Chapter 2

Human papillomavirus type 16 genetic variants: phylogeny and classification based on E6 and LCR

- Iris Cornet
- Tarik Gheit
- Silvia Franceschi
- Jerome Vignat
- Robert D. Burk
- Bakary S. Sylla
- Massimo Tommasino
- Gary M. Clifford
- IARC HPV Variant Study Group

Journal of Virology, 2012 Jun;86(12):6855-6861.

Abstract

Naturally occurring genetic variants of HPV16 are common and have previously been classified into 4 major lineages; European-Asian (EAS), including the sub-lineages European (EUR), and Asian (As), African 1 (AFR1), African 2 (AFR2) and North-American/Asian-American (NA/AA). We aimed to improve the classification of HPV16 variant lineages by using a large resource of HPV16-positive cervical samples collected from geographically diverse populations in studies on HPV and/or cervical cancer undertaken by the International Agency for Research on Cancer. In total, we sequenced the entire E6 genes and long control regions (LCRs) of 953 HPV16 isolates from 27 different countries worldwide. Phylogenetic analyses confirmed previously described variant lineages and subclassifications. We characterized two new sublineages within each of the lineages AFR1 and AFR2 that are robustly classified using E6 and/or the LCR. We could differentiate previously identified AA1, AA2 and NA sublineages, although they could not be distinguished by E6 alone, requiring the LCR for correct phylogenetic classification. We thus provide a classification system for HPV16 genomes based on 13 and 32 phylogenetically distinguishing positions in E6 and the LCR, respectively, that distinguish nine HPV16 variant sublineages (EUR, As, AFR1a, AFR1b, AFR2a, AFR2b, NA, AA1, and AA2). Ninety-seven percent of all 953 samples fitted this classification perfectly. Other positions were frequently polymorphic within one or more lineages but did not define phylogenetic subgroups. Such a standardised classification of HPV16 variants is important for future epidemiological and biological studies of the carcinogenic potential of HPV16 variant lineages.

Introduction

Human papillomaviruses (HPV) are circular double-stranded DNA viruses that are highly prevalent in the general population. HPVs can be divided into two groups, mucosal and cutaneous, according to their tissue tropism. Several mucosal HPV types have been identified as the etiological agents for cervical cancer and are termed high-risk (HR) HPV types [17]. HPV type 16 (HPV16) is the most prevalent HR HPV type worldwide and is found in the majority of cervical cancer cases [7,21].

The HPV16 genome is about 7,900 bp long and consists of 8 protein-coding genes (L1, L2, E1, E2, E4, E5, E6 and E7) and 2 noncoding regions (the noncoding region [NCR] and the long control region [LCR]) [20]. E6 and E7 are the major oncoproteins, which are involved in tumorigenesis and are highly expressed in tumors. The LCR, adjacent to E6 downstream, contains the early promoter and regulatory elements involved in viral DNA replication and transcription. The NCR is a short region localized between E5 and L2.

The reference HPV16 genome was first sequenced by Seedorf *et al.* in 1985 [19] and was revised by Myers *et al.* in 1995 [18]. Many naturally occurring variants have since been found. The first worldwide study of HPV16 variants was done in 1993 by Ho *et al*. [11], who reported that HPV16 LCR variants segregate robustly into a phylogenetic tree with five major variant lineages: European (EUR), Asian (As), Asian-American (AA), and two African lineages, African-1 and -2 (AFR1 and AFR2) [11,12]. Lineage names derive from the geographical origin of the populations in which they are most prevalent [11,12]. Yamada *et al*. subsequently sequenced several genes and described one additional lineage, North American (NA) [23,26,27].

HPVs mutate very slowly because they are double-stranded DNA viruses that use the excellent DNA polymerase proofreading ability of their host. Nevertheless, nucleotide polymorphisms can occur through random mutation and can become established in a population. This genetic drift has been observed among HPV16 variants, suggesting their coevolution with humankind[4,11].

An HPV variant is a genome defined by a unique combination of single nucleotide polymorphisms (SNPs). A recent proposal would define major variant lineages by an approximately 1.0% difference between full genomes of the same HPV type, with differences of 0.5-0.9% defining sublineages [3,5,6].

Several studies, based principally on the sequencing of E6 and/or LCR in studies from Europe and the Americas, have suggested that HPV16 variants can influence viral persistence and the development of cervical cancer[9,24,25,28,30]. However, for future large-scale epidemiological studies, it is important to have a complete and standardized classification of HPV16 variant sublineages worldwide. Hence, we have sequenced the entire E6 genes and LCRs for 953 patients from 27 different countries, including a major focus on HPV16 isolates from Africa and Asia, in order to examine the practical classification of HPV16 variant sublineages by using a frequently analyzed region.

Materials and Methods

Origin of clinical specimens

The International Agency for Research on Cancer (IARC) has coordinated cervical cancer case series, cervical cancer case-control studies, and population-based HPV prevalence surveys in a large number of countries around the world [2,8,16]. Cervical samples (exfoliated cells or tissue biopsy specimens) derived from these studies have been comprehensively genotyped for 37 HPV types by using a standardised and well-validated protocol (General Primer GP5+/6+ PCR-based enzyme-based immunosorbent assay [ELISA]) [14] in one centralized laboratory (P.J.F. Snijders, Department of Molecular Pathology, Vrije University, Amsterdam, The Netherlands). HPV16-positive cervical samples in the IARC biobank were selected for the current analysis and were categorized into the following regions: North Africa, sub-Saharan Africa, the Americas, Eastern Asia, Western Asia, and Europe. Country-specific details are noted in Table 1.

Table 1. Distribution of 953 E6/LCR sequences by country and region.

Region	Country	No. of sequences
Africa, North	Algeria	103
	Morocco	118
Africa, Sub-saharan	Benin	3
	Guinea	23
	Kenya	46
	Mali	38
	Nigeria	23
	South Africa	27
	Tanzania	13
	Uganda	18
Asia, Eastern	Philippines	1
	Thailand	173
	Korea	1
Asia, Western	India	137
Europe	Georgia	2
	Poland	105
	Spain	6
The Americas	Argentina	65
	Brazil	9
	Bolivia	2
	Canada	1
	Chile	5
	Colombia	4
	Cuba	5
	Panama	4
	Paraguay	9
	Peru	12

The samples included represent the full spectrum of HPV16 infection from normal cytology to cervical cancer. However, the comparison of samples by cyto/histopathological diagnosis is not relevant to phylogenetic analysis and is not the focus of the current paper.

