ACCSCIENCE PUBLISHING

Microbes & Immunity

ORIGINAL RESEARCH ARTICLE

Diversity of human papillomavirus-16 L1 in the Asian region: A comparative analysis of sequences

Rana Ozdogan[1], Muharrem Okan Cakir[2], Gholam Hossein Ashrafi[2], and Ugur Bilge[3]*

[1]Department of Molecular Biology and Genetics, Faculty of Life and Natural Sciences, Abdullah Gul University, Kayseri, Turkey

[2]Department of Biomolecular Sciences, School of Life Sciences, Pharmacy and Chemistry, Kingston University London, United Kingdom

[3]Department of Biostatistics and Medical Informatics, Faculty of Medicine, Akdeniz University, Antalya, Turkey

Abstract

Human papillomavirus (HPV) infection is a leading cause of several cancers, with types 16 and 18 classified as high-risk. This study investigates the L1 capsid protein of HPV16, a crucial target for Food Drug and Administration-approved vaccines, by analyzing its nucleic acid and amino acid sequences to identify phylogenetic relationships, sequence variations, and conserved functional elements. The study utilizes sequences obtained from the National Center for Biotechnology Information virus database to assess geographical clustering and evolutionary trends. Key regions, including structural loops, nuclear localization signals, and viral attachment sites, were examined for variability. Phylogenetic analysis demonstrated distinct clustering patterns, particularly among sequences from Japan and Pakistan, indicating potential regional influences on HPV-16 evolution. Despite observed sequence variability, critical residues essential for viral entry remained conserved, suggesting evolutionary constraints on functionally significant domains. These findings offer insights into HPV-16 diversity in Asia and highlight the importance of continuous genomic surveillance for vaccine development and epidemiological assessments.

Keywords: Human papillomavirus; HPV-16 L1; Phylogenetic analysis

***Corresponding author:**
Ugur Bilge
(ubilge@akdeniz.edu.tr)

Citation: Ozdogan R, Cakir MO, Ashrafi GH, Bilge U. Diversity of HPV-16 L1 in the Asian region: A comparative analysis of sequences. *Microbes & Immunity*. 2025;2(2):53-63.
doi: 10.36922/mi.8410

Received: January 4, 2025

Revised: February 17, 2025

Accepted: February 19, 2025

Published online: March 6, 2025

Copyright: © 2025 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Human papillomavirus (HPV) infection is a significant risk factor for both men and women, particularly women, worldwide. HPV is responsible for more than 300,000 cervical cancer-related deaths annually and is one of the most common viral infections categorized under sexually transmitted infections by the World Health Organization (WHO).[1] As of January 2024, there are 451 distinct references to HPV genome entries on papillomavirus episteme, a web-based database serving as an integrated resource for papillomavirus genome sequences.[2,3] Among these, only 12 types (HPV-16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59) have been classified as carcinogenic by the International Agency for Research on Cancer.[4] Of these, HPV-16 and 18 are the most common types,

accounting for nearly seven out of ten cervical cancers reported globally.[5] Further analysis reveals that this ratio remains consistently similar regionally. HPV infection remains a significant public health concern, particularly in Asia, where cervical cancer accounts for a substantial proportion of female cancer-related mortality. According to the WHO reports, over 60% of cervical cancer cases in Asia are linked to HPV-16 and 18, with variations in prevalence across different countries. In Southern and Western Asia, HPV-16 and 18 infections contribute to more than 80% of cervical cancer cases.[5] In addition, HPV-related anogenital and oropharyngeal cancers are increasingly reported among men, with incidence rates varying from 17 to 43 cases per million depending on geographical location.[5] In the Asian region, the rate of cervical cancer due to HPV-16 and 18 infections is 68% among infected women, and in Southern and Western Asia, this rate exceeds 80%.[5] In Asia, HPV-related anogenital and oropharyngeal cancers affect both men and women, with incidence rates ranging from 17 to 43 cases per million. However, the prevalence of HPV-driven cancers remains significantly higher in women due to the strong association between HPV and cervical cancer, which accounts for the majority of HPV-related malignancies.[5]

HPV is a double-stranded DNA virus belonging to the *Papillomaviridae* family, with a genome size of approximately 8 kb.[6] Structurally, HPV consists of 72 capsomeres in an icosahedral formation, lacks an envelope, and has a diameter of 55 nm.[6] Figure 1 shows a schematic representation of the HPV-16 double-stranded circular DNA genome which eight proteins: E1, E2, E4, E5, E6, E7, L1, and L2. The E1, E2, E4, E5, E6, and E7 proteins are categorized as "early" proteins essential for the viral life cycle, including genome replication, cell cycle regulation, and cell growth in specific HPV types. Notably, the E6 and E7 proteins play critical roles in oncogenesis by disrupting the normal functions of p53 and pRb, thereby interfering with key cellular processes.[7,8] The "late" proteins, L1 and L2, form the major and minor capsid proteins, respectively, constructing the virus's capsid structure.[9]

The major capsid protein L1 is important for understanding the antigenicity, diversity, and classification of HPV.[9] HPV types are classified based on nucleotide similarity in their L1 sequences. To name a novel HPV type, its L1 sequence similarity with other virus types must be less than 90%; for novel subtypes, it must be below 98%.[10,11] Despite its rather conservative characteristic, specific sequences in the L1 protein can vary in loops. These surface loops are named BC, CD, DE, EF, FG, and HI loops based on the beta strands that they connect.[3,12,13] Moreover, the lysine residues on FG and HI loops have been identified as regions involved in the initial stages of

