The Papillomavirus Episteme: a major update to the papillomavirus sequence database

Koenraad Van Doorslaer[1], Zhiwen Li[2], Sandhya Xirasagar[2], Piet Maes[3], David Kaminsky[2], David Liou[2], Qiang Sun[2], Ramandeep Kaur[2], Yentram Huyen[2] and Alison A. McBride[1,*]

[1]DNA Tumor Virus Section, Laboratory of Viral Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 209892, USA, [2]Bioinformatics and Computational Biosciences Branch, Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 209892, USA and [3]KU Leuven, Department of Microbiology and Immunology, Laboratory for Clinical Virology, Rega Institute for Medical Research, 3000 Leuven, Belgium

ABSTRACT

The Papillomavirus Episteme (PaVE) is a database of curated papillomavirus genomic sequences, accompanied by web-based sequence analysis tools. This update describes the addition of major new features. The papillomavirus genomes within PaVE have been further annotated, and now includes the major spliced mRNA transcripts. Viral genes and transcripts can be visualized on both linear and circular genome browsers. Evolutionary relationships among PaVE reference protein sequences can be analysed using multiple sequence alignments and phylogenetic trees. To assist in viral discovery, PaVE offers a typing tool; a simplified algorithm to determine whether a newly sequenced virus is novel. PaVE also now contains an image library containing gross clinical and histopathological images of papillomavirus infected lesions. Database URL: https://pave.niaid.nih.gov/.

INTRODUCTION

Episteme is derived from the ancient Greek noun ἐπιστμη (knowledge or science) and verb ἐπσταμαι (to know). This definition symbolizes our goal to provide accurate information about papillomaviruses (PVs) in order to understand their genomic structure, perform comparative genomics analyses and to assist in the understanding and treatment of papillomavirus-associated diseases. Thus, the database was named The Papillomavirus Episteme (or PaVE) (1).

The *Papillomaviridae* are a family of circular viruses with a double-stranded DNA genome of about 7.9 kb. A typical papillomavirus genome encodes seven to eight proteins.

Most highly conserved are the E1 and E2 regulatory proteins (2,3) and the L1 and L2 structural proteins (4,5). In addition to this core set of proteins, most viruses encode for E6 and E7 proteins, which optimize the cellular milieu for the viral life cycle (6,7). Certain viruses also express small, hydrophobic proteins that are designated E5 (8) or E10 (9). The highly divergent, yet highly expressed, E4 protein is imprinted within the E2 open reading frame (ORF) (10). The Upstream Regulatory Region (URR), otherwise known as the Long Control Region (LCR), is located between the L1 and E6 ORFs, and contains cis- responsive elements involved in transcription and replication (11,12). Taxonomic classification of papillomaviruses is based on nucleotide similarity across the L1 ORF (13). The family *Papillomaviridae* contains 49 genera, each of which is further divided into several species. To be designated as a new type, an individual PV type cannot share >90% similarity to any other known PV type (14,15). Individual types have been shown to be highly species and tissue specific. To date, papillomaviruses have been described in fish (16), reptiles (17,18), birds (19) and mammals (20). Based on the suggestion that papillomaviruses have coevolved with their hosts (21), it appears that papillomaviruses have been an evolutionary success for over 500 million years.

Since its inception, PaVE has become highly respected within the papillomavirus community. An important aim of the PaVE database is to provide clinicians, epidemiologists and bench scientists with a uniform data source, thereby facilitating cross pollination among these different, yet complementary research areas (1). Here, we describe major additions and updates to the PaVE database, new analysis tools, and an enhanced web-based user interface.

IMPROVEMENTS

Software details

The PaVE utilizes open platform web technologies, and is deployed on a server running Apache Tomcat and a MySQL database. The underlying code is written in Java, and uses Hibernate database queries with BioJava. The web graphical user interface is developed using the Google Web Toolkit and jQuery. Details including versions of these platforms and visualization and analysis tools are described in Supplementary Table S1.

Enhanced interface design

Since its initial deployment, numerous new tools and resources have been added to PaVE. To accommodate these additions, and simplify navigation, the web page interface was redesigned (Figure 1). Users can navigate the database via 'Search', 'Analyse' or 'Explore' pages. Integration of advanced analysis tools (multiple sequence alignment, BLAST, phylogenetic tools) with visualization tools (sequence tree, protein structure and image viewers) in the Search feature, enables users to perform highly targeted searches and analysis for DNA and protein sequences. The 'Analyse' page gives direct access to the Multiple Sequence Alignment, Protein Structure Viewer and L1 Taxonomy tools. 'Explore' lists extensive information related to the functional genomics of papillomaviruses, as shown in Figure 1.

