HHS Public Access

Author manuscript

Methods Mol Biol. Author manuscript; available in PMC 2016 January 01.

Published in final edited form as:

Methods Mol Biol. 2015 ; 1249: 3–26. doi:10.1007/978-1-4939-2013-6_1.

Evolution and classification of oncogenic human papillomavirus types and variants associated with cervical cancer

Zigui Chen, **Luciana Bueno de Freitas**, and **Robert D. Burk**[1]

Albert Einstein College of Medicine, Jack and Pearl Resnick Campus, 1300 Morris Park Avenue, Ullmann Building, Room 515, Bronx, NY, 10461, USA

Abstract

The nomenclature of human papillomavirus (HPV) is established by the International Committee on Taxonomy of Virus (ICTV). However, the ICTV does not set standards for HPV below species levels. This chapter describes detailed genotyping methods for determining and classifying HPV variants.

Keywords

Papillomavirus; Evolution; Taxonomy; Genus; Species; Type; Variant

1 Introduction

Papillomaviruses (PVs) are a heterogeneous group of viruses with circular double-stranded DNA genomes about 8,000 nucleotides in size. PV genomes include three general regions: (1) an upstream regulatory region (URR), which contains sequences that control viral transcription and replication; (2) an early region, which contains open reading frames (ORFs; e.g., E1, E2, E4, E5, E6, and E7) involved in multiple functions including trans-activation of transcription, transformation, replication, and viral adaptation to different cellular milieus, and (3) a late region, which codes for the L1 and L2 capsid proteins which form the structure of the virion and facilitate viral DNA packaging and maturation. All PVs described to date contain an E1, E2, E4, L1, L2 and some E6/E7-like functions.

Papillomavirus nomenclature is established by the International Committee on Taxonomy of Viruses (ICTV) based on recommendations from the Study Group of Papillomavirus [1–3]. PV taxa are defined based on L1 nucleotide sequence identities and their topological position within a PV phylogenetic tree. Based on global multiple sequence alignment and a matrix of pairwise comparisons, the distribution of L1 identities shows a bimodal pattern consistent with the genus (<60 % identity) and species (60–70 % identity) nomenclature (*see* Figure 1 in Bernard 2010). PV genera are designated using the Greek alphabet (e.g., *Alphapapillomavirus*). The prefix “dyo” (i.e., Greek “a second time”) is added to the Greek

[1]Corresponding author: burk@aecom.yu.edu.

letter to encompass the expanding genera of PVs. Species within genera are named by a number (e.g., *Alphapapillomavirus 9*).

The ICTV does not set standards below the species level. Papillomavirus researchers evolved a “community” nomenclature that has been widely embraced and extremely useful in epidemiological studies. A distinct papillomavirus “type” is established when the nucleotide sequence of the L1 gene of a cloned virus differs from that of any other characterized types by >10 % [1, 2]. Papillomavirus types are named based on the scientific name of the host from which the PV genome was isolated, using the host genus and species designation. In case of overlaps, a third letter is added to give each PV type a unique name (e.g., the *Caretta caretta* PV became CcPV1, while the *Capreolus capreolus* PV became CcaPV1). Nevertheless, some historical names are maintained, e.g., “HPV” (with H standing for human or Homo without incorporating the species designation “sapiens”) [2]. A Bayesian tree inferred from alignment of protein and nucleotide sequences of six concatenated ORFs (E6, E7, E1, E2, L2, and L1) of the mucosal/genital alpha- HPVs shows the relationships of species and types associated with cervical neoplasia. For example, human papillomavirus 16, abbreviated as HPV16, is a type within species *Alphapapillomaviruses* 9 (designated alpha-9 or α9) of the genus *Alphapapillomaviruses* (*see* Fig. 1).

For a papillomavirus to be recognized as a distinct/novel type by the HPV reference center (http://www.hpvcenter.se/), the full genome should be cloned and the sequence of the L1 gene should be no more than 90 % similar to previously curated and named PV types [1, 2]. Isolates of the same HPV type are referred to as variants when the nucleotide sequence of the L1 gene of a cloned virus differs from that of any other characterized types by less than 10 %. The L1-based classification of (genital) HPVs correctly groups HPV types into species and genera even with phylogenetic tree incongruence between trees inferred from different ORFs/regions (*see* Fig. 2). However , a single gene/region (e.g., L1 ORF) does not always contain sufficient sequence information for unambiguously distinguishing closely related HPV variants. A common nomenclature for HPV variants for the multiplicity of HPV types is being implemented using the complete genome [4–6]. Distinct variant lineages are defined by a nucleotide sequence difference of approximately 1.0–10.0 % between two or more variants of the same type. Similarly, differences across the genome of 0.5–1.0 % are used to identify sublineages. Each variant lineage is classified and named with an alphanumeric value. The prototype sequence (i.e., the cloned genome selected as the original type) is always designated variant lineage A and/or sublineage A1 (*see* Fig. 3). A comprehensive clas- sification system will significantly facilitate understanding the clinical role sequence variations play in genotype–phenotype associations. An established variant lineage nomenclature allows investigators to classify isolates based on a limited amount of sequence information from almost anywhere within the genome. Since there are many highly correlated sets of single nucleotide polymorphisms (SNPs) within each lineage/ sublineage, this allows HPV researchers to classify HPV variant lineages without having to refer to specific changes at nucleotide positions. Furthermore, this classification defines a group of HPV variants that can be combined with other studies, sequencing different regions of the viral genome. Nevertheless, as with the human genome, rare variants accumulating in

the HPV genome with the rapid expansion of the human population are not completely captured by variant lineages, which represent changes that occurred over tens to hundreds of thousands of years ago [7].

A freely accessible, web-based tool, Papillomavirus Episteme (PaVE, http://pave.niaid.nih.gov), provides an integrated resource for the analysis of papillomavirus (PV) genome sequences. A detailed schematic for determining whether a viral isolate represents a novel PV type or is a variant of an existing PV type is available in PaVE (http://pave.niaid.nih.gov/#prototypes?type=submission and http://pave.niaid.nih.gov/#prototypes?type=variant%20genomes, respectively).

The ICTV does not deal with taxonomic nomenclature below the species level (e.g., serotypes, strains, and subtypes). In addition, HPV46, 55, and 64 did not meet the updated criteria as unique HPV types; they were found to be variants of HPV20, 44, 34, respectively, and their numbers left vacant [1]. PV types cloned from PCR products are now accepted for curation and classification and the briefly used term “candidate type” for these genomes has been eliminated [2].

Taking HPV16 as an example, in this chapter we provide the methods of how to identify variants by PCR amplifying and Sanger sequencing a partial region (e.g., URR and/or E6). In some instances, based on isolates containing unique polymorphisms in a limited genomic region, can have their complete genome sequenced for further classification (i.e., viral variant lineages and sublineages are only designated from differences in the complete genome) (*see* Fig. 4). Algorithms of phylogenetic analysis to classify variant lineages using complete genomes are also described. Alternatively, if a potential novel PV type is observed, the procedure to clone the complete genome, submit the cloned genome for official naming, and classify the phylogenetic position are described.

2 Materials

2.1 Reagents for PCR

2.1.1 Clinical Specimens and DNA Preparation—Appropriate samples include exfoliated cervical cells collected with a Dracon swab, cytobrush, or obtained by cervicovaginal lavage. Cervix lesion biopsy material, either fresh or formalin fixed can also be used, but will not be described here. The exfoliated cervical cells are usually collected in commercial buffers which allow both cytological examination and HPV testing, e.g., the methanol-based PreservCyt® (Hologic Inc., Marlborough, MA) or ethanol-based SurePath® (Becton Dickinson, Franklin Laker, NJ). Alternatively, if only HPV testing will be preformed they can be collected in normal saline (9.0 g NaCl/L), or Digene® specimen transport media (STM) (Qiagen, Valencia, CA). Other transport media are also available. Depending on time of storage and other considerations, samples are usually maintained frozen at −20 or −80 °C, until processing occurs; freezing and thawing should be avoided.

