Transmission between Archaic and Modern Human Ancestors during the Evolution of the Oncogenic Human Papillomavirus 16

Ville N. Pimenoff,*,1,2 Cristina Mendes de Oliveira,1,3 and Ignacio G. Bravo1,4

1Infections and Cancer Laboratory, Cancer Epidemiology Research Programme, Catalan Institute of Oncology, Barcelona, Spain
2Unit of Biomarkers and Susceptibility, Bellvitge Institute of Biomedical Research (IDIBELL), Barcelona, Spain
3Virology Laboratory, Institute of Tropical Medicine, University of São Paulo, São Paulo, Brazil
4MIVEGEC (UMR CNRS 5290, IRD 224, UM), National Center for Scientific Research (CNRS), Montpellier, France

*Corresponding author: E-mail: ville.pimenoff@gmail.com.
Associate editor: Beth Shapiro

Abstract

Every human suffers through life a number of papillomaviruses (PVs) infections, most of them asymptomatic. A notable exception are persistent infections by Human papillomavirus 16 (HPV16), the most oncogenic infectious agent for humans and responsible for most infection-driven anogenital cancers. Oncogenic potential is not homogeneous among HPV16 lineages, and genetic variation within HPV16 exhibits some geographic structure. However, an in-depth analysis of the HPV16 evolutionary history was still wanting. We have analyzed extant HPV16 diversity and compared the evolutionary and phylogeographical patterns of humans and of HPV16. We show that codivergence with modern humans explains at most 30% of the present viral geographical distribution. The most explanatory scenario suggests that ancestral HPV16 already infected ancestral human populations and that viral lineages co-diverged with the hosts in parallel with the split between archaic Neanderthal-Denisovans and ancestral modern human populations, generating the ancestral HPV16A and HPV16BCD viral lineages, respectively. We propose that after out-of-Africa migration of modern human ancestors, sexual transmission between human populations introduced HPV16A into modern human ancestor populations. We hypothesize that differential coevolution of HPV16 lineages with different but closely related ancestral human populations and subsequent host-switch events in parallel with introgression of archaic alleles into the genomes of modern human ancestors may be largely responsible for the present-day differential prevalence and association with cancers for HPV16 variants.

Key words: divergence, evolutionary medicine, Hominin evolution, host-switch, human papillomavirus, infection and cancer, sexually transmitted infection, variant, virus-host coevolution.

Introduction

Virus lifestyle shapes viral population structure through essential virus life traits, such as infectivity, immunogenicity, latency, transmission rate, mutation rate, virion productivity, and duration of the infection (Sharp 2002). The interaction between the virus and the host’s immune system strongly influences viral evolution: immune memory reduces the number of susceptible hosts, and differential immune recognition and response against different pathogen phenotypes may favour selection of certain viral lineages (Holmes 2008). Finally, the evolutionary history of the host constrains further the population structure of the virus. Thus, for viruses and hosts with a long shared evolutionary history, the virus population is shaped by population-level processes of genetic drift and natural selection, in both the virus and the host populations (Holmes 2008).

Papillomaviruses (PVs) have a long common history with amniotes (Bravo and Félez-Sánchez 2015). PVs are double-stranded DNA viruses, with a circular genome of about 8 kb. More than 300 PVs have been described and above 200 of them have been retrieved from humans (Bzhalava et al. 2015). Human PVs (HPVs) infect dividing epithelia, and virtually all humans are hosts to a variable number of HPV infections at cutaneous (Antonsson et al. 2003) but often also at mucosal sites (Ma et al. 2014). The viral infection does not kill the target cell and in most cases PVs can complete their life cycle and be maintained as a chronic, asymptomatic infection with viral genomes remaining episomally as plasmids in the infected cell (Doorbar et al. 2012). In some cases, they can cause productive, wart-like lesions or even be involved in certain cancers (Doorbar et al. 2012). The fine balance between viral replication and host immune tolerance suggests a long coexistence between PVs and their hosts, and indeed ancestral PVs probably infected the ancestral amniotes (García-Vallvé et al. 2005) and possibly all bony vertebrates (López-Bueno et al. 2016). The evolution of PVs in mammals likely started with an initial adaptive radiation event that generated a handful of viral crown groups, linked to the evolution of mammalian fur and skin glands (Gottschling et al. 2011; Bravo and Félez-Sánchez 2015). This diversification was followed by limited host-linked evolution and encompassed also
incomplete lineage sorting and host-switch events (Gottschling et al. 2007; Gottschling et al. 2011). At shallower evolutionary timescale, it has been commonly assumed that HPV16 has co-diverged with modern human populations, based on certain geographical structure of HPV16 variant distribution (Bernard 1994; Cornet et al. 2012), but this hypothesis has never been rigorously tested.