PCR and DNA sequencing

DNA extraction was performed using the QIAmp DNA minikit (Qiagen), according to the manufacturer's protocols: for cells in phosphate-buffered saline (PBS), the protocol for "DNA purification from blood of body fluids" was used, and for frozen biopsy specimens the protocol for "DNA purification from tissues" was used. Several precautionary measures were taken to minimize the risk of contamination.

HPV16 E6 open reading frame (ORF) was amplified by polymerase chain reaction (PCR) as described by Gheit *et al.* in 2011 [9]. The primers used, flanking the coding region of HPV16 E6 (nucleotides [nt] 52 to 575) were 5'-CGAAACCGGTTAGTATAA-3' and 5'-GTATCTCCATGCATGATT-3' [29]. The HPV16 LCR region was amplified with primers LCR_Fw (5'-ACCTCCAGCACCTAAAGAAG-3') and LCR_Rv (5'-GTCCAGAAACAT TGCAGTTCT-3'), spanning the region between nt 6934 and 115. Forty amplification cycles were run in the GeneAmp PCR System 2400 with a 94°C denaturation step (1 min), an annealing step at 50°C for E6 or 55°C for the LCR (1 min) and a 72°C extension step (1 min), including an initial denaturation step of 15 min and a final extension step of 10 min, resulting in a 524-bp E6 PCR product and a 1087-bp LCR PCR product. The PCR mixture contained 1X PCR buffer, 200 µM of each deoxynucleoside triphosphate (dNTP), 0.2 µM of each primer, and 0.625 U HotStarTaq DNA polymerase in a final volume of 25 µl (Qiagen, Hilden, Germany). Portions (5 µl) of the PCR products were checked by GelRed (FluoProbes) agarose gel electrophoresis. Samples that were negative for the PCR amplification of E6 or LCR were not used for further analysis.

After enzymatic purification with 0.4 µl of exonuclease I (10 U/µl; New England BioLabs) and 0.2 µl of dhrimp alkaline phosphatase (1 U/µl; USB) in 10 µl of PCR product at 37°C for 15 min, and an inactivation step at 80°C for 15 min, the HPV16 PCR products were sequenced by the fluorescent dye dideoxy termination method using an ABI Prism 377 DNA sequencer (Perkin-Elmer Applied Biosystems) according to the manufacturer's protocol. For the

sequencing reaction, the same primers were used as for the PCR. For the LCR, an additional sequencing reaction was carried out using internal primers 5'-GAATCACTATGTACATTGTGTC-3' and 5'-GCTTGTGTAACTATTGTGTCA-3', to make an internal control with overlapping sequences.

The sequences were analyzed using the Blast function from PubMed, and SNPs were visually interpreted in comparison to the HPV16R prototype sequence described by Myers *et al.* in 1995 [18].

Data quality assurance

In this study, we have identified approximately 200 novel SNPs, which were not reported in the literature previously. Independent PCR and sequencing reactions were performed to distinguish between mutations introduced by PCR amplification and natural HPV16 variants. Other steps were also put in place to reduce errors in the final E6/LCR variant database. Firstly, all E6 data were validated by comparing the SNPs reported by visual inspection in BLAST with the output files from the DNA sequencer, using an *ad hoc*-designed computer program. Secondly, following the classification of all HPV16 isolates into sublineages (see below), all isolates that were not classified perfectly into one of the lineages, as well as any E6 and/or LCR sequences that were found in only one sample, were carefully reevaluated. Last, an algorithm was designed for rechecking any SNP that was present in fewer than 5 samples within one sublineage, if it also appeared in at least one other sublineage. Each of these steps resulted in the resolution of a number of errors in the database and potential false links in the phylogenetic tree.

Phylogenetic tree construction and HPV16 variant classification

Unrooted consensus-UPGMA (unweighted-pair group method using average linkages) trees, with 100 bootstrapped replicates, were built using Phylip software, version 3.69. For the principal tree analysis, we included only E6/LCR sequences found in at least two samples (n = 99). In a subsequent analysis, we included all 353 unique E6/LCR sequences from this study, supplemented by

other entire E6/LCR sequences reported in the literature, including those from the studies of Kammer *et al.* 2002 (Finland; n = 9), Bhattacharjee *et al.* 2008 (India; n = 18) (S. Sengupta, personal communication) and Smith *et al.* 2011 (Costa Rica; n = 46) [1,6,15,20].

Results

The E6 gene and LCR were sequenced in a total of 985 HPV16-positive cervical samples. Seven samples coinfected with more than one HPV16 variant and 25 samples with deletions in the LCR were excluded, leaving 953 samples in the subsequent analysis. The samples included were from North Africa (n = 221), sub-Saharan Africa (n = 191), the Americas (n = 116), Western Asia (n = 137), Eastern Asia (n = 175), and Europe (n = 113) (Table 1).

In total, we identified 49 variable nucleotides in E6 (from nt 104 to nt 559), which occurred in 68 unique combinations, and 169 variable nucleotides in the LCR (from nt 7157 to nt 83), occurring in 288 unique combinations. For E6 and LCR combined, there were 353 unique variants. These included 99 variants found at least twice (constituting 73% of all samples), which were included in the principal phylogenetic tree analysis (Figure 1). The phylogenetic tree segregated into four major branches (i.e., lineages) that could be recognised by the nomenclature of previous studies: (i) EAS, including the European (EUR) and Asian (As) sublineages, (ii) African 1 (AFR1), including two sublineages that we tentatively defined as AFR1a and AFR1b, (iii) African 2 (AFR2), including two sublineages that we tentatively defined as AFR2a and AFR2b, and (iv) AA/NA, including the North-American (NA), Asian-American-1 (AA1) and Asian-American-2 (AA2) sublineages.

