Human Papillomavirus Genomics: Past, Present and Future

Ariana Harari\(^a\) · Zigui Chen\(^b\) · Robert D. Burk\(^a-d\)

Departments of \(^a\)Microbiology and Immunology, \(^b\)Pediatrics, \(^c\)Obstetrics, Gynecology and Women’s Health, and \(^d\)Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, N.Y., USA

Abstract

Human papillomaviruses (HPV) are a group of divergent DNA viruses, of which a select few evolutionarily related HPVs have emerged to be highly oncogenic and of significant medical importance. Essentially all cases of cervical cancer, as well as a subset of other anogenital and oral cancers are caused by this limited set of HPV types. At present, over 150 HPV types have been identified and may be classified into genera, species and types based upon comparison of the viral genome. Established nucleotide phylogenies sort the highly pathogenic HPV types to the genus Alphapapillomavirus (α-PV). A species group includes viral types with 60–70% genomic nucleotide similarity that share a most-recent common ancestor; for example the species group’s alpha-9 (HPV16-related) and alpha-7 (HPV18-related), contain the majority of known oncogenic HPV types. Genomes from the same HPV type with 1–10% nucleotide differences designate HPV variant lineages. The established nucleotide variations observed in extant HPV genomes have been fixed through evolutionary processes prior to human population expansion and global dissemination. To characterize viral types and variants associated with pathology for clinical applications (e.g. screening), molecular epidemiological studies have proven essential for identifying links between HPV natural history and carcinogenicity. This chapter presents a historical account of HPV genomics in the context of major discoveries and advances over the past 2 thousand years.

© 2014 S. Karger AG, Basel

Papillomaviruses (PVs) are ubiquitous, highly diverse DNA viruses that have been isolated from all 4 classes of Tetrapoda, including most mammals as well as birds, turtles and snakes [1–3]; their origin predates the existence of modern humans [4–7]. Observational accounts of warts from ancient Greece and Rome describe condylomatous lesions on the skin and genitals, and it was presumed that genital warts were associated with promiscuous sexual behavior. Warts in general were surmised to be transmissible [8]. Manifestations of animal PVs have historically been documented in myths and paintings, particularly of the ‘jackalope’. This animal does not exist, but likely represents a case of mistaken identity as a result of a PV infection that produced
cornified growths (i.e. horns) on jackrabbits. Historical records and myths provide evidence for the antiquity of PVs and the diseases they cause. In 1842, the Italian physician Dr. Rigoni-Stern was the first to hypothesize that cervical cancer might be linked with sexual behavior. He observed that cervical cancer frequency was disproportionately higher in prostitutes and married women than nuns and virgins, implying that the causative agent was likely sexually transmitted [9, 10]. Rigoni-Stern’s observation would be validated almost 150 years later; today, human PVs (HPVs) are known to be among the most commonly sexually transmitted infections, and infections by specific high-risk (HR) HPV types are known to be the etiological agents of cervical cancer [11–14].

In 1911, Francis Peyton Rous famously demonstrated that filtered tumor cell extract obtained from chicken carcinomas and transplanted to naïve chickens promoted sarcoma growth of a virulent nature, identifying the first oncogenic virus, i.e. Rous sarcoma virus. Almost 25 years later, Dr. Rous and Dr. Richard Shope identified the first PV, known as Shope PV or cottontail rabbit PV, from warty growths on cottontail rabbits. They went on to demonstrate that transmission of cottontail rabbit PV exhibited neoplastic potential in domestic rabbits [15, 16]. Approximately 75 years later, Prof. zur Hausen was awarded a Nobel Prize for the discovery of HPVs causing cervical cancer [17, 18]. The discovery that HPVs are major contributors to cancer and represent highly adaptive, carcinogenic viral pathogens causing essentially all of cervical cancer and approximately 20% of head and neck cancers has energized the research and medical communities [14] (fig. 1).

Cervical cancer ranks 3rd amongst cancers affecting women worldwide and 2nd in developing countries [14]. Women in developing countries account for 85% of the global incidence of cervical cancer. Incidence rates are nearly double in developing compared to developed countries, 17.8 and 9.0%, respectively. This difference is thought to be largely due to the implementation of early diagnostic screening methods, which have reduced the risk of cervical cancer associated with persistent HPV infection. Yet, cervical cancer is still responsible for 275,000 deaths/year [14].