Infections by anogenital HPVs are very common: >80% of all sexually active adults are infected by one of these HPVs at least once in their lifetime (Einstein et al. 2009; Chesson et al. 2014), and the time point prevalence of cervical HPVs infections in sexually active healthy women is estimated to be around 12% (Bruni et al. 2010) or even higher in younger women (Molden et al. 2016). At least 12 evolutionarily related HPVs are carcinogenic to humans (Bouvard et al. 2009), causing different fractions of anogenital and oropharyngeal cancers (Moscicki et al. 2012). HPV16 is not only the most prevalent HPV in infection-associated cancers of the cervix, vulva, vagina, anus, penis, and oropharynx (Moscicki et al. 2012) but also the most prevalent HPV infection of the genital tract occurring asymptomatically in healthy individuals (Bruni et al. 2010). Within the HPV16 lineage, distribution of genetic diversity presents certain geographic structure, with different HPV16 variants showing differential prevalence in different geographical regions (Cornet et al. 2012). Furthermore, the oncogenic potential is not homogeneous for all HPV16 variants (Schiffman et al. 2010) and varies in association with different human populations (Villa et al. 2000; Berumen et al. 2001; Burk et al. 2003; Xi et al. 2006; Zuna et al. 2011; Cornet et al. 2013). Such differential connection between viral phylogeography and oncogenic potential might indicate that the evolution of HPV16 lineage has been at least partly shaped by the differential host immune response.

In this study, we have analyzed human and virus genomic data to infer the evolutionary history of HPV16. The common understanding about the evolution of HPVs is that the modern repertoire of diverse HPVs has coevolved with modern humans as a host population, so that the current geographic distribution of HPVs reflects modern human migration dispersal patterns (Ho et al. 1993; Bernard 1994). Here, we show for the first time that differential coevolution of HPV16 lineages with different but closely related ancestral human populations together with recent host-switch events may be largely responsible for the present-day differential prevalence and association with cancers for HPV16 variants. The most parsimonious interpretation of our results requires that the ancestor of HPV16 already infected the ancestral human populations >500 thousands years ago (kya). The split between Neanderthals/Denisovans and modern human ancestors was mirrored by a split in the viral populations, namely HPV16A, carried by ancestral human populations, and HPV16BCD, carried by the populations of modern human ancestors in Africa. After the last out-of-Africa migration event 60–120 kya, the modern human ancestor populations that left Africa carried a reduced sample of the viral diversity in the continent: the HPV16B lineage remained in African populations but was lost by lineage sorting in those leaving the continent. The viral lineage HPV16CD gave rise in allopatriy to the HPV16C variant in Africa and to the HPV16D variant outside Africa. Later, the interbreeding events between Neanderthal and Denisovan populations with modern human ancestor populations lead to a host-switch through sexual transmission of the HPV16A virus lineage from archaic populations into the modern human ancestors. The HPV16A lineage thus transmitted expanded rapidly in the new host populations and became dominant in Eurasia and in the Americas.

**Results**

**HPV16 Phylogeography**

We inferred the phylogenetic relationship among all available, non-recombinant HPV16 genomes (n = 118) under maximum likelihood (ML) framework for both the full-genome alignment (supplementary material S1, Supplementary Material online) and for the coding region alignment (supplementary material S2, see also supplementary table S1, Supplementary Material online). Phylogenetic inference recovered all previously reported HPV16 variant lineages, and we retained for subsequent analysis HPV16A1-3, A4, B, C, and D lineages, all supported by >98% bootstrap values for the full-genome alignment (fig. 1A and supplementary fig. S1, Supplementary Material online). Pairwise distances under ML were significantly larger (2.28 times, 95% confidence interval [CI], 1.88–2.69; Wilcoxon–Mann–Whitney test $P = 2.2e−16$) for the full-genome than for the selection-filtered alignment (supplementary fig. S2, Supplementary Material online).

To assess HPV16 phylogeography, we analyzed a total of 1,719 HPV16 isolate cases (alignments available as supplementary materials S3–S4, Supplementary Material online) by complementing the 118 HPV16 full-genome sequences with 1,601 partial HPV16 isolate sequences spanning the LCR, the E6 oncogene, and the L2 capsid gene. A total of 1,680 global HPV16 isolate sequences with known geographical origin could be unambiguously assigned to a precise HPV16 lineage using an evolutionary placement algorithm (supplementary table S2, Supplementary Material online). Phylogeography of the 1,680 isolate cases showed that the HPV16A1-3 lineage predominated in Europe, South Asia, and Central/South America, and was also present in all other continental subgroups, albeit with very low prevalence in sub-Saharan Africa (fig. 2). HPV16A4 was the most prevalent lineage in East Asia and was also present in North America, but was virtually absent elsewhere. Variants B and C were largely restricted to Africa and were especially prevalent in Sub-Saharan Africa, although they were also observed in North America. Variant D was present in all continents, displaying low prevalence in Sub-Saharan Africa, and the highest frequency in Central/South America.