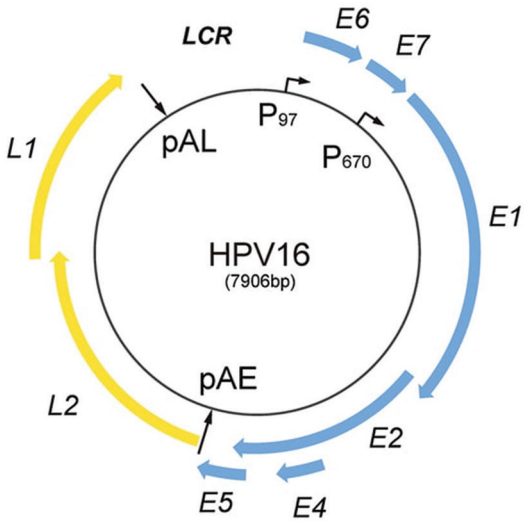

Figure 1. Schematic representation of the human papillomavirus-16 double-stranded circular DNA genome with 7,908 bp long. Early genes, which are E1, E2, E4, E5, E6, and E7, are represented in blue. Late genes, L1 and L2, are represented in yellow. P97 and P670 are promoters, and pAE and pAL are polyadenylation signals, respectively, for early and late regions. LCR is the long control region.[7]

the viral infection cycle, where they interact with heparan sulfate.[14,15] The residues contained within these loops hold significant functional importance.[14]

The L1 major capsid protein self-assembles into virus-like particles (VLPs), and Food Drug and Administration (FDA)-approved HPV vaccines, such as Cervarix®, Gardasil®, and Gardasil®9, are developed based on VLPs.[3,16-19] The currently available U.S. FDA-approved HPV vaccines, including Cervarix®, Gardasil®, and Gardasil®9, are widely used across Asia, though their approval status and distribution policies vary by country. While Japan, China, and South Korea have fully approved all three vaccines, some countries, such as Pakistan, have limited access due to regulatory delays and cost-related challenges. In addition, national immunization programs in certain regions prioritize specific vaccines based on availability and cost-effectiveness considerations. HPV vaccination is recommended primarily for adolescents aged 9 – 14 years as part of national immunization programs, with catch-up vaccinations available for individuals up to 26 years old in some countries. However, vaccine accessibility varies; for instance, Japan and South Korea offer free HPV vaccination under government-funded programs, whereas in countries like Pakistan and India, vaccination is often available only through private healthcare providers, making affordability a significant barrier to widespread uptake. HPV vaccine coverage varies significantly across Asia. Japan and South Korea report high coverage rates exceeding 70%, while coverage remains below 30% in

countries such as Pakistan and India due to logistical and financial constraints.[5] The disparity in vaccination rates underscores the need for enhanced public health efforts to improve accessibility and awareness, particularly in lower-income regions. These particles trigger the immune systems of host organisms by mimicking the actual virus capsid without the viral genome.[19] The quadrivalent Gardasil®, the first approved of these vaccines as 4vHPV, received FDA approval in 2006. This vaccine, formulated with VLPs of HPV-6, 11, 16, and 18 types, targets the most common cancer-causing and genital wart-causing types.[16,20,21] The second HPV vaccine approved by the FDA was the bivalent (2vHPV) Cervarix® vaccine, which protects against HPV-16 and 18.[17] The latest vaccine, Gardasil®9, was approved by the FDA in 2014 and targets nine HPV types, including 6, 11, 16, 18, 31, 33, 45, 52, and 58, expanding on the coverage provided by the quadrivalent Gardasil®.[18]

Given the significant role of L1 proteins in the development of HPV vaccines, this study aims to analyze the L1 gene and protein sequences of HPV-16 in Asia, a region where data availability from the National Center for Biotechnology Information (NCBI) Virus database enables comprehensive analysis. The focus on the Asian region also considers the diverse genetic background of HPV-16, which may influence vaccine efficacy and regional disease prevalence. By investigating these sequences, the study seeks to identify phylogenetic relationships, sequence variations, and conserved elements critical for viral function. Such insights are essential to bridge gaps in understanding HPV16 genetic variability and its implications for vaccine design and global prevention strategies.

2. Materials and methods

This study employed a computational bioinformatics approach, where publicly available nucleotide and protein sequences of HPV16 L1 were retrieved from the NCBI Virus database. These sequences were processed using multiple alignment and phylogenetic analysis tools in the R software environment to assess sequence variability and clustering patterns.

2.1. Selecting and obtaining sequences

On the NCBI Virus page, searching filters were determined; first, for the virus search bar, "Human papillomavirus type 16, taxid: 333760" was entered. Sequence length is limited by nucleotide completeness criteria selected as "complete" to eliminate incomplete sequences. To reach all L1 protein-related entries, paying attention to case sensitivity, these strings were entered in the protein name section: "L1," "L1 Protein," "major capsid protein L1," "major capsid L1 protein," "late protein L1," "L1 capsid protein," "L1 major capsid protein," "late major capsid protein L1," and "Major capsid protein L1." For selection of host organisms, "Homo sapiens (human), taxid: 9606" was applied to the Host bar. "From December 31, 2014, to November 28, 2023" was specified for reaching the sequences obtained during the collection date. Nucleotide and protein sequences of 25 results were downloaded with their year and accession names in FASTA format. The primary reason for selecting the Asian region was the availability of sufficient data in the NCBI Virus database. The studied sequences of HPV-16 (NCBI: txid333760) from human host organisms (NCBI: txid9606) have been collected from the NCBI Virus public database from the Asian region, with selection criteria referencing the year of the latest vaccine approval.[22,23]

The sequences in the NCBI Virus database were analyzed based on continent and country. It was determined that the most appropriate data for the R software code used in the analysis of the sequences were data sourced from Asia. Therefore, for the geographic region, "Asia" was selected.

2.2. Data preprocessing

First, an alignment was performed, and sequences that were significantly shorter and exhibited lower similarity than others were identified. As a result, three sequences with accession codes BDO24687, BEU33870, and BEU33846 were removed from the dataset due to their inability to be effectively analyzed or compared with the remaining sequences.

Attributes of the chosen nucleotide sequences were compiled, as shown in Table 1. This table contains the accession numbers, collection dates of isolates, submitted locations, name of the host organism, name of the isolate, and length of the selected sequences. These pieces of information were obtained from NCBI Virus.

Paralleling the approach utilized in Table 1, relevant information regarding each selected protein sequence was compiled. This information obtained from NCBI Virus included accession number, collection date, location, host organism, isolate name, and sequence length, as presented Table 2.