HPV infection by any of the 150 identified HPV types is not sufficient to cause cancer. All genital oncogenic types belong to the genus Alphapapillomavirus, which is currently comprised of 62 known HPV genotypes that infect the mucosal epithelium. Further classification beyond the genus level is required to distinguish the 13–15 HR HPV types that are associated with oncogenic risk. It is well established that phylogenetic analyses cluster HPV taxa by host cell tropism (e.g. skin vs. genitalia), degree of oncogenic potential and morphology of clinical lesions [1, 19]. Progression to cancer is rare. The malignant potential of specific HPVs likely results from niche adaptation (i.e. evolution of an organism/virus to a specific biological/anatomical ecosystem on the body) of PVs, as cancer is undesirable for both the virus and its host. The majority of HPV infections are cleared within 6–10 months. Persistent infections with HR HPVs are a critical risk factor for the development of HPV-associ-
ated precancer and cancer [20–22]. Delineating the differences intrinsic to these HR HPV genotypes, compared to the majority of HPV types that lack oncogenic potential, will help to elucidate the genetic basis of such carcinogenic properties, thus contributing to a better understanding of the biological mechanisms exploited by the virus to facilitate cancer development. Epidemiology studies provide a platform for obtaining viral isolates used to investigate the dynamic relationship of HPV genotype differences and clinical disease. The recent explosion in DNA sequencing tech-
nologies will continue to revolutionize methods of HPV detection [23] contributing to a better understanding of HPV biology and the development of new therapies against HPV-associated cancers.

**Human Papillomavirus Genome Characteristics**

HPV is a circular, nonenveloped, double-stranded DNA virus approximately 8 kb in size that infects basal keratinocytes. Upon infection, the virus exists as an autonomous episome in the host cell nucleus. The viral life cycle is mediated by a series of virus-host interactions, which govern viral transcription, virion production and eventual clearance in the majority of infections [24, 25] (fig. 2).

The structure and function of the HPV genome are conserved throughout the Papillomaviridae and are broadly divided into 3 general components. (1) The early gene region, denoted by ‘E’, consists of 6 open reading frames (ORFs): E6, E7, E1, E2, E4, and E5. The early genes E1, E2, E6 and E7 are generated as a polycistronic transcript. Several additional early ORFs E3, E5 and E8 have also been identified, but their expression is not uniformly observed throughout the Papillomaviridae. Viral transcripts can undergo extensive alternative splicing, contributing to the intricate balance between viral and host-regulated transcription [for a review, see 26]. The early genes code for nonstructural proteins that function in viral replication, adaptation of the cellular milieu for viral activities, trans-activation of viral transcription and cellular transformation and proliferation. (2) The late gene region, denoted by ‘L’, consists of the L2 and L1 ORFs. L1 and L2 encode the structural proteins, the major and minor capsid proteins, respectively. The L2 ORF encodes for group-specific epitopes whereas the L1 ORF contains type-specific protein domains. (3) The upstream regulatory region (URR) is the noncoding region between the end of the L1 ORF and the E6 start codon, comprising approximately 10% of the genome. The URR contains DNA recognition sites for both viral and host transcription factors and regulates early gene transcription, viral amplification and cellular tropism. The URR contains a keratinocyte-specific enhancer region proximal to the early gene promoter (p97), which highlights the significance of host cell tropism to viral gene expression and life cycle. A smaller noncoding region located between the E5 stop and L2 start codons harbors a highly conserved early polyadenylation signal required for gene expression from the early promoter, including alternatively spliced early transcripts and their gene products [26]. Both the early gene region and the URR display variability, useful in assessing genetic heterogeneity [27]. Increased understanding of the inherent genomic differences between HPV species that contribute to viral function is predicated on biochemical techniques such as DNA sequencing and polymerase chain reaction (PCR) which have contributed greatly to understanding the molecular pathogenesis of HPV. Great strides have been made since the inception of clinical HPV molecular biology in the 1970s [7, 28–30].
The viral proteins E1 and E2 function in viral genome replication and are dependent on the host DNA polymerase and replication machinery. E1 functions as an ATP-dependent helicase capable of melting double-stranded DNA for strand separation, prior to DNA polymerization. The E2 ORF encodes for a viral-DNA binding transcription factor. E1 and E2 proteins form heterodimers at the viral origin of replication to initiate bidirectional genome synthesis. Recently, E1 and E2 have been shown to function in the early induction of the DNA damage response pathway con-