To explain the evolution and diversity of HPV16 and to infer its origins, we explored the differential fit to the data of two alternative scenarios: 1) the recent-out-of-Africa (ROOA) model of exclusive codiversgence for HPV16 with modern human populations and 2) the Hominin-host-switch (HHS)
model, including a viral transmission between archaic and ancestral modern human populations. On the one hand, the ROOA scenario required exclusive codivergence of HPV16 lineages with modern human dispersals. On the other hand, the HHS scenario implied an ancestral virus-host codivergence episode during the split between Homo sapiens and H. neanderthalensis lineages, followed by more recent transmission events between Neanderthals/Denisovans and modern human ancestors (fig. 3). Both scenarios are plausible given the topology of the phylogenetic relationships between HPV16 lineages (fig. 1) and the impossibility to root the HPV16 tree using extant HPVs sequences.

To test the explanatory power of the ROOA model, we analyzed the correlation between geography-based population structure of modern humans and of HPV16 by comparing genetic distances between geographically defined human metapopulations with the genetic distances between HPV16 isolates. We used nine previously described metapopulation groups (supplementary table S3, Supplementary Material online), excluding North America due to lack of human genome data, and tested the correlation between the $F_{ST}$ distance matrices for humans and for HPV16 (supplementary fig. S3, Supplementary Material online). We observed indeed correlation between both matrices, but geographical structure

\[ \text{FIG. 1.} \text{ Phylogenetic tree of the 118 HPV16 complete genomes. (A) Unrooted ML phylogenetic tree inferred using the complete coding region alignment after excluding a total of 26 positively selected and 115 negatively selected codons. The final alignment encompassed 118 sequences, 6,093 bp and 316 distinct alignment patterns. Bootstrap support after 500 replicates is given for branches leading to the A1–3, A4, B, C, and D variant lineages. Scale bar in substitutions per nucleotide site. (B) Maximum clade credibility tree inferred for selection-filtered HPV16 genome coding region alignment under the HHS scenario, 9.6 \times 10^{-9} [5.5 \times 10^{-9}–13.6 \times 10^{-9}] \text{ subs/site/yr as substitution rate prior and enforcing two time point calibrations: archaic Hominin divergence at 500 kya (95\% CI, 400–600 kya) and recent modern human migration out-of-Africa at 90 kya (95\% CI, 60–120 kya). Scale bar in millions of years ago. Node bars indicate the 95\% probability intervals for the corresponding node age.} \]

\[ \text{FIG. 2.} \text{ Phylogeographic distribution of the 1,680 HPV16 sequences encompassing the LCR, E6 and L2 genome loci. Each sequence was assigned to a specific HPV16 variant lineage (see the color coding for A1–3, A4, B, C, and D variants). The size of the pie charts is proportional to the number of sequences from the corresponding geographic region (supplementary table S3, supplementary material online).} \]
across different human genome loci explained <30% of the worldwide HPV16 geographical structure (table 1, see also supplementary table S4, Supplementary Material online).

Notably, most of the European and sub-Saharan African ascertained autosomal human genome loci variability showed significant but limited (<30%) correlation with the HPV16...
LCR-E6 variability. Very interestingly, after excluding the sites under selection in the HPV16 E6 gene only the correlation with the mitochondrial genome variability and most of the sub-Saharan African but not of the European ascertained autosomal genome data remained significant, and accounted for barely 15% of the global distribution of HPV16 diversity. Nevertheless, it is important to note that after controlling for false discovery rate none of the human genome and of HPV16 correlations showed to be significant.