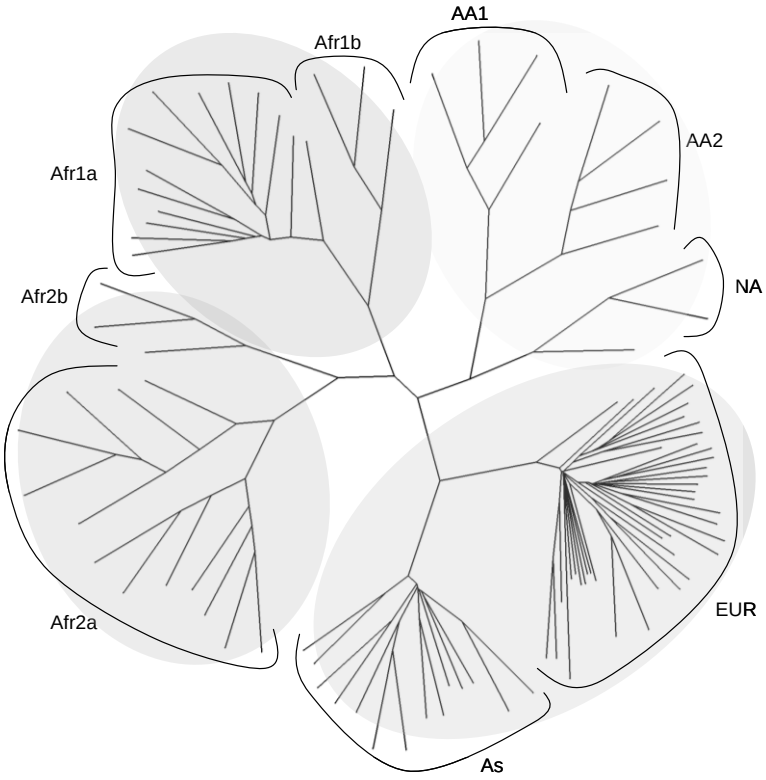

Figure 1. HPV 16 phylogenetic tree based on 99 unique E6/LCR sequences found in at least 2 samples. This is a bootstrapped (100 replicates) consensus UPGMA tree.

A second, larger, phylogenetic tree was generated including all 353 unique E6/LCR variants from the present study, as well as additional novel published E6/LCR sequences retrieved from the work of Kammer *et al.* (2002) (n = 9), Bhattacharjee *et al.* (2008) (n = 18) and Smith *et al.* (2011) (n = 46) [1,15,20] (Figure 2). This larger tree had a similar structure, with the same nine sub-lineages (EUR, As, AFR1a, AFR1b, AFR2a, AFR2b, NA, AA1, AA2), even if the distinction between AA1 and AA2 became less clear, demonstrating the robustness of the tree.

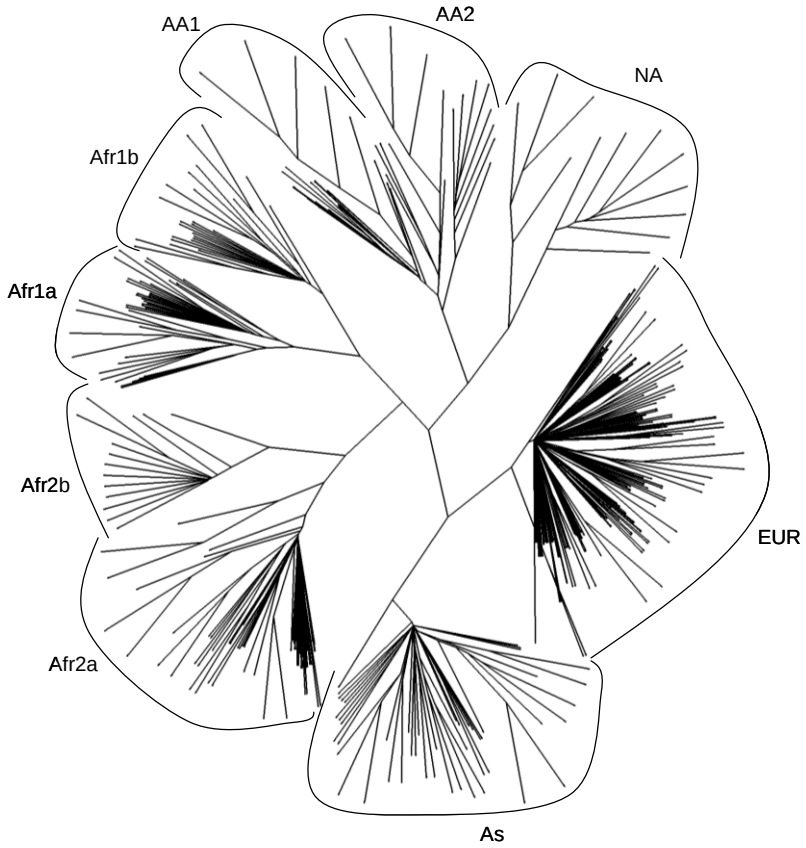

Figure 2. HPV16 phylogenetic tree based on 353 unique E6/LCR sequences from the present study plus additional E6/LCR sequences from the work of Smith *et al.*, Bhattacharjee *et al.* and Kammer *et al.* [1,15,20]. This is a bootstrapped (100 replicates) consensus UPGMA tree.

Sequences clustering in each of these nine sublineages showed a specific "core" pattern of SNPs. Table 2 shows a classification based upon SNPs (13 and 32 in E6 and LCR, respectively) that distinguished at least two of the nine sublineages from each other. Certain SNPs were "diagnostic" for a given sublineage (see below), meaning that they were unique to a specific sublineage. No such diagnostic SNPs were available for NA or AFR2b (Table 2). Nomenclature equivalent to our sublineage designations, based on full-genome sequencing (R. D. Burk, personal communication), is also given.

Of the total of 953 samples, 96.5% perfectly fitted the classification of Table 2 (Table 3). Of the 33 samples that did not fit perfectly, 30 differed by only one SNP from one of the variant sublineages (all fell clearly into one of the major branches of the phylogenetic tree in Figure 2). Variant sublineage classifications based on E6 correlated highly with those based on yhe LCR (Table 3). The only exceptions were two isolates that were classified as As in the LCR but were missing T178G/C, leading to the misclassification of these two isolates as EUR based on E6, and one isolate that was classified as AFR2b in the LCR but had G132C, leading to misclassification as AFR1am based on E6.