---

**Fig. 2.** Schematic of the HPV-16R genome. The HPV-16 genome displayed shows the general organization of the HPV open reading frames (ORFs) and regulatory regions. Early (E) genes and related transcriptional regulatory motifs are shown in green; late (L) genes and related transcriptional regulatory motifs are shown in blue; the 2 noncoding regions are shown in gray; the smaller region between E5 and L2 is termed the noncoding region (NCR), and the larger region between L1 and E6 is known as the upstream regulatory region (URR). The URR contains regulatory elements including the viral DNA origin of replication (Ori) depicted by a yellow oval and the early gene promoter p97 depicted by a green arrow. The blue arrow in the E7 ORF depicts late gene expression regulated by the late promoter p670. Polyadenylation sites termed early (pAE) or late (pAL) are shown as green or blue triangles, respectively. Arrowheads at the tip of each ORF depict if they overlap. The thick solid line represents genomic regions commonly overexpressed in cancer, the thin black line represents genomic regions frequently disrupted by integration, and the dashed line represents the genomic region rarely expressed in cancer.
tributing to a permissive environment for viral genome amplification [31]. E1 can induce beaks in the host double-stranded DNA that activates the ataxia-telangiectasia DNA damage response pathway, signaling cell cycle arrest [31]. The URR contains 4 highly conserved E2 binding sites that differentially regulate viral replication and early gene transcription [32, 33]. E2-dependent downregulation of early promoter activity maintains low-copy numbers of viral genomes prior to differentiation dependent activation of the late promoter and genome amplification. In high-grade neoplastic lesions/cancer the E2 ORF can be disrupted by viral integration into the host genome. Integration results in the loss of E2-dependent early promoter regulation, a ramifications of which can be overexpression of E6 and E7 [24, 25, 34]. The viral proteins E6 and E7 function as oncogenes in the HR α-HPVs. HR HPV E6 and E7 proteins disrupt cell cycle regulation in upper epithelial cells (stratum spinosum), which normally exit the cell cycle to terminally differentiate. Virally mediated inactivation of key cell cycle regulators, the known tumor suppressors, p53 and pRb (by E6 and E7, respectively), are distinct among certain α-PV types, although not specific to those causing cancer [35]. The E5 ORF encodes a transmembrane protein that probably contributes to cell signaling [36]. The E5 ORF as defined by the presence of a true start and stop codon is found in select members of the human α-PVs: the HR HPV species, the α-10 species and other PVs including bovine PV 2 implicated in urinary cancer (cattle). At the nucleotide level, E5 is not highly conserved. The E5 protein is associated with late gene viral life cycle events and interacts with epidermal growth factor and platelet-derived growth factor to influence cellular proliferation [37]. The integrity of the E5 ORF provides another example of the differences revealed by studying HPV genotypes and phylogeny [38]. The late genes L1 and L2 function in virion maturation and orchestrate virion self-assembly that packages the genome for release in the upper epithelia. During late stages of precancer/cancer, L1 and L2 proteins are not expressed. This initially made early studies on HPV identification difficult and required the use of extracted DNA from warts to be obtained for analysis. Virus-like particles made from HR HPV L1 readily self-assemble and induce a neutralizing antibody response. This is the basis for the two prophylactic vaccines currently available.

Viral Identification and Classification: Phylogeny and Taxonomy

HPV genome heterogeneity was apparent during the initial studies examining viral isolates obtained from dysplastic lesions (cutaneous or genital). By the late 1970s, an unknown viral agent was known to be the causative agent implicated in the development of multiple wart types, condylomata acuminata (genital warts) and possibly cervical cancer. Attempts to identify the viral agent from lesions were hampered by the fact that DNA probes often did not cross-react with DNA extracts obtained from different wart types, implicating the presence of diverse viruses [39]. Furthermore, early attempts to study the infectious nature of the virus obtained from warts was restricted
by the highly specific host epithelial-cell specificity required for a productive HPV life
cycle [30, 40]. Understanding the delicate balance of HPV-host interactions proved
fundamental to addressing the clinical significance of HPV-related disease.