2.3. Processing data with R

Data processing was carried out in the R software environment, and the code was used by Toparslan *et al.*[24] The Fasta sequences obtained from NCBI Virus were used in the analysis. The outputs of the R script were presented in the form of figures.

3. Results

The results section of this study presents the output images generated during the processing of selected sequences in the R software environment. Residues differing in amino

Table 1. Information of selected nucleotide sequences from NCBI virus

Accession	Collection_Date	Geo_Location	Tissue/Specimen/Source	Isolate	Length
MT783409	2016	China	-		7908
MT783410	2016	China	-		7906
OQ911727	2017	Pakistan	-	HNC49	7171
MW320358	2017	China	-		7909
MH892050	2017	China	-		7905
MK484705	2018	China	-	xuca1916	7912
MZ447800	2018	Pakistan	-	C50	7909
MZ447801	2019	Pakistan	Oronasopharynx	C122	7155
LC718899	2020	Japan	-	21-20-P-002	7905
LC786753	2021	Japan	-	SW0127	7905
LC786755	2021	Japan	-	SW0129	7905
LC786756	2021	Japan	-	SW0131	7905
LC786758	2021	Japan	-	SW0138	7909
LC718895	2021	Japan	-	K3131	7905
LC718897	2021	Japan	-	K5048	7905
LC718898	2021	Japan	-	K5060	7909
LC718900	2021	Japan	-	21-21-P-001	7905
LC718901	2021	Japan	-	21-21-P-007	7904
LC718902	2021	Japan	-	21-21-P-008	7905
LC718903	2021	Japan	-	21-21-P-011	7904
LC786759	2022	Japan	-	SW0142	7905
LC786760	2022	Japan	-	SW0152	7905

Notes: The table contains the nucleotide sequences and their information after the dataset's preprocessing step: NCBI accession number, collection date, geographic location, host, name of isolate, and length of nucleotide sequences, respectively.

acid sequences compared to the reference sequence were identified and discussed in detail. Key residues involved in primary viral attachment were analyzed comprehensively, with a focus on loop structures, critical residues, and nuclear localization signals (NLS) within the L1 sequence. These comparisons facilitated the identification of residue differences among the analyzed sequences.

An unrooted and scaled rectangular phylogram was created for the phylogenetic tree, as shown in Figure 2. Here, the aim was to show the relationships between sequences, highlighting how closely related or divergent the sequences are without implying a common ancestor. At first glance, the most noticeable inference of this phylogram is the presence of two main separate clusters. Within these clusters, the data in the group shown in the lower cluster of Figure 2C appear to be separated with shorter branches from the roots, compared to the upper cluster, as shown in Figure 2B. Available data support the contention that they share a close phylogenetic affiliation. Notably, these clusters exhibited a marked geographic association, with sequences from the same countries predominantly clustering. On reviewing the data presented in Table 1, it is discernible that the sequences featured in this section were submitted by various institutions located in Japan.

The cluster dendrogram in Figure 3 was generated using the distance matrix of sequences. In contrast to the clustering pattern shown in Figure 2, the cluster dendrogram reveals significant differences in branching topology and root positions, especially for the Pakistani sequences (2017-01-|OQ911727, 2018-11-|MZ447800, and 2019-01-|MZ447801). Despite their close relationship with other sequences in the phylogenetic tree, these sequences form a distinct clade. The phylogenetic tree analysis further emphasizes that isolates such as 2016 |MT783409, 2016 |MT783410, 2017-06-05 |MH892050, 2021 |LC786758, 2021 |LC718895, 2021 |LC718897, and 2021 |LC718901 share a common root with the Pakistani sequences. While these sequences appear closely related in the phylogenetic tree (Figure 2B), they distinctly cluster when visualized using the distance matrix-based cluster dendrogram. However, further analysis of the phylogenetic tree revealed that isolates such as 2016 |MT783409, 2016

Table 2. Information of selected protein sequences from NCBI virus

Accession	Collection_Date	Geo_Location	Tissue/Specimen/Source	Isolate	Length
QOI17574	2016	China	-		533
QOI17579	2016	China	-		533
WKC12512	2017	Pakistan	-	HNC49	531
QQL88061	2017	China	-		531
AYV61481	2017	China	-		531
QEG53826	2018	China	-	xuca1916	531
UNF16173	2018	Pakistan	-	C50	531
UNF16181	2019	Pakistan	Oronasopharynx	C122	531
BDO24711	2020	Japan	-	21-20-P-002	531
BEU33838	2021	Japan	-	SW0127	531
BEU33854	2021	Japan	-	SW0129	531
BEU33862	2021	Japan	-	SW0131	531
BEU33878	2021	Japan	-	SW0138	531
BDO24681	2021	Japan	-	K3131	531
BDO24695	2021	Japan	-	K5048	531
BDO24703	2021	Japan	-	K5060	531
BDO24719	2021	Japan	-	21-21-P-001	531
BDO24727	2021	Japan	-	21-21-P-007	531
BDO24735	2021	Japan	-	21-21-P-008	531
BDO24743	2021	Japan	-	21-21-P-011	531
BEU33886	2022	Japan	-	SW0142	531
BEU33894	2022	Japan	-	SW0152	531

Notes: The table contains the nucleotide sequences and their information after the dataset's preprocessing step: NCBI accession number, collection date, geographic location, host, name of isolate, and length of nucleotide sequences, respectively.

|MT783410, 2017-06-05 |MH892050, 2021 |LC786758, 2021 |LC718895, 2021 |LC718897, and 2021 |LC718901 shared a common root with the Pakistani sequences.

In addition, the visualization of the alignment of studied nucleic acid sequences based on the color code for each nucleotide was conducted in the R software environment, as shown in Figure 4. In this alignment, certain conserved guanine/cytosine-dense regions were identified around positions 3979 and 5795 in Figure 4, along with adenine-dense regions around position 1024. Notably, two Pakistani nucleotide sequences, submitted in 2017 and 2019, exhibit extensive terminal gaps within their alignments, as shown in Figure 4. Cross-checking the sequence length data from Table 1 revealed that these two sequences are 742 ± 8 base pairs shorter than the others. These factors likely explain why the 2017-01|OQ911727 and 2019-01|MZ447801 data are located in a different branch than the 2018-11|MZ447800.1 data, as shown in Figure 2B.

