Infection, Genetics and Evolution 18 (2013) 357–361

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

# Taxonomy and phylogeny of papillomaviruses: An overview and recent developments

Hans-Ulrich Bernard *

*University of California Irvine, Dept. of Molecular Biology and Biochemistry and Program in Public Health, Irvine, CA 92617, United States*

ARTICLE INFO

*Article history:*
Received 10 September 2012
Received in revised form 4 March 2013
Accepted 7 March 2013
Available online 20 March 2013

*Keywords:*
Papillomavirus
Evolution
Phylogeny
Taxonomy

ABSTRACT

For more than 30 years, papillomaviruses are standing in the center of medical and molecular interest as they cause several important cancers in humans. Research of the sheer unlimited number of different papillomavirus genomes, their host specificity and slow mutation rate is an important a branch of these efforts and has led to fascinating insight into the phylogeny of a virus family that can be traced back for several 100 million years.



## 1. Biology and genome organization of papillomaviruses

Papillomaviruses were originally detected as the cause of skin warts in various mammals including humans (zur Hausen, 2009). The isolation of certain papillomaviruses from cancers of the cervix uteri and other types of anogenital and oral malignancies in the early 1980s (Dürst et al., 1983) led to the rapid expansion of the study of these viruses, which culminated in recent years in the development of clinical DNA tests for cervical cancer and vaccines against papillomavirus infection. However, papillomaviruses are not always inducing neoplastic growth. Many papillomaviruses are preferentially found in clinically asymptomatic infections, and even infections with the carcinogenic papillomaviruses often lead to associations in the absence of any histologic alterations.

Papillomaviruses infect nearly exclusively the skin, squamous mucosal epithelia of the mouth, the anus and the female genitals, and the columnar epithelia of the endocervix, but not other simple epithelia such as the colon or the lung. The precise nature of this restriction is not fully understood, but is apparently based in cis-responsive elements, which activate the papillomavirus transcription only in these cell types but not elsewhere (Gloss et al., 1987).

Papillomaviruses have double-stranded circular DNA genomes with sizes of slightly less than 8 kb. This combination of properties is unique and defines this group of viruses, which has been ranked by the International Committee on Taxonomy of Viruses (ICTV) as a "family", *Papillomaviridae* (de Villiers et al., 2004). The genome of most papillomaviruses contains eight genes. Among these, E6 and E7 encode the principal transforming proteins, E1 and E2 regulators of replication and transcription, and L1 and L2 the major and minor capsid proteins (Fig. 1). Transcription and replication are mostly regulated via cis-responsive elements in a long control region (LCR) located between the L1 and E6 genes, while some other regulatory elements are positioned within genes. For the purpose of this article the L1 gene and the LCR are of eminent importance, as the papillomavirus taxonomy is based on comparisons of the L1 nucleotide sequences, and many studies of minor phylogenetic diversity of some papillomaviruses have been based on the study of LCR sequences.

## 2. The history of central concepts of papillomavirus nomenclature

The first papillomaviruses were identified in mammals such as rabbits (Shope and Hurst, 1933) and cattle and in large human skin warts in the form of viral particles and later as clonable DNA. The viruses were designated using the name of their host species, e.g. cottontail rabbit papillomavirus (CRPV), bovine papillomavirus (BPV), and human papillomavirus (HPV). If more than one papillomavirus was found in the same host, a number was added to the name and abbreviation, e.g. HPV1 and HPV2. As these efforts predated a formal taxonomy and a concept of what a virus species is, each of these viruses was designated as a "papillomavirus type". As most papillomaviruses could never be grown in cell culture and animal models, the concept of a papillomavirus type became "a DNA clone of a full papillomavirus genome". The novelty of a pap-

* Tel.: +1 949 824 5162; fax: +1 949 824 8551.
*E-mail address:* hbernard@uci.edu

1567-1348/$ - see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.meegid.2013.03.011

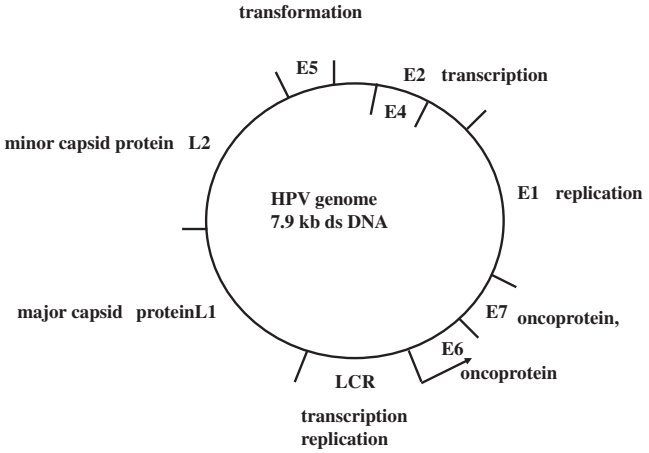


**Fig. 1.** The genome of HPV16 has a size and gene arrangement typical for most papillomavirus. All papillomaviruses have homologous E1, E2, E4, L2 and L1 genes. Some few papillomaviruses have rearrangements or replacements of E6 and E7, and the E5 gene is missing in some types.

illomavirus type was originally decided through restriction digestion or hybridization experiments, but is today defined as a complete papillomavirus genome, whose L1 gene nucleotide sequence is at least 10% different from that of any other known papillomavirus type. New papillomavirus types are in today's nomenclature designated with the scientific name of the host species, e.g. MfPV1, for *Macaca fascicularis papillomavirus*-1 (Joh et al., 2009; Bernard et al., 2010). Papillomavirus "types" have always remained the primary objects of isolation and biological studies, while concepts such as "species" and "genera" became used at taxonomic levels above the papillomavirus types.