HPVs are extremely ancient viruses, and related PVs have been identified in most
nonhuman primates, including humans’ most recent common ancestors [41–43]. Se-
quence and phylogenetic analysis of nonhuman primate PVs isolated from the cervi-
covaginal area revealed genomic similarity to the α-HPVs. Several nonhuman primate
PVs identified from rhesus macaques (Macaca mulatta), cynomolgus macaques (Ma-
caca fascicularis), and olive baboons (Papio hamadryas anubis) comprise the α-12
species. These phylogenetically related α-PV types can induce epithelial dysplasia of
varying degrees, resembling the high-grade cervical intraepithelial neoplasia (CIN)
associated with persistent HR mucosal-HPV infections, specifically within the α-9
species, which includes HPV-16 [5, 6, 44]. Experimental transmission of M. fascicu-
laris PV-3 from a naturally infected female to naïve female macaques was associated
with the development of CIN [45]. These nonhuman primate PVs are more similar to
the α-9 species, than more distantly related HPV types, such as the α-10 species, sug-
gest that the mechanisms governing cellular transformation result from a common
ancestral trait that predates human/primate divergence.

Diversification of HPVs occurred prior to the emergence of Homo sapiens approxi-
mately 200,000–150,000 years ago [46]. The extensive diversification and demo-
graphic range of HPVs reflect human dispersal and population expansion [4], and
may intuitively suggest that HPVs have had to undergo a high rate of mutations to
adapt to such diverse hosts; however, this is not the case. Rates of nucleotide muta-
tions are remarkably low in the virus, observed at a rate of approximately 10^{-8}–10^{-7}
nucleotides/year [47].

HPV characterization in population-based studies affords the unique opportunity
to study hominid evolution through a viral lens to better understand virus-host inter-
actions as they pertain to immune system surveillance and cancer progression [5, 6,
38]. The unique coupling of the PV-host-dependent life cycle has selected for special-
ized viral adaptation to specific host niches, coincident with the ability of the virus to
evade a host immune response. These features have enabled PVs to capitalize on their
hosts, and exploit global diversity to thrive throughout evolutionary time. Genital
HPV infections are commonly dependent on sexual transmission. This exemplifies
one way the virus invests in its existence; they have hijacked the most fundamental
aspect of host species success, reproductive fitness. After infection, viral gene expres-
sion is tightly regulated such that it is dependent on host epithelial differentiation for
near exclusive activation of either the early or late gene promoters driving gene ex-
pression. Such adaptation tactics suggests a commensal virus-host relationship, yet
HPV-dependent malignancies defy this viewpoint. Molecular epidemiological studies
assessing the phylogenetic association of HPVs based on oncogenic risk support spe-
cific biological and pathological traits distinct to HPV genera, species and types [38].
HPV-16 is unique in its ability to establish persistence that is highly associated with
neoplastic progression, both of the cervix and in head and neck carcinogenesis. The α-HPVs exhibit agreement between (clinical) natural history studies and HPV phylogeny and taxonomy, providing evidence to support that carcinogenicity is an evolved trait most probably related to niche adaptation [38, 48–50].

**Human Papillomavirus Classification**

Methods to culture HPV in vitro or produce infectious virus through xenotropic models are not efficient, and do not provide a robust method for identification and characterization of HPV types [51]. Furthermore, a lack of a robust, consistent antibody response in infected individuals limits the use of antibody titer and serology for HPV taxonomy [51]. Historically, HPV has been identified from biopsies of warts or lesions and classified by comparison to known types by restriction endonuclease cleavage patterns and/or DNA-DNA or DNA-RNA blot hybridization. Such methods had innate flaws for characterizing PVs; specifically there was no quantifiable means of comparison amongst HPV types from different lesions, the virus titer in lesions was not available, and cross-hybridization was difficult to explain except for association by anatomical site [30, 40]. The appreciation for a genomic, DNA-sequence-based classification system was agreed upon by the PV research community by the late 1980s. This system has relied on the rapid advances in DNA technologies from PCR to the Next-Gen sequencing era upon us.