Database content

The initial description of the PaVE database included 241 completely annotated viral genomes. The PaVE database is updated four times per year. An automated GenBank search identifies putative novel papillomavirus types not currently present in PaVE. The current PaVE database contains 347 annotated papillomavirus genomes, 3562 genes and regions, 3215 protein sequences and 57 protein structures, which users can explore, analyse or download. To be recognized as a novel papillomavirus type, a viral genome has to fulfill a strict set of requirements (for details see (14,15)): (i) The entire viral genome must be cloned; (ii) the L1 sequence cannot share >90% nucleotide sequence identity with its closest neighbour and (iii) the cloned genome must be submitted to, and reviewed by, the International Human Papillomavirus Reference Centre (22). However, because of recent advances in Next-Generation sequencing, several novel genomes have been described (for a review see (13)) that do not meet all these requirements, and will therefore not be recognized as novel viral types by the International Human Papillomavirus Reference Centre. In order to reflect the true diversity of papillomaviruses, PaVE includes viruses that meet the '90% sequence identity' rule, even when the additional requirements are not met. At the time of publication, 20 of the 347 viral genomes in PaVE are considered to be non-reference genomes. Throughout the PaVE database, the official named viral types are indicated with the suffix 'REF'. Non-reference viruses are identified by the appendix 'nr' (e.g. HPV-mXXXXnr; where XXXX reflects the original name given by the submitting authors).

All PaVE tools contain a filter to enable inclusion or exclusion of these 'non-reference' viruses. When a non-reference genome becomes recognized by the reference centre (i.e. meets all inclusion criteria), the non-reference genome will be transferred to the reference genome category in PaVE.

Manual editing of viral genomes

For most viral types, the automated annotation process has proven to be highly accurate. The resulting annotations are manually verified to ensure that they include all expected ORFs, based on papillomavirus biology and phylogenetic classification of the virus. Finally, a multiple sequence alignment of all homologous viral ORFs is used to verify that the proteins are of similar length, and that no frame-shifts or deletions are present within the ORF. In certain cases, viral genome sequences are manually edited to restore a viral ORF (see (1) for details). Genomes that have been to date are shown in Supplementary Table S2. Edited genomes are noted in the PaVE database and links to the original GenBank record are provided.

Expanded annotation of viral genomes

As was described in the initial description of the PaVE database (1), viral ORFs are identified by a BLASTp search against a custom database. In addition, current scientific knowledge is applied into the annotation process and genomes are manually curated, if necessary. For example, the L1 protein is usually translated from a spliced mRNA, with a highly conserved splice acceptor motif just upstream of the L1 ORF [YYNYAG(A)TG; Y represents T or C]. In addition to ORFs and associated proteins, PaVE also annotates conserved spliced transcripts (see next paragraph for more details).

Splice site prediction is performed using the validated ASSP tool (23). The potential list of splice donor/acceptor pairs are filtered based on certain assumptions. These assumptions are the result of extensive comparative genomics among well studied papillomaviruses. For example, the E4 portion of the E1^E4 mRNA is transcribed from the +1 reading frame of E2. Similarly, the E8 fragment of the E8^E2 protein is contained within the E1 +1 frame. Both E1^E4 and E8^E2 utilize the same splice acceptor site located within E2.

The E8^E2 spliced transcript

Many papillomaviruses express a spliced transcript that encodes an E8^E2 fusion protein (3,24). This protein consists of a short peptide from an ORF designated E8 (overlapping the E1 ORF) fused to the C-terminal domain of E2. The resulting proteins function as repressors of viral transcription and replication (25). In the latest version of the annotation pipeline, we have incorporated the algorithm described by Puustusmaa and Abroi, to more accurately identify the E8 portion of the E8^E2 coding regions (24).

The E6* transcript

The E6*I transcript is produced by internal splicing within the E6 ORF. Within the so-called high-risk Alphapapillomaviruses the splice donor site is highly conserved. The
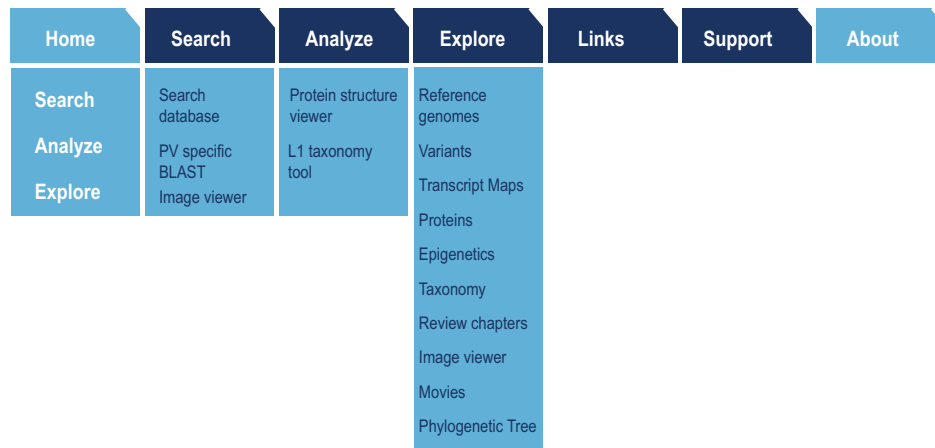
Figure 1. Information Architecture of the PaVE Web Resource. Architecture of the redesigned PaVE web interface, showing the three main ways to access the data and tools using the concepts of 'Search, 'Analyse' and 'Explore'.

spliced transcript encodes a protein that contains the first CxxC zinc-binding motif of E6 with the addition of a few amino acids of variable non-conserved sequence encoded by sequence following the acceptor site. To date, the E6*I spliced mRNA has not been described in other viral species or genera (reviewed in (7)). Currently, PaVE has annotated only experimentally validated E6*I proteins (26). Efforts to use sequence homology to aid in the prediction of additional E6*I proteins are underway.