A variety of methods and commercial kits can be used to isolate nucleic acids from cervicovaginal epithelial cells for PCR amplification. In general, DNA material for HPV PCR can be isolated from cellular material after digestion with proteinase K and detergent

(e.g., Laureth-12 is compatible with PCR), followed by phenol–chloroform extraction and ethanol precipitation, or by one of many proprietary commercial DNA extraction kits.

2.1.2 Cervical HPV Genotyping—Most laboratories use a pair of degenerate general primers or consensus general primers to amplify a short segment of target DNA in the L1 gene to detect the large spectrum of HPV types associated with cervix cancer. The genotyping of the HPV-positive samples can be carried out using a variety of methods including dot blot hybridization employing biotinylated type-specific oligonucleotide probes that recognized type-specific sequences within the amplified fragments. Alternatively, the HPV fragments can be amplified with biotinylated primers and reverse hybridized to a strip or bead for type determination. Commonly used PCR primers include the degenerate MY09/MY11 (MY) primer pair or the modified version, PGMY09/PGMY11, in addition to other primer pairs targeting slightly different regions within the highly conserved L1 ORF, e.g., GP5+/GP6+ and SPF [8–11]. The MY assay includes a control primer set (PC04/GH20), which simultaneously amplified a 268-bp cellular beta-globin DNA fragment and serves as an internal control for amplification. Controls for HPV detection should include a positive control, e.g., 100-cell copy and a 2-cell copy of SiHa DNA, a HPV negative control, e.g., 100-cell copy of HuH7 DNA, and a water blank. The preparation of sample DNA should take place in a physically isolated region (i.e., pre-PCR) from where the PCR and typing will occur.

2.1.3 HPV Variant Detection: Primer Design and Selection—All primers used to classify HPV variants are type-specific. The primers amplifying partial genome regions should be located within a region of the genome that is variable and informative (e.g. non-coding regions) and should be as small as possible (*see* **Note 1**). The short fragment amplification will help increase the PCR efficiency. Table 1 shows the one-tube nested PCR primers used to amplify a fragment of the URR (267 bp) and E6 (158 bp) regions of HPV16 (the outer and inner primers use different annealing temperatures, *see* **Note 2**). The PCR products are purified for direct Sanger sequencing; sequences will be aligned with the prototype sequence to determine identity of single nucleotide polymorphisms (SNPs) and/or insertions and deletions (indels).

Specimens containing a possible unique variant or major variation patterns will be chosen to amplify and sequence the complete genomes [4, 7]. Type-specific primer sets are designed based on the prototype sequences for nested PCR to amplify the complete genomes in 2–3 overlapping fragments (as shown in Table 2) (*see* **Note 3**). The outer PCR product serves as

[1]For defining HPV variants, use the most heterogeneous genomic regions, for example, the URR or non-codon region between E2 and L2. Conserved regions should provide major lineage information but might not properly classify isolates into sublineages (e.g., E6 region of HPV16 does not distinguish some sublineages [13]).

[2]One-tube nested PCR for HPV variant analyses increases amplification yield and prevents potential cross-contamination. The outer primers are present at low concentrations and have common annealing temperature (~55 °C) in the first 30 cycles to yield limited product that serves as the temperate of the inner PCR amplification by the primers with lower annealing temperatures (~45 °C) in the second 35 cycles. A second region, even if not as informative (e.g., E6), can be tested for specimens that did not yield data for the URR region. The second line testing should amplify a smaller fragment for increased sensitivity.

[3]Overlapping nested PCR of 2–4 fragments is suggested to represent the complete genome. However, additional primers targeting smaller sizes are sometimes necessary to amplify the genomes that do not amplify all fragments. Alternatively, rolling circle amplification (RCA) could be used to amplify the complete genome with specific primers. However, this method is rarely successful when overlapping PCR does not work. RCA requires high copy numbers of intact circular viral genomes.

the template for the inner PCR. In order to minimize PCR errors, a proofreading high fidelity DNA Taq polymerase is recommended. Variations that appear to disrupt an ORF, splice site and/or are seen only once should be validated by an independent PCR and sequencing experiment. Purified products are either directly sequenced or cloned into TOPO TA pCR2.1 vectors (Invitrogen, Carlsbad, CA) and then sequenced. Subsequent sequencing is performed using primer walking; the complete genomic sequences are then assembled from the sequences of overlapping fragments using the prototype as a scaffold. Figure 5 shows the protocol to amplify HPV16 complete genomes in three overlapping fragments (2,988 bp, 2,843 bp, and 2,933 bp, respectively).

2.1.4 Molecular Biology Reagents

1. AmpliTaq Gold® DNA Polymerase (Cat. N8080247, Life technologies, Foster City, CA).
2. Platinum® Taq DNA Polymerase High Fidelity (Cat. 11304011, Life technologies).
3. Oligonucleotide primers via IDT (http://www.idtdna.com/).
4. PCR Nucleotide Mix (Cat. 11581295001, Roche, Nutley, NJ).
5. QuickStep™2 PCR Purification Kit (Cat. 33617, Edge BioSystems, Gaithersburg, MD).
6. QIAquick PCR Purification Kit (Cat. 28104, Qiagen).
7. QIAquick Gel Extraction kit (Cat. 28704, Qiagen).
8. TOPO® TA Cloning (Cat. K2040-01, Life technologies).
9. illustra TempliPhi 100/500 Amplification Kits (Cat. 25-640010, GE Healthcare, Little chalfont, UK).

2.2 Equipment for DNA Amplification and Sequencing

1. Polymerase Chain Reaction Thermocycler (ABI 9700, Life technologies).
2. Microcentrifuge.
3. Agarose gel electrophoresis apparatus.
4. UV Image analysis and capture system.
5. Incubator sets.
6. Sanger sequencing facility (usually an academic core or commercial facility).

2.3 Software for Genomic Analysis

1. Genome analysis software (Geneious, BioEdit).
2. Sequence alignment software (ClustalW, Muscle, MAFFT).
3. Genomic diversity software (MEGA5).
4. Phylogenetic analysis software (PAUP, MrBayes, RAxML, PhyML).

5. Tree illustration software (FigTree, TreeView).

3 Methods

PCR products, generated by nested PCR, can either be cloned or used directly for Sanger sequencing (*see* **Note 4**). The following procedure, when used for HPV variant testing and/or complete genome analysis, involves nested PCR, cloning, Sanger sequencing, sequence alignment and phylogenetic analysis. PCR master mixes should be set up in the sample room, and moved to the PCR room for thermocycler reaction.

3.1 HPV Variant Determination

3.1.1 One-Tube Nested PCR for a Partial Genome

- Prepare 1–94 HPV16 DNA samples previously genotyped, plus one positive (HPV16) control and one negative control in the sample preparation room.
- Prepare one-tube PCR master mix in a 15 ml tube by adding the following:

PCR master mix	One reaction	96-well mix
dd water	16.25	1625
10× Buffer (15mM $MgCl_2$)	2.50	250
25 mM $MgCl_2$	2.50	250
dNTP mixture (10 mM each)	0.50	50
Primer 16URR.Fout (0.2 μM)	0.50	50
Primer 16URR.Rout (0.2 μM)	0.50	50
Primer 16URR.Fin (20 μM)	0.50	50
Primer 16URR.Rin (20 μM)	0.50	50
Ampli Taq Gold (5 U/μl)	0.25	25
Sample DNA	1.00	
Total volume	25 μl	24 μl/each

- Vortex PCR master mix and centrifuge briefly to collect all fluid to the bottom of the tube.
- Aliquot 24 μl of PCR master mix into each well of the 96-well plate.
- Add 1 μl of 94 DNA samples into each sample well. Gently mix with the master mix by moving pipette tip up and down.
- Add 1 μl of water into the negative control well, and add 1 μl of HPV16 DNA positive sample or plasmid into the control well. Gently mix with the master mix.
- Seal or cap the 96-well plate firmly, and centrifuge briefly to bring all liquids in each well to the bottom.