Today, PCR-based amplification of DNA obtained from clinical samples is common. As the realization that phylogeny and genotyping methods validated one another, the notion that only a limited set of HPV types were associated with cancer spurred the need for highly specific assays that could discriminate HPV types. Consensus PCR primers targeting highly conserved regions within the L1 ORF, such as the MY09/11 PCR assay or the GP5+/GP6+ PCR assays, are typically used for HPV identification [52]. A review of HPV detection methods has recently been published [53]. In addition, sequencing of the PCR amplicons for alignment to known HPV type(s) facilitates classification of genotype by nucleotide identity and identification of novel types.

PVs belong to the family *Papillomaviridae* and were given this status by recognition of a genome-based, DNA sequence system for classification [54]. DNA sequencing provides a quantifiable means to catalog nucleotide heterogeneity, affording the classification and taxonomy of HPV to genera, species, types and variant lineages [27, 55]. Classification of HPV genera and species based on DNA sequence was recognized by the PV Working Group at the 14th International Papillomavirus conference in Quebec in 1995, later adapted by the International Committee on the Taxonomy of Viruses [1, 19, 51, 56]. Variant lineage classification is a more recent development within the PV community that will become increasingly more relevant as high-resolution techniques, such as next-generation sequencing, generate a plethora of PV sequencing data that need to be coherently analyzed, named and correlated with phenotype and geographic locations [55].
Human Papillomavirus Genome ‘Typing’

The L1 gene is highly conserved and provides the basis for HPV genotyping. An HPV ‘type’ is designated when the nucleotide sequence of the L1 ORF from the cloned viral genome is more than 10% dissimilar to all known types [1].

Nucleotide differences across HPV genotypes are correlated with viral lineages based on evolution without significant, if any, recombination, and these correlated changes are the result of lineage fixation [57]. Whole genome sequencing established L1 sequence identity as representative of complete genome variation due to its high conservation [1, 50]. Nucleotide sequences are used to build phylogenetic trees used for HPV taxonomy. Phylogenetic analysis suggests the underlying relationship between the biological observations identified within this heterogeneous group of viruses: including host cell tropism (mucosal or cutaneous), carcinogenic risk and associated pathology [38]. Currently, over 150 HPV types have been formally identified and predominantly cluster to 3 main genera. (1) Alphapapillomavirus are primarily isolated from the genital, mucosal epithelium and are the overwhelming cause of anogenital cancer; (2) Betapapillomavirus are primarily isolated from skin lesions. The β-PVs include HPV types frequently associated with the rare genetic disorder epidermodysplasia verruciformis (EV), which predisposes individuals to develop HPV-associated cutaneous, scaly wart-like squamous cell carcinomas. Many β-HPV types were originally identified in isolates obtained from EV patients, previously termed HPV-EV types and include HPV-5 and HPV-8, both members of the β-1 species, identified in approximately 90% of EV-related cutaneous squamous cell carcinomas. Less prevalent EV types extend to the β-2 species (HPV-38) and β-3 species (HPV-49). These types are also found associated with malignancy in immunocompromised hosts, and are less prevalent in the general population [58]. (3) Gammapapillomavirus are primarily isolated from cutaneous epithelia, some from cutaneous lesions histologically defined by the presence of homogeneous intracytoplasmic inclusion bodies [1, 59]. Both Gammapapillomaviruses and Betapapillomaviruses have been identified in oral samples suggesting they have an expanded tropism including the oral cavity [60] (fig. 3). Thus, the tropism of these later genera has to be reconsidered in light of the new information. This also demonstrates that not testing a specific anatomical site (e.g. the oral cavity) for HPV (using methods to detect the gamut of types) does not mean the virus is not there.