Time Inference for HPV16 Diversification
To estimate divergence times for HPV16 diversification under the two alternative evolutionary scenarios, we applied a Bayesian Markov chain Monte Carlo (MCMC) inference. First, we tested whether we could accurately identify the root of the HPV16 genome diversity by performing phylogenetic inference incorporating the most closely related HPVs, members of Alphapapillomaviruses species 9. The long evolutionary distances between these viruses and HPV16 prevented unambiguous identification of the root of HPV16 lineages, with either ML (RAxML) or Bayesian (Phylobayes) approaches (supplementary fig. S4, Supplementary Material online). In all reconstructions, the preferred topology was (A,(B,(C,D))), which corresponded to the HHS scenario. Nevertheless, the likelihood values for the HPV16 trees obtained after forcing our two alternative scenarios, namely, the HHs and ROOA models, differed by <0.01%, and neither the Shimodaira–Hasegawa test nor the Robinson–Foulds split method showed significant differences between both topologies. When introducing time into the Bayesian (BEAST) phylogenetic analyses without imposing any prior on tree topologies, the same (A,(B,(C,D))) topology (i.e., the HHs model) was systematically the preferred one for both HPV16 complete and HPV16 selection-filtered coding region alignments (supplementary fig. S4, Supplementary Material online). The combined difference in topology and branch length between trees obtained with the full-length and with the selection-filtered alignments was very small (as assessed by a K-score <0.004) (supplementary table S5, Supplementary Material online). We resorted then to dated Bayesian MCMC inference for identifying the more plausible between the two alternative evolutionary scenarios. For the HHs scenario, we used the archaic Hominin divergence at 500 kya (95% CI, 400–600 kya, see references in Materials and Methods) to calibrate the time for the most recent common ancestor (tmrca) of extant HPV16 lineages, as well as the recent modern human migration out-of-Africa at 90 kya (95% CI, 60–120 kya, see references in Materials and Methods) to calibrate the tmrca of lineages HPV16C and D. For the ROOA scenario, we used the recent migration

### Table 1. Mantel’s Permutation Test for Similarity of FST Distance Matrices between HPV16 and Human Genome Sequence Data Sets.

