Article

Not peer-reviewed version

Diversity of HPV 16 L1 in the Asian Region: A Comparative Analysis of Sequences

Rana Ozdogan , Muharrem Okan Cakir , Gholam Hossein Ashrafi , Ugur Bilge *

Posted Date: 3 March 2024

doi: 10.20944/preprints202403.0074.v1

Keywords: Human Papillomavirus, HPV 16, L1, Phylogenetic analysis

Article

Diversity of HPV 16 L1 in the Asian Region: A Comparative Analysis of Sequences

Rana Özdoğan [1], Muharrem Okan Çakır [2], G Hossein Ashrafi [2] and Uğur Bilge [3,*]

[1] Department of Molecular Biology and Genetics Faculty of Life and Natural Sciences, Abdullah Gul University, Kayseri, Turkey
[2] Kingston University, London, United Kingdom
[3] Department of Biostatistics and Medical Informatics, Faculty of Medicine, Akdeniz University, Antalya, Turkey
* Correspondence: ubilge@akdeniz.edu.tr

Abstract: It has been shown that Human papillomavirus (HPV) infection is associated with several forms of cancer. Additionally, although there are more than 100 types of HPV, more than ten types of it, including types 16 and 18, are considered high-risk. The two proteins that make up the viral capsid of HPV 16 are L1, the major capsid protein, and L2, the minor capsid protein. FDA (The Food and Drug Administration) approved HPV vaccines mostly target L1 and L2 capsid proteins, which facilitate intracellular entry during infection. In this study, it was set out to analyze the creation of phylogenetic relations of L1 nucleic acid sequences of viral isolates. This study also analyzed sequences of three key elements within the amino acid sequences of HPV 16 L1: loops critical for structural stability, nuclear localization signal sequences, and residues implicated in viral attachment. As highlighted by the literature, these elements represent pivotal aspects of viral function. Furthermore, the report highlighted the particular inter-sequence variability within the HPV 16 L1 protein sequences. Overall analysis revealed clustering of sequences primarily due to geographical and characterized by more phylogenetic relatedness within location-specific clusters.

Keywords: human papillomavirus; HPV 16; L1; phylogenetic analysis

BACKGROUND

Human papillomavirus (HPV) infection is a major risk factor for both men and women, especially women, around the world. HPV, which is the cause of more than 300,000 cervical cancer-related deaths every year, is one of the most common viral infections in the sexually transmitted infections (STI) category of the World Health Organization (WHO) [1]. As of January 2024, there exist 451 distinct references to Human papillomavirus genome entries on Papillomavirus Episteme (PaVE), a web-based database that serves as an integrated resource for papillomavirus genome sequences [2,3]. Altogether there are some XX types are documented, and some of them are probably carcinogenic, only twelve of these types (HPV 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59) have been highlighted as carcinogenic by the International Agency for Research on Cancer [4]. Two of them, HPV-16 and 18 are the most common types, and almost seven out of ten cervical cancers reported worldwide develop as a result of infections of these two types [5]. Upon conducting a thorough analysis of more specified locations, it has been observed that this ratio remains consistently similar. The rate of cervical cancer due to HPV 16 and 18 infections was stated as 68% in women in the Asian region, and even in the regions of Southern and Western Asia, this rate was recorded as more than 80% [5]. It has also been reported that HPV-related anal, laryngeal, oropharyngeal, and oral cavity cancer types also have a fair incidence among men [5].

Human papillomavirus (HPV) is a dsDNA virus belonging to the papillomaviridae family and its genome size is indicated about 8kb [6]. Structurally, it is a virus with 72 capsomeres in an icosahedral structure, without an envelope, 55 nm tall. Its genome encodes 8 proteins; E1, E2, E4, E5, E6, E7, L1, and L2 [6]. The E1, E2, E4, E5, E6, and E7 proteins, which are referred to as "Early" proteins, are involved in vital cycle functions such as genome replication, cell cycle and regulation, cell growth

 doi:10.20944/preprints202403.0074.v1

and differentiation [7]. Also, the oncogenicity of specific HPV types relies on the act of E6 and E7 proteins on the p53 and pRb by interfering with the normal functions [8]. The "late" proteins, on the other hand, are involved in the formation of the major capsid protein, L1, and the minor capsid protein, L2, which forms the capsid structure of the virus [9].

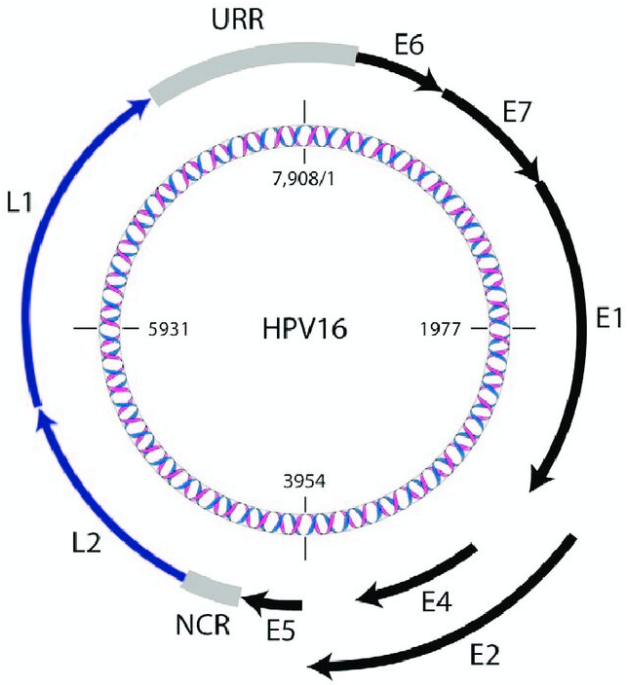


Figure 1. Schematic representation of the HPV-16 double-stranded circular DNA genome with 7,908 bp long. Early genes are represented with black which are E1, E2, E4, E5, E6, E7. Late genes are represented with dark blue which are L1 and L2. URR refers to the Upstream Regulatory Region and NCR refers to Non-Coding Region. [7].