[4]For discrepancies between cloned sequences, we suggest the sequence of the PCR product be considered the valid genome sequence, since the sequenced PCR product represents the composite genome.

- Move the plate to the PCR room. Place the plate into the PCR thermocycler, programmed with the following parameters (Program A):

Program A steps		°C	Time
Initial denaturation		95	3 min
30 cycles	Denaturation	94	30 s
	Annealing	57	30 s
	Extension	72	60 s
35 cycles	Denaturation	94	30 s
	Annealing	45	30 s
	Extension	72	60 s
Final extension		72	10 min
Maintenance		4	∞

3.1.2 Gel Electrophoresis

- Prepare a 2–3 % agarose gel containing ethidium bromide (final concentration 0.5 μg/ml).
- Place a freshly prepared gel slab in the electrophoresis chamber and cover with electrophoresis buffer.
- Pipette 3 μl of the molecular size marker (e.g., 100 bp DNA ladder) into the first well of every gel slab row being used.
- Pipette 5 μl of the PCR product mixed with 1 μl loading dye (×6) into each corresponding well in the gel.
- Set the voltage to 100, and the timer to 40 min. The time will depend on the size of the gel box, thus the run should be followed by eye, visualizing the migration of the gel-loading buffer, which is blue. The run should be finished when the bands of the loading buffer dye reach 1.5 cm from the end of the gel.
- After electrophoresis is finished, place the gel on the viewing tray in the UV light box to visualize the DNA fragments, capture the image and print (if available).
- Confirm if the sizes of the PCR products in the gel are exactly the predicted size using the DNA ladder as a guide.

3.1.3 PCR Product Purification—If there is a single amplicon of predicted size it can be directly purified by a column; if there are multiple bands, one of which appears to be the band of interest, it will have to be cut out of the agarose gel and purified. Follow the protocol of QuickStep™2 PCR Purification Kit (Single Cartridges or Ultra Plates) (Edge BioSystems) or QIAquick PCR Purification Kit (Qiagen) to purify single band amplicons. If the amplicon contains multiple bands, purify the cut out band in the gel using QIAquick Gel Extraction kit (Qiagen) following the manufacturer's protocol.

3.1.4 Sanger Sequencing—The purified PCR products are sequenced using the Sanger sequencing method. Use the inner PCR primers (e.g., 16URR.Fin and 16URR.Rin; *see* Table 1) as the sequencing primers (the sequencing primers need to be diluted as directed by the Sanger sequencing facility). Sequencing both strands verifies the sequence and is useful when one strand does not give good sequence information. Contact the Sanger sequencing facility for their requirements for sample preparation. After Sanger sequencing, an electron file is provided with the sequence and usually a figure of the peaks for each sample.

3.1.5 Sequence Alignment and Variant Determination

- Import sequence files in ABI format in Geneious software.
- For each sample, assemble sequences using the forward and reverse sequencing primers mapped to the prototype sequence.
- Trim bases at the ends with low sequence quality, and validate SNPs that differ from the prototype sequence. If there is disagreement between SNPs using the forward and reverse sequences, the sample should be repeated. It should be noted that if the sample contains multiple HPV types, the forward and reverse sequencing primers can sometimes generate different sequences (*see* **Note 5**).
- Extract the consensus nucleotide sequence of each sample.
- Align all extracted consensus sequences plus the prototype sequence using MAFFT with default parameters.
- Generate a maximum likelihood tree using an integrated tree software program, for example, PHYLP, PhyML, etc. All variants should cluster into different groups based on sequence similarity.
- Export the alignment in FASTA format that can be opened in MEGA5.
- Within MEGA5, highlight all variable sites and save as an excel sheet. Note the position of each variation using the nucleotide numbering positions of the prototype.
- Review potential distinct variants or variation patterns by compiling the data in Excel.
- Based on the tree topology and variant patterns, choose samples containing distinct variants (i.e., SNPs and/or indels) or major variant patterns for complete genome analysis (*see* **Note 6**).

[5]Since the HPV genomes are evolving through nucleotide changes and not gross rearrangement or recombination, nucleotide changes in one region (e.g., URR) are highly correlated with and inseparable from changes in other regions (e.g., URR) within genomes from the same lineage. Multiple variant infections, although rare, might be indicated if the DNA sequence traces have multiple nucleotide peaks at SNP positions.

[6]Samples containing major variant patterns, unique variations/ indels or distinct variants within the partial region(s) should be chosen for the complete genome analysis. If available, at least two or more samples with the same variations should have their complete genomes sequenced. Moreover, isolates from different ethnicities, geographical regions or histological categories may help to capture a greater extent of the genomic diversity of a specific type.

3.2 Complete Genome Amplification and Sequencing

3.2.1 Nested PCR to Amplify the Complete Genome in Three Overlapping Fragments

- Based on the SNPs and/or indels observed within the initially sequenced genomic fragment, prepare samples representing potential distinct variants or containing major variant patterns in the sample preparation room.
- Prepare the “outer” PCR master mix for Fragment 1 (with Fragment 1 outer PCR primers, *see* Table 2 and Fig. 5) by adding the following. Separate “outer” PCR master mixes should be prepared for Fragments 2 and 3.

PCR master mix	One reaction	96-well mix
dd water	17.25	1725
10× Hi-Fi Buffer (25 mM $MgSO_4$)	2.50	250
$MgSO_4$ (50 mM)	1.00	100
dNTP mixture (10 mM each)	0.50	50
BSA (×100)	0.25	25
Primer 16.3Fr1.Fout (20 μM)	0.50	50
Primer 16.3Fr1.Rout (20 μM)	0.50	50
Platinum Taq (5 U/μl)	0.25	25
Ampli Taq Gold (5 U/μl)	0.25	25
Sample DNA	2.00	
Total volume	25 μl	23 μl/each

- Vortex PCR master mix and centrifuge briefly to collect all fluid at the bottom of the tube.
- Aliquot 23 μl of PCR master mix into 0.2 ml Thermo-Tube or 96-well plate wells.
- Add 2 μl of each sample DNA into each tube or well. Gently mix with the master mix.
- Include one negative control tube by adding 2 μl of water. Gently mix with the master mix.
- Cap the tube(s) or plate, and centrifuge briefly to bring all liquids to the bottom.
- Move to the PCR room. Place samples in the PCR thermocycler using the following parameters (Program B):

Program B steps		°C	Time
Initial denaturation		94	2 min
40 cycles	Denaturation	94	30 s
	Annealing	57	30 s

Program B steps	°C	Time
Extension	68	4 min
Final extension	68	10 min
Maintenance	4	∞

- When the outer PCR is complete, prepare inner PCR master mixes in the sample room by adding the following for each fragment separately. Each inner PCR reaction corresponds to the same outer PCR and uses 1 μl from the outer PCR mix (be careful not to confuse the outer PCR products).

PCR master mix	One reaction	96-well mix
dd water	18.25	1,825
10× Hi-Fi Buffer (25 mM $MgSO_4$)	2.50	250
$MgSO_4$ (50 mM)	1.00	100
dNTP mixture (10 mM each)	0.50	50
BSA (×100)	0.25	25
Primer 16.3Fr1.Fin (20 μM)	0.50	50
Primer 16.3Fr1.Rin (20 μM)	0.50	50
Platinum Taq (5 U/μl)	0.25	25
Ampli Taq Gold (5 U/μl)	0.25	25
Outer PCR product	1.00	
Total volume	25 μl	24 μl/each

- Vortex PCR master mix and centrifuge briefly to collect all fluid to the bottom of the tube.
- Aliquot 24 μl of PCR master mix into 0.2 ml Thermo-Tube or 96-well plate.
- Move to the PCR room. Add 1 μl of outer PCR product into each tube or well. Gently mix the DNA sample with the master mix in the tube or well.
- Include one negative control tube by adding 2 μl of water. Gently mix with the master mix.
- Start the inner PCR amplification following the PCR Program B as shown above (Subheading 3.2.1).