Furthermore, a notable observation from Figure 4 involves the sequences named 2016|MT783410.1 and 2016|MT783409.1, submitted from China in 2016. The L1 sequences of these two isolates are distinguished from others by their length. The nucleic acid sequences of these isolates are 7908 bp and 7906 bp, respectively, similar in length to other sequences. However, the protein sequences are 533 amino acids in length, two amino acids longer than the other 20 sequences, which are 531 amino acids in length. These additional methionine and leucine residues at the N-terminus differentiate these sequences.

Sequences with extended terminal gaps, deposited by the same Japanese institution between 2020 and 2022, suggest a common ancestor or experimental origin. These sequences, originating from unpublished studies, exhibit unique alignment features that warrant further investigation.

The circular tree in Figure 5 was created based on the nucleic acid sequences of the samples. In the circular tree, the first noticeable feature is the branch length of 2017-06-15_|MW320358, 2021_|LC718901, and 2021_|LC718895, which are highlighted in blue. Furthermore, as noted in Figure 2C, the data deposited in 2021 (specifically 2021_|LC718900, 2021_|LC786753, 2021_|LC786755, 2022_|LC786759, 2020_|LC718899,

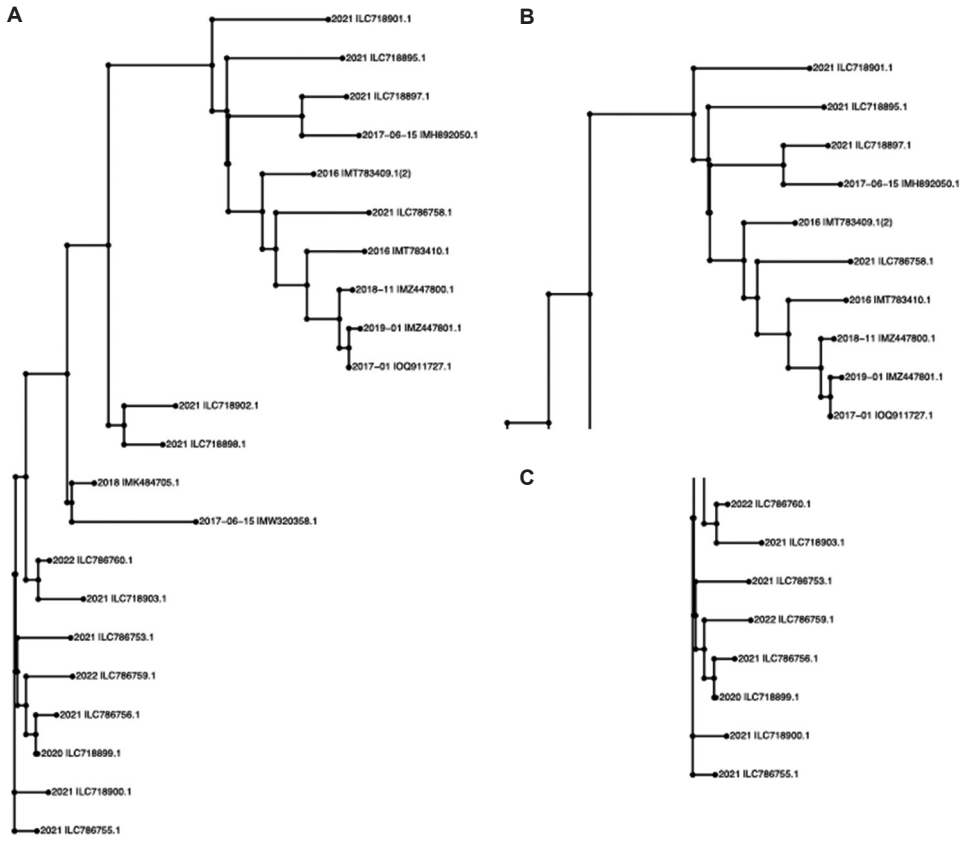

Figure 2. The phylogenetic tree of sequences based on L1 nucleotide sequences. (A) The unrooted rectangular phylogram here is created according to the nucleotide sequences of isolates on the R software. (B) The upper section is the first leading group or cluster of the phylogram. (C) The lower section is the phylogram's second leading group or cluster.

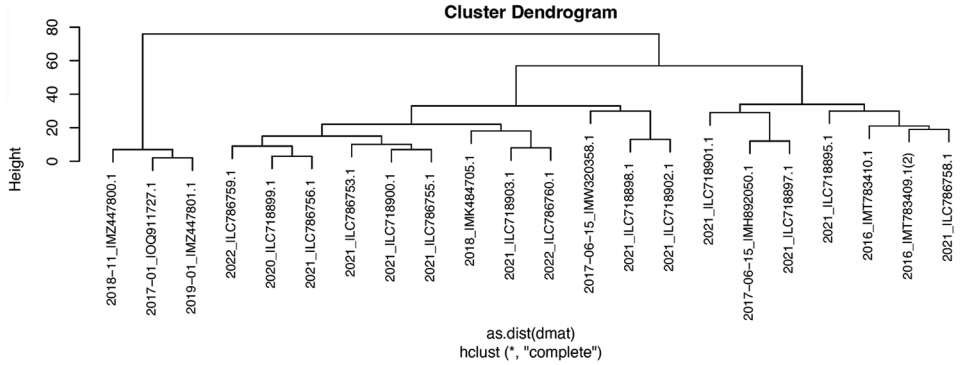

Figure 3. The cluster dendrogram. Cluster dendrogram graph of the nucleotide sequences. All sequences are named CollectionDate_|AccessionNumber. On the upper left of the panel, the height scale of the dendrogram was given. In the bottom-middle of the panel, "as.dist(dmat)" and "hclust (*, "complete")" refer to the clustering algorithms.

2021_|LC718903, 2021_|LC786756, and 2022_|LC786760) exhibit stronger phylogenetic affinities with each other. This observation is supported by the branching lengths and provided color coding.