The major capsid protein, L1, is a protein with a highly conserved sequence of about 55 kDa [9]. Due to this characteristic, HPV types are classified according to the similarity rates in L1 sequences. In the process of assigning a name to a novel HPV type, it is required that the L1 sequence similarity be less than 90% and this score rises to 98% for novel subtypes [10,11]. Despite its rather conservative characteristic, there are certain sequences in the L1 protein where there are also variable sequences. These surface loops are named BC, CD, DE, EF, FG, and HI loops based on the beta strands they connect [12–14]. Moreover, the lysine residues located on FG and HI loops have been identified as regions involved in the initial stages of the viral infection cycle, where they interact with heparan sulfate. [15,16]. The residues contained within these loops hold immense significance.

The L1 major capsid protein self-assembles into virus-like particles (VLPs) and FDA-approved HPV vaccines such as Cervarix®, Gardasil®, and Gardasil®9, are developed based on VLPs [12,17–20]. These particles trigger the immune systems of host organisms by mimicking the actual virus capsid without the viral genome [20]. The quadrivalent Gardasil®, which was 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 whose infection results in the common cancerous and genital wart-causing types, is used for prevention against these four types [17,21,22]. The second HPV vaccine approved by the FDA was the bivalent (2vHPV) Cervarix® vaccine which provided protection against HPV types 16 and 18 [17]. The last one, Gardasil®9 was approved for use by the FDA in 2014. This vaccine was developed against nine human papillomavirus types, including 6, 11, 16, 18, 31, 33, 45, 52, and 58, unlike the quadrivalent Gardasil® [19].

This study aims to analyze the L1 gene and protein sequences, which are the major capsid proteins of the HPV16 type. The studied sequences of HPV type 16 (NCBI:txid333760) from human host organism (NCBI:txid9606) have been collected from the NCBI Virus public database from the Asian region and selected referring to the year when the last vaccine was approved [23,24].

 doi:10.20944/preprints202403.0074.v1

METHOD

Selecting and Obtaining Sequences

In the NCBI-Virus page, searching filters was determined, first for the virus search bar "Human papillomavirus type 16, taxid: 333760" was entered. In order to eliminate incomplete sequences, sequence length is limited by nucleotide completeness criteria which were selected as "complete". To reach all L1 protein-related entries, paying attention to the case sensitivity, these strings were entered on 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 the geographic region "Asia" was selected. For selection of host organisms, "Homo sapiens (human), taxid: 9606" was applied to the Host bar. For reaching the date of sequences obtained, to the collection date, "**From** Dec 31, 2014 **To** Nov 28, 2023" is specified. Nucleotide and protein sequences of 25 results were downloaded with their year and accession names in fasta format.

Data PreProcessing

First alignment was performed and sequences that were meaningless compared to other sequences and had data that could not be analyzed and compared with them, 3 sequences with BDO24687, BEU33870, and BEU33846 accession codes for the protein, were removed from the data set. The remaining sequences with their information are entered in a table format.

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

Table 1. Information of Selected Nucleotide Sequences from NCBI-Virus. The table contains the nucleotide sequences and their information after the preprocessing step of the dataset. NCBI accession number, collection date, geographic location, host, name of isolate, and length of nucleotide sequences, respectively.

Accession	Collection_Date	Geo_Location	Host	Isolate	Length	Accession
MT783409	2016	China	Homo sapiens		7908	MT783409
MT783410	2016	China	Homo sapiens		7906	MT783410
OQ911727	2017	Pakistan	Homo sapiens	HNC49	7171	OQ911727
MW320358	2017	China	Homo sapiens		7909	MW320358
MH892050	2017	China	Homo sapiens		7905	MH892050
MK484705	2018	China	Homo sapiens	xuca1916	7912	MK484705
MZ447800	2018	Pakistan	Homo sapiens	C50	7909	MZ447800
MZ447801	2019	Pakistan	Homo sapiens	C122	7155	MZ447801
LC718899	2020	Japan	Homo sapiens	21-20-P-002	7905	LC718899
LC786753	2021	Japan	Homo sapiens	SW0127	7905	LC786753
LC786755	2021	Japan	Homo sapiens	SW0129	7905	LC786755
LC786756	2021	Japan	Homo sapiens	SW0131	7905	LC786756
LC786758	2021	Japan	Homo sapiens	SW0138	7909	LC786758
LC718895	2021	Japan	Homo sapiens	K3131	7905	LC718895
LC718897	2021	Japan	Homo sapiens	K5048	7905	LC718897
LC718898	2021	Japan	Homo sapiens	K5060	7909	LC718898

 doi:10.20944/preprints202403.0074.v1

LC718900	2021	Japan	Homo sapiens	21-21-P-001	7905	LC718900
LC718901	2021	Japan	Homo sapiens	21-21-P-007	7904	LC718901
LC718902	2021	Japan	Homo sapiens	21-21-P-008	7905	LC718902
LC718903	2021	Japan	Homo sapiens	21-21-P-011	7904	LC718903
LC786759	2022	Japan	Homo sapiens	SW0142	7905	LC786759
LC786760	2022	Japan	Homo sapiens	SW0152	7905	LC786760

Paralleling the approach utilized in Table 1, a second table was constructed to compile relevant information regarding each selected protein sequence. This information obtained from NCBI-Virus included accession number, collection date, location, host organism, isolate name, and sequence length, referring to Table 2.

