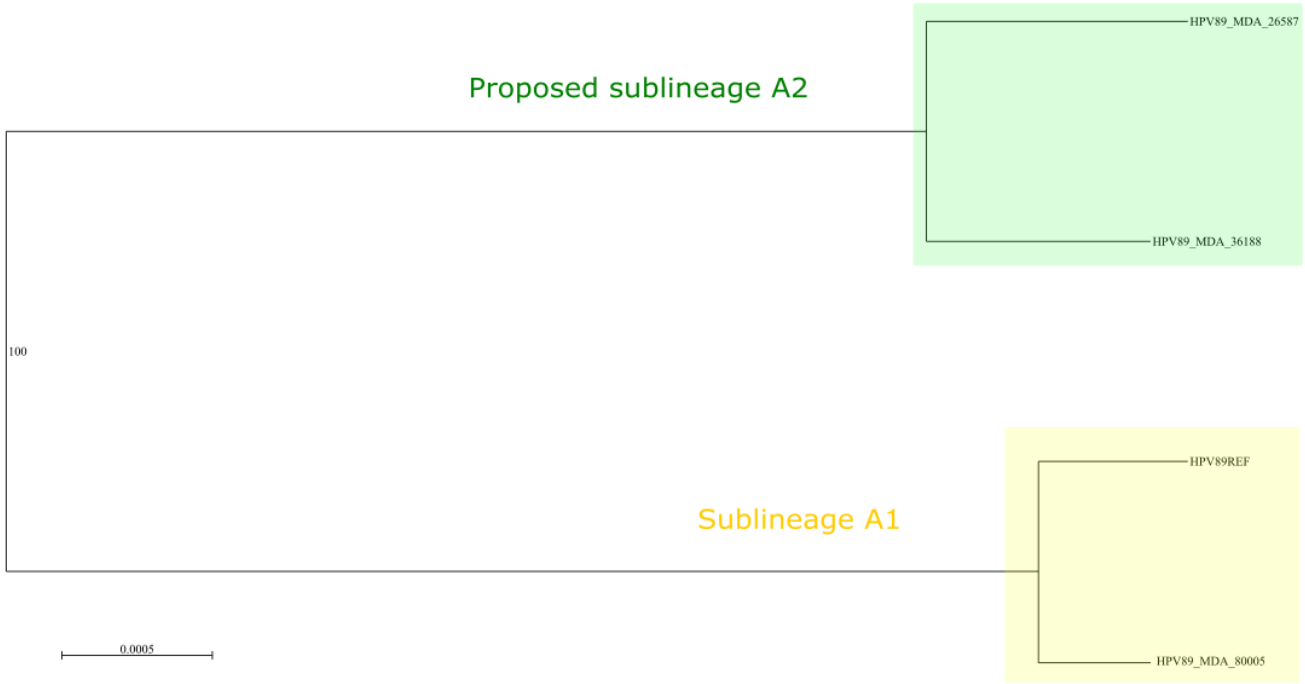

Figure S28. Maximum likelihood phylogenetic tree based on complete HPV89 genome sequences, including study-derived samples and the PAVE reference. One sample clusters closely with the reference, forming sublineage A1, while the two others cluster together and are 0.8% divergent from A1, supporting the designation of a new sublineage, A2.

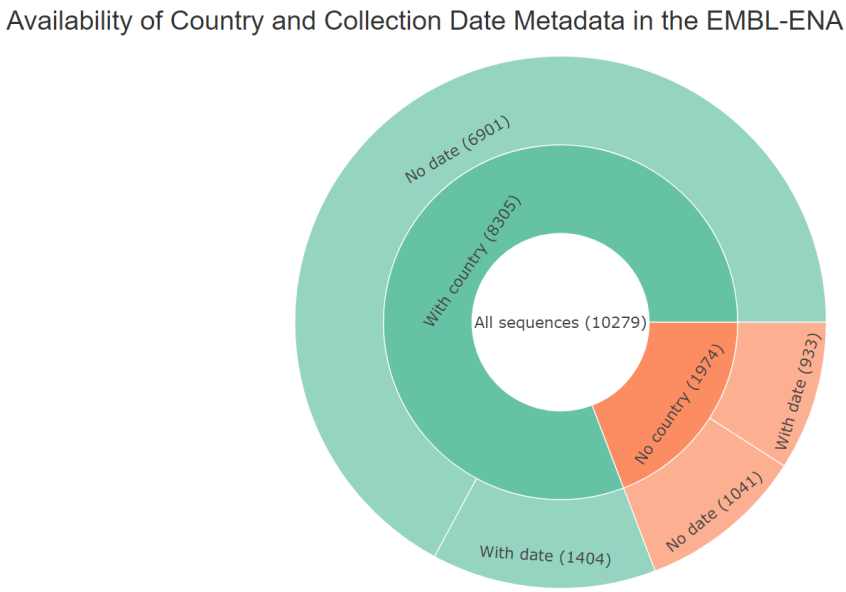

Figure S29. Availability of country and collection date metadata for HPV genome sequences in the EMBL-ENA database. The inner ring shows the number of sequences with and without country annotation (77% have country information). The outer ring further divides these categories by the presence or absence of collection date metadata. Only 1,404 out of 10,279 sequences (~14%) contain both country and date information, highlighting the limitations of the dataset for spatio-temporal analyses