3.2.2 Gel Electrophoresis and Product Purification

Follow Subheading 3.1.2 to visualize both outer and inner PCR products by gel electrophoresis. For samples positive by both outer and inner PCRs, the specific amplicons with better yields will be purified as in Subheading 3.1.3 . Purified products will be used for direct Sanger sequencing or cloned in the vector and then submitted for sequencing.

3.2.3 TOPO TA Cloning

- Set up a 3′ “A-overhangs post-amplification” for blunt-ended fragments, since the proofreading polymerases (Platinum Taq) remove the 3′ A-overhangs necessary for TA cloning. Incubate at 72 °C for 30 min (do not cycle).

10× Buffer (15 mM $MgCl_2$)	1.00 μl
dATP (2 mM)	1.00 μl
Water	5.75 μl
Purified PCR product	2.00 μd
Ampli Taq Gold (5 U/μl)	0.25 μl
Total volume	10 μl

- Set up the TOPO Cloning reaction for eventual transformation into chemically competent DH5α-T1 *E. coli*. Mix reaction gently and incubate for 5–30 min at room temperature. Then place the reaction on ice.

3′ A-overhangs product	3 μl
Water	1 μl
Salt Solution	1 μl
TOPO vector (pCR2.1)	1 μl
Total volume	6 μl

- Thaw on ice chemically competent DH5α-T1 *E. coli* that should have been stored at −80 °C.
- Add 3 μl of the TOPO Cloning reaction into 50 μl chemically competent DH5α-T1 *E. coli* and mix gently. Do not mix by pipetting up and down.
- Incubate on ice for 5–30 min.
- Heat-shock the cells for 30–45 s at 42 °C without shaking.
- Immediately transfer the tubes to ice.
- Add 250 μl of S.O.C. medium at room temperature , cap the tube(s) tightly and shake the tube horizontally (200 rpm) at 37 °C for 1 h.
- Spread 10–50 μl from each transformation on a prewarmed agar plate with the appropriate antibiotic. To ensure even spreading of small volumes, add 20 μl of S.O.C. medium.
- Incubate plates at 37 °C. If using ampicillin selection, visible colonies should appear within 8 h, and blue/white screening can be performed after 12 h.

- Circle ~10 white or light blue colonies and pick up half of each using a sterile tip and place the bacteria into a 0.2 ml Thermo-Tube, usually by touching the medium, containing 10 μl PCR master mix with M13F and M13R primers.

dd water	7.3 μl
10× Buffer (15mM $MgCl_2$)	1.0 μl
25 mM $MgCl_2$	1.0 μl
dNTP mixture (10 mM each)	0.2 μl
M13 Forward (20 μM)	0.2 μl
M13 Reverse (20 μM)	0.2 μl
Ampli Taq Gold (5 U/μl)	0.1 μl
Bacteria/colony	-
Total volume	10 μl

- Place tubes into the PCR thermocycler using the following program (Program C):

Program C steps		**°C**	**Time**
Initial denaturation		94	2 min
30 cycles	Denaturation	94	30 s
	Annealing	55	30 s
	Extension	72	4 min
Final extension		72	10 min
Maintenance		4	∞

- Electrophoresis the amplified products in 1 % agarose gels to visualize the PCR products. The anticipated size of the insertion from the colony should be around 3,000 bp.
- Pick 3–5 white or light blue colonies with the gel-verified insertion for FastPlasmid Miniprep (Eppendorf, Hamburg, Germany) to harvest plasmid DNA following the manufacturer's protocol.

3.2.4 Complete Genome Sanger Sequencing—Either purified PCR products or plasmid DNA can be used for Sanger sequencing. Because each Sanger sequencing reaction can read ~ 500 bp with good quality, 6–7 sequencing primers are required to “walk” across each overlapping fragment, as listed in Table 3 and should generate the complete sequence ~3,000 bp in length of the HPV fragment.

3.3 Complete Genome Sequence Analysis

3.3.1 Sequence Assemble and Alignment

- Import sequence files in ABI format into Geneious software.
- For each sample, assemble sequences mapping to the prototype sequence.

- Validate SNPs or indels differing from the prototype sequence. For discrepancies between cloned sequences, we used the sequence of the PCR product as the valid "consensus" sequence.
- Compile the complete genome for each sample.
- Align complete genome nucleotide sequences of all samples plus the prototype sequence using MAFFT.
- Export the sequence alignment in FASTA or PHYLIP format for further analysis.

3.3.2 Phylogenetic Tree Construction Using RAxML, a Maximum Likelihood Method

- RAxML (linux) should be installed on a computer.
- RAxML recognizes the nucleotide sequence alignment in PHYLIP format.
- Type the following command in the Terminal windows: $ raxmlHPC -m [MODEL] -s [INPUT_SEQUENCE.phy] -n [OUTPUT] -f a -k -N autoMRE -x 12345 <enter> where [MODEL] could be GTRCAT if choose GTR + Optimization of substitution rates + Optimization of site-specific, [INPUT_SEQUENCE.phy] is the sequence alignment in PHY format, and [OUTPUT] is the specified name of the output file. We suggest use of "-f a" for rapid Bootstrap analysis and search for best-scoring ML tree in one program run, "-k" to print branch lengths to the bootstrapped trees, "-N autoMRE" for the majority-rule tree based criteria, and "-x 12345" to turn on rapid bootstrapping with random seeds. Due to large analytical requirements, it can take several minutes to hours to create the consensus tree. Alternatively, some high-throughput computers (e.g., CIPRES Science Gateway) offer biological sequence analysis including sequence alignment, tree creation, and divergence estimation.
- When the run is complete, several files ended with OUTPUT will be created. RAxML_bipartitions.OUTPUT is the Maximum Likelihood (ML) tree with bootstrap values and branch distances.
- Open the tree file using FigTree, and edit or label the tree.
- Several other programs, such as PhyML, PAUP, MrBayes, offer different algorithms to generate Maximum Likelihood (ML), Maximum Parsimony (MP), Neighbor-Joining (NJ), or Markov Chain Monte Carlo (MCMC) Bayesian trees (*see* **Note 7**). Alternatively, you may also access CIPRES Science Gateway (http://www.phylo.org/index.php/portal/) for inference of large phylogenetic trees.

3.3.3 Complete Genome Nucleotide Sequence Comparisons

- Follow the tree topology to sort the order of sequences exported by Geneious in FASTA format.

[7]Different phylogenetic trees using multiple algorithms can be constructed and compared to infer a comprehensive topology of HPV variants.

- Open the sorted nucleotide sequence alignment in MEGA5.
- Within MEGA5, choose the icon “Distance,” then “Computer Pairwise Distance...”.
- Choose “p-distance” in “Model/Method,” “Pairwise deletion” in “Gaps/Missing Data Treatment.” Click “Continue.”
- Export the p-distance table to Excel format.
- Open the exported p-distance file using Excel to create a chart with pairwise nucleotide sequence differences.

3.3.4 Complete Genome SNP Characterizations

- Open the sorted sequence alignment in FASTA format using MEGA5. Make sure the prototype sequence was sorted on the top of the alignment.
- Within MEGA5, export the active data.
- Select the icon “Use identical symbol.” All identical sites against the reference sequence are replaced with dots. Mark all variable sites.
- Export all highlighted variable sites to Excel or CSV format. Set “For each site” for “Writing site numbers.”
- Edit the SNP patterns for each sample using Excel. Be aware, the site numbers are the sequence alignment positions. If gaps were introduced to the prototype sequence of the alignment, the actual variation site based on the nucleotide numbering of the prototype reference sequence must be adjusted (*see* **Note 8**).