Table 2. Information of Selected Protein Sequences from NCBI-Virus. The table contains the protein sequences and their information after the preprocessing step of the dataset. NCBI accession number, collection date, geographic location, host, name of isolate, and length of protein sequences, respectively.

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

 doi:10.20944/preprints202403.0074.v1

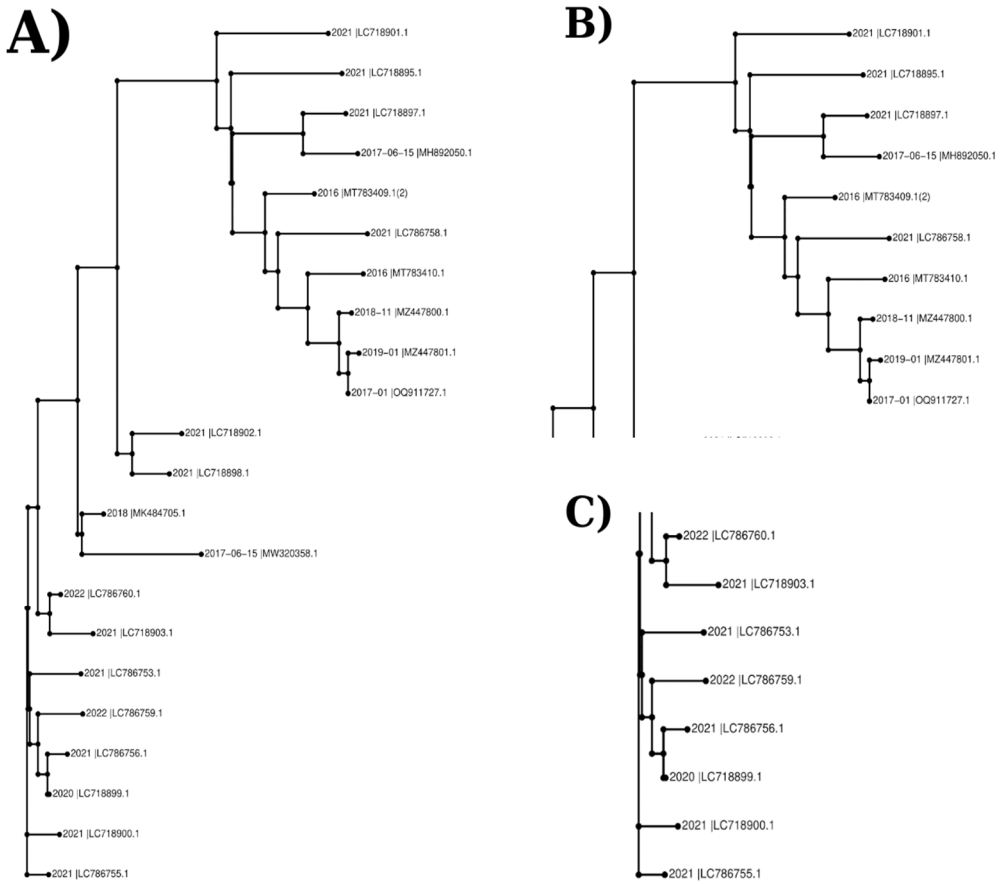
Processing Data with R

Data processing was carried out in R software environment and the code was used from Toparslan et. al. [25]. The fasta sequences obtained from NCBI-Virus were used in the analysis. The resulting images were obtained as the output of the R script. These visuals include cluster dendrogram, heat-map created based on distance matrix, phylogenetic tree, B-tree circular tree with branch length scale. In addition, it outputs colored alignment images created based on the loaded nucleic acid and protein sequences in PDF format. In the alignment image of nucleic acid sequences, when a sequence does not have a corresponding nucleotide, it is indicated in black.

RESULTS

The result section of the study will provide output images from the processing of selected sequences in the R software environment. Residues that changed in amino acid sequences compared to the reference sequence were presented in a table. In addition, the crucial residues that facilitate primary viral attachment were also scrutinized, it was also aimed to compare loop structures and their associated critical residues, and nuclear localization signal sequences located in the L1 sequence. This comparison was conducted to identification of potential residue dissimilarities between the studied sequences.

An unrooted and scaled rectangular phylogram was created for the phylogenetic tree, as shown in Figure 2. 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 Figure 2C, which can be called the lower cluster, appear to be separated from each other with the shorter branches from the roots, compared to the upper cluster in Figure 2B. Available data supports 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 together. Upon reviewing the data presented in Table 1, it is discernible that the sequences featured in this section were submitted by various individuals and institutions located in Japan.



 doi:10.20944/preprints202403.0074.v1

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)** Upper section, first main group or cluster, of the phylogram. **C)** Lower section, second main group or cluster, of the phylogram.

The cluster dendrogram in Figure 3 was created using the distance matrix of sequences. The cluster dendrogram of nucleotide sequences was also revealed mainly divided into 2 clusters. In contrast to the clustering pattern observed in Figure 2, Figure 3 reveals a notable discrepancy regarding the branching topology and root position of specific sequences. While the Pakistani sequences (*2017-01-_|OQ911727*, *2018-11-_|MZ447800*, and *2019-01-_|MZ447801* submitted to NCBI from Pakistan in 2017, 2018, and 2019), depicted in Figure 2B appear proximal to other sequences within the phylogenetic tree, they form a distinct clade when visualized through the distance matrix-based cluster dendrogram. However, when the phylogenetic tree was analyzed, the isolates named *2016 |MT783409, 2016 |MT783410, 2017-06-05 |MH892050, 2021 |LC786758, 2021 |LC718895, 2021 |LC718897 and 2021 |LC718901* possessed a common root with the samples submitted from Pakistan.