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Loci</th>
<th>Region/SNPs</th>
<th>n</th>
<th>Correlation (%)</th>
<th>P-value</th>
<th>FDR</th>
<th>Correlation (%)</th>
<th>P-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>mtDNA</td>
<td>1</td>
<td>16 kb genome</td>
<td>875</td>
<td>12.4</td>
<td>0.037</td>
<td>0.017</td>
<td>15.8</td>
<td>0.026</td>
<td>0.018</td>
</tr>
<tr>
<td>NRY</td>
<td>1</td>
<td>5,000 kb</td>
<td>551</td>
<td>8.5</td>
<td>0.062</td>
<td>0.028</td>
<td>11.8</td>
<td>0.044</td>
<td>0.036</td>
</tr>
<tr>
<td>Europe</td>
<td>chr1</td>
<td>6,353</td>
<td>773</td>
<td>8.8</td>
<td>0.062</td>
<td>0.029</td>
<td>12.7</td>
<td>0.033</td>
<td>0.026</td>
</tr>
<tr>
<td>Europe</td>
<td>chr2</td>
<td>5,918</td>
<td>773</td>
<td>6.3</td>
<td>0.091</td>
<td>0.040</td>
<td>9.6</td>
<td>0.054</td>
<td>0.040</td>
</tr>
<tr>
<td>Europe</td>
<td>chr3</td>
<td>5,325</td>
<td>773</td>
<td>8.8</td>
<td>0.063</td>
<td>0.032</td>
<td>12.7</td>
<td>0.037</td>
<td>0.030</td>
</tr>
<tr>
<td>Europe</td>
<td>chr4</td>
<td>4,211</td>
<td>773</td>
<td>8.9</td>
<td>0.063</td>
<td>0.030</td>
<td>12.7</td>
<td>0.036</td>
<td>0.029</td>
</tr>
<tr>
<td>Europe</td>
<td>chr5</td>
<td>4,743</td>
<td>773</td>
<td>5.6</td>
<td>0.104</td>
<td>0.047</td>
<td>8.7</td>
<td>0.061</td>
<td>0.047</td>
</tr>
<tr>
<td>Europe</td>
<td>chr6</td>
<td>4,736</td>
<td>773</td>
<td>6.3</td>
<td>0.095</td>
<td>0.041</td>
<td>9.5</td>
<td>0.055</td>
<td>0.042</td>
</tr>
<tr>
<td>Europe</td>
<td>chr7</td>
<td>4,162</td>
<td>773</td>
<td>5.2</td>
<td>0.112</td>
<td>0.048</td>
<td>8.1</td>
<td>0.062</td>
<td>0.048</td>
</tr>
<tr>
<td>Europe</td>
<td>chr8</td>
<td>3,750</td>
<td>773</td>
<td>8.0</td>
<td>0.070</td>
<td>0.036</td>
<td>11.6</td>
<td>0.045</td>
<td>0.037</td>
</tr>
<tr>
<td>Europe</td>
<td>chr9</td>
<td>3,222</td>
<td>773</td>
<td>8.8</td>
<td>0.064</td>
<td>0.033</td>
<td>12.7</td>
<td>0.038</td>
<td>0.033</td>
</tr>
<tr>
<td>Europe</td>
<td>chr10</td>
<td>4,223</td>
<td>773</td>
<td>9.2</td>
<td>0.059</td>
<td>0.026</td>
<td>13.1</td>
<td>0.035</td>
<td>0.028</td>
</tr>
<tr>
<td>Europe</td>
<td>chr11</td>
<td>3,850</td>
<td>773</td>
<td>5.9</td>
<td>0.099</td>
<td>0.046</td>
<td>9.2</td>
<td>0.058</td>
<td>0.046</td>
</tr>
<tr>
<td>Europe</td>
<td>chr12</td>
<td>3,339</td>
<td>773</td>
<td>5.1</td>
<td>0.116</td>
<td>0.049</td>
<td>8.0</td>
<td>0.066</td>
<td>0.049</td>
</tr>
<tr>
<td>Europe</td>
<td>chr13</td>
<td>2,234</td>
<td>773</td>
<td>5.8</td>
<td>0.098</td>
<td>0.045</td>
<td>9.2</td>
<td>0.057</td>
<td>0.045</td>
</tr>
<tr>
<td>Europe</td>
<td>chr14</td>
<td>2,307</td>
<td>773</td>
<td>5.9</td>
<td>0.097</td>
<td>0.043</td>
<td>9.3</td>
<td>0.056</td>
<td>0.043</td>
</tr>
<tr>
<td>Europe</td>
<td>chr15</td>
<td>2,346</td>
<td>773</td>
<td>6.2</td>
<td>0.096</td>
<td>0.042</td>
<td>9.3</td>
<td>0.056</td>
<td>0.041</td>
</tr>
<tr>
<td>Europe</td>
<td>chr16</td>
<td>2,636</td>
<td>773</td>
<td>7.5</td>
<td>0.078</td>
<td>0.038</td>
<td>11.1</td>
<td>0.045</td>
<td>0.038</td>
</tr>
<tr>
<td>Europe</td>
<td>chr17</td>
<td>2,109</td>
<td>773</td>
<td>6.8</td>
<td>0.084</td>
<td>0.039</td>
<td>10.1</td>
<td>0.045</td>
<td>0.039</td>
</tr>
<tr>
<td>Europe</td>
<td>chr18</td>
<td>2,243</td>
<td>773</td>
<td>9.8</td>
<td>0.056</td>
<td>0.025</td>
<td>13.9</td>
<td>0.033</td>
<td>0.025</td>
</tr>
<tr>
<td>Europe</td>
<td>chr19</td>
<td>1,175</td>
<td>773</td>
<td>3.9</td>
<td>0.139</td>
<td>0.050</td>
<td>6.5</td>
<td>0.085</td>
<td>0.050</td>
</tr>
<tr>
<td>Europe</td>
<td>chr20</td>
<td>2,067</td>
<td>773</td>
<td>9.9</td>
<td>0.055</td>
<td>0.023</td>
<td>13.8</td>
<td>0.031</td>
<td>0.021</td>
</tr>
<tr>
<td>Europe</td>
<td>chr21</td>
<td>973</td>
<td>773</td>
<td>10.0</td>
<td>0.055</td>
<td>0.024</td>
<td>14.1</td>
<td>0.031</td>
<td>0.022</td>
</tr>
<tr>
<td>Europe</td>
<td>chr22</td>
<td>1,141</td>
<td>773</td>
<td>7.7</td>
<td>0.073</td>
<td>0.037</td>
<td>11.6</td>
<td>0.042</td>
<td>0.034</td>
</tr>
</tbody>
</table>

**Note.**—HPV16 sequence data (1,192 bp) encompassing LCR (735 bp) and E6 (457 bp) regions for 1,101 worldwide isolates were used in the correlations. mitochondrial DNA (mtDNA); nonrecombining part of the Y chromosome (NRY); autosomal SNP data stratified by chromosome and ascertained using a European (Europe) ancestry data (autosomal SNP data ascertained using a sub-Saharan ancestry data is provided in supplementary table S8, supplementary material online). For both datasets nine geographic metapopulations were defined: North-East/West/Southern Africa, South Asia, East Asia, Europe, Central and South America (supplementary table S2 and S5, supplementary material online).