One of the outputs of the studied sequences is a heat map. The heat map uses a color gradient ranging from dark red (representing maximum distance) to light yellow (indicating decreasing distance), illustrating the pairwise relationships between the sequences. Consistent with observations from Figures 2, 3, and 5, the heat map reveals the presence of two distinct clusters. In addition, sequences originating from Pakistan exhibit

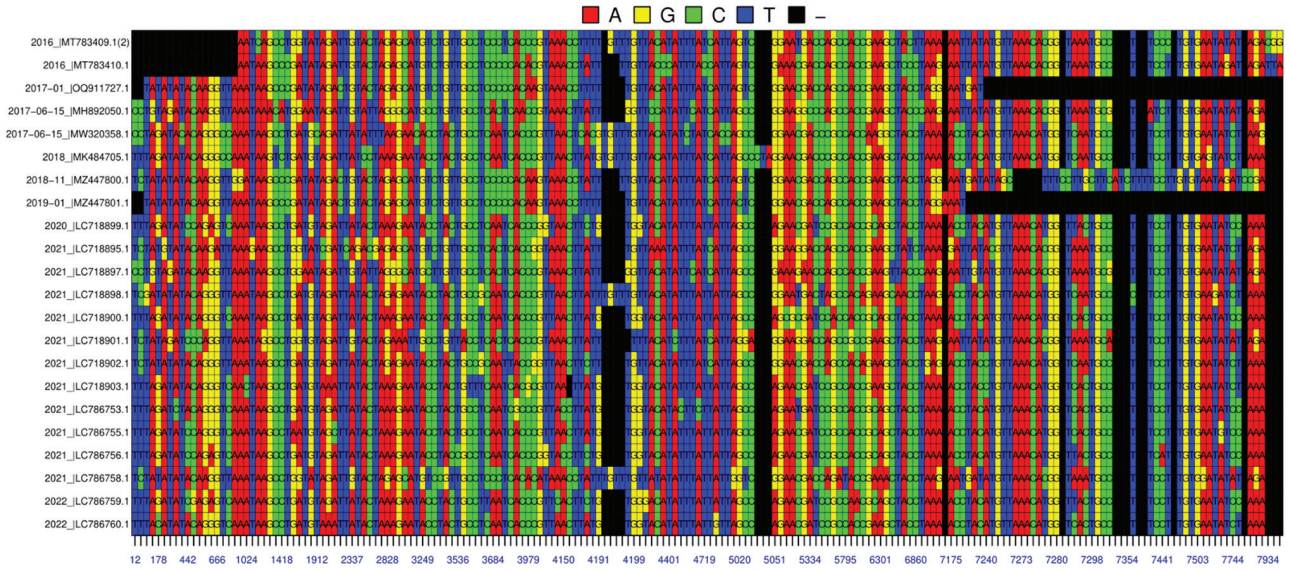

Figure 4. Alignment of the L1 nucleic acid sequences of selected isolates. The graph was colored based on the nucleotides; red = adenine, yellow = guanine, green = cytosine, blue = thymine, and black = missing. Numbers below the alignment table represent the positions of the nucleotides in the sequence. The visualization was performed using custom R scripts to highlight sequence conservation and divergence among isolates from different Asian region in NCBI Virus database.

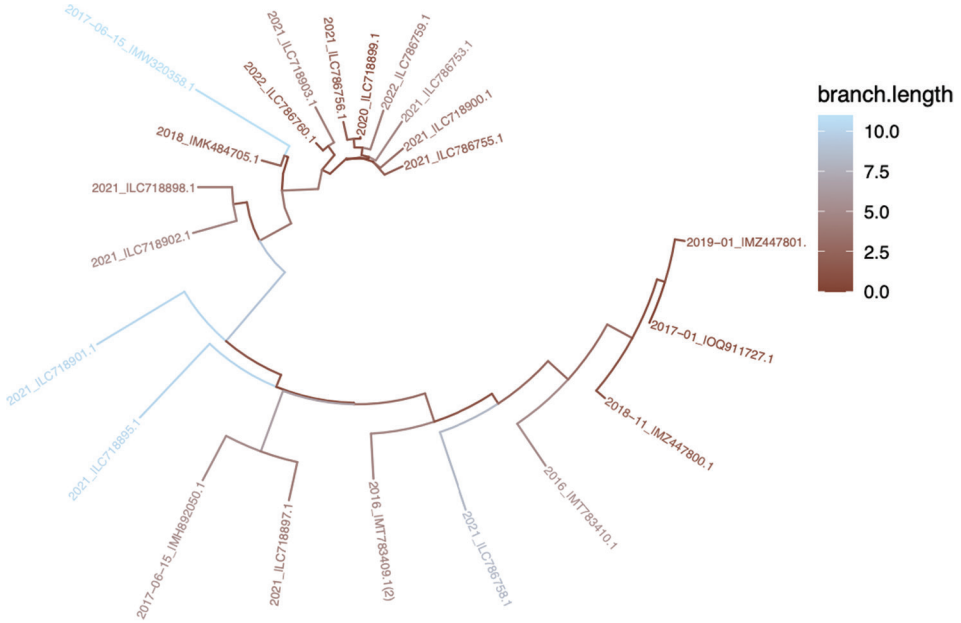

Figure 5. Circular tree based on nucleic acid sequences of L1 protein. Branch lengths are color-scaled, where blue refers to further distance, and brown refers to closer sequences.

a significant degree of similarity among themselves, as evidenced in Figures 2 and 3. Within the two main clusters, sequences demonstrating close phylogenetic affinities include 2021_|LC786755, 2021_|LC718900, 2021_|LC786753, 2021_|LC718903, 2022_|LC786760, 2021_|LC786756, 2020_|LC718899, and 2022_|LC786759. These eight sequences, all uploaded from Japan as indicated in Table 1, highlight the possibility of phylogenetic clustering based on shared geographical origins, similar to the sequences submitted from Pakistan.