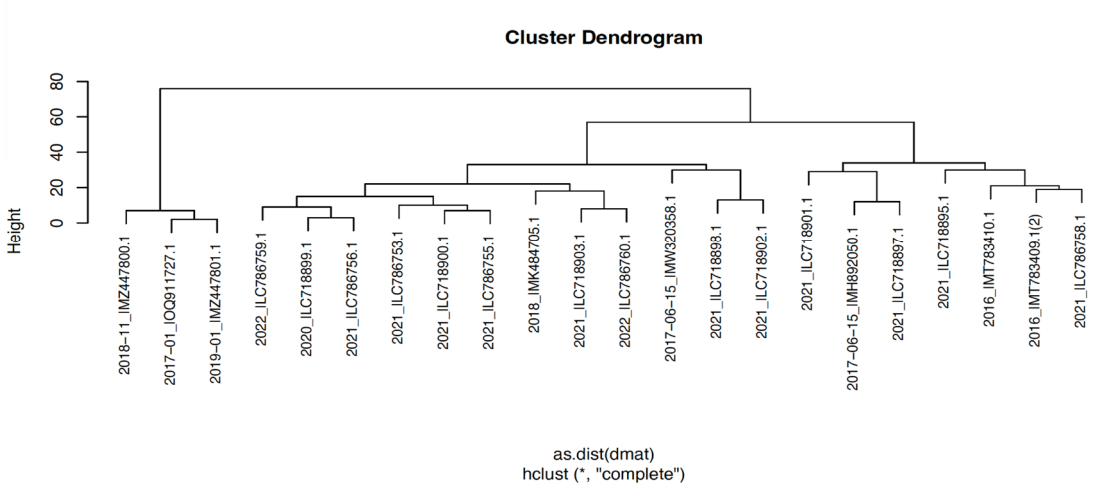


Figure 3. The Cluster Dendrogram. Cluster dendrogram graph of the nucleotide sequences. All sequences are named *CollectionDate_|AccessionNumber*. On the upper left of the panel, the height scale of the dendrogram was given. In the bottom-middle of the panel "as.dist(dmat)" and "hclust (*, "complete")"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

Additionally, 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, referring to Figure 4. In this alignment, there are certain conserved guanine/cytosine-dense, around the indicated position at 3979 and 5795 in Figure 4, and adenine-dense, around the indicated position at 1024, regions within sequences. It is noteworthy that two Pakistani nucleotide sequences, submitted in 2017 and 2019, present extensive terminal gaps within their alignments, Figure 4. Upon cross-checking of guided-by sequence length data presented in Table 1, it has been revealed that these two particular sequences are 742 ±8 base pairs shorter than the others. These factors could offer an explanation for why the *2017−01_|OQ911727* and *2019−01_|MZ447801* data are located in a different branch than the *2018−11 |MZ447800.1* data, as data clearly shown in Figure 2B.

Furthermore, a relevant observation from Figure 4 was made on sequences that were named *2016 |MT783410.1* and *2016 |MT783409.1* were submitted from China in 2016. There is one point that separates the L1 sequences of these two isolates from the other; the length of the sequences. The nucleic acid sequences of isolates 7908bp and 7906bp are of similar length to the other sequences, respectively. However, because the protein sequences are 533 amino acid lengths, they are 2 amino acids longer, with additional methionine and leucine residues at the N-terminus, than the other 20 sequences, which are 531 amino acid lengths.

 doi:10.20944/preprints202403.0074.v1

The sequences under discussion, the presence of extended terminal missing in two specific sequences, were deposited from the same institution in Japan, between 2020 and 2022. Notably, these sequences originated from studies associated with multiple unpublished articles and could have suggested a potential shared source or experimental origin.

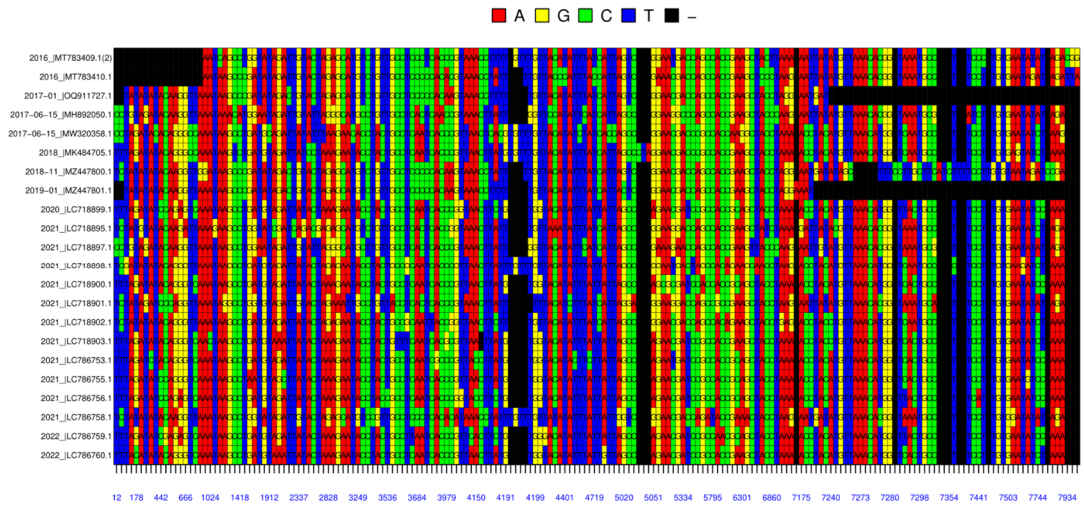


Figure 4. Alignment of the L1 Nucleic Acid Sequences of Selected Isolates. The graph was colored based on the nucleotides; the Red - Adenine, Yellow– Guanine, Green – Cytosine, Blue – Thymine, and Black indicate a missing. Numbers below the alignment table represent the positions of the nucleotides in the sequence.