## 3. The family papillomaviridae

Papillomaviruses were originally lumped together with polyomaviruses in a family called "*Papovaviridae*" based on the shared properties that these were the only viruses with circular double-stranded DNA genomes and with a non-enveloped icosahedral capsid. Polyoma- and papillomaviruses became split into two families when it became clear that the genomes were very different from one another. The genomes have sizes of 5 and 8 kb, respectively, early and late genes are transcribed in polyomaviruses in opposite orientations, but in the same direction in papillomaviruses. And there are no pairs of homologous genes among the two virus families. Even the capsid proteins that give rise to virions with similar morphology do not share any sequence similarities. A single exception is known for decades but is rarely discussed: the papillomavirus E1 protein and the polyomavirus large-T-antigen each have a 200 amino acid domain functioning as DNA helicase, and about 25% of these amino acids are identical or chemically related (Clertant and Seif, 1984) making a strong case for homology of these domains. This observation invites the speculation that both virus families had a common replication gene as a common ancestor, which may be even shared with single stranded DNA viruses and episomes of certain worms (Rebrikov et al., 2002).

## 4. Papillomavirus genera, species, and types

L1 genes of different papillomavirus types show clusters of sequence similarity throughout the whole length of the genes, permitting an unequivocal alignment of their nucleotide sequences. It was originally an arbitrary decision to define a papillomavirus type through these alignments, namely that L1 nucleotide sequences must be at least 10% different from that of any other papillomavirus type. The large body of research that exists today confirmed that this definition was a very satisfactory choice, as nearly every new isolate has genomic distances larger than this limit. For isolates with 2–10% sequence diversity the term "subtype" has once been used, but studies suggest that genomic variation on this subtype level is extremely rare and therefore probably of little biological and clinical interest (Calleja-Macias et al., 2005) in contrast to genomic variation of up to 2%, termed "variants" (see below).

The reliable genomic distinction between papillomavirus types and the lack of intermediate genomes, which must have existed during evolution or that might have come into being by DNA recombination, could make papillomavirus types candidates to be considered "virus species". Unfortunately, a central rule for taxonomic decisions by the ICTV did not allow this conclusion. A virus can only be considered a species if it is biologically and clinically distinct from any other virus (van Regenmortel et al., 1991). Numerous papillomavirus types do not fulfill this criterion. For example, we do not know any biological difference between HPV2, 27, and 57, the cause of common skin warts, and HPV6 and HPV11, the cause of genital warts, although these types are genomically well separated. As a resolution of this problem, it was decided to consider a papillomavirus species a phylogenetic cluster of related papillomavirus types with indistinguishable biology, but being functionally distinct from related clusters.

The phylogenetic evaluation of L1 alignments became the foundation for the modern papillomavirus taxonomy (de Villiers et al., 2004). Fig. 2, taken from Bernard et al., 2010, represents the relationship between all 189 papillomavirus types known at that time, the last update of the official taxonomy. Among these 189 types, 120 were isolated from humans, 64 from mammals, three from birds and two from reptiles, and these 189 types form 70 species belonging to 30 genera. Biological and clinical properties of the viruses formed a foundation for these groupings, but taxonomic decisions were ultimately based on tree topology, with genera and species forming major and minor phylogenetic branches, respectively. These tree topologies are supported by L1 nucleotide sequence identities, which decrease from close to 90% to about 45% in intraspecies, interspecies and intergeneric comparisons. Papillomavirus genera are termed by a Greek letter and papillomavirus species by addition of a number to the letter. For example, the type HPV16, the cause of most cervical cancers, is a member of the type-rich genus *Alphapapillomaviruses* and forms with six other HPV types the species *Alphapapillomavirus* 9.

## 5. Genomic variation within papillomavirus types

Repeated isolates of the same papillomavirus are not as diverse from one another as one normally encounters in RNA viruses. They are not completely identical either, but their nucleotide sequences often differ by up to 2% in the non-coding LCR and somewhat less in genes. These genomic isolates became designated as "variants" of specific papillomavirus types. Surprisingly, there is not an unlimited amount of different variants of each papillomavirus type but only a moderate number (not formally documented, but based on partial genomic sequences maybe about 100, to give a ballpark figure), and these variants form simple phylogenetic trees with about 5–9 separate branches and sublineages. The variants of some of these branches are specific or at least more prevalent in certain parts of the world, e.g. Africa, and East-Asia, most likely since they evolved in ethnic groups that originally occurred only in this part of the world (Ho et al., 1993; Ong et al., 1993). These original observations have led to a standardized classification of HPV16 variants that is important for future epidemiological and biological