3.3.5 Variant Lineage Classification—Based on the complete genome nucleotide sequences, phylogenetic trees are generated to cluster variants into groups; sequence difference between variants is calculated. Distinct variant lineages have approximately 1.0–10 % nucleotide sequence difference; sublineages have 0.5–1.0 % nucleotide sequence difference (*see* **Note 9**). Each variant lineage is named with an alphanumeric value. The prototype or reference genome is always in variant lineage A and sublineage A1 (*see* ref. 4 for examples).

3.4 Novel HPV Designation

Cervical samples HPV positive by, for instance, the MY09/MY11 PCR, but negative by all type-specific probes may potentially contain a novel HPV type. The following steps involve the designation of a novel HPV and include complete genome cloning and official name assignment.

---

[8]Based on the concept of a single ancestor for each type, a unique genome size is assigned to each HPV type based on the global alignment. Variations in genome size of isolated variants are ascribed to insertions and deletions (indels). Each indel is counted as one event. The assignment of position numbers for each nucleotide is based on the nucleotide numbering of the prototype reference sequence.

[9]The suggested sequence differences defining variant lineages/ sublineages are an approximation of the natural process of evolution, since there are overlaps between lineage and sublineage for some isolates. Thus, as with PV species and genera assignment, variant lineage and sublineages are determined based on a combination of data (*see* ref. 2 for discussion).

3.4.1 Novel HPV Identification

- Purify MY PCR product and submit for Sanger sequencing using both MY09 and MY11 primers.
- Blast the consensus MY sequence for HPV genotyping via NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) or PaVE (http:// pave.niaid.nih.gov/#prototypes?type=classification). The PaVE database contains the PV prototype sequences only, and is more specific but less sensitive than the NCBI database. The NCBI is a more exhaustive database, including sequences of PV (prototype and variants) and non-PV sequences and genomes.
- If using NCBI blast, open the website through the above link. Choose “nucleotide blast,” then in the new webpage, paste or upload the query sequence. Choose “Database” by clicking on “Others,” and type “ *Papillomaviridae* ” in “Organism.” We suggest choosing “Highly similar sequences (megablast)” in the “Program Selection” for a more specific search. Click on the “BLAST” bar.
- If choosing PaVE blast, paste or upload the query sequence in FASTA format. Click on the “submit” bar.
- Both tools will align the query sequence with the sequence of all PVs in the database.
- The species and genotype of your pasted sequence, along with percentage match to established database genotypes will appear on the screen. The closest type will be listed at the top, sorted by the “Max score.” We suggest comparing the blast results using both tools.
- If the alignment with the max score has ≥90 % identity, the virus is nearly certain a variant of an existing PV.
- If the alignment with the max score has <90 % identity (usually >50 % by NCBI blast), the virus is potentially novel. Its complete L1 ORF and complete genome must be amplified and sequenced.

3.4.2 Rolling Circle Amplification (RCA) of Novel HPV DNA—Several methods can be used to amplify the complete genome of novel HPV. Overlapping primers based on the MY sequence and the conserved sequences within the E1 or other regions of the closest type can be used to PCR amplify a large region of the PV genome, similar to the above protocol to amplify the complete genome of HPV variants. Random primers for rolling circle amplification of the PV DNA using the “illustra TempliPhi 100 Amplification Kit” (GE) can also be attempted, as described below.

- Transfer 5 μl of TempliPhi Sample Buffer to an appropriate reaction tube.
- Transfer 0.2–0.5 μl virus DNA (10 ng) to the dispensed TempliPhi Sample Buffer.
- Denature the sample by heating at 95 °C for 3 min. Cool to room temperature or 4 °C.

- In a separate tube, combine 5 μl of TempliPhi Reaction Buffer and 0.2 μl TempliPhi Enzyme Mix.
- Transfer 5 μl of the TempliPhi Premix prepared above to the cooled, denatured sample from **step 3** .
- Incubate at 30 °C for 4–18 h.
- Heat-inactivate the enzyme by heating the sample to 65 °C for 10 min. Cool sample to 4 °C.

3.4.3 RCA Cloning and Sequencing

- Digest RCA product with several 6 bp recognition site restriction enzymes (RE) to identify an enzyme, which linearizes the genome at a single site. Test RE previously shown to cleave PV genomes once (e.g., *BamH* I, *EcoR* I, *Pst* I). An enzyme that cleaves the RCA product once should generate a single band at exactly the size of the complete genome.
- Gel extract one or more digested fragments representing the linearized complete genome.
- Subclone fragments into pUC18/19 pre-digested with the same restriction enzyme.
- Sequence the insertion using M13 forward and reverse primers.
- Design new primers based on the sequenced region to finish sequencing the complete cloned genome using the primer walking technique [12].

3.4.4 Novel HPV Assignment

- Blast the complete L1 nucleotide sequence as in Subheading 3.4.1 . Validate the sequence with similarity <90 % with curated and named existing PVs.
- Submit DNA and sequence to HPV reference center (http://www.hpvcenter.se/) to assign official name. Priority is given to the first lab to submit the cloned sequence to the reference center.
- Submit the complete genome sequence to NCBI/GenBank. If the complete genome sequence is unable to be released immediately, the complete L1 nucleotide sequence should be accessible in NCBI/GenBank.

References

1. de Villiers EM, Fauquet C, Broker TR, Bernard HU, zur Hausen H. Classification of papillomaviruses. Virology. 2004; 324(1):17–27. [PubMed: 15183049]
2. Bernard HU, Burk RD, Chen Z, van Doorslaer K, Hausen H, de Villiers EM. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. Virology. 2010; 401(1):70–79. doi: 10.1016/j.virol.2010.02.002. [PubMed: 20206957]
3. Fauquet, CM.; Mayo, MA.; Maniloff, J.; Desselberger, U.; Ball, LA. Virus taxonomy. The eight report of the international committee on taxonomy of viruses. Elsevier; Amsterdam: 2005. Family pap illomaviridae.; p. 239-255.
4. Chen Z, Schiffman M, Herrero R, Desalle R, Anastos K, Segondy M, Sahasrabuddhe VV, Gravitt PE, Hsing AW, Burk RD. Evolution and taxonomic classification of human papillomavirus 16

(HPV16)-related variant genomes: HPV31, HPV33, HPV35, HPV52, HPV58 and HPV67. PLoS One. 2011; 6(5):e20183. doi: 10.1371/journal.pone.0020183. [PubMed: 21673791]

5. Chen Z, Schiffman M, Herrero R, Desalle R, Anastos K, Segondy M, Sahasrabuddhe VV, Gravitt PE, Hsing AW, Burk RD. Evolution and taxonomic classification of alp hapapillomavirus 7 complete genomes: HPV18, HPV39, HPV45, HPV59, HPV68 and HPV70. PLoS One. 2013; 8(8):e72565. doi: 10.1371/journal.pone.0072565. [PubMed: 23977318]
6. Burk RD, Harari A, Chen Z. Human papillomavirus genome variants. Virology. 2013 doi: 10.1016/j.virol.2013.07.018.
7. Chen Z, DeSalle R, Schiffman M, Herrero R, Burk RD. Evolutionary dynamics of variant genomes of human papillomavirus types 18, 45, and 97. J Virol. 2009; 83(3):1443–1455. doi: 10.1128/JVI.02068-08. [PubMed: 19036820]
8. Bauer, HM.; Greer, CE.; Manos, MM. Determination of genital HPV infection using consensus PCR.. In: Herrington, CS.; McGee, JOD., editors. Diagnostic molecular pathology: a practical approach. IRL Press; Oxford: 1992. p. 131-152.
9. Gravitt PE, Peyton CL, Alessi TQ, Wheeler CM, Coutlee F, Hildesheim A, Schiffman MH, Scott DR, Apple RJ. Improved amplifi cation of genital human papillomaviruses. J Clin Microbiol. 2000; 38(1):357–361. [PubMed: 10618116]
10. Kleter B, van Doorn LJ, ter Schegget J, Schrauwen L, van Krimpen K, Burger M, ter Harmsel B, Quint W. Novel short-fragment PCR assay for highly sensitive broad-spectrum detection of anogenital human papillomaviruses. Am J Pathol. 1998; 153(6):1731–1739. [PubMed: 9846964]
11. van den Brule AJ, Pol R, Fransen-Daalmeijer N, Schouls LM, Meijer CJ, Snijders PJ. GP5+/6+ PCR followed by reverse line blot analysis enables rapid and high-throughput identification of human papillomavirus geno-types. J Clin Microbiol. 2002; 40(3):779–787. [PubMed: 11880393]
12. Terai M, Burk RD. Complete nucleo tide sequence and analysis of a novel human papillomavirus (HPV 84) genome cloned by an overlapping PCR method. Virology. 2001; 279(1):109–115. [PubMed: 11145894]
13. Smith B, Chen Z, Reimers L, van Doorslaer K, Schiffman M, Desalle R, Herrero R, Yu K, Wacholder S, Wang T, Burk RD. Sequence imputation of HPV16 genomes for genetic association studies. PLoS One. 2011; 6(6):e21375. doi: 10.1371/journal.pone.0021375. [PubMed: 21731721]
14. Narechania A, Chen Z, Desalle R, Burk RD. Phylogenetic incongruence among oncogenic genital alpha human papillomavi ruses. J Virol. 2005; 79(24):15503–15510. [PubMed: 16306621]