The circular tree, in Figure 5, was constructed based on the nucleic acid sequences of the samples. In the circular tree, the first noticeable point is the branch length of *2017-06-15_|MW320358*, *2021 |LC718901*, and *2021 |LC718895*, which also were indicated with blue color. Moreover, In Figure 5, as mentioned above in the Figure 2C, the data deposited in 2021 (*specifically 2021_|LC718900, 2021_|LC786753, 2021_|LC786755, 022_|LC786759, 2020_|LC718899, 2021_|LC718903, 2021_|LC786756, 2022_|LC786760)* were observed to have stronger phylogenetic affinities with each other. This observation was made in the light of branching lengths and provided colors.

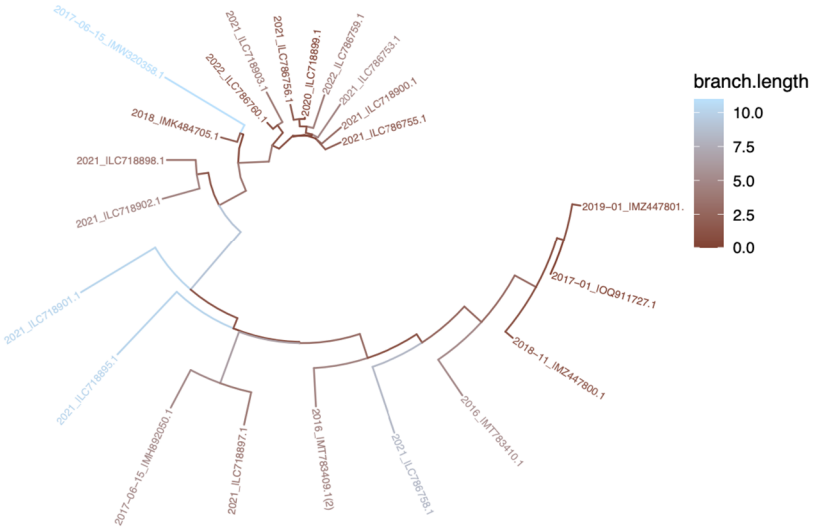


Figure 5. Circular Tree Based on Nucleic Acid Sequences of L1 Protein. Branch lengths are colour-scaled where blue refers to further distance, and brown refers to closer sequences.

 doi:10.20944/preprints202403.0074.v1

Visual output of the studied sequences yielded a heat map visualization, constructed through R code processing. The heat map employed a color gradient ranging from dark red (representing maximum distance) to light yellow (indicating decreasing distance), effectively illustrating the pairwise relationships between the sequences. Consistent with previous observations that can be demonstrated from Figure 2, Figure 3, and Figure 5 so far appeared similarly from the heat-map; the presence of two distinct clustering. In addition to these two main clusters, the sequences that originated from Pakistan also have a significant degree of similarity. among themselves, as evidenced by Figures 2 and 3. Among the two main clusters, the sequences whose phylogenetic affinity to each other can be clearly observed indicate the same sequences as the previously indicated in Figure 2C which were *2021_|LC786755, 2021_|LC718900, 2021_|LC786753, 2021_|LC718903, 2022_|LC786760, 2021_|LC786756, 2020_|LC718899,* and *2022_|LC786759*. The only common point of all these 8 sequences showing a close phylogenetic relationship with each other is that they were uploaded from Japan according to the information checked from Table 1, so again, the possibility of phylogenetic clustering based on the shared geographical origin of sequences can be highlighted as in the sequences submitted from Pakistan.