The E5 proteins

The E5 proteins are short, hydrophobic, transmembrane proteins (8). E5 genes are expressed from the region located between the E2 and L2 ORFs, however sometimes more than one hydrophobic E5-like protein can be encoded by this region (27). Different E5 protein share little sequence similarity. E5 proteins derived from Alphapapillomaviruses have been named in PaVE as E5-alpha, E5-beta, E5-gamma and E5-delta according to the scheme suggested by Bravo and Alonso (27). Using a similar naming scheme, PaVE has added E5-epsilon and E5-zeta.

The E10 proteins

In addition to the E5 proteins expressed from the region located between the E2 and L2, a few viruses encode hydrophobic 'E5 like' proteins expressed from an ORF that either overprints the E6 ORF, or is encoded in this region in the absence of an E6 gene. These proteins have been named either E5, or E8, which has led to confusion because of the existence of the E8 exon (derived from the E1 ORF) that generates the widely encoded E8^E2 protein. In July 2016, key members of the papillomavirus community decided to name hydrophobic 'E5-like' peptides encoded in the 5′ early region, E10. These changes have been implemented in the PaVE database, and an informational page describing the E10 protein has been added. E10 proteins encoded by the bovine Xipapillomaviruses were originally designated E8 (28), and are now renamed E10. SfPV1 contains an ORF within E6 that is similar to the HPV and BPV1 E5

proteins and was previously designated E8 (29); it is also renamed E10. Finally, members of Gammapapillomavirus species 6 (HPV101, 103 and 108) lack a canonical E6 ORF, but potentially encode a highly hydrophobic (E5-like) protein from an overlapping open reading frame, now also designated E10 (9).

Cis-elements

In addition to viral ORFs, the PaVE annotates cis-regulatory elements, such as the Upstream Regulatory Region (URR) and DNA binding sites for the viral E2 regulatory protein [the relaxed ACCN(6)GGT consensus (3) is used].

Circular genome browser

The PaVE locus viewer shows a linear representation of the viral genome by default. However, PaVE also offers the ability to visualize the circular genome (Figure 2). This interactive feature allows powerful navigation and analysis of individual viral features. The genes encoding for spliced transcripts can be toggled on, or off.