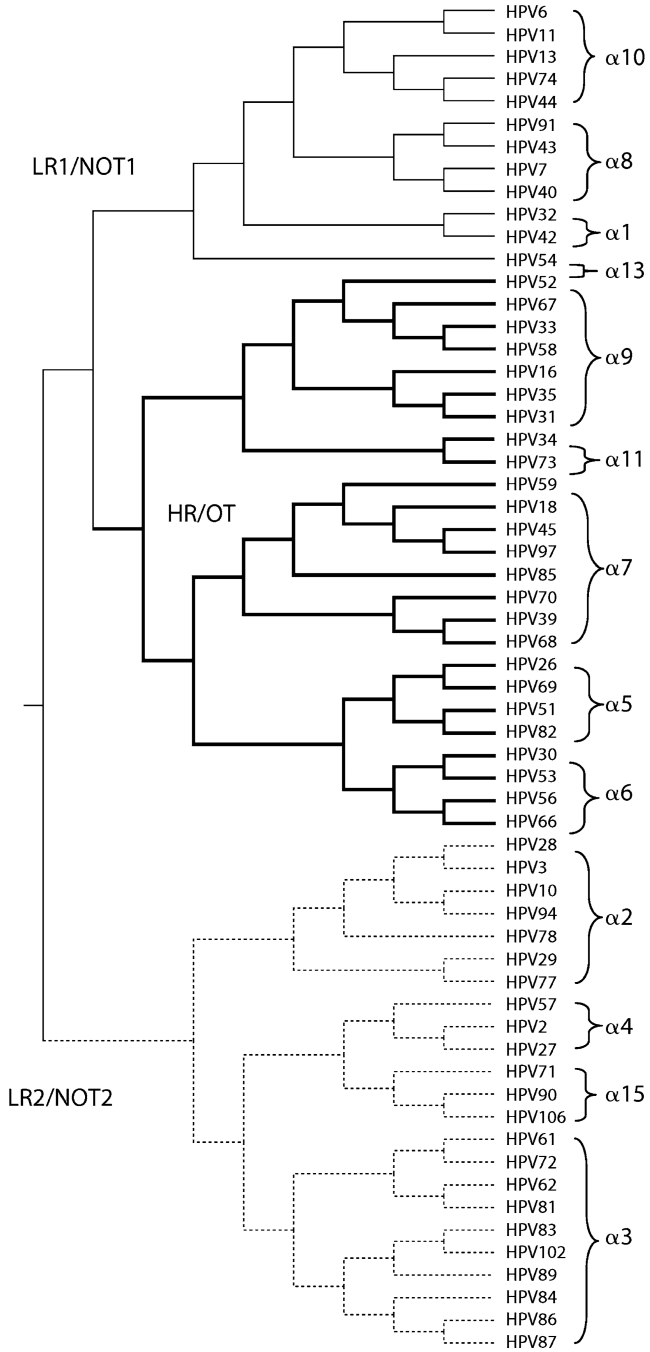

Fig. 1.

Phylogenetic tree of the mucosal/genital human *Alphapapillomaviruses*. The tree shown is from a Bayesian analysis inferred from alignment of protein and nucleotide sequences of six concatenated ORFs (E6, E7, E1, E2, L2, and L1). Bovine PV type 1 was used as the outgroup taxa. The *numbers* to the *right* represent the species group (e.g. “alpha-9” contains HPVs 16, 31, 35, 58, 33, 67, and 52). At least three ancestral papillomaviruses are responsible for the current heterogeneous groups of genital HPV genomes including LR1/NOT1 (α10, α8, α1, and α13), LR2/NOT2 (α2, α3, α4, and α15) and HR/OT (α5, α6, α7,

α9, and α11), the later joined by bold lines represents the clade that contains all known HPV types associated with cervix cancer. *HR* high-risk; *OT* oncogenic type; *LR* low-risk; *NOT* non-oncogenic type. (Cited from Public Health Genomics 2009;12:281–290 with permission)

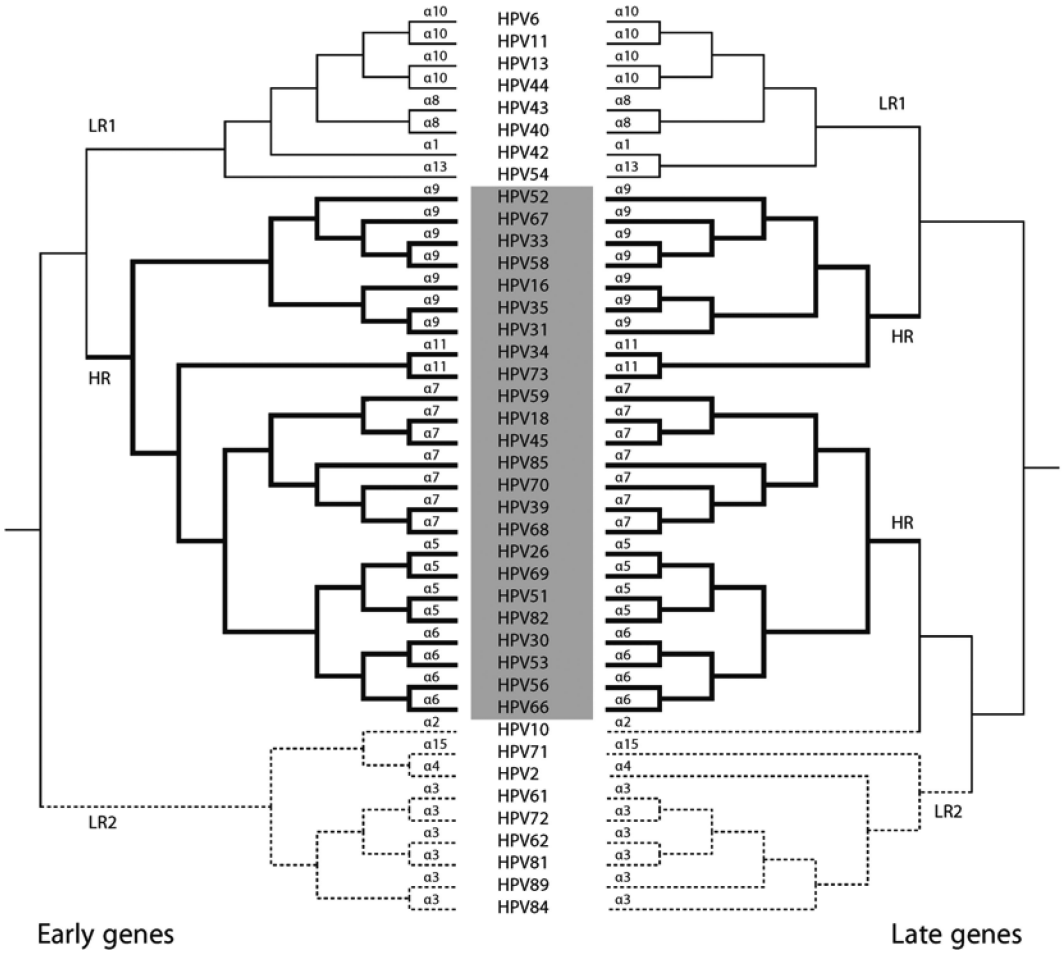

Fig. 2.