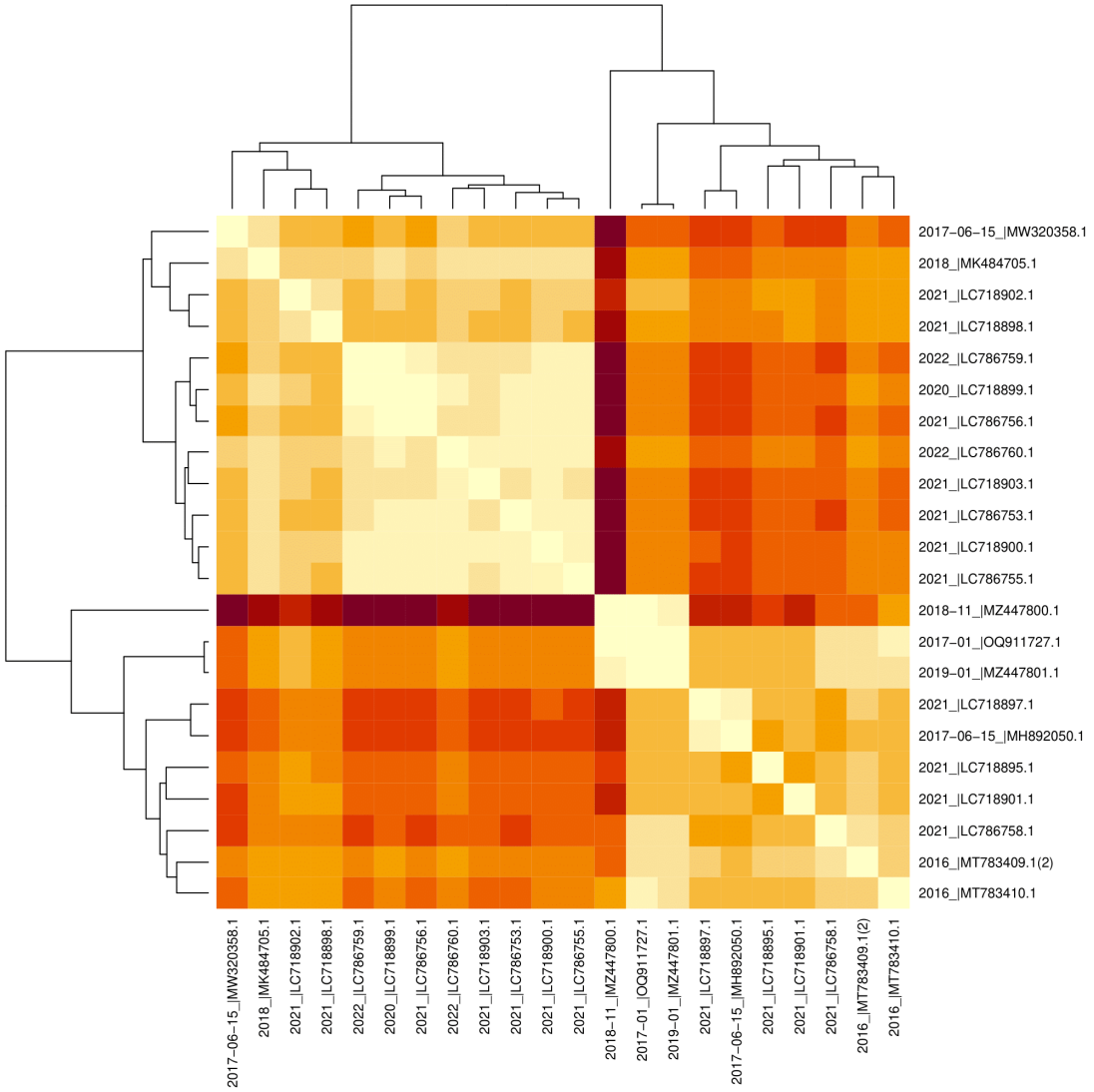


Figure 6. Heatmap of the Studied Sequences. The heatmap graph of the studied nucleic acid sequences with their clustering that each branch of the clusters represents related sequences, nearer sequences are light-colored, and further distance is indicated with darker colors Each branch of the clusters represents related sequences.

 doi:10.20944/preprints202403.0074.v1

In Table 3, residue variabilities between the studied protein sequences were combined in a table. From this table, it was found that Asparagine was the most common amino acid at position 209, occurring in 19 out of 24 sequences (79.2%). Threonine was the second most common amino acid, occurring in the remaining 5 sequences (20.8%).

In the amino acid residue at position 76, a change from Proline to Threonine was observed in the protein sequence with BEU33878 accession, unlike the other 23 sequences. In the same sequence, again unlike the other 23 sequences, Serine was observed instead of Proline at the 79 locations. Assuming, the Pro-to-Thr substitution is equivalent to the value of -1 found in the difference matrix. Similarly, if we apply the theory of substitution, the Pro-to-Ser difference would also be -1. [26]. The protein sequence with accession number QQL88061 shows asparagine instead of aspartic acid at positions 127 and 153, which is unlike the other 23 sequences. Using the BLOSUM62 matrix, it can be seen that this change from Asp-to-Asn has a substitution that would have a value of +1 [26]. With a ratio of Asn209:Thr209 at 19:5, five isolates exhibit threonine protein location at 209. This particular difference is valued at zero when assuming a substitution from Asn-to-Thr according to the BLOSUM62 [26].

Table 3. Presence Inter-sequence Variabilities of Protein Sequences. In this table, protein sequences were compared with each other, and amino acids were represented with their one-letter codes. 5 residues were detected in terms of their differences compared to the majority and these are; 75, 77, 127, 153, and 209.

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**

 doi:10.20944/preprints202403.0074.v1

BEU33894	P	P	D	D	N

In addition, amino acid substitutions were not observed in the nuclear localization signal region, located at the C terminus of the L1 protein sequence include about 22 amino acid, which is involved in the recruitment of karyopherin that mediates taking the viral genome into the host cell [27,28].

In our research, we also conducted a thorough comparison of the L1 protein sequence with the HPV 16 type reference sequence obtained from NCBI (NCBI Reference Sequence: NP_041332.2) [29]. Our main area of focus was on the amino acid residues at specific locations that play a crucial role in initiating attachment which are at 54, 278, 356, and 361, in addition, each of these specific points, except 361K, is located in the hypervariable loops mentioned in the introduction section. 54K in BC loop, 278K in FG loop, and 356K in HI loop [15,30]. Upon conducting an alignment analysis of the reference sequence, it was discerned that the vital residues at positions 278, 356, and 361 retained their original Lysine amino acid composition. Hence, it can be inferred that Lysine remained present at these crucial locations.

DISCUSSION

This research delved into the L1 sequences of HPV 16, specifically focusing on the major capsid protein, at both the protein and nucleic acid levels. The sequences were sourced from a public database that covered the period since the approval of HPV vaccines in Asia. The primary objective was to discern and investigate any changes between these sequences. Furthermore, a comparative analysis was conducted to check the significant positions within the reference sequence that play a crucial role in the entry of virions and subsequently bind with heparan within the host organism.