The last output, Figure 6 presents a heatmap of the studied L1 sequences with hierarchical clustering. The sequences

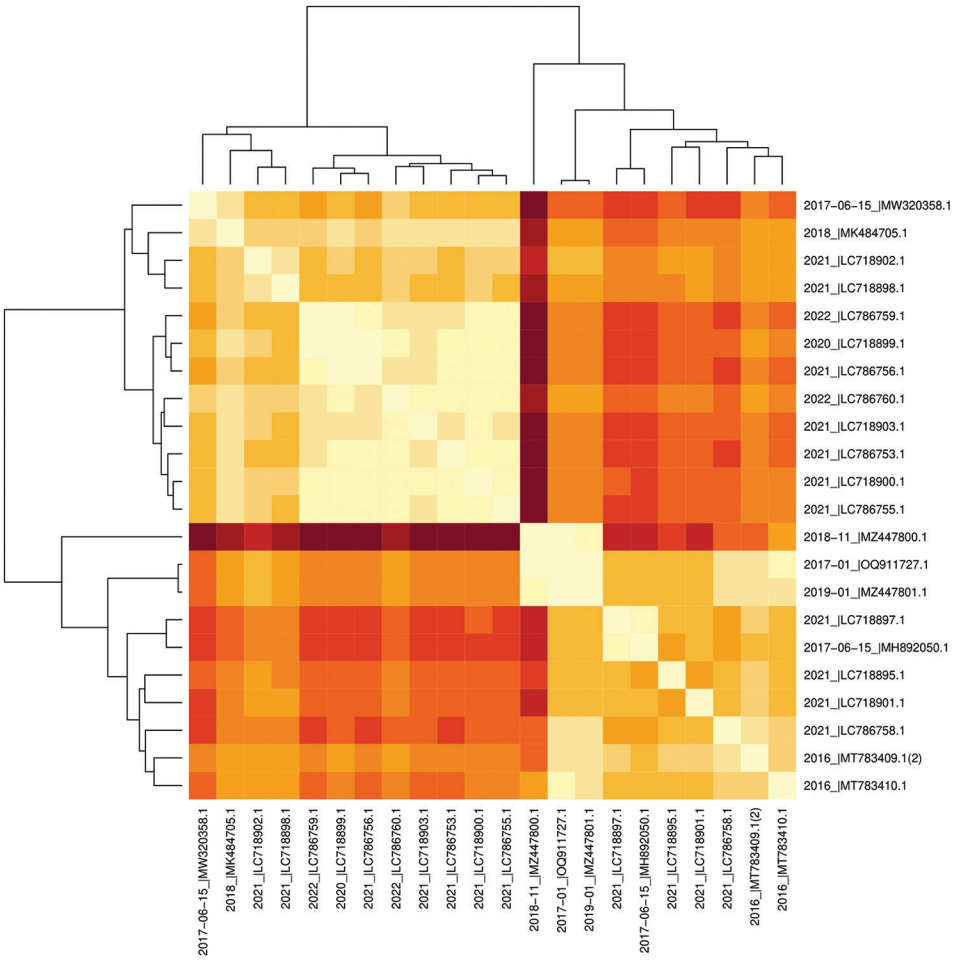

Figure 6. Heatmap of the studied L1 sequences. The heatmap graph of the studied nucleic acid sequences with their clustering shows that each branch of the clusters represents related sequences, nearer sequences are light-colored, and sequences with further distance are indicated with darker colors. Each branch of the clusters represents related sequences. The visualization was performed using custom R scripts to highlight sequence conservation and divergence among isolates from different Asian region in NCBI Virus database.

are organized in dendrograms on both axes, revealing their evolutionary relationships. Two major sequence clusters are evident as observed in the other output results. The upper cluster includes sequences from 2017-06-15_MW320358.1, 2018_IMK484705.1, and multiple 2021-2022 samples (LC718902.1, LC718898.1, LC786759.1, LC718899.1, LC786756.1, LC786760.1, LC718903.1, LC786753.1, LC718900.1, and LC786755.1). The lower cluster contains 2018-11_MZ447800.1, 2017-01_OQ911727.1, 2019-01_MZ447801.1, 2021_LC718897.1, 2017-06-15_MH892050.1, 2021_LC718895.1, 2021_LC718901.1, 2021_LC786758.1, 2016_IMT783409.1, and 2016_IMT783410.1.

Table 3 summarizes residue variability observed among the analyzed protein sequences. Asparagine was identified as the predominant amino acid at position 209, appearing in 19 out of 24 sequences (79.2%). Threonine was present in the remaining five sequences (20.8%). At position 76, a unique substitution from proline to threonine was observed in the protein sequence with accession number BEU33878, distinguishing it from the other 23 sequences. In addition, in the same sequence, a substitution of serine for proline at position 79 was noted. Based on the difference matrix, the proline-to-threonine substitution corresponds to a score of −1, and similarly, the proline-to-serine substitution also has a score of −1.[25] In the protein sequence with accession number QQL88061, a unique substitution of asparagine for aspartic acid was observed at positions 127 and 153, deviating from the consensus of the other 23 sequences. Using the BLOSUM62 substitution matrix, this aspartic acid-to-asparagine substitution has a value of +1.[25] The ratio of Asn209 to Thr209 was found to be 19:5, with five isolates exhibiting threonine at position 209. According to the BLOSUM62 matrix, the substitution of asparagine to threonine at this position is neutral, with a score of 0.[25]

Table 3. Presence inter-sequence variabilities of L1 protein sequences

	Residues				
Accession	76	79	127	153	209
QOI17574	P	P	D	D	N
QOI17579	P	P	D	D	N
WKC12512	P	P	D	D	N
QQL88061	P	P	N*	N*	N
AYV61481	P	P	D	D	N
QEG53826	P	P	D	D	N
UNF16173	P	P	D	D	N
UNF16181	P	P	D	D	N
BDO24711	P	P	D	D	T
BEU33838	P	P	D	D	T
BEU33854	P	P	D	D	T
BEU33862	P	P	D	D	T
BEU33878	T*	S*	D	D	N
BDO24681	P	P	D	D	N
BDO24695	P	P	D	D	N
BDO24703	P	P	D	D	N
BDO24719	P	P	D	D	T
BDO24727	P	P	D	D	N
BDO24735	P	P	D	D	N
BDO24743	P	P	D	D	N
BEU33886	P	P	D	D	T*
BEU33894	P	P	D	D	N

Notes: In this table, protein sequences are compared, and amino acids are represented with one-letter codes. *Indicates residues that are detected in terms of their differences compared to the majority: 75, 77, 127, 153, and 209.