Trees from early and late genes show phylogenetic incongruence. Phylogenetic trees were inferred using Bayesian methods [14]. The early gene tree (*left*) was calculated from E1, E2, E6 and E7 concatenated nucleotide alignments, while the late gene tree (*right*) was derived from combined L1 and L2 nucleotide sequence data. The human *Alphapapillomavirus* group designations are shown on their respective leaf branches adjacent to the name of the HPV type as shown in the *center*. All types within the HR/OT clade are shown and representative viruses were chosen from each of the other alpha-HPV species groups. (Cited from Public Health Genomics 2009;12:281–290 with permission)

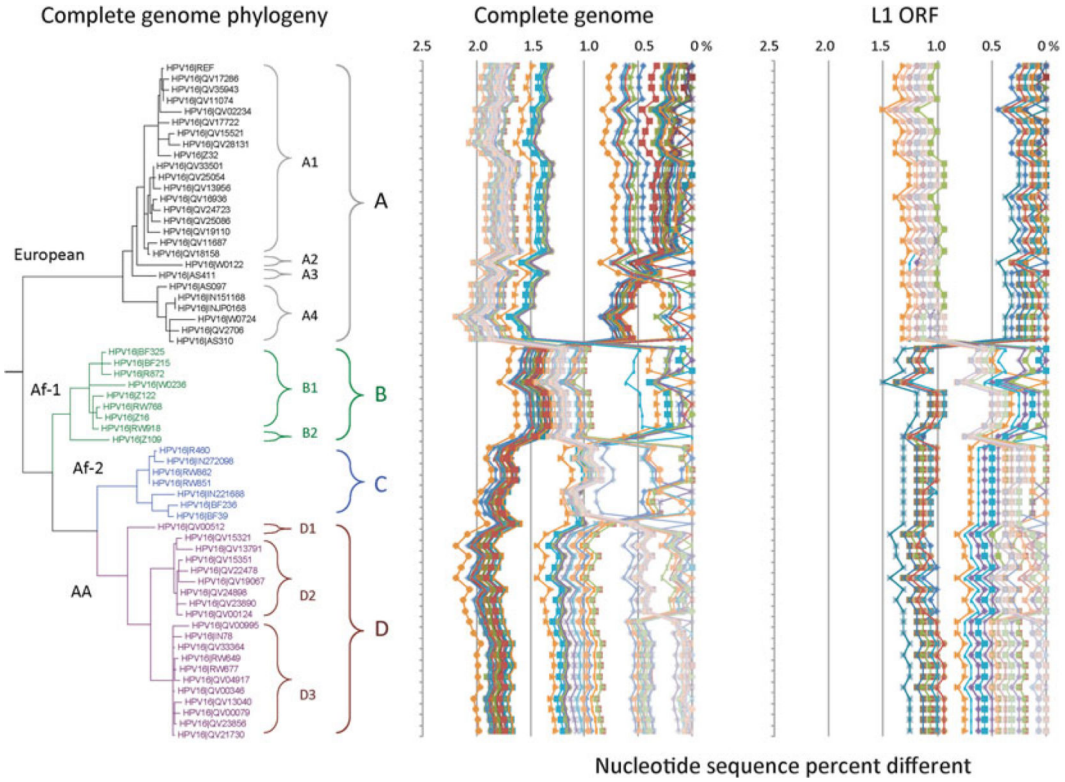

Fig. 3.

HPV16 variant tree topologies and pairwise comparisons of individual complete genomes. Maximum likelihood tree using RAxML was inferred from global alignment of complete genome nucleotide sequences. Distinct variant lineages (i.e., termed A, B, and C) are classifi ed according to the topology and nucleotide sequence differences from 1 to 10 %. The percent nucleotide sequence differences were calculated for each isolate compared to all other isolates of the same type based on the complete genome nucleotide sequences and are shown in the *panel to the right* of each phylogeny. Values for each comparison (1 × 1) of a given isolate are connected by lines and the comparison to self is indicated by the 0 % difference point. *Symbols* and *lines* used are different for each distinct variant lineage to facilitate visual comparisons. *Af-1* African-1, *Af-2* African-2, *AA* Asian-American (This figure is taken from Burk, R.D., Harari, A. and Chen, Z. Human papilloma-virus genome variants. *Virology* 445:232–243, 2013)