HR vs LR blue				**LR black vs LR blue**		
variable	**p_value**	**p_adjusted**		**variable**	**p_value**	**p_adjusted**
charged_apolar	0.04479406	0.1642449		percent_charged_nis	0.0018589140	0.02044806
binding_affinity_kcal_mol	0.02967401	0.1642449		percent_apolar_nis	0.006990955	0.03845025
dissociation_constant_M	0.03357667	0.1642449		charged_polar	0.158411028	0.58084044

apolar_polar	0.11943446	0.2237475		charged_charged	0.34989841	0.69200819
percent_apolar_nis	0.10056150	0.2237475		charged_apolar	0.377459011	0.69200819
percent_charged_nis	0.12204411	0.2237475		apolar_apolar	0.252905461	0.69200819
apolar_apolar	0.14467297	0.2273432		polar_polar	0.575791696	0.90481552
intermolecular_contacts	0.19205327	0.2640732		intermolecular_contacts	0.716473743	0.98515140
charged_polar	0.34424414	0.4207428		apolar_polar	1	1
polar_polar	0.63805078	0.7018559		binding_affinity_kcal_mol	1	1
HR vs all LR (blue+black)				**HR+ LR (blue) vs LR (black)**		
variable	**p_value**	**p_adjusted**		**variable**	**p_value**	**p_adjusted**
percent_apolar_nis	4.580531×10^{-7}	5.038584×10^{-6}		percent_apolar_nis	5.36×10^-9	5.89×10^-8
percent_charged_nis	1.190937×10^{-5}	6.550151×10^{-5}		percent_charged_nis	1.36×10^-7	7.45×10^-7
binding_affinity_kcal_mol	6.922328×10^{-5}	2.163771×10^{-4}		binding_affinity_kcal_mol	3.15×10^-3	9.25×10^-3
dissociation_constant_M	7.868259×10^{-5}	2.163771×10^{-4}		dissociation_constant_M	3.36×10^-3	9.25×10^-3
apolar_polar	3.076371×10^{-4}	6.768016×10^{-4}		apolar_polar	9.62×10^-3	2.12×10^-2
charged_apolar	4.808884×10^{-3}	8.816288×10^{-3}		polar_polar	3.34×10^-2	6.13×10^-2
intermolecular_contacts	2.852888×10^{-2}	4.483110×10^{-2}		charged_charged	7.12×10^-2	1.12×10^-1
charged_charged	7.029558×10^{-2}	9.665642×10^{-2}		charged_apolar	1.49×10^-1	2.05×10^-1
polar_polar	1.692560×10^{-1}	1.861816×10^{-1}		charged_polar	1.89×10^-1	2.32×10^-1
apolar_apolar	1.644472×10^{-1}	1.861816×10^{-1}		intermolecular_contacts	2.67×10^-1	2.93×10^-1

Table A5. Wilcoxon rank-sum test results for interfacial parameters in four comparisons for the E6-E6AP complex. Top 10 structural features extracted from E6–E6AP complexes using PRODIGY for (i) HR types vs. HR-clade LR types, (ii) HR-clade LR types vs.

phylogenetically distant LR types, (iii) HR types vs. all LR types, and (iv) HR + HR-clade LR types vs. distant LR types. For each comparison, the table lists the feature name, the raw p-value, and the Benjamini–Hochberg (BH) adjusted p-value.

upper HR vs lower HR				all HR vs all LR		
variable	p_value	p_adjusted		variable	p_value	p_adjusted
intermolecular_contacts	0.0008582637	0.005199886		charged_polar	0.004900030	0.03926605
charged_apolar	0.0012308708	0.005199886		apolar_apolar	0.007139282	0.03926605
percent_apolar_nis	0.0014660182	0.005199886		intermolecular_contacts	0.615273969	0.99251373
percent_charged_nis	0.0018908676	0.005199886		charged_charged	0.992513733	0.99251373
polar_polar	0.0365551374	0.080421302		charged_apolar	0.679584653	0.99251373
apolar_apolar	0.3857993533	0.707298814		polar_polar	0.679431649	0.99251373
charged_charged	0.9336300813	0.933630081		apolar_polar	0.517971469	0.99251373
charged_polar	0.8652672670	0.933630081		percent_apolar_nis	0.955590732	0.99251373
apolar_polar	0.8858945812	0.933630081		percent_charged_nis	0.710188432	0.99251373
binding_affinity_kcal_mol	0.7824109923	0.933630081		binding_affinity_kcal_mol	0.947987629	0.99251373