- **a**R² of the correlation coefficient.
- **b**Mantel-test after 10,000 permutations.
- **c**False discovery rate controlling procedure (q = 0.05).
out-of-Africa with the same 90-kya settings as above but in this case to calibrate the tmrca of all HPV16 lineages. Both strict and relaxed clock were tested for all MCMC calibrations and the best log-likelihood value using Bayes Factor (2lnBF) was obtained for relaxed clock (2lnBF > 38). The demographic models of exponential growth or of Bayesian skyline for the population function of the coalescent tree prior yielded in both cases growth rates consistently above zero, excluding thus constant population size as a model. In addition, Bayesian skyline plot estimating changes in effective population size through time showed a significant increase of the HPV16 population in the last 12 kya (supplementary fig. S5, Supplementary Material online). For final analyses, we used the evolutionary models and priors showing the best performance, namely, relaxed log-linear model (2lnBF > 83). Comparison of divergence time estimates between the HHS and ROOA scenarios for HPV16 evolution is presented in table 2. For the analyses based on the positions evolving neutrally, the log-marginal likelihoods showed the highest value and strongest support (2lnBF = 8.9) for the HHS scenario with a substitution rate $1.2 \times 10^{-9}$ (95% CI, $1.43 \times 10^{-9} - 22.1 \times 10^{-9}$) subs/site/yr. For the analyses that included the positions under selection, the HHS model was again the preferred model albeit with a substitution rate two times higher. The log-marginal likelihood value for the HHS scenario including the positions under selection did not significantly differ from that for the ROOA model (2lnBF = 1.2) although the inferred evolutionary rate in the case of ROOA model needed to be around five times higher. Maximum clade credibility tree inferred for selection-filtered HPV16 genome coding region alignment under the HHS scenario and enforcing the two HHS model time point calibrations is presented in fig 1B. Furthermore, to estimate whether our time point calibrations had an impact on the divergence time estimates, we also performed MCMC analysis either by imposing the HHS model topology or without imposing any prior on tree topology and without using any time point calibrations (table 3). Depending on substitution rate priors, divergence times of extant HPV16 lineages showed to be between 260 kya and 4.8 Ma. In all cases, the preferred model was the HHS scenario. Although HPV16A was thus always the basal clade, divergence within HPV16A showed to be lower than within HPV16B (supplementary fig S6, Supplementary Material online).

Variation within the HPV16 E6 Oncogene
To investigate geographical differences in HPV16 genetic diversity, we estimated summary statistics within each geographic HPV16 population using all 1,123 available HPV16 nearly complete E6 sequences (table 4). Overall, HPV16 E6 genetic diversity was highest outside sub-Saharan Africa. Both Tajima’s D and Fu’s-Li’s D statistics as well as the more robust R2 index showed statistically large negative values for European isolates, suggesting past selective sweeps and/or population expansion, for HPV16 population in Europe only. HPV16 E6 haplotype diversity was similar in sub-Saharan Africa, Europe, and South America and significantly higher in North Africa and Central America than in sub-Saharan Africa (table 4). East Asian and Central American HPV16 isolates showed higher average number of pairwise differences compared with sub-Saharan African isolates, even after accounting for intralineage diversity. Furthermore, to estimate the geographical origin of the HPV16 E6 gene haplotype diversity, we constructed a median-joining haplotype network for the E6 gene sequences (table 5, see also supplementary fig. S7, Supplementary Material online). Two of the most common HPV16A1-3 lineage E6 haplotypes were observed in Eurasian isolate background, and only 1–3% of these common haplotypes originated from sub-Saharan African isolates. In contrast, the most frequent HPV16B and C variant E6 haplotypes were observed in >92% of African isolate background. The most common HPV16D E6 haplotype displayed the highest frequency (58%) in the American isolate background and the lowest (4–5%) in North and sub-Saharan African isolates. The second most common HPV16D E6 haplotype showed the highest frequency (58%) in North African isolates, intermediate frequencies in Eurasian and in American isolates (18% and 21%, respectively) and very small (3%) of the sub-Saharan African isolate background.

Discussion
HPV16 is the most oncogenic virus for humans, and epidemiological evidence suggests that its oncogenic potential is associated with high viral loads and with infection persistence (Doorbar et al. 2012; Moscicki et al. 2012). Viral genotypic variation may underlie strong phenotypic variation so that different HPV16 variants may display differential persistence and viral load and therefore differential carcinogenicity. Several reports have communicated differential oncogenic potential for different HPV16 variant lineages (Villa et al. 2000; Schiffman et al. 2010; Zuna et al. 2011; Freitas et al. 2014). However, the differential risks for different variants seem to depend on the host population (Cornet et al. 2013; Qmichou et al. 2013), and to vary with the specific individual genetic background (Xi et al. 2006; Lopera et al. 2014). In the present study, we have addressed the origin, diversification, and dispersal of HPV16. We have used the largest available collection of HPV16 full-length genomes and partial sequences to generate estimates of the worldwide phylogeography and divergence time of extant HPV16 lineages. Our results support a scenario of codivergence of HPV16 with separate but closely related ancestral Hominin populations with subsequent host-switch events, likely through sexual transmission between archaic and modern human ancestral populations in the recent human evolutionary history. We have further compared the genetic diversity patterns of HPV16 with those estimated for modern human populations using genome data. Altogether, our results suggest that virus-host coevolution and host switch events have been fundamental factors that have shaped HPV16 evolution.