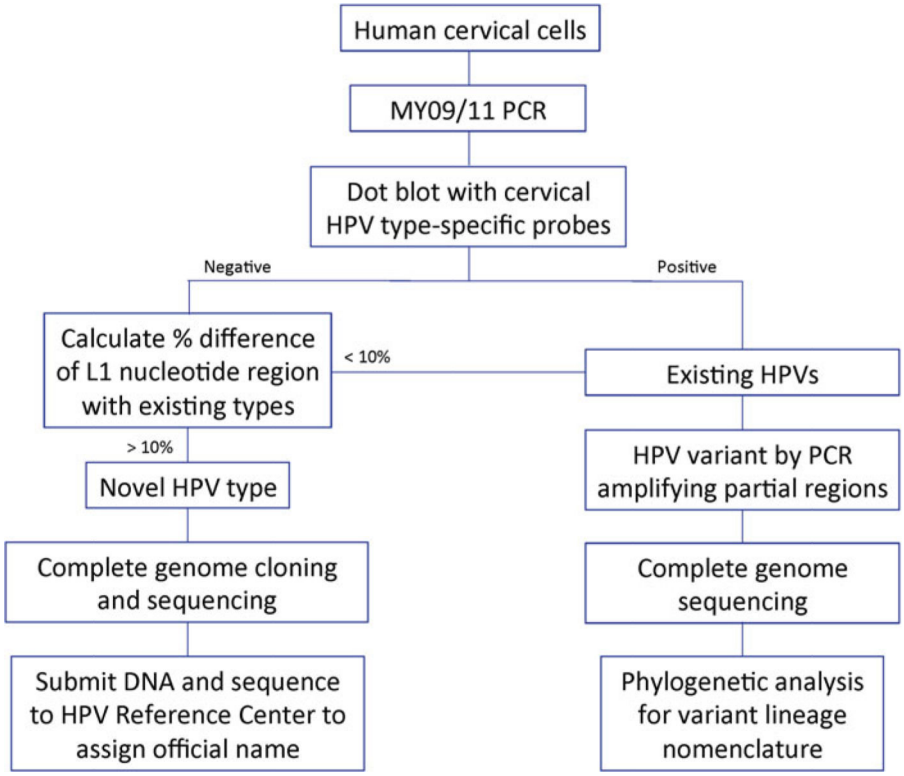

Fig. 4.
The workfl ow of cervical HPV type and variant nomenclature

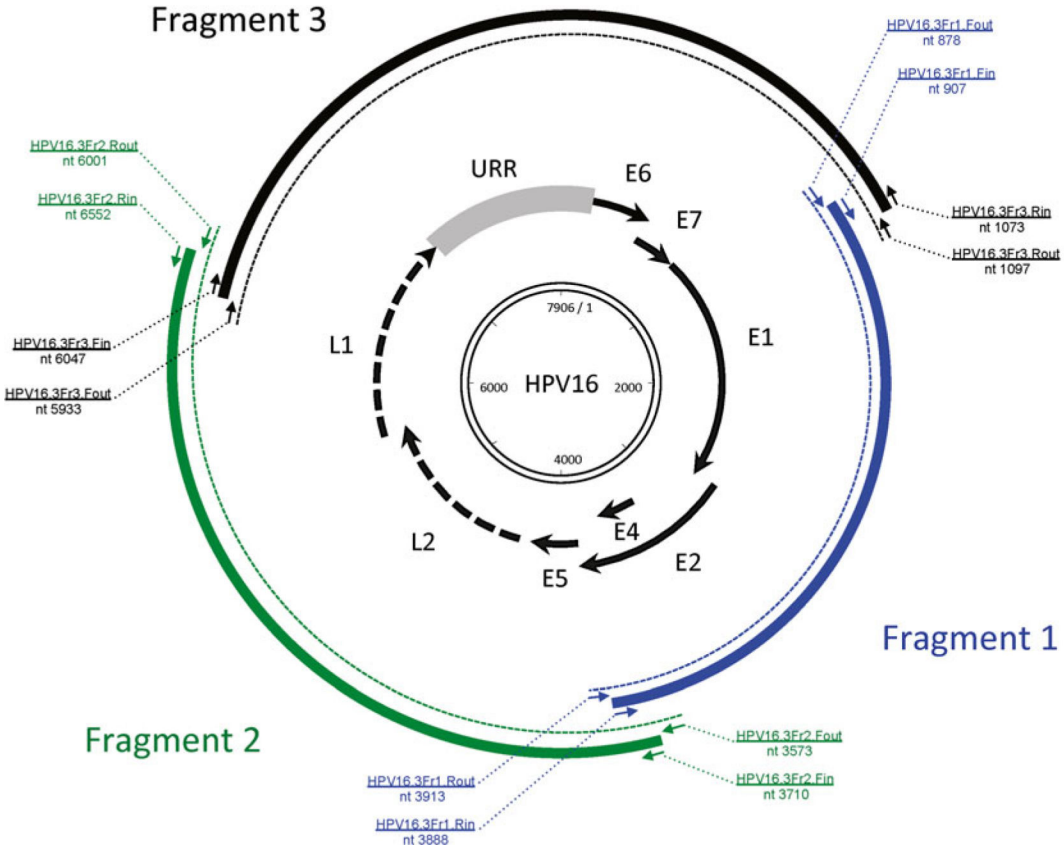

Fig. 5.

Scheme of overlapping PCR to amplify and obtain the complete genome sequence of HPV16

Table 1

One-tube nested PCR primers to amplify a partial region (URR or E6) of HPV16

Primer name	Region	Start position	Direction	Nested PCR	Length	Tm	Sequence (5′-3′)
16URR.Fout	URR	7589	Forward	Outer PCR	20	59.7	GCCAACCATTCCATTGTTTT
16URR.Rout	URR	43	Reverse	Outer PCR	20	58.7	GATTTCGGTTACRCCCTTAG
16URR.Fin	URR	7621	Forward	Inner PCR	18	45.6	ATGTGCAACTACTGAATC
16URR.Rin	URR	7887	Reverse	Inner PCR	18	44.3	TGCTTGTAAATKTGTAAC
16E6.Fout	E6	208	Forward	Outer PCR	20	59.0	ACAGTTACTGCGACGTGAGG
16E6.Rout	E6	438	Reverse	Outer PCR	20	59.7	GGACACAGTGGCTTTTGACA
16E6.Fin	E6	226	Forward	Inner PCR	18	44.6	GGTATATGACTTTGCTTT
16E6.Rin	E6	383	Reverse	Inner PCR	18	44.8	TGTTGTATTGCTGTTCTA

Table 2

Nested PCR primers to amplify HPV16 complete genomes in three overlapping fragments

Primer name	Region	Start position	Direction	Nested PCR	Length	Tm	Sequence (5′-3′)
16.3Fr1.Fout	Fragment 1	878	Forward	Outer PCR	20	59.4	CAGGTACCAATGGGGAAGAG
16.3Fr1.Rout	Fragment 1	3913	Reverse	Outer PCR	20	60.0	GCACACAAAAGCAAAGCAAA
16.3Fr1.Fin	Fragment 1	901	Forward	Inner PCR	20	59.0	ACGGGATGTAATGGATGGTT
16.3Fr1.Rin	Fragment 1	3888	Reverse	Inner PCR	20	60.0	CGCCAGTAATGTTGTGGATG
16.3Fr2.Fout	Fragment 2	3573	Forward	Outer PCR	20	60.1	GCTCACACAAAGGACGGATT
16.3Fr2.Rout	Fragment 2	6601	Reverse	Outer PCR	20	60.0	ATTATTGTGGCCCTGTGCTC
16.3Fr2.Fin	Fragment 2	3710	Forward	Inner PCR	20	59.9	TGGCATTGGACAGGACATAA
16.3Fr2.Rin	Fragment 2	6552	Reverse	Inner PCR	20	60.3	TGGGCATCAGAGGTAACCAT
16.3Fr3.Fout	Fragment 3	5933	Forward	Outer PCR	20	59.9	GTTTGGGCCTGTGTAGGTGT
16.3Fr3.Rout	Fragment 3	1097	Reverse	Outer PCR	20	59.5	TSTTTTGCTTCCTGTGCAGT
16.3Fr3.Fin	Fragment 3	6047	Forward	Inner PCR	20	59.6	GCAAATGCAGGTGTGGATAA
16.3Fr3.Rin	Fragment 3	1073	Reverse	Inner PCR	20	59.5	AACGCATGTGCTGTCTCTGT

Table 3

Sequencing HPV16 complete genomes by overlapping PCR and primer walking

Primer name	Start position	Direction	Sequence (5′-3′)
Fragment 1			
HPV16.3Fr1.R1	1570	Reverse	CAATCGCAACACGTTGATTT
HPV16.3Fr1.F1	1388	Forward	GTGGGGGAGAGGGTGTTAGT
HPV16.3Fr1.F2	1785	Forward	AGAGCCTCCAAAATTGCGTA
HPV16.3Fr1.F3	2291	Forward	ATGGTGCAGCTAACACAGGT
HPV16.3Fr1.F4	2703	Forward	CTCAAGGACGTGGTCCAGAT
HPV16.3Fr1.F5	3261	Forward	TGCAGTTTAAAGATGATGCAGA
Fragment 2			
HPV16.3Fr2.R1	4537	Reverse	AAGGGCCCACAGGATCTACT
HPV16.3Fr2.F1	3710	Forward	TGGCATTGGACAGGACATAA
HPV16.3Fr2.F2	4298	Forward	CATGCAAACAGGCAGGTACA
HPV16.3Fr2.F3	4608	Forward	TTCCATTCCCCCAGATGTAT
HPV16.3Fr2.F4	5096	Forward	TTGCTTTACATAGGCCAGCA
HPV16.3Fr2.F5	5437	Forward	CCTTTTGGTGGTGCATACAA
HPV16.3Fr2.F6	5933	Forward	GTTTGGGCCTGTGTAGGTGT
Fragment 3			
HPV16.3Fr3.R1	6622	Reverse	TTGGTTACCCCAACAAATGC
HPV16.3Fr3.F1	6425	Forward	AGGGCTGGTACTGTTGGTGA
HPV16.3Fr3.F2	6833	Forward	TTGGAGGACTGGAATTTTGG
HPV16.3Fr3.F3	7200	Forward	TTTGTATGTGCTTGTATGTGCTTG
HPV16.3Fr3.F4	7768	Forward	GGCCAACTAAATGTCACCCTA
HPV16.3Fr3.F5	422	Forward	CAAAAGCCACTGTGTCCTGA
HPV16.3Fr3.F6	825	Forward	AATTGTGTGCCCCATCTGTT