European-Asian lineage

The major European-Asian branch can be specifically diagnosed by three nucleotide positions in E6 (145G, 286T and 289A) and three in the LCR (7489G, 7764C and 7786C) (Table 2).

The EUR sublineage cannot be specifically diagnosed by any nucleotide position. Several nucleotide positions that phylogenetically distinguish other HPV16 sublineages, however, can be polymorphic within the European sublineage (positions 109, 131, 178, 335, and 350 in E6; 7233, 7507, 7730, and 24 in the LCR), of which the most frequently observed SNP is T350G (54.5 %), giving rise to the amino acid change L83V. These SNPs can be found alone or in combination with each other. However, they do not appear to define any phylogenetic branches within the EUR sublineage. In the LCR, T7193G and G7521A are present in 77.6% and 80.5% of EUR isolates (compared to 100% of isolates from all other lineages). The only other common (>10%) SNP in EUR isolates was T7450C in the LCR (25.8 %).

The As sublineage can be specifically diagnosed by two SNPs on nucleotide 178 in E6: T178G, which gives rise to the amino acid change D25E, and T178C, a silent SNP. The As sublineage shows a specific combination of 8 SNPs in the LCR, of which six are diagnostic (T7177C, T7201C, C7270T, A7287C, G7842A/T, and C24T).

Table 2. Classification of HPV16 variant sublineages based on distinguishing positions in E6 and the LCR[a]

			E6 nucleotide position													LCR nucleotide position					
Lineage[b]	sub-lineage	Nomenclature[c]	109	131	132	143	145	178	286	289	295	335	350	403	532	7175	7177	7201	7232	7233	7270
HPV16 reference			T	A	G	C	G	T	T	A	T	C	T	A	A	A	T	T	A	A	C
EAS	EUR	A1/2	-/C	-/G	-	-	-	-/A	-	-	-	-/T	-/G	-	-	-	-	-	-	-/C	-
	As	A3	-	-	-	-	-	**G/C**	-	-	-	-	-	-	-	C	**C**	**C**	-	-	**T**
AFR1	AF1a	B1	-	-	**C**	G	T	-	A	G	-	T	-	-	-	-	-	-	-/G	-	-
	AF1b	B2	-	G	-	G	T	-	A	G	**G**	T	G	-	-	-/C	-	-	C	-	-
AFR2	AFR2a	C1	C	-	**T**	G	T	-	A	G	-	T	-	**G**	-	-	-	-	-	-/C	-
	AFR2b	C2	-	-	-	G	T	-	A	G	-	T	-	-	-	-	-	-	-	C	-
NA/AA	NA	D1	-	-	-	-	T	-	A	G	-	T	G	-	-	-	-	-	-	C	-
	AA1	D2	-	-	-	-	T	-	A	G	-	T	G	-	G	-	-	-	-	C	-
	AA2	D3	-	-	-	-	T	-	A	G	-	T	G	-	-/G	-	-	-	-	C	-

(Continued on next page)

Table 2. (Continued)

LCR nucleotide position																												
7287	7339	7348	7394	7395	7435	7485	7489	7507	7669	7689	7729	7730	7743	7764	7786	7826	7834	7837	7839	7842	7876	7886	24	25	31	83
A	A	A	C	C	G	A	G	A	C	C	A	A	T	C	C	G	G	A	A	G	C	C	C	T	C	A
-	-	-	-	-	-	-	-	-/C	-	-	-	-/C	-	-	-	-	-	-	-	-	-	-	-/G	-	-	-
C	-	-	-	-	-	-	-	-	-	-	-	C	-	-	-	-	-	-	-	**A/T**	-	-	**T**	-	-	-
-	-		-	-	-	-	A	-	-	A	-	-	-	T	T	-	T	-	-	-	-/A	-	-	-	T	**C**
-	-	**C**	-	-	-	-	A	-	-	A	-	-	-	T	T	-	T	-	-	-	A	-	-	**C**	T	-
-	-	-	-	-	**A**	C	A	-	T	-/A	-	-	-	T	T	A	-/T	C	**G**	-	-	-	-	-	T	-
-	-	-	-	-	-	C	A	-	T	A	-	-	-	T	T	A	T	C	-	-	-	-	-	-	T	-
-	-	-	-	-	-	C	A	-	T	A	- / C	-	-	T	T	-	T	-	-	-	-	-	-	-	-	-
-	T	-	T	T	-	C	A	-	T	A	C	-	**G**	T	T	-	-	-	-	-	-	G	-	-	-	-
-	T	-	T	-/T	-	C	A	**G**	T	A	C	-	-	T	T	-	-	-	-	-	-	G	-	-	-	-

[a] Includes only nucleotide positions that can distinguish at least two of the nine identified sublineages from each other.
Dashes indicate no nucleotide exchange from the HPV16 reference sequence. Nucleotides separated by a slash are different nucleotides that can be in the given position for the given sublineage. Diagnostic SNPs are noted boldfaced in a square.
[b] EAS, European-Asian; EUR, European; As, Asian; AFR, African; NA, North-American; AA, Asian-American.
[c] Nomenclature equivalent to our sublineage designations, based on full-genome sequencing (R. D. Burk, personal communication).

Table 3. Distribution of 953 E6/LCR sequences based on the classification of E6 and the LCR given in Table 2.

E6 classification	No. of sequences with the following LCR classification										
	EUR	As	AFR1a	AFR1b	AFR2a	AFR2b	AA1	AA2	NA	Other*	Total
EUR	410	2	-	-	-	-	-	-	-	1	413
EAS	-	129	-	-	-	-	-	-	-	2	131
AFR1a	-	-	77	-	-	1	-	-	-	3	81
AFR1b	-	-	-	30	-	-	-	-	-	2	32
AFR2a	-	-	-	-	146	-	-	-	-	3	149
AFR2b	-	-	-	-	-	26	-	-	-	2	28
NA/AA	-	-	-	-	-	-	60	10	30	8	108
Other[a]	-	1	4	1	-	2	1	1	-	1	11
Total	410	132	81	31	146	29	61	11	30	22	953

[a] "Other" sequences differed by at least one SNP from all sublineage-specific classifications in E6 or the LCR.