Multiple sequence alignment and derivation of phylogenetic trees

The analysis of a Multiple Sequence Alignment (MSA) can be used to predict functional residues and conserved motifs, aid in primer design and 3D structure prediction (30). Sequence alignments can be used to infer evolutionary history and visualized in a phylogenetic tree. The PaVE user interface can be used to select homologous protein sequences derived from viral types of interest (Figure 3). At the click of a button, PaVE will calculate either a multiple sequence alignment, or a neighbour-joining phylogenetic tree. While several high quality multiple sequence aligners are available, the COBALT aligner is used to construct the alignments (31). COBALT leverages information present in several databases (conserved domain database (32) and
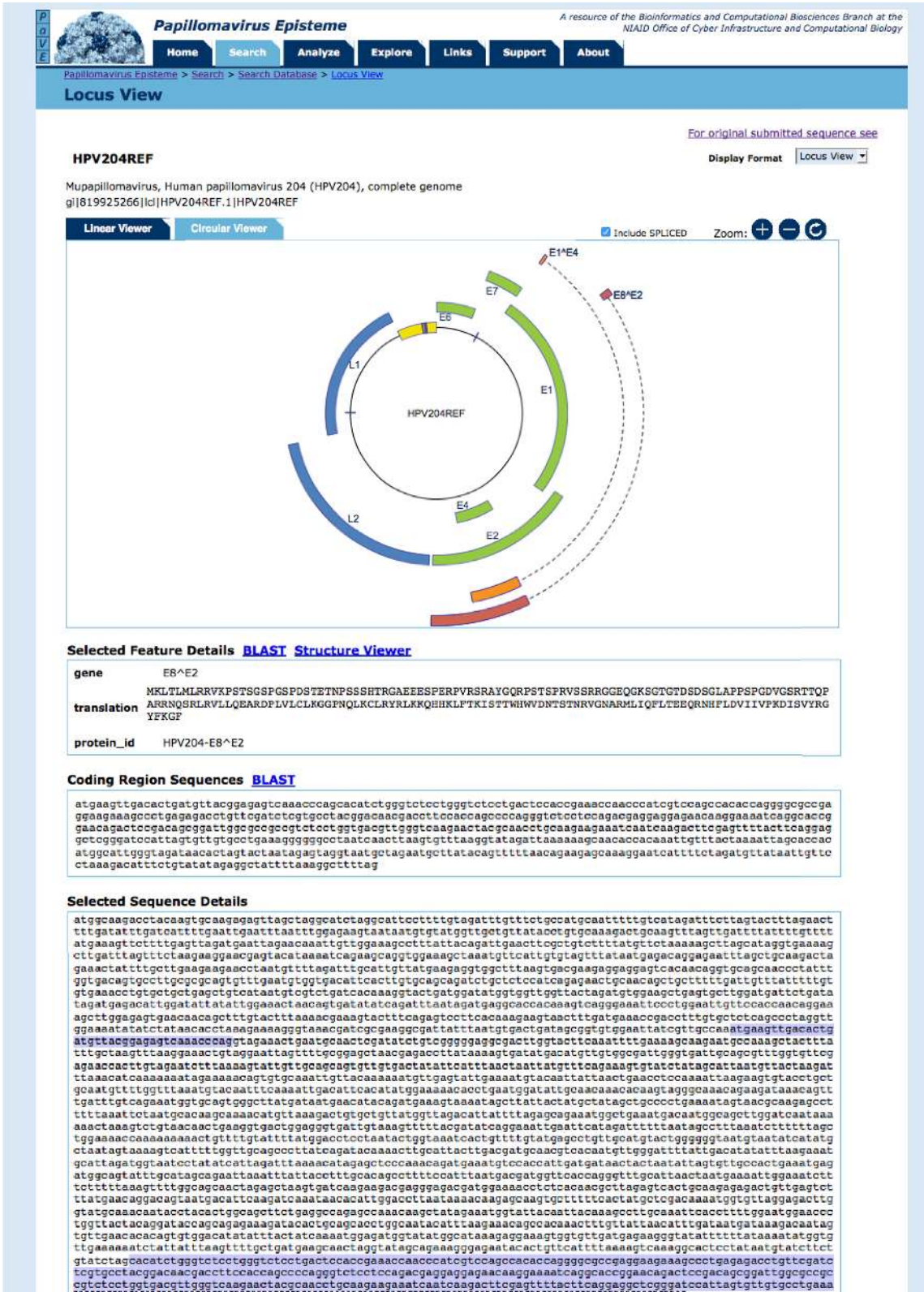
Figure 2. Circular genome browser. The PaVE locus view allows users to toggle between a linear and circular (shown here) representation of the viral genome. Viral features annotated to HPV204 are shown and can be selected. The 'selected feature details' window displays additional information about the ORF. The 'coding region sequences' window shows the nucleotide sequence for the selected ORF. Sequences can be further analysed using BLAST or the PaVE structure viewer (1). Finally, the relative position of the selected feature is displayed.
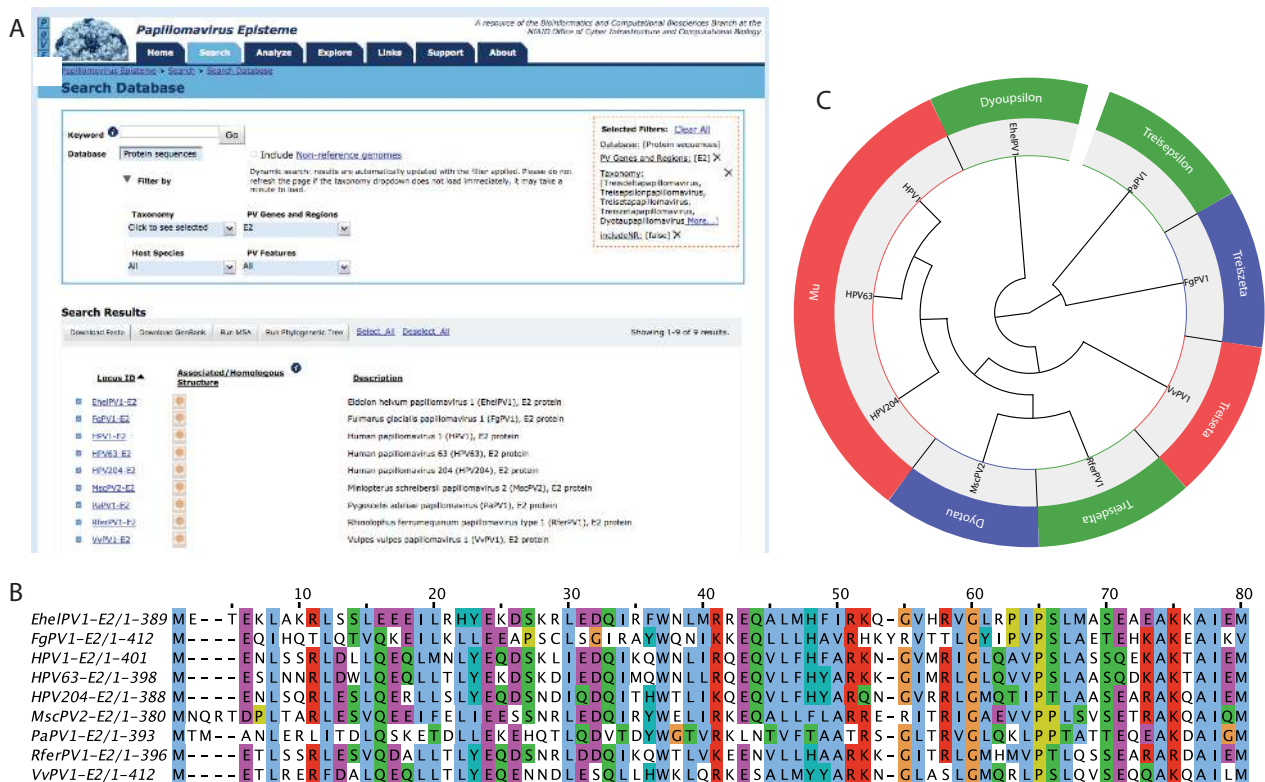
Figure 3. Integrated calculation of multiple sequence alignments and phylogenetic trees. An example of a typical analysis is shown. (**A**) Filters and pull-down menus on the PaVE search page are used to select the protein sequences of interest. The applied filters are indicated in the orange box. Sequences can be selected and used for the construction of a multiple sequence alignment or phylogenetic tree. (**B**) The first 80 amino-acids of a Cobalt based multiple sequence alignment of the E2 proteins of nine highly divergent papillomaviruses is shown as displayed within Jalview. Residues are coloured according to Clustal X. (**C**) The Neighbour-joining phylogenetic tree is displayed using JsPhyloSVG (39). If available, viral genera are included.