The phylogenetic relationships of the sequences in the section in Figure 2C are closer to each other, and all of these sequences were uploaded from Japan for the years 2020-2021-2022. The sequences uploaded to NCBI may have a common infectious connection since they are from the same country and have a phylogenetic common ancestor. However, this will stay a quite weak argument because of the expansion of geographic location. Also, it is hard to evaluate and discuss this argument numerically, and it is impossible to be sure without obtaining detailed information from the individuals from whom the isolates were collected.

Table 1 indicates that substitutions from Proline to Threonine or Serine, which carry a score of -1 in BLOSUM62, do not result in a loss of hydrophilic properties conferred by the R groups. According to the Kyte & Doolittle Scale, Proline, Threonine, and Serine exhibit hydropathy indexes of -1.6, -0.7, and -0.6, respectively [31]. Also, because of their negative BLOSUM62 score, it could be stated that these substitutions are unlikely to transpire during evolutionary events [26,30]. Another observation regarding Table 1 is the substitution from Asp-to-Asn, which boasts a BLOSUM62 score of +1, indicating its more expected or high likelihood occurrence [26,32]. Moreover, the two amino acids possess identical hydropathy indices of -3.5 on the Kyte & Doolittle scale and are highly hydrophilic [31].

Table 1 reveals that all five isolates, with accession BDO24711, BEU33838, BEU33838, BEU33862, and BDO24719, displayed a substitution Asn-to-Thr at position 209. Although this change falls under the category of expected change with a BLOSUM62 matrix score of 0, the significant variance in the hydropathy index between Asn (-3.5) and Thr (-0.7) on the hydropathy index scale implies that the substitution may have a discernible effect [26,32]. Despite both amino acids having hydrophilic properties, this distinction in the hydropathy index indicates that the substitution may have an impact.

CONCLUSION

In conclusion, the residues that have been required for the primary attachment and entry mechanisms of the L1 protein, the predominant capsid protein of the HPV 16 virus, have been found to be preserved in the sequences we have examined. Analysis of amino acid variations revealed no noteworthy changes at critical amino acid residues within the studied sequences.

 doi:10.20944/preprints202403.0074.v1

The study was conducted in a wide area, the Asian region. Nevertheless, owing to the complexity of comprehending the interactions between the individuals from whom these sequences were obtained and interpreting the infection line of the virus is challenging. The sole detectable observation is that the sequences from Japan have a shared ancestor, as depicted in Figure 2. While a comparable deduction can be drawn for isolates from Pakistan, it cannot be substantiated with only three isolates. Notably, clustering primarily reflects geographical proximity, suggesting potential regional influences on sequence similarity in nucleic acid and protein sequences. It is important to note that the conclusion stems from theoretical investigations involving protein and nucleic acid sequence data.

References

1. Sexually transmitted infections (STIs). (2023, July 10). World Health Organization (WHO). https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis)
2. The Papillomavirus Episteme. (Accessed November 2023). Papillomavirus Episteme. Retrieved from https://pave.niaid.nih.gov/index
3. Van Doorslaer, K., Li, Z., Xirasagar, S., Maes, P., Kaminsky, D., Liou, D., … McBride, A. A. (2016). The Papillomavirus Episteme: a major update to the papillomavirus sequence database. Nucleic Acids Research, 45(D1), D499–D506. doi:10.1093/nar/gkw879
4. List of classifications. (2023, May 12). IARC Monographs on the Identification of Carcinogenic Hazards to Humans – INTERNATIONAL AGENCY FOR RESEARCH ON CANCER. https://monographs.iarc.who.int/list-of-classifications
5. Bruni L, Albero G, Serrano B, Mena M, Collado JJ, Gómez D, Muñoz J, Bosch FX, de Sanjosé S. ICO/IARC Information Centre on HPV and Cancer (HPV Information Centre). Human Papillomavirus and Related Diseases in the World. Summary Report 10 March 2023. [November 2023]
6. Graham, S. V. (2010). Human papillomavirus: gene expression, regulation and prospects for novel diagnostic methods and antiviral therapies. Future Microbiology, 5(10), 1493–1506. doi:10.2217/fmb.10.107
7. Smith, B., Chen, Z., Reimers, L., van Doorslaer, K., Schiffman, M., DeSalle, R., … Burk, R. D. (2011). Sequence Imputation of HPV16 Genomes for Genetic Association Studies. PLoS ONE, 6(6), e21375. doi:10.1371/journal.pone.0021375
8. Mu¨nger, K., Baldwin, A., Edwards, K. M., Hayakawa, H., Nguyen, C. L., Owens, M., Grace, M., & Huh, K. (2004). Mechanisms of human papillomavirus-induced oncogenesis. Journal of Virology, 78(21), 11451-11460. https://doi.org/10.1128/jvi.78.21.11451-11460.2004
9. Finnen, R. L., Erickson, K. D., Chen, X. S., & Garcea, R. L. (2003). Interactions between Papillomavirus L1 and L2 Capsid Proteins. Journal of Virology, 77(8), 4818–4826. doi:10.1128/jvi.77.8.4818-4826.2003
10. Burk, R. D., Harari, A., & Chen, Z. (2013). Human papillomavirus genome variants. Virology, 445(1-2), 232–243. doi:10.1016/j.virol.2013.07.018
11. Burd E. M. (2003). Human papillomavirus and cervical cancer. *Clinical microbiology reviews*, *16*(1), 1–17. https://doi.org/10.1128/CMR.16.1.1-17.2003
12. Nelson, C. W., & Mirabello, L. (2023). Human papillomavirus genomics: Understanding carcinogenicity. *Tumour Virus Research*, *15*, 200258. https://doi.org/10.1016/j.tvr.2023.200258
13. Chen, X. S., Garcea, R. L., Goldberg, I., Casini, G., & Harrison, S. C. (2000). Structure of Small Virus-like Particles Assembled from the L1 Protein of Human Papillomavirus 16. Molecular Cell, 5(3), 557–567. doi:10.1016/s1097-2765(00)80449-9
14. Liu, X., Chen, J., Wang, Z., Wang, D., He, M., Qian, C., Song, S., Chi, X., Kong, Z., Zheng, Q., Wang, Y., Yu, H., Zhao, Q., Zhang, J., Li, S., Gu, Y., & Xia, N. (2019). Neutralization sites of human papillomavirus-6 relate to virus attachment and entry phase in viral infection. *Emerging Microbes & Infections*, *8*(1), 1721-1733. https://doi.org/10.1080/22221751.2019.1694396
15. Knappe, M., Bodevin, S., Selinka, H.-C., Spillmann, D., Streeck, R. E., Chen, X. S., … Sapp, M. (2007). Surface-exposed Amino Acid Residues of HPV16 L1 Protein Mediating Interaction with Cell Surface Heparan Sulfate. Journal of Biological Chemistry, 282(38), 27913–27922. doi:10.1074/jbc.m705127200
16. Bissett, S. L., Godi, A., & Beddows, S. (2016). The DE and FG loops of the HPV major capsid protein contribute to the epitopes of vaccine-induced cross-neutralising antibodies. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep39730
17. Food and Drug Administration. (2016). Cervarix (human papillomavirus bivalent [types 16 and 18] vaccine, recombinant): Prescribing information [package insert]. Silver Spring, MD: US Department of Health and Human Services, Food and Drug Administration. https://www.fda.gov/media/78013/download
18. Food and Drug Administration. (2015). Gardasil (human papillomavirus quadrivalent [types 6, 11, 16, and 18] vaccine, recombinant): Prescribing information [package insert]. Silver Spring, MD: US Department of