Interestingly, the T178A SNP (n = 4), which has previously been misclassified in the As sublineage on the basis of E6 alone, is clearly shown to be specific to the EUR sublineage when the LCR is taken into account.

Other common (>10%) SNPs in As isolates were A7289C (93.1%), T7384G (11.4 %), G7429A (23.7 %), T7781C (16.0 %), and C7874G (22.1 %).

African lineages

The major AFR branch can be specifically diagnosed by the presence of C143G in E6 and C31T in the LCR (Table 2). All AFR isolates show a common pattern of five SNPs in E6, namely C143G, G145T, T286A, A289G and C335T, which give rise to two amino acid changes, Q14D and H78Y.

The two previously described African lineages, AFR1 and AFR2, were confirmed, and could be specifically distinguished from each other based on 4 nucleotide positions in the LCR (7485, 7669, 7826 and 7837). No SNPs in E6 could distinguish AFR1 from AFR2. AFR2 could be specifically diagnosed by the presence of G7826A and A7837C in the LCR. No SNPs could specifically diagnose AFR1.

The AFR1 and AFR2 branches each showed an additional split into two sublineages, which we tentatively named African-1a, African-1b, African-2a and African-2b. Even though these splits were less stable than other lineages in the phylogenetic tree (Figures 1 and 2), the classification based on the distinguishing SNPs in E6 and the LCR correlated very well (Table 3). These four African sublineages could be distinguished from each other by the combination of six SNPs in E6 (at positions 109, 131, 132, 295, 350, and 403) and eleven in the LCR (at positions 7232, 7233, 7435, 7485, 7669, 7826, 7837, 7839, 7876, 25, and 83). Diagnostic SNPs appear to exist in E6 and the LCR for AFR1a (G132C and A83C), AFR1b (T295G, A7438C, and T25C), and AFR2a (G132T, A403G, G7435A, and A7839G) but not for AFR2b.

Furthermore, within the AFR2a branch of the phylogenetic tree, an additional branching could be observed (but is not included in the classification

in table 2). This is due to two SNPs in the LCR, T7282G and A7372C, that are always present in one branch but always absent in the other.

Other common SNPs (>10% of any sub-lineage) in AFR isolates were T7293G (12.3 %), A7611G (17.3 %) and T7714A (43.2 %) in AFR1a; G7868A (37.5 %) in AFR1b; T7282G (45.9%), G7372C (46.6 %), G7387C (45.6 %) and G7868A (35.6 %) in AFR2a; T7282G (10.4 %), A7348G (10.4 %), T7450G (17.2 %), T7643G (24.1 %), A7688T (24.1 %) and A7688G (13.8 %) in AFR2b.

North-American and Asian-American lineages

The NA/AA lineages cluster together in a major branch of the phylogenetic tree (Figures 1 and 2). However, no SNP in E6 or the LCR can specifically recognize the major NA/AA branch (Table 2). All NA/AA isolates show a common pattern of five SNPs in E6, namely G145T, T286A, A289G, C335T and T350G (which give rise to three amino acid changes, Q14H, H78Y, and L83V), and 7 SNPs in the LCR, namely A7233C, A7485C, G7489A, C7669T, C7689A, C7764T, and C7786T.

The NA/AA lineage splits into two branches, one for the two AA sublineages and one for the NA sublineage (Figures 1 and 2). The NA, AA1 and AA2 sublineages can be distinguished based upon the combination of six SNPs in the LCR (at positions 7339, 7394, 7507, 7743, 7834, and 7886). Of these, A7507C and T7743G are diagnostic for the AA2 and AA1 sublineages, respectively. However, the NA, AA1 and AA2 sublineages could not be distinguished based on E6 alone. While A532G was always present in AA1 and always absent in NA, it could be absent or present in AA2.

Of note, in our samples, 100% of AA2 isolates contained A7894C. However, as this appears often not to be the case in other published sample sets [20], we did not include this position in our classification. Other common SNPs (>10% of any sub-lineage) occurred only in NA isolates; T183G (78.6 %) (resulting in an amino acid change I27R), T271C (21.4 %) (a silent SNP), G7359A (70.0 %), G7360A (70.0 %), T7441G (70.0 %) and C7784T (70.0 %).

Geographical distribution of HPV16 sublineages

Figure 3 shows the distribution of the HPV16 sublineages by geographical region. Samples with the EUR lineage were well distributed among the different geographical regions. The As and AFR variant lineages predominated in samples from Eastern Asia and Africa, respectively. However, the AFR1a lineage was found mainly in sub-Saharan Africa and the AFR1b lineage in North Africa (with no differences for the AFR2 lineages). The AA1 and AA2 lineages were both commonly seen in samples from South/Central America, but the AA1 lineage was more likely to be found in samples from Asia. Lastly, the NA lineage was found to be particularly frequent in samples from North Africa.

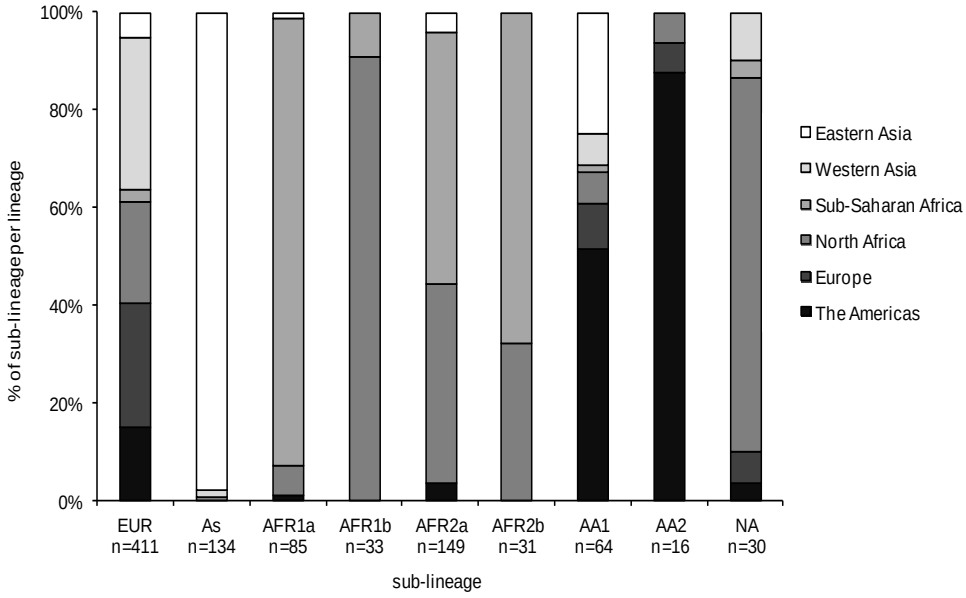

Figure 3. Distribution of HPV16 variant sublineages by region.