PROSITE protein-motif database (33) to define a collection of pairwise constraints. These constraints combined with sequence similarity are incorporated into a progressive multiple alignment. The MSA can be visualized using Jalview (version 2.8, (34)), or can be downloaded for further downstream analysis. For the phylogenetic tree construction, the neighbour-joining method (35,36) as implemented in PAUP* (version 4.0b10, (37)) is used. The neighbour-joining method is able to handle large datasets, making it feasible to rapidly construct a single phylogenetic tree of multiple homologous sequences present in the PaVE database (37). The phylogenetic tree is displayed using Js-PhyloSVG (38). While the PaVE generated phylogenetic trees provide a representation of the evolutionary history of the selected viruses, these trees are meant to be exploratory, and therefore we recommend advanced tree building methods for detailed evolutionary analysis.

A static phylogenetic tree generated from multiple alignments using MAFFT (39) of all L1 nucleotide sequences of all papillomavirus types currently in PaVE is also available. The tree was built using a maximum likelihood tree in PhyML (40) implementing a gamma model allowing for among-site rate variation and variable substitution rates [GTR+I+G; model selected using jModeltest (41) and displayed using JsPhyloSVG (38)]. Within the graphical representation, the names are linked back to the L1 sequence.

Enhanced structure viewer

The structure viewer allows the user to structurally align any protein sequence from the PaVE database with other related structures in the PDB. To rectify alignment inaccuracies resulting from inconsistent numbering of residues in PDB entries (often arising from mutations in constructs used to obtain structures or because only the structure of a single domain had been solved), the PaVE includes a pipeline to standardize the numbering. This has significantly improved the user experience and allows the seamless browsing of homology structures without the need to verify the alignments manually to adjust for these inconsistencies.

The table of available PV structures also includes links to 3D prints on the NIH 3D Print Exchange. NIH 3D Print Exchange is a website that allows users to download, edit and print models of biomolecules, such as proteins. 3D print files have been generated for all protein structures in PaVE.

Image viewer

Distinct papillomavirus types have been isolated from diverse hosts. With few exceptions, these viral types are also highly tropic to certain tissues in the host. Papillomavirus infections can result in a wide range of clinical outcomes; in many cases, infections are asymptomatic, while others result in flat macules or papular warts, verrucas, and condylomas (42). PaVE now incorporates an extensive image library
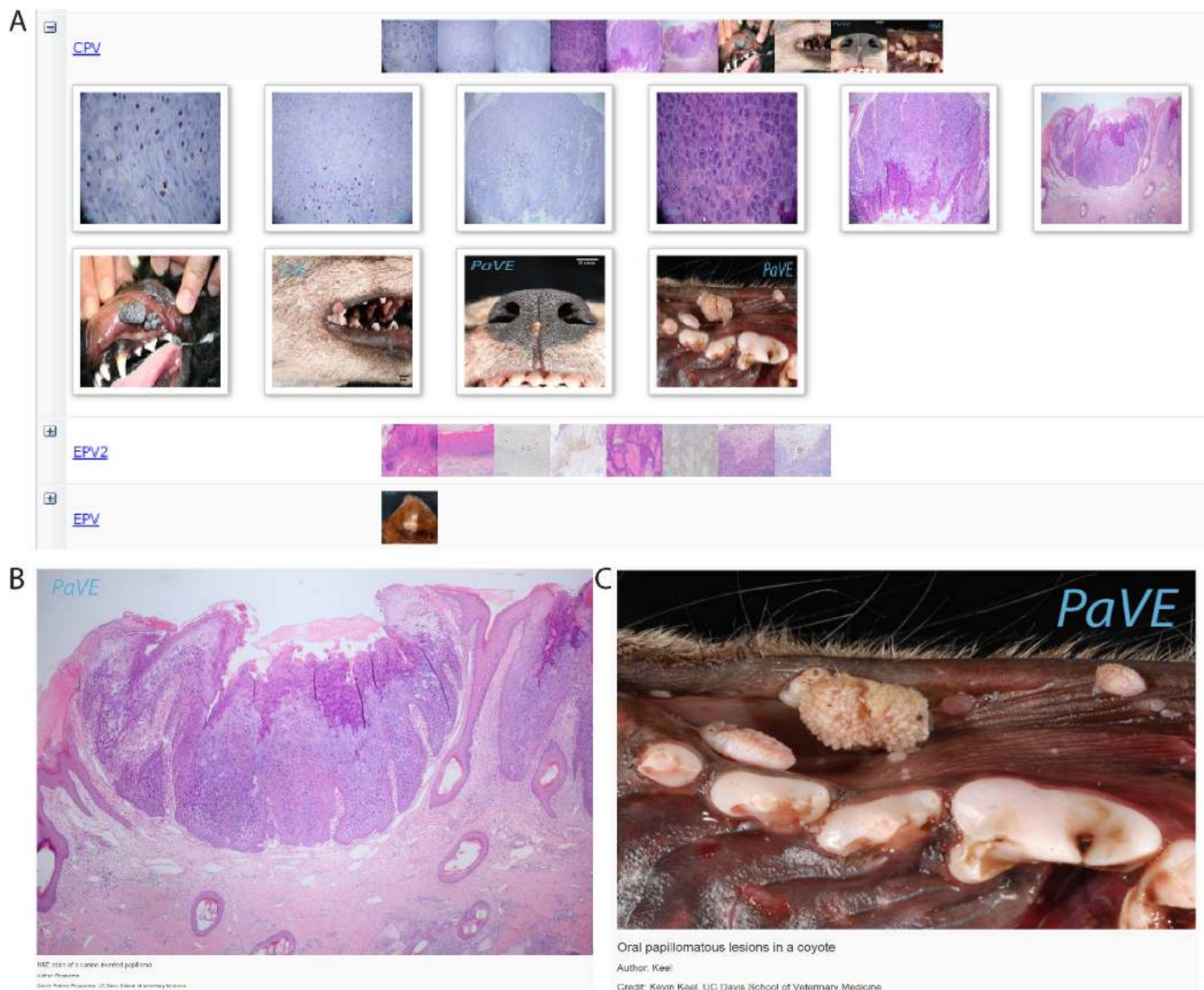
Figure 4. PaVE image database. PaVE has collected images related to papillomavirus infections. These can be browsed using the PaVE Image Viewer. (**A**) A screenshot of thumbnails in a portion of the image gallery. (**B**) Oral papillomatous lesions in a coyote. Courtesy of Kevin Keel, UC Davis School of Veterinary Medicine. (**C**) H&E stain of a canine inverted papilloma. Courtesy of Patricia Pesavento, UC Davis School of Veterinary Medicine.