Purifying selection can mask the ancient origin of recently sampled pathogens (Wertheim and Kosakovsky Pond 2011), and our results point in the same direction. Results based on only HPV16 genome positions evolving neutrally showed significantly better fit to the data for our HHS model while
Table 2. Substitution Rate and Divergence Time Estimates with Highest Probability Density (HPD) in Thousands of Years Ago (kya) Inferred for HPV16 Complete (Full) or Selection-Filtered Genome Alignment (Neutral), under the HHS or the ROOA Scenarios, and Considering Different Priors.

<table>
<thead>
<tr>
<th>Model</th>
<th>Calibration</th>
<th>Rate uniform prior</th>
<th>Data</th>
<th>(\times 10^{-9}) (sub/site/yr)</th>
<th>Log marginal</th>
<th>2lnBF</th>
<th>tmrca A–4 (kya)</th>
<th>tmrca BCD (kya)</th>
<th>tmrca CD (kya)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHS HHS 2 calib</td>
<td>PVs estimate [13.2–24.7] (^{a})</td>
<td>Neutral</td>
<td>18.4 (14.5–22.1)</td>
<td>–12476.94 Ref</td>
<td>461 (365–561)</td>
<td>87 (48–133)</td>
<td>197 (121–293)</td>
<td>105 (82–129)</td>
<td></td>
</tr>
<tr>
<td>HHS HHS 2 calib</td>
<td>Human mtDNA [17.6–32.3] (^{b})</td>
<td>Neutral</td>
<td>20.3 (16–25.1)</td>
<td>–12450.34 6.8</td>
<td>456 (358–553)</td>
<td>78 (42–123)</td>
<td>184 (109–273)</td>
<td>101 (76–124)</td>
<td></td>
</tr>
<tr>
<td>HHS HHS 2 calib</td>
<td>PVs estimate [5.5–13.6] (^{c})</td>
<td>Neutral</td>
<td>12.8 (11.1–14.3)</td>
<td>–12446.91 15.9</td>
<td>484 (396–580)</td>
<td>130 (80–190)</td>
<td>250 (158–360)</td>
<td>122 (100–144)</td>
<td></td>
</tr>
<tr>
<td>HHS HHS 2 calib</td>
<td>Mammal genome [2.0–2.4] (^{d})</td>
<td>Neutral</td>
<td>3.6 (3.1–4.1)</td>
<td>–12567.19 180.5</td>
<td>571 (488–655)</td>
<td>423 (284–578)</td>
<td>408 (267–575)</td>
<td>180 (158–201)</td>
<td></td>
</tr>
<tr>
<td>HHS HHS 2 calib 90 kya</td>
<td>PVs estimate [13.2–24.7] (^{a})</td>
<td>Full</td>
<td>22.9 (20.3–25.3)</td>
<td>–19313.69 4.7</td>
<td>476 (387–568)</td>
<td>147 (101–200)</td>
<td>246 (164–343)</td>
<td>127 (106–148)</td>
<td></td>
</tr>
<tr>
<td>HHS HHS 2 calib 90 kya</td>
<td>Human mtDNA [17.6–32.3] (^{b})</td>
<td>Full</td>
<td>27.5 (23.5–31.0)</td>
<td>–19311.33 Ref</td>
<td>459 (364–551)</td>
<td>120 (79–166)</td>
<td>216 (140–300)</td>
<td>116 (95–138)</td>
<td></td>
</tr>
<tr>
<td>HHS HHS 2 calib 90 kya</td>
<td>Mammal genome [2.0–2.4] (^{d})</td>
<td>Full</td>
<td>4.6 (4.0–5.2)</td>
<td>–19502.91 383.1</td>
<td>650 (571–727)</td>
<td>616 (518–705)</td>
<td>411 (320–498)</td>
<td>214 (189–232)</td>
<td></td>
</tr>
<tr>
<td>HHS HHS 2 calib 4.5–4.5E–11</td>
<td>PVs estimate [13.2–24.7] (^{a})</td>
<td>Full</td>
<td>33.9 (24.4–43.4)</td>
<td>–19314.23 5.8</td>
<td>446 (346–544)</td>
<td>95 (53–142)</td>
<td>186 (110–271)</td>
<td>105 (80–132)</td>
<td></td>
</tr>
<tr>
<td>HHS HHS 2 calib 500 kya</td>
<td>PVs estimate [13.2–24.7] (^{a})</td>
<td>Full</td>
<td>22.0 (13.2–31.2)</td>
<td>–19311.87 1.1</td>
<td>490 (390–586)</td>
<td>156 (78–240)</td>
<td>320 (176–473)</td>
<td>209 (113–319)</td>
<td></td>
</tr>
<tr>
<td>HHS HHS 2 calib 90 kya</td>
<td>PVs estimate [13.2–24.7] (^{a})</td>
<td>Full</td>
<td>55.0 (31.8–82.9)</td>
<td>–19311.82 1.2</td>
<td>199 (99–322)</td>
<td>63 (31–100)</td>
<td>130 (73–193)</td>
<td>84 (54–115)</td>
<td></td>
</tr>
<tr>
<td>HHS HHS 2 calib 90 kya</td>
<td>4.5–4.5E–11</td>
<td>Full</td>
<td>130.6 (71.8–206.0)</td>
<td>–19311.94 1.2</td>
<td>84 (54–115)</td>
<td>27 (12–44)</td>
<td>55 (27–88)</td>
<td>36 (17–58)</td>
<td></td>
</tr>
</tbody>
</table>