Discussion

This updated HPV16 phylogenetic analysis confirmed previously reported variant lineages [4,11,12,23,26,27] and was able to identify additional levels of HPV16 phylogenetic stratification. Based on a total of 953 E6/LCR sequences isolated worldwide, with a high proportion of isolates from Africa and Asia, we were able to produce an updated phylogenetic tree that clearly identified nine sublineages: EUR, As, AA1, AA2, NA, AFR1a, AFR1b, AFR2a, and AFR2b. This tree structure was robust, irrespective of whether all unique sequences or only those found in at least two samples were considered.

Our analysis was particularly informative with respect to elaborating the African branches of the phylogenetic tree, since African HPV16 isolates were better represented than in previous studies. We newly identified two branches within both AFR1 and AFR2, which we tentatively identified as African-1a, African-1b, African-2a and African-2b. Each of these four sublineages showed a specific combination of SNPs in both E6 and the LCR. The characterisation of these sublineages shows that there is no single SNP in E6 that can distinguish AFR1 from AFR2, as had been suggested previously. Of note, the strong representation of AFR1b (and NA) in the present updated analysis was due largely to the inclusion of samples from North Africa (Algeria and Morocco), suggesting that this geographical region represents a previously understudied branch of HPV16 evolution.

This updated analysis also revealed that E6 alone does not allow the distinction between the three closely related sublineages AA1, AA2 and NA, as had been suggested previously [13]. Although the A532G SNP is always present in AA1 and always absent in NA, it can be absent or present in AA2. The inability of E6 to distinguish these sublineages is important for epidemiological studies, since it is particularly the AA1 sublineage that has been suggested to be associated with cervical intraepithelial neoplasia grade 3 or worse (CIN3+) risk based upon full-genome analysis [20].

Previous classifications based on E6 alone have situated isolates containing T178A in the As lineage (due to the common presence of the similar

T178G SNP in the As lineage)[13]. However, the four such isolates present in our analysis were clearly classified as EUR on the basis of the LCR.

The LCR was confirmed to contain much more phylogenetic information than E6 and distinguished all nine proposed sublineages without the requirement for E6. Certain parts of the LCR were denser in phylogenetic information than others. The shortest fragment of the LCR that allowed the distinction of all nine sublineages was the region from nt 7743 to nt 25 (~300bp).

For epidemiological studies based on the detection of SNPs in E6 and/or the LCR, we thus propose a practical classification of variant lineages using 45 nucleotide positions (13 and 32 in E6 and the LCR, respectively) that can each distinguish at least two of the nine sublineages described above from each other. Nevertheless, there is much redundancy in this classification, so that not all 45 positions in E6 and the LCR are required for classification into one of the nine sublineages. Indeed, there are a smaller number of diagnostic SNPs that are specific for a given sublineage. However, no diagnostic SNPs exist in E6/LCR for EUR, NA, or AFR2b. Furthermore, many E6/LCR diagnostic SNPs previously proposed by Smith *et al.* [20] (based on 62 complete HPV16 genomes), proved not to be truly sublineage specific in our wider analysis. In E6, for example, C335T is not diagnostic for non-EAS lineages, and T109C, G132T, and A403G are not diagnostic for AFR2 [20]. This indicates that with an expansion of samples from around the world, such as those presented in this study, the number of unique SNPs diagnostic for specific lineages or sublineages will decrease.

The robustness of the classification was confirmed by the almost-perfect correlation of the pattern of SNPs in E6 and the LCR, confirming previous findings that patterns of SNPs correlate throughout the whole HPV16 genome [6,20]. Given this strong correlation, epidemiological analyses of the natural history and carcinogenetic potential of HPV16 genetic variants will not be able to distinguish causal lineage-specific SNPs from noncausal lineage-

specific SNPs [20] and hence should focus on establishing robust relative risks at the level of HPV16 sublineages.

We also identified a large number of non-lineage-specific SNPs, mostly in the LCR. Only a few non-lineage-specific SNPs occurred in E6 in >10% of the samples in any one sublineage, namely, the well-characterised T350G SNP in the EUR lineage (see below), as well as T183G, T271C, and A532G in the AA2 lineage. However, these SNPs did not show evidence of defining phylogenetic subgroups. Thus, epidemiological studies should compare these non-lineage-specific SNPs within a given lineage only, following the example of studies that have suggested that the EUR-350G and EUR-350T sublineages differ in their risk for viral persistence [9], and/or cervical cancer [9,10,22,28]. However, this approach is unlikely to be statistically feasible for rarer SNPs, and care should also be taken not to overinterpret their importance [13].

Our newly utilized E6/LCR classification is consistent with previous lineage definitions based on complete HPV16 genome sequencing, whilst highlighting some finer stratification [6,20]. Indeed, we have included the translation to the most recent equivalent nomenclature based on full-genome sequencing (R. Burk, personal communication) in Table 2. However, whereas recent whole-genome analyses included very few isolates representing the NA, AFR1b and AFR2b sublineages, our analysis (whilst based on E6 and the LCR only), included approximately 30 isolates of each of these sublineages [20]. Thus, a selection of most informative E6/LCR-characterised isolates would warrant sequencing across their whole genomes in order to strengthen the full picture of HPV16 genetic evolution.

In summary, we made use of the wide geographical representation of the IARC cervical sample biobank to provide the most complete and practical classification for HPV16 variant sublineages to date. This work can help the standardisation of future epidemiological studies of the natural history and carcinogenicity of HPV16 genetic variants, as well enabling the pooling of data from different studies to overcome issues of sample size limitations from individual studies.