containing gross clinical and histopathological images from a variety of lesions (Figure 4). If the specific causative viral type was identified, the lesions can be searched and sorted by associated virus.

Depending largely on community contributions, we welcome additionally contributions to enrich this repository. Contributors can reach us via the PaVE contact page. Contributors and users should carefully review the images sources to determine whether copyright or permission for further use is warranted. We ask that images not be altered, and that credit be given to both original authors and to the PaVE website.

Taxonomic classification of papillomavirus isolates; the L1 typing tool

Nucleotide identity across the L1 ORF is the basis for the taxonomy of the family *Papillomaviridae*. If a novel isolate shares less than 60% sequence identity with a known type it is considered to belong to a new genus, 70% identity demarcates different species within a genus. To be considered a new type, the isolate cannot share >90% sequence identity to a known virus (14). This approach requires the calculation of hundreds of pairwise sequence comparisons. Furthermore, the use of different alignment algorithms and/or parameters could affect the conclusion of this analysis. To streamline this process, an L1 specific typing tool has been developed and incorporated in the PaVE (Figure 5). As implemented, the typing process consists of several subsequent steps. Initially the query sequence is aligned to all L1 sequences in the PaVE database using the profile alignment functions of MAFFT (39). Next, a distance matrix is constructed based on pairwise alignments (Needleman–Wunsch algorithm; (43)). Finally, a neighbour-joining phylogenetic tree is built using PAUP* (Phylogenetic Analysis Using Parsimony*; (44)). Both the phylogenetic tree and
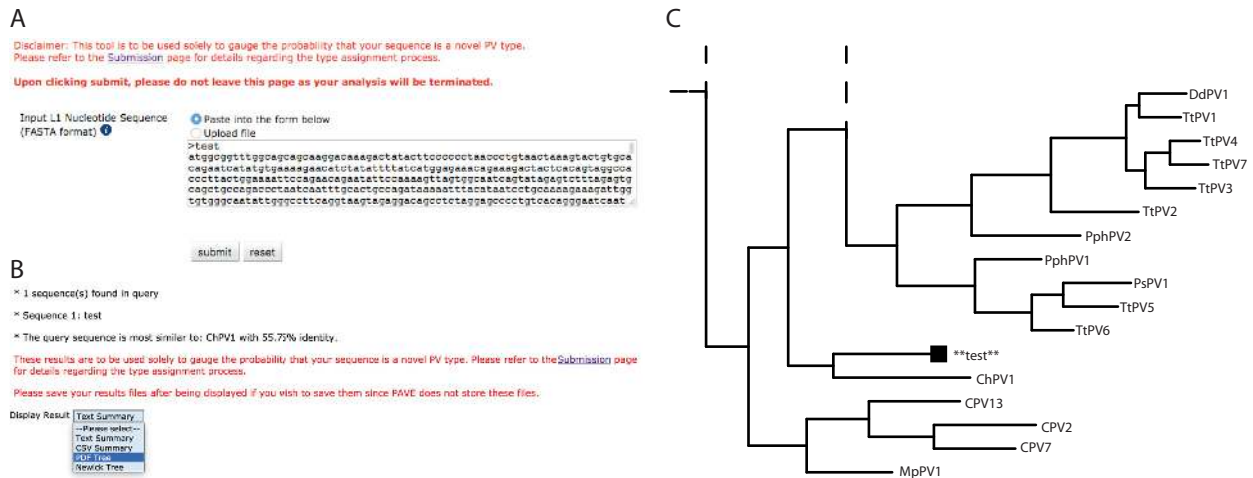
Figure 5. L1 based taxonomic classification. The PaVE taxonomy tool is designed to aid researchers in preliminary classification of novel papillomaviruses. (**A**) The tool accepts the L1 nucleotide sequence in FastA format. The sequence can be uploaded, or pasted into the search box. (**B**) Results are summarized, and more detailed information is provided, in multiple formats. (**C**) The phylogenetic position of the submitted L1 sequence is displayed. For representation purposes, the tree was cropped from the full tree. Dotted lines indicate that the tree continues along the indicated branch.

the nucleotide identity cut-off values are used to determine whether the query sequence should be considered a new type. The reliability of clustering in the neighbour-joining tree is assessed using 100 bootstrap replicates, using a 70% cut-off value.