α-HPV Phylogenetic Association with Cancer Risk

Phylogenetic analysis based on the sequence of the HPV L1 ORF has been the standard for genomic analysis and type classification [1, 19, 54]. Observations of phylogenetic incongruence within the α-PVs are shown by comparison of trees built utilizing either the early or late regions of the genome. This results in differences regarding the monophyletic origin of the oncogenic α-HPV types. Trees generated from the early genes or the complete genome cluster α-PVs by associated oncogenic risk as a mono-
Phylogenetic incongruence is suggestive of genomic distinctions within the α-PVs that are exemplified by examining differences inherent to the species α9 (HPV-16-related) and α7 (HPV-18-related) that differ in biology and pathological outcome. HPV-16 and HPV-18 are highly prevalent, oncogenic types implicated in 70% of cervical cancer, and as members of different species groups they exhibit different biological niches and cellular tropisms, and manifest as precancerous and/or cancerous lesions differently. HPV-16 targets the cutaneous squamous epithelia abundant in the...
ectocervix and predominantly causes cervical squamous cell carcinoma that evolves through differing grades of squamous intraepithelial neoplasia observed by histological/cytological screens [61]. HPV-16 also infects the squamous epithelia of the oropharynx and is identified in oropharynx cancer (targeting the regions of the throat including the base of the tongue, the soft palate and the tonsils). Additionally, HPV-16 is implicated as causal agent for cancers of the vulva, vagina, penis and anus [18]. Conversely, HPV-18 disproportionately targets glandular, mucin-secreting columnar epithelial cells found primarily in the endocervix, and is overrepresented in the development of cervical adenocarcinoma [62, 63].

Comparison of phylogenetic trees generated from either early or late gene sequences results in different phylogenies for distinguishing higher-level taxa [48]; the clades defining species groups remain intact. Yet, the nodes defining a clades’ most recent common ancestor appear differently depending on the genes used for the phylogenetic tree construction, reflecting a degree of phylogenetic incongruence. This hints at the occurrence of early selection events likely driven by ecological niche adaptation such that early and late genes are regulated through distinct promoters that are strictly governed by the availability of host proteins differentially expressed within the stratified epithelia [25, 48]. Incongruence may also result from disproportionate selective pressures on the two genomic regions. Incongruence does not likely result from an early viral recombination event, as similar genome characteristics are maintained across diverse hosts throughout evolutionary time, such as humans and non-human primates, supporting the clonal nature of viral expansion as opposed to recombination [38, 48].

**Variant Lineages: Model for Recent Human Papillomavirus Speciation**

Classification below the species level is not formally recognized by the International Committee on the Taxonomy of Viruses [19, 55]. Evidence from large epidemiological studies identifying HPV genomic heterogeneity and associated pathologies supports the need for distinction of taxa below the HPV type level and will likely be updated soon [55]. The term HPV subtype has become obsolete. It previously referred to isolates that exhibited 2–10% differences in nucleotide identity compared to its closest known type and/or differences in restriction enzyme cleavage patterns [1]. As better systems for identification and classification of HPVs emerged, subtypes are now considered variant lineages.

**Variant Lineages and Sublineages: Updates on Current Nomenclature Guidelines**

HPV variant lineages represent viral isolates that exhibit genomic nucleotide differences of 1–10%, as compared to their prototype, or reference genome [5, 6, 27, 50, 55, 64]. Viral variants share a most recent common ancestor that is unique to the specific HPV type. HPV variants are common and may differ in the risk for development of
cervical precancer (CIN2/3) or cancer. Initial studies examining HPV-16 and HPV-18 intratypic variation aimed to discern the contribution of variant lineages to geographic differences in virus distribution [5, 6, 65]. Variants of both HPV-16 and HPV-18 were initially classified by sequencing the URR, as this noncoding region exhibits greater nucleotide variability than protein-coding regions and is a valuable tool for identifying stable nucleotide polymorphisms, the basis for HPV classification. However, unlike type identification that is sufficiently designated by sequencing the L1 ORF, complete genome sequencing is required to identify HPV variants since the distribution of differences is not evenly spaced across the genome. In addition, a variant nomenclature system based on the complete genome permits the identification and quantification of nucleotide polymorphisms using different regions of the genome [55]. Evidence supports that certain genomic regions exhibit greater heterogeneity than others.

HPV replication depends on host DNA replication machinery; proofreading capabilities by host DNA polymerases maintain a low rate of mutation within the viral genome at approximately $10^{-8} - 10^{-7}$ nucleotide substitutions/year [4]. Intratypic HPV genetic variation results from random nucleotide polymorphisms or insertions/deletions (indels) acquired through genetic drift or natural selection that become fixed over time. Stable acquisition over time, of these nucleotide changes, eventually leads to PV speciation through a process termed lineage fixation [57] and is further supported by a lack of evidence for viral recombination events [49]. Furthermore, the stable acquisition of polymorphisms among isolates of the same HPV type (sharing a most recent common ancestor) eventually leads to type speciation, characterized by genomic nucleotide identity that is less than 90% and occurs over millions of years [46]. To this end, evidence of prehistoric human population bottlenecks is reflected through inter- and intratypic PV genomic diversity. Distinction below the species level (PV type) is common within the PV research community, and is useful for physicians, researchers and epidemiologists investigating HPV variants for association with geographical host population and viral genetic changes that confer variable phenotypic outcomes, including varying pathological manifestations such as cancer. At present, guidelines for variant HPV lineage classification are beginning to be formally established [55]. A formal classification and nomenclature system to describe intratypic HPV variants, at the lineage or sublineage level, will undoubtedly become increasingly useful as the future of HPV genomics expands to include data obtained from next-generation sequencing and metagenomic studies, and will facilitate a better system for cataloguing genotype-phenotype changes.