No amino acid substitutions were detected in the NLS region located at the C-terminus of the L1 protein sequence. This region, comprising approximately 22 amino acids, is critical for recruiting karyopherin, which facilitates the entry of the viral genome into the host cell.[26,27]

To further elucidate amino acid variability, the L1 protein sequence was thoroughly compared with the HPV16 reference sequence available on NCBI (NCBI Reference Sequence: NP_041332.2).[28] Particular attention was given to amino acid residues at positions 54, 278, 356, and 361, which play pivotal roles in attachment initiation. Except for 361K, these residues are located within hyper-variable loops as highlighted in the introduction: 54K in the BC loop, 278K in the FG loop, and 356K in the HI loop.[14]

Alignment analysis of the reference sequence revealed that the critical residues at positions 278, 356, and 361 retained their original lysine (K) composition. Thus, it can be concluded that lysine is conserved at these essential positions, emphasizing their functional importance.

4. Discussion

This study analyzed the L1 sequences of HPV-16, focusing on the major capsid protein at both the protein and nucleic acid levels. Sequences were sourced from a public database covering the period following the approval of HPV vaccines in Asia. The primary objective was to investigate sequence variations and assess their potential implications for viral entry mechanisms and host interactions, particularly in binding with heparan sulfate.

The phylogenetic analysis revealed that sequences in Figure 2C, uploaded from Japan between 2020 and 2022, exhibited close relationships, suggesting a potential common infectious source or a shared evolutionary lineage. These findings are supported by the clustering of sequences with phylogenetic similarities. However, this hypothesis is limited by the geographic expansion of isolates and the lack of detailed metadata on the individuals from whom the isolates were collected. Therefore, while these observations suggest a shared evolutionary trajectory; further, validation through epidemiological and molecular studies is required.

Analysis of residue variability provided additional insights. Substitutions from proline to threonine or serine, observed in specific sequences, were found to retain the hydrophilic properties of the R groups, as indicated by their Kyte and Doolittle hydropathy indices (−1.6 for proline, −0.7 for threonine, and −0.6 for serine).[29] Despite the negative substitution scores in the BLOSUM62 matrix (−1), which suggest these changes are less likely to occur during evolutionary events, their presence indicates potential structural or functional adaptations.[25] Conversely, the substitution of aspartic acid to asparagine, with a positive BLOSUM62 score (+1), was more common and expected. This substitution maintained identical hydropathy indices of −3.5, preserving the hydrophilic nature of the residues.[25,29]

The substitution from asparagine to threonine at position 209, observed in five isolates (accessions BDO24711, BEU33838, BEU33838, BEU33862, and BDO24719), further highlighted evolutionary dynamics. Although this substitution is neutral in the BLOSUM62 matrix (score = 0), the significant difference in hydropathy indices (−3.5 for asparagine and −0.7 for threonine) suggests a potential functional impact on protein interactions or stability. While both residues are hydrophilic, the reduced hydrophilicity of threonine may influence its interaction with surrounding molecules or the overall capsid structure.

Migratory movements, particularly from conflict-affected and resource-limited regions, may contribute to the spread and genetic diversification of HPV strains. Large-scale population displacements can introduce distinct viral variants into new geographical areas, potentially altering HPV epidemiology. Studies on viral transmission dynamics suggest that increased migration-driven genetic exchange may impact HPV stability, necessitating continuous genomic surveillance to monitor emerging sequence variations that could affect vaccine efficacy.

Overall, the findings from this study underscore the importance of specific residues in maintaining the structural and functional integrity of the L1 protein. The observed variations, particularly those with significant hydropathy index differences or positive substitution scores, may have implications for viral entry mechanisms and host interactions. Future studies incorporating structural modeling and functional assays are essential to validate these findings and elucidate their biological significance. In addition, expanding metadata collection for uploaded sequences would enhance the ability to correlate phylogenetic relationships with clinical and epidemiological data, providing a more comprehensive understanding of HPV-16 evolution and its implications for vaccine efficacy and disease prevention.

5. Conclusion

This study highlights the genetic diversity and evolutionary trends of the HPV-16 L1 capsid protein in Asia. Despite regional sequence variations, critical residues necessary for viral entry remain highly conserved, reinforcing their functional importance. Phylogenetic clustering revealed distinct grouping among sequences from Japan and Pakistan, emphasizing the need for region-specific genomic surveillance. Our findings contribute to a deeper understanding of HPV-16 evolution and have potential implications for vaccine efficacy and public health strategies. Future research should focus on functional assays and structural modeling to further assess the impact of identified variations.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Ugur Bilge, Hossein Ashrafi, Muharrem Okan Cakir
Formal analysis: Ugur Bilge, Rana Ozdogan
Investigation: Ugur Bilge, Rana Ozdogan, Muharrem Okan Cakir
Methodology: Ugur Bilge, Rana Ozdogan, Muharrem Okan Cakir
Writing – original draft: Rana Ozdogan
Writing – review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data are available on request from corresponding author.