Importantly, while the typing tool informs users whether their new isolate is different from other named viruses in PaVE, it cannot determine whether the additional criteria for a new type (cloned genome, submitted to the HPV Reference Center) are met. The papillomavirus study group of the International Committee on the Taxonomy of Viruses (ICTV)) is the only recognized authority for nomenclature of new papillomaviruses. The L1 typing tool is intended only to aid the researcher with the initial steps in submitting a putative novel viral type.

User support

PaVE has a Support page that contains Frequently Asked Questions and a User Manual. Submission of the Request Support form (pavesupport@collabmail.niaid.nih.gov) yields quick responses by the NIAID PaVE group. Information about PaVE is disseminated through the PaVE Facebook site https://www.facebook.com/PapillomavirusEpisteme and by outreach efforts at major HPV conferences.

Review chapters

The PaVE database is accompanied by a special issue of the Elsevier journal 'Virology', entitled Functional Genomics of Papillomaviruses (45). The goal of these accompanying papers was to provide an unbiased, encyclopedic overview of different aspects of papillomavirus genomics, proteomics, and related disease. Each article in the special issue is open access and indexed in PubMed. Full text articles are available directly through the PaVE website, or from the Elsevier website. The information embedded within these chapters

serves as an important knowledge resource for the scientific community.

CONCLUSIONS

The goal of the PaVE is to assist the study of papillomavirus biology and to aid in the development of therapeutics and diagnostics. This paper describes updates and improvements to the PaVE database, which was first described in 2013 (1). Since the previous description, a steadily increasing number of papillomaviruses has been described using high-throughput, metagenomic approaches. A recent metagenomic study describes 226 putative novel HPV types amplified from different human lesions (46). Furthermore, there is an increasing interest in animal virology and it is expected that the virome of non-human animals will extend the papillomavirus sequence universe with thousands of novel types. PaVE will be able to assist in the herculean task of annotating, cataloguing and classifying these viral genomes. Moreover, the ever-expanding array of bioinformatics tools integrated within the PaVE database elevates PaVE to a papillomavirus knowledgebase rather than a compilation of genomic data.

In conclusion, PaVE provides the papillomavirus community with access to uniformly curated Reference genomes, thereby ensuring that scientists across several disciplines study the same prototype genome, thus improving communication, data reporting and reproducibility.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