**Variant Lineages and Sublineages: Identification and Clinical Relevance**

Variant lineage classification is based on isolates of a known type that have had their complete genome sequenced and reveal genetic heterogeneity of at least 1% and less than 10%, based on multiple sequence alignments to the prototype (first characterized genome of a given type). Parameters for variant classification have recently been established by phylogenetic analysis on isolates from the $\alpha_9$-species (including
HPV-16, HPV-31, HPV-33, HPV-35, HPV-52, HPV-58 and HPV-67), types from the α-7 species and two types from the α-10 species, HPV6 and HPV11 [27, 55, 64]. A divergence rate of 1% conservatively designates variant lineages. Similarly, pairwise nucleotide identity differences in the range of 0.5–1% define a type sublineage. Nomenclature for lineage and sublineage is based on an alphanumeric system wherein the prototype ‘reference’ genome is always designated with an ‘A’. If sublineages for the given type are present, the prototype reference sequence is designated as ‘A1’ (table 1).

Variant lineages from many clinically relevant α-HPVs are known. Variants have been characterized for the majority of the HR mucosal HPV types: HPV-16 [5, 65, 66], HPV-18 [6, 46, 64], HPV types 31, 33, 35, 52, 58, and 67 [27]; other α-HPV variants that have been identified include members of the α-4 species, i.e. HPV-2, HPV-27 and HPV-57 [67], and two types from the low-risk α-10 species, specifically HPV-6 and HPV-11 [55], associated with genital warts have been extensively examined.

Phylogenetic analysis of HPV-16 variation revealed variants reflective of human dispersal out of Africa and divergence into the 3 major human races, Africans, Caucasians and Asians. Initial phylogenetic studies identified 4 major HPV-16 variants representative of host geographic origin. The major HPV-16 variants were broadly termed ‘European’ now lineage ‘A’ or ‘non-European’ now lineages ‘B, C and D’. The non-European lineages are more heterogeneous and included 2 African HPV-16 lineages, African-1 and African-2 (B and C, respectively), and Asian-American variants (lineage D) [5, 57, 62, 68]. These variants display phylogenetic congruence with host ethnicity and geographic origin [5]. However, with the recent update to nomenclature the designation of variant lineage simply (for the case of the HPV-16 example) as non-European or European is misleading and overlooks details within the viral genomes (fig. 4). Previously termed HPV-16 variants have thus been renamed according to the

---

Table 1. HPV taxa definition and nomenclature overview

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Defining genomic characteristics</th>
<th>Region</th>
<th>Example (HPV-16 reference genome)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
<td>genome size approx. 8 kb, distinct genome organization</td>
<td>L1</td>
<td>Papillomaviridae</td>
</tr>
<tr>
<td>Genus</td>
<td>&gt;40% difference between genera</td>
<td>L1</td>
<td>Alphapapillomavirus</td>
</tr>
<tr>
<td>Species</td>
<td>30–40% difference to other types</td>
<td>L1</td>
<td>Alphapapillomavirus 9</td>
</tr>
<tr>
<td>Type</td>
<td>&gt;10% difference to known types</td>
<td>L1</td>
<td>HPV-16</td>
</tr>
<tr>
<td>Lineage</td>
<td>1–10% difference to same type</td>
<td>CG</td>
<td>variant lineage A</td>
</tr>
<tr>
<td>Sublineage</td>
<td>0.5–1% difference to same type</td>
<td>CG</td>
<td>variant sublineage A1</td>
</tr>
</tbody>
</table>