Table A6: Wilcoxon rank-sum test results for interfacial parameters in two comparisons of E7-pRB complex: (left) upper HR clade vs lower HR clade from the phylogenetic tree, and (right) all HR variants vs all LR variants. For each variable, the raw p-value and the Benjamini–Hochberg–adjusted p-value are shown. Variables describe physicochemical properties of the E7–pRB interaction interface, including residue contact types, proportion of apolar/charged/polar residues, and predicted binding affinity

Scripts and Source Code

Several preprocessing steps and analyses were performed to manage the data, including file handling, sorting, and statistical analysis in R, which are not shown here. The main scripts and larger workflows are available in our tool on GitHub (https://github.com/ElmYassine29/HPVxHunter). As an example, I provide here the script used to analyze the interaction between E7 and pRB across all HPV types, using structures predicted by the AlphaFold Server and processed with PRODIGY.

After obtaining all model in cif we convert them in pdb with the following script '*convert.py*' :

```
from Bio.PDB import MMCIFParser, PDBIO
import sys

cif_file = sys.argv[1]
pdb_file = sys.argv[2]

parser = MMCIFParser()
structure = parser.get_structure("structure", cif_file)

io = PDBIO()
io.set_structure(structure)
io.save(pdb_file)
```

Then apply it with :

```
for f in *.cif; do python convert.py "$f" "${f%.cif}.pdb"; done
```

Then apply "*run_prodigy.sh*":

```
#!/bin/bash

# Folder with PDB files

PDB_DIR="."
# output file
OUTPUT_FILE="prodigy_results.txt"

# delete outpur file if already existed
> "$OUTPUT_FILE"

# Loop on all pdb files
for pdb_file in "$PDB_DIR"/*.pdb; do
```

```
    echo "Processing $pdb_file"
    echo "==== $(basename "$pdb_file") - Chaînes A et B ====" >> "$OUTPUT_FIL
E"
    prodigy "$pdb_file" --selection A B >> "$OUTPUT_FILE" 2>> "$OUTPUT_FILE"
    echo -e "\n" >> "$OUTPUT_FILE"
done

echo " All results are in the file: $OUTPUT_FILE"
```

And finally, '*make.py*' was used to generate a summary table that can be directly opened in Excel o R for statistics analysis:

```
import re
import csv

input_file = "prodigy_E7vspRB_results.txt"
output_file = "prodigy_E7vspRB_summary.csv"

headers = [
    "sample",
    "intermolecular_contacts",
    "charged_charged",
    "charged_polar",
    "charged_apolar",
    "polar_polar",
    "apolar_polar",
    "apolar_apolar",
    "percent_apolar_nis",
    "percent_charged_nis",
    "binding_affinity_kcal_mol",
    "dissociation_constant_M"
]

rows = []
current_sample = None
current_data = {}

patterns = {
    "intermolecular_contacts": r"No\. of intermolecular contacts:\s+(\d+)",
    "charged_charged": r"No\. of charged-charged contacts:\s+([\d.]+)",
    "charged_polar": r"No\. of charged-polar contacts:\s+([\d.]+)",
    "charged_apolar": r"No\. of charged-apolar contacts:\s+([\d.]+)",
    "polar_polar": r"No\. of polar-polar contacts:\s+([\d.]+)",
    "apolar_polar": r"No\. of apolar-polar contacts:\s+([\d.]+)",
    "apolar_apolar": r"No\. of apolar-apolar contacts:\s+([\d.]+)",
    "percent_apolar_nis": r"Percentage of apolar NIS residues:\s+([\d.]+)",
```

```
    "percent_charged_nis": r"Percentage of charged NIS residues:\s+([\d.]+)",
    "binding_affinity_kcal_mol": r"Predicted binding affinity \(kcal\.mol-1\)
:\s+(-?[\d.]+)",
    "dissociation_constant_M": r"dissociation constant \(M\).*?:\s+([^\s]+)"
}

with open(input_file) as f:
    for line in f:
        line = line.strip()
        if line.startswith("==== "):
            if current_sample and current_data:
                rows.append(current_data)
            filename = re.search(r"==== (.*?) ====", line).group(1)
            sample = re.search(r"fold_(.*?)_cute", filename)
            current_sample = sample.group(1) if sample else filename
            current_data = {"sample": current_sample}
        else:
            for key, pattern in patterns.items():
                match = re.search(pattern, line)
                if match:
                    current_data[key] = match.group(1)

    if current_sample and current_data:
        rows.append(current_data)

#  convert to “;” for Excel :
with open(output_file, "w", newline='') as csvfile:
    writer = csv.DictWriter(csvfile, fieldnames=headers, delimiter=';')
    writer.writeheader()
    for row in rows:
        writer.writerow({key: row.get(key, "") for key in headers})

print(f" Résumé exporté dans : {output_file}")
```