Acknowledgements

This work was supported by grants from The Association for International Cancer Research, UK (project grant number 08–0213), the Institut National du Cancer, France (collaboration agreement 07/3D1514/PL-89–05/NG-LC), the Fondation Innovations en Infectiologie (FINOVI) (Project No AO1-project 2), and the European Commission, grant HPV-AHEAD (FP7-HEALTH-2011-282562).

Thanks to Sharmila Sengupta for sharing raw data from the 2008 paper of Bhattacharjee *et al.*[1]. Thanks also go to Sophie Pallardy, Annie Arslan, Vanessa Tenet, Sophie Guillot, Annick Rivoire and Veronique Chabanis for technical and/or administrative support to the project and to the very many IARC and local collaborators that contributed to the epidemiological studies from which the samples for this study were sourced.

In addition to the authors of this article, the members of the IARC HPV Variant Study Group include previous IARC staff (N. Muñoz, R. Herrero, X. Bosch) and local study coordinators in the following countries: Algeria (D. Hammouda); Argentina (D. Loria, E. Matos); Benin (E. Alihonou); Bolivia (J.L. Rios-Dalenz); Brazil (J. Eluf-Neto); Canada (P. Ghadirian); Chile (C. Ferreccio, A. Luzoro, J.M. Ojeda, R. Prado); Colombia (N. Aristizabal, L.A. Tafur, M. Molano, H. Posso); Cuba (M. Torroella); Georgia (T. Alibegashvili, D. Kordzaia); Guinea (N. Keita, M. Koulibaly); India (T. Rajkumar, R. Rajkumar); South Korea (D-H. Lee, H.R. Shin); Mali (S. Bayo); Morocco (N. Chaouki); Nigeria (J.O. Thomas, C. Okolo, I. Adewole); The Netherlands (C.J.L.M. Meijer, P.J.F. Snijders); Panama (E. de los Rios); Paraguay (P.A. Rolon); Peru (E. Caceres, C. Santos); the Philippines (C. Ngelangel); Poland (W. Zatonski); South Africa (D. Moodley); Kenya (P. Gichangi, H. de Vuyst); Spain (S. de Sanjose, X. Castellsague); Tanzania (J.N. Kitinya); Thailand (S. Chichareon, S. Sukvirach, S. Tunsakul) and Uganda (H.R. Wabinga).

Reference List

1. **Bhattacharjee, B., N. R. Mandal, S. Roy, and S. Sengupta**. 2008. Characterization of sequence variations within HPV16 isolates among Indian women: prediction of causal role of rare non-synonymous variations within intact isolates in cervical cancer pathogenesis. Virology **377**:143-150.

2. **Bosch, F. X., M. M. Manos, N. Munoz, M. Sherman, A. M. Jansen, J. Peto, M. H. Schiffman, V. Moreno, R. Kurman, and K. V. Shah**. 1995. Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. International biological study on cervical cancer (IBSCC) Study Group [see comments]. J. Natl. Cancer Inst. **87**:796-802.

3. **Burk, R. D., Z. Chen, A. Harari, B. C. Smith, B. J. Kocjan, P. J. Maver, and M. Poljak**. 2011. Classification and nomenclature system for Human Alphapapillomavirus variants: general features, nucleotide landmarks and assignment of HPV6 and HPV11 isolates to variant lineages. Acta Dermatovenerol. Alp Panonica. Adriat. **20**:113-123.

4. **Chan, S. Y., L. Ho, C. K. Ong, V. Chow, B. Drescher, M. Durst, M. J. ter, L. Villa, J. Luande, H. N. Mgaya, and .** 1992. Molecular variants of human papillomavirus type 16 from four continents suggest ancient pandemic spread of the virus and its coevolution with humankind. J. Virol. **66**:2057-2066.

5. **Chen, Z., M. Schiffman, R. Herrero, R. Desalle, K. Anastos, M. Segondy, V. V. Sahasrabuddhe, P. E. Gravitt, A. W. Hsing, and R. D. Burk**. 2011. Evolution and taxonomic classification of human papillomavirus 16 (HPV16)-related variant genomes: HPV31, HPV33, HPV35, HPV52, HPV58 and HPV67. PLoS. One. **6**:e20183.

6. **Chen, Z., M. Terai, L. Fu, R. Herrero, R. Desalle, and R. D. Burk**. 2005. Diversifying selection in human papillomavirus type 16 lineages based on complete genome analyses. J. Virol. **79**:7014-7023.

7. **Clifford, G., S. Franceschi, M. Diaz, N. Munoz, and L. L. Villa**. 2006. Chapter 3: HPV type-distribution in women with and without cervical neoplastic diseases. Vaccine **24 Suppl 3**:S3-26-S3/34.

8. **Clifford, G. M., S. Gallus, R. Herrero, N. Munoz, P. J. Snijders, S. Vaccarella, P. T. Anh, C. Ferreccio, N. T. Hieu, E. Matos, M. Molano, R. Rajkumar, G. Ronco, S. S. de, H. R. Shin, S. Sukvirach, J. O. Thomas, S. Tunsakul, C. J. Meijer, and S. Franceschi**. 2005. Worldwide distribution of human papillomavirus types in cytologically normal women in the International Agency for Research on Cancer HPV prevalence surveys: a pooled analysis. Lancet **366**:991-998.

9. **Gheit, T., I. Cornet, G. M. Clifford, T. Iftner, C. Munk, M. Tommasino, and S. K. Kjaer**. 2011. Risks for persistence and progression by human papillomavirus type 16 variant lineages among a population-based sample of Danish women. Cancer Epidemiol. Biomarkers Prev. **20**:1315-1321.

10. **Grodzki, M., G. Besson, C. Clavel, A. Arslan, S. Franceschi, P. Birembaut, M. Tommasino, and I. Zehbe**. 2006. Increased risk for cervical disease progression of French women infected with the human papillomavirus type 16 E6-350G variant. Cancer Epidemiol. Biomarkers Prev. **15**:820-822.