CG refers to the complete genome sequence used for identification and nomenclature. The table lists the current taxonomic distinctions used for classification of PVs illustrated by the HPV-16 reference genome. The defining genomic characteristics refer to the aligned nucleotide sequence identity differences based on either the sequence of the L1 ORF or complete genome; these values are used to distinguish HPV taxa. The proper taxonomic orthography is shown for the HPV-16R genome. The inclusion of lineage and sublineage designation is not currently recognized by the International Committee on the Taxonomy of Viruses, but represents important taxa distinction recognized by the PV community.
updated variant nomenclature guidelines, whereby European variants are classified as the ‘A’ lineage, further resolved to 4 sublineages (A1, A2, A3, A4). The HPV-16 non-European variants are now recognized as 3 distinct lineages with the appropriate sub-lineage designation: HPV16-B (previously Af-1) contains sublineages B1 and B2; HPV-16C (previously Af-2); HPV-16D (previously NA1, AA1, AA2) contains sublineages D1, D2 and D3. This update helps to resolve and term the viral distinctions below the type level and is important since many of the nucleotide changes are correlated within taxa (fig. 4).
The Future of Human Papillomavirus Genomics

Biomedical Technology Advancements and Biomarker Prediction
Global epidemiological studies now permit access to vast numbers of clinical HPV samples to better understand the natural history of HPV infections [50] and the biological and clinical ramifications [50, 69]. The risk associated with a persistent HR infection, such as HPV-16, is significantly associated with neoplastic progression. However, determinants of viral persistence and clearance are not well understood. Recent evidence suggests differences in the duration of persistence for different HPV types and/or lineages associated with clinical outcomes [50, 70, 71]. Understanding of HPV genomics and classification will help further characterize host/viral components involved in viral persistence and/or clearance [72, 73]. This may contribute to the design of new treatments and therapeutics targeting those most at risk for cancer. On a broader scale, it may also help delineate which infections require treatment versus those that naturally regress, having implications on the financial burden associated with cancer prevention and treatment globally.

Rapid technology advances during the past decade have, and will continue to, advance our understanding of PV genomic heterogeneity that contributes to pathological clinical outcomes. Next-generation sequencing provides increased resolution for the detection and classification of viral DNA, at the single molecule level, and will enhance methods for identification of new HPV taxa that may have previously fallen under the limit of detection by current methods [74]. This will facilitate a more complete view of Papillomaviridae diversity and should contribute to an improved understanding of the mechanisms of HPV-associated malignancy [75, 76]. Increased use of novel genotyping methods will improve the efficiency of identifying HPV DNA from an array of samples obtained by large cohort studies. The explosion in data generated by next-generation sequencing techniques reinforces the need for current, widely accepted methods to characterize the results. This is highlighted by recent reports describing an extensive array of novel Betapapillomavirus and Gammapapillomavirus types, which are currently difficult to classify given the extensive diversity of these types and the lack of a simple method for their characterization in a standard clinical assay.

New Codes in the HPV Genome
The appreciation of an additional ‘epigenetic code’ (i.e. CpG sites that can be methylated) in the HPV genome has energized new and future studies on understanding the significance of information. HPV epigenetics is a burgeoning field [for a review, see 77]. Accounts of CpG methylation of viral DNA, predominately from HPV-16 and HPV-18 were initiated at the turn of the 21st century and identified regions of viral CpG methylation by methylation-specific restriction endonuclease maps. Different cleavage patterns from viral isolates obtained from women with different stages of HPV-associated CIN have been observed. The onset of technological advances within the past 2 decades has generated methods for more accurately identifying and quantitating DNA methylation. CpG sites are highly conserved amongst HPV species. CpG sites appear
throughout the genome and exhibit varied methylation states and likely play a physiological role in the viral life cycle. Large epidemiological studies that aim to elucidate the oncogenic properties of HPV have demonstrated that methylation at specific CpG sites within the viral genome is predictive of clinical outcome, such as precancer (CIN2/3) or cancer [77–79]. Specifically, CpG sites found within the HPV-16 L1 ORF are highly predictive of CIN2/3 [78]. Interestingly, the α-7 species contain higher numbers of CpG sites relative to the α9-species. The development of high-resolution methods for HPV identification and classification should lead to the discovery of additional coded information in the HPV genome beyond nucleotide, amino acid, CpG and DNA-binding protein regions.

References