References

1. World Health Organization (WHO). *Sexually Transmitted Infections* (*STIs*). Available from: https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis) [Last accessed on 2023 Nov 28].
2. *The Papillomavirus Episteme*. Available from: https://pave.niaid.nih.gov/index [Last accessed on 2023 Nov 23].
3. Nelson CW, Mirabello L. Human papillomavirus genomics: Understanding carcinogenicity. *Tumour Virus Res.* 2023;15:200258.
   doi: 10.1016/j.tvr.2023.200258
4. International Agency for Research on Cancer (IARC). *List of Classifications. IARC Monographs on the Identification of Carcinogenic Hazards to Humans*. Available from: https://monographs.iarc.who.int/list-of-classifications [Last accessed on 2023 Nov 28].
5. Bruni L, Albero G, Serrano B, *et al*. *Human Papillo-mavirus and Related Diseases in the World. ICO/IARC Information Centre on HPV and Cancer* (*HPV Information Centre*); 2023. Available from: https://hpvcentre.net [Last accessed on 2023 Nov 28].
6. Goetschius DJ, Hartmann SR, Subramanian S, Bator CM, Christensen ND, Hafenstein SL. High resolution Cryo EM analysis of HPV16 identifies minor structural protein L2 and describes capsid flexibility. *Sci Rep.* 2021;11:3498.
   doi: 10.1038/s41598-021-83076-5
7. Kajitani N, Schwartz S. Role of viral ribonucleoproteins in human papillomavirus type 16 gene expression. *Viruses*. 2020;12(10):1110.
   doi: 10.3390/v12101110

8. Pešut E, Đukić A, Lulić L, *et al*. Human papillomaviruses-associated cancers: An update of current knowledge. *Viruses*. 2021;13(11):2234.

doi: 10.3390/v13112234

9. El Aliani A, El Abid H, Kassal Y, *et al*. HPV16 L1 diversity and its potential impact on the vaccination-induced immunity. *Gene*. 2020;747:144682.

doi: 10.1016/j.gene.2020.144682

10. Arroyo Mühr LS, Lagheden C, Hassan SS, Eklund C, Dillner J. The international human papillomavirus reference center: Standardization, collaboration, and quality assurance in HPV research and diagnostics. *J Med Virol*. 2023;95(12):e29332.

doi: 10.1002/jmv.29332

11. Cosper PF, Bradley S, Luo L, Kimple RJ. Biology of HPV mediated carcinogenesis and tumor progression. *Semin Radiat Oncol*. 2021;31(4):265-273.

doi: 10.1016/j.semradonc.2021.02.006

12. 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. *Mol Cell.* 2000;5:557-567.

doi: 10.1016/s1097-2765(00)80449-9

13. Liu X, Chen J, Wang Z, *et al*. Neutralization sites of human papillomavirus-6 relate to virus attachment and entry phase in viral infection. *Emerg Microbes Infect*. 2019;8:1721-1733.

doi: 10.1080/22221751.2019.1694396

14. Mukherjee AG, Wanjari UR, Gopalakrishnan AV, *et al*. Exploring the molecular pathogenesis, pathogen association, and therapeutic strategies against HPV infection. *Pathogens*. 2022;12(1):25.

doi: 10.3390/pathogens12010025

15. Bissett SL, Godi A, Beddows S. The DE and FG loops of the HPV major capsid protein contribute to the epitopes of vaccine-induced cross-neutralizing antibodies. *Sci Rep.* 2016;6:39730.

doi: 10.1038/srep39730

16. Food and Drug Administration. *Cervarix* (*Human Papillomavirus Bivalent* [*Types 16 and* 18 *Vaccine, Recombinant*)*: Pre-scribing Information* [*Package Insert*]. US Department of Health and Human Services, Food and Drug Administration; 2016. Available from: https://www.fda.gov/media/78013/download [Last accessed on 2023 Nov 23].

17. Food and Drug Administration. *Gardasil* (*Human Papillomavirus Quadrivalent* [*Types 6, 11, 16, and* 18 *Vaccine, Recombinant*)*: Prescribing Information* [*Package Insert*]. US Department of Health and Human Services, Food and Drug Administration; 2015. Available from: https://www.fda.gov/media/74350/download [Last accessed on 2023 Nov 23].

18. Food and Drug Administration. *Gardasil 9* (*Human Papillomavirus 9-Valent Vaccine, Recombinant*)*: Prescribing Information* [*Package Insert*]. US Department of Health and Human Services, Food and Drug Administration; 2018. Available from: https://www.fda.gov/media/90064/download [Last accessed on 2023 Nov 23].

19. Markowitz LE, Schiller JT. Human papillomavirus vaccines. *J Infect Dis*. 2021;224:S367-S378.

doi: 10.1093/infdis/jiaa621

20. Cheng L, Wang Y, Du J. Human papillomavirus vaccines: An updated review. *Vaccines* (*Basel*). 2020;8(3):391.

doi: 10.3390/vaccines8030391

21. Akhatova A, Azizan A, Atageldiyeva K, *et al*. Prophylactic human papillomavirus vaccination: From the origin to the current state. *Vaccines* (*Basel*). 2022;10:1912.

doi: 10.3390/vaccines10111912

22. Schoch CL, Ciufo S, Domrachev M, *et al*. NCBI taxonomy: A comprehensive update on curation, resources and tools. *Database* (*Oxford*). 2020;2020:baaa062.

doi: 10.1093/database/baaa062

23. *NCBI Virus*. Bethesda, MD: National Library of Medicine (US), National Center for Biotechnology Information; 2004. Available from: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?viruslineage_ss=viruses,taxid:10239&seqtype_s=nucleotide [Last accessed on 2024 Oct 28].

24. Toparslan E, Karabağ K, Bilge U. A workflow with R: Phylogenetic analyses and visualizations using mitochondrial cytochrome b gene sequences. *PLoS One*. 2020;15(12):e0243927.

doi: 10.1371/journal.pone.0243927

25. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*. 1992;89:10915-10919.

doi: 10.1073/pnas.89.22.10915

26. Zhou J, Doorbar J, Sun XY, Crawford LV, McLean CS, Frazer IH. Identification of the nuclear localization signal of human papillomavirus type 16 L1 protein. *Virology*. 1991;185:625-632.

doi: 10.1016/0042-6822(91)90533-h

27. Lai KY, Rizzato M, Aydin I, Villalonga-Planells R, Drexler HCA, Schelhaas MA Ran-binding protein facilitates nuclear import of human papillomavirus type 16. *PLoS Pathog*. 2021;17(5):e1009580.

doi: 10.1371/journal.ppat.1009580

28. National Center for Biotechnology Information (NCBI). Available from: https://www.ncbi.nlm.nih.gov [Last accessed on 2024 Dec 08].

29. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol.* 1982;157:105-132.

doi: 10.1016/0022-2836(82)90515-0