Health and Human Services, Food and Drug Administration. https://www.fda.gov/media/74350/download
19. Food and Drug Administration. (2018). Gardasil 9 (human papillomavirus 9-valent vaccine, recombinant): Prescribing information [package insert]. Silver Spring, MD: US Department of Health and Human Services, Food and Drug Administration. https://www.fda.gov/media/90064/download
20. Markowitz, L. E., & Schiller, J. T. (2021). Human Papillomavirus Vaccines. *The Journal of infectious diseases*, *224*(12 Suppl 2), S367–S378. https://doi.org/10.1093/infdis/jiaa621
21. Braaten, K. P., & Laufer, M. R. (2008). Human Papillomavirus (HPV), HPV-Related Disease, and the HPV Vaccine. *Reviews in obstetrics & gynecology*, *1*(1), 2–10.
22. Akhatova, A., Azizan, A., Atageldiyeva, K., Ashimkhanova, A., Marat, A., Iztleuov, Y., Suleimenova, A., Shamkeeva, S., & Aimagambetova, G. (2022). Prophylactic human papillomavirus vaccination: From the origin to the current state. *Vaccines*, *10*(11), 1912. https://doi.org/10.3390/vaccines10111912
23. *Schoch CL, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). 2020: baaa062.* PubMed: 32761142 PMC: PMC7408187.
24. Hatcher, E. L., Zhdanov, S. A., Bao, Y., Blinkova, O., Nawrocki, E. P., Ostapchuck, Y., Schäffer, A. A., & Brister, J. R. (2017). Virus Variation Resource - improved response to emergent viral outbreaks. *Nucleic acids research*, *45*(D1), D482–D490. https://doi.org/10.1093/nar/gkw1065
25. REF Toparslan E., KARABAĞ K., BİLGE U., A workflow with R: Phylogenetic analyses and visualizations using mitochondrial cytochrome b gene sequences, PLOS ONE, vol.15, pp.12, 2020 https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0243927 15.12.2020
26. Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(22), 10915–10919. https://doi.org/10.1073/pnas.89.22.10915
27. Zhou, J., Doorbar, J., Xiao Yi Sun, Crawford, L. V., McLean, C. S., & Frazer, I. H. (1991). Identification of the nuclear localization signal of human papillomavirus type 16 L1 protein. Virology, 185(2), 625–632. doi:10.1016/0042-6822(91)90533-h
28. Nelson, L. M., Rose, R. C., & Moroianu, J. (2002). Nuclear Import Strategies of High Risk HPV16 L1 Major Capsid Protein. Journal of Biological Chemistry, 277(26), 23958–23964. doi:10.1074/jbc.m200724200
29. National Center for Biotechnology Information (NCBI)[Internet]. National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2024 Dec 08]. Available from: https://www.ncbi.nlm.nih.gov/
30. Dasgupta, J., Bienkowska-Haba, M., Ortega, M. E., Patel, H. D., Bodevin, S., Spillmann, D., … Chen, X. S. (2010). Structural Basis of Oligosaccharide Receptor Recognition by Human Papillomavirus. Journal of Biological Chemistry, 286(4), 2617–2624. doi:10.1074/jbc.m110.160184
31. Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology, 157(1), 105–132. doi:10.1016/0022-2836(82)90515-0
32. Mount, D. W. (2008). Using BLOSUM in Sequence Alignments. Cold Spring Harbor Protocols, 2008(6), pdb.top39–pdb.top39. doi:10.1101/pdb.top39