11. **Ho, L., S. Y. Chan, R. D. Burk, B. C. Das, K. Fujinaga, J. P. Icenogle, T. Kahn, N. Kiviat, W. Lancaster, P. Mavromara-Nazos, and .** 1993. The genetic drift of human papillomavirus

type 16 is a means of reconstructing prehistoric viral spread and the movement of ancient human populations. J. Virol. **67**:6413-6423.

12. **Ho, L., S. Y. Chan, V. Chow, T. Chong, S. K. Tay, L. L. Villa, and H. U. Bernard**. 1991. Sequence variants of human papillomavirus type 16 in clinical samples permit verification and extension of epidemiological studies and construction of a phylogenetic tree. J. Clin. Microbiol. **29**:1765-1772.

13. **Huertas-Salgado, A., D. C. Martin-Gamez, P. Moreno, R. Murillo, M. M. Bravo, L. Villa, and M. Molano**. 2011. E6 molecular variants of human papillomavirus (HPV) type 16: an updated and unified criterion for clustering and nomenclature. Virology **410**:201-215.

14. **Jacobs, M. V., J. M. Walboomers, P. J. Snijders, F. J. Voorhorst, R. H. Verheijen, N. Fransen-Daalmeijer, and C. J. Meijer**. 2000. Distribution of 37 mucosotropic HPV types in women with cytologically normal cervical smears: the age-related patterns for high-risk and low-risk types. Int. J. Cancer **87**:221-227.

15. **Kammer, C., M. Tommasino, S. Syrjanen, H. Delius, U. Hebling, U. Warthorst, H. Pfister, and I. Zehbe**. 2002. Variants of the long control region and the E6 oncogene in European human papillomavirus type 16 isolates: implications for cervical disease. Br. J. Cancer **86**:269-273.

16. **Munoz, N., F. X. Bosch, S. de Sanjose, R. Herrero, X. Castellsague, K. V. Shah, P. J. Snijders, and C. J. Meijer**. 2003. Epidemiologic classification of human papillomavirus types associated with cervical cancer. N. Engl. J. Med. **348**:518-527.

17. **Munoz, N., X. Castellsague, A. B. de Gonzalez, and L. Gissmann**. 2006. Chapter 1: HPV in the etiology of human cancer. Vaccine **24 Suppl 3**:S3-1-S310.

18. **Myers, G. Delius H. Icenogle J. Bernard H. U. Baker C. Halpern A. and Wheeler C**. 1995. Human Papillomaviruses 1995: A compilation and analysis of nucleic acid and amino acid sequences. Los Alamos National Library 1995, Los Alamos, NM.

19. **Seedorf, K., G. Krammer, M. Durst, S. Suhai, and W. G. Rowekamp**. 1985. Human papillomavirus type 16 DNA sequence. Virology **145**:181-185.

20. **Smith, B., Z. Chen, L. Reimers, D. K. van, M. Schiffman, R. Desalle, R. Herrero, K. Yu, S. Wacholder, T. Wang, and R. D. Burk**. 2011. Sequence Imputation of HPV16 Genomes for Genetic Association Studies. PLoS. One. **6**:e21375.

21. **Smith, J. S., L. Lindsay, B. Hoots, J. Keys, S. Franceschi, R. Winer, and G. M. Clifford**. 2007. Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: a meta-analysis update. Int. J. Cancer **121**:621-632.

22. **Tornesello, M. L., M. L. Duraturo, I. Salatiello, L. Buonaguro, S. Losito, G. Botti, G. Stellato, S. Greggi, R. Piccoli, S. Pilotti, B. Stefanon, P. G. De, S. Franceschi, and F. M. Buonaguro**. 2004. Analysis of human papillomavirus type-16 variants in Italian women with cervical intraepithelial neoplasia and cervical cancer. J. Med. Virol. **74**:117-126.

23. **Wheeler, C. M., T. Yamada, A. Hildesheim, and S. A. Jenison**. 1997. Human papillomavirus type 16 sequence variants: identification by E6 and L1 lineage-specific hybridization. J. Clin. Microbiol. **35**:11-19.

24. **Xi, L. F., L. A. Koutsky, D. A. Galloway, J. Kuypers, J. P. Hughes, C. M. Wheeler, K. K. Holmes, and N. B. Kiviat**. 1997. Genomic variation of human papillomavirus type 16 and risk

for high grade cervical intraepithelial neoplasia [see comments]. J. Natl. Cancer Inst. **89**:796-802.

25. **Xi, L. F., L. A. Koutsky, A. Hildesheim, D. A. Galloway, C. M. Wheeler, R. L. Winer, J. Ho, and N. B. Kiviat**. 2007. Risk for high-grade cervical intraepithelial neoplasia associated with variants of human papillomavirus types 16 and 18. Cancer Epidemiol. Biomarkers Prev. **16**:4-10.

26. **Yamada, T., M. M. Manos, J. Peto, C. E. Greer, N. Munoz, F. X. Bosch, and C. M. Wheeler**. 1997. Human papillomavirus type 16 sequence variation in cervical cancers: a worldwide perspective. J. Virol. **71**:2463-2472.

27. **Yamada, T., C. M. Wheeler, A. L. Halpern, A. C. Stewart, A. Hildesheim, and S. A. Jenison**. 1995. Human papillomavirus type 16 variant lineages in United States populations characterized by nucleotide sequence analysis of the E6, L2, and L1 coding segments. J. Virol. **69**:7743-7753.

28. **Zehbe, I., G. Voglino, H. Delius, E. Wilander, and M. Tommasino**. 1998. Risk of cervical cancer and geographical variations of human papillomavirus 16 E6 polymorphisms [letter]. Lancet **352**:1441-1442.

29. **Zehbe, I., E. Wilander, H. Delius, and M. Tommasino**. 1998. Human papillomavirus 16 E6 variants are more prevalent in invasive cervical carcinoma than the prototype. Cancer Res. **58**:829-833.

30. **Zuna, R. E., W. E. Moore, R. P. Shanesmith, S. T. Dunn, S. S. Wang, M. Schiffman, G. L. Blakey, and T. Teel**. 2009. Association of HPV16 E6 variants with diagnostic severity in cervical cytology samples of 354 women in a US population. Int. J. Cancer **125**:2609-2613.