REFERENCES

1. Van Doorslaer,K., Tan,Q., Xirasagar,S., Bandaru,S., Gopalan,V., Mohamoud,Y., Huyen,Y. and McBride,A.A. (2013) The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res.*, **41**, D571–D578.
2. Bergvall,M., Melendy,T. and Archambault,J. (2013) The E1 proteins. *Virology*, **445**, 35–56.
3. McBride,A.A. (2013) The papillomavirus E2 proteins. *Virology*, **445**, 57–79.
4. Buck,C.B., Day,P.M. and Trus,B.L. (2013) The papillomavirus major capsid protein L1. *Virology*, **445**, 169–174.
5. Wang,J.W. and Roden,R.B. (2013) L2, the minor capsid protein of papillomavirus. *Virology*, **445**, 175–186.
6. Roman,A. and Munger,K. (2013) The papillomavirus E7 proteins. *Virology*, **445**, 138–168.
7. Vande Pol,S.B. and Klingelhutz,A.J. (2013) Papillomavirus E6 oncoproteins. *Virology*, **445**, 115–137.
8. DiMaio,D. and Petti,L.M. (2013) The E5 proteins. *Virology*, **445**, 99–114.
9. Van Doorslaer,K. and McBride,A.A. (2016) Molecular archeological evidence in support of the repeated loss of a papillomavirus gene. *Sci. Rep.*, **6**, 33028.
10. Doorbar,J. (2013) The E4 protein; structure, function and patterns of expression. *Virology*, **445**, 80–98.
11. Schwartz,S. (2013) Papillomavirus transcripts and posttranscriptional regulation. *Virology*, **445**, 187–196.
12. Bernard,H.U. (2013) Regulatory elements in the viral genome. *Virology*, **445**, 197–204.
13. de Villiers,E.M. (2013) Cross-roads in the classification of papillomaviruses. *Virology*, **445**, 2–10.
14. de Villiers,E.M., Fauquet,C., Broker,T.R., Bernard,H.U. and zur,H.H. (2004) Classification of papillomaviruses. *Virology*, **324**, 17–27.
15. Bernard,H.U., Burk,R.D., Chen,Z., van Doorslaer,K., zur Hausen,H. and de Villiers,E.M. (2010) Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology*, **401**, 70–79.
16. Lopez-Bueno,A., Mavian,C., Labella,A.M., Castro,D., Borrego,J.J., Alcami,A. and Alejo,A. (2016) Concurrence of iridovirus, polyomavirus and a unique member of a new group of fish papillomaviruses in lymphocystis disease affected gilthead seabream. *J. Virol.*, **90**, 8768–8779.
17. Lange,C.E., Favrot,C., Ackermann,M., Gull,J., Vetsch,E. and Tobler,K. (2011) Novel snake papillomavirus does not cluster with other non-mammalian papillomaviruses. *Virol. J.*, **8**, 436.
18. Herbst,L.H., Lenz,J., Van Doorslaer,K., Chen,Z., Stacy,B.A., Wellehan,J.F. Jr, Manire,C.A. and Burk,R.D. (2009) Genomic characterization of two novel reptilian papillomaviruses, Chelonia mydas papillomavirus 1 and Caretta caretta papillomavirus 1. *Virology*, **383**, 131–135.
19. Van Doorslaer,K., Sidi,A.O., Zanier,K., Rybin,V., Deryckere,F., Rector,A., Burk,R.D., Lienau,E.K., van Ranst,M. and Trave,G. (2009) Identification of unusual E6 and E7 proteins within avian papillomaviruses: cellular localization, biophysical characterization, and phylogenetic analysis. *J. Virol.*, **83**, 8759–8770.
20. Rector,A. and Van Ranst,M. (2013) Animal papillomaviruses. *Virology*, **445**, 213–223.
21. Van Doorslaer,K. (2013) Evolution of the papillomaviridae. *Virology*, **445**, 11–20.
22. Bzhalava,D., Eklund,C. and Dillner,J. (2015) International standardization and classification of human papillomavirus types. *Virology*, **476**, 341–344.
23. Wang,M. and Marin,A. (2006) Characterization and prediction of alternative splice sites. *Gene*, **366**, 219–227.
24. Puustusmaa,M. and Abroi,A. (2016) Conservation of the E8 CDS of E8^E2 protein among mammalian papillomaviruses. *J. Gen. Virol.*, **97**, 2333–2345.
25. Stubenrauch,F., Hummel,M., Iftner,T. and Laimins,L.A. (2000) The E8E2C protein, a negative regulator of viral transcription and replication, is required for extrachromosomal maintenance of human papillomavirus type 31 in keratinocytes. *J.Virol.*, **74**, 1178–1186.
26. Mesplede,T., Gagnon,D., Bergeron-Labrecque,F., Azar,I., Senechal,H., Coutlee,F. and Archambault,J. (2012) p53 degradation activity, expression, and subcellular localization of E6 proteins from 29 human papillomavirus genotypes. *J. Virol.*, **86**, 94–107.
27. Bravo,I.G. and Alonso,A. (2004) Mucosal human papillomaviruses encode four different E5 proteins whose chemistry and phylogeny correlate with malignant or benign growth. *J. Virol.*, **78**, 13613–13626.
28. Jackson,M.E., Pennie,W.D., McCaffery,R.E., Smith,K.T., Grindlay,G.J. and Campo,M.S. (1991) The B subgroup bovine papillomaviruses lack an identifiable E6 open reading frame. *Mol. Carcinogenesis*, **4**, 382–387.
29. Harry,J.B. and Wettstein,F.O. (1996) Transforming properties of the cottontail rabbit papillomavirus oncoproteins Le6 and SE6 and of the E8 protein. *J.Virol.*, **70**, 3355–3362.
30. Mullan,L.J. (2002) Multiple sequence alignment–the gateway to further analysis. *Brief. Bioinform.*, **3**, 303–305.
31. Papadopoulos,J.S. and Agarwala,R. (2007) COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, **23**, 1073–1079.
32. Marchler-Bauer,A., Derbyshire,M.K., Gonzales,N.R., Lu,S., Chitsaz,F., Geer,L.Y., Geer,R.C., He,J., Gwadz,M., Hurwitz,D.I. *et al.* (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res.*, **43**, D222–D226.
33. Sigrist,C.J., de Castro,E., Cerutti,L., Cuche,B.A., Hulo,N., Bridge,A., Bougueleret,L. and Xenarios,I. (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, D344–D347.
34. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
35. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
36. Studier,J.A. and Keppler,K.J. (1988) A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.*, **5**, 729–731.
37. Van de Peer,Y. (2009) In: Lemey,P, Salemi,M and Vandamme,A-M (eds). *The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press, Cambridge.
38. Smits,S.A. and Ouverney,C.C. (2010) jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web. *PLoS one*, **5**, e12267.
39. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
40. Guindon,S., Dufayard,J.-F., Lefort,V., Anisimova,M., Hordijk,W. and Gascuel,O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.
41. Darriba,D., Taboada,G.L., Doallo,R. and Posada,D. (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods*, **9**, 772–772.
42. Cubie,H.A. (2013) Diseases associated with human papillomavirus infection. *Virology*, **445**, 21–34.
43. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
44. Swofford,D.L. (2003) Sinauer Associates, Sunderland, Vol. **4**.
45. Lambert,P.F., McBride,A.A. and Ulrich Bernard,H. (2013) Special issue: the Papillomavirus Episteme. *Virology*, **445**, 1.
46. Bzhalava,D., Muhr,L.S., Lagheden,C., Ekstrom,J., Forslund,O., Dillner,J. and Hultin,E. (2014) Deep sequencing extends the diversity of human papillomaviruses in human skin. *Sci. Rep.*, **4**, 5807.