Note.—Likelihood values in bold are the best-supported models, taken as reference (Ref) for the path sampling Bayes factor (2lnBF) comparison.

\(^{a}\)Rector et al. (2007).
\(^{b}\)Fu et al. (2014).
\(^{c}\)Shah et al. (2010).
\(^{d}\)Kumar and Subramanian (2002).
Table 3. Substitution Rate and Divergence Time Estimates with Highest Probability Density (HPD) in Thousands of Years Ago (kya) Inferred for HPV16 Complete (Full) or Selection-Filtered Genome Alignment (Neutral) Either under the HHS Scenario or without Imposing Any Prior on Tree Topology, and without Using Any Time Point Calibrations, and Using Instead Only Substitution Rate Priors.

<table>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>Calibration</th>
<th>Uniform prior</th>
<th>Substitution rate, (10^{-9}), (sub/site/yr)</th>
<th>(10^{-9}), (sub/site/yr)</th>
<th>(\times 10^{-9}) (sub/site/yr)</th>
<th>Inferred parameters</th>
<th>Mean (kya) (95% HPD)</th>
<th>Mean (kya) (95% HPD)</th>
<th>Mean (kya) (95% HPD)</th>
<th>Mean (kya) (95% HPD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>No topology</td>
<td>No calib</td>
<td>PVs estimate</td>
<td>18.0 (12.5–23.8)</td>
<td>581 (298–911)</td>
<td>188 (109–278)</td>
<td>391 (221–589)</td>
<td>261 (146–392)</td>
<td>261 (146–392)</td>
<td>261 (146–392)</td>
<td>261 (146–392)</td>
</tr>
<tr>
<td>Neutral</td>
<td>HHS</td>
<td>No calib</td>
<td></td>
<td>18.2 (12.4–24.2)</td>
<td>361 (178–592)</td>
<td>92 (50–147)</td>
<td>228 (123–353)</td>
<td>145 (81–222)</td>
<td>145 (81–222)</td>
<td>145 (81–222)</td>
<td>145 (81–222)</td>
</tr>
<tr>
<td>Full</td>
<td>No topology</td>
<td>No calib</td>
<td>Human mtDNA</td>
<td>23.6 (16.4–31.1)</td>
<td>440 (233–693)</td>
<td>142 (85–214)</td>
<td>296 (167–442)</td>
<td>198 (112–296)</td>
<td>198 (112–296)</td>
<td>198 (112–296)</td>
<td>198 (112–296)</td>
</tr>
<tr>
<td>Neutral</td>
<td>No topology</td>
<td>No calib</td>
<td>[17.6–32.3]</td>
<td>24.0 (16.4–31.7)</td>
<td>258 (124–420)</td>
<td>68 (37–107)</td>
<td>175 (94–269)</td>
<td>118 (61–186)</td>
<td>118 (61–186)</td>
<td>118 (61–186)</td>
<td>118 (61–186)</td>
</tr>
</tbody>
</table>

Table 4. Summary Statistics of the HPV16 E6 Gene Sequences (n = 1,123) with Geographic Origin of the Sample.

<table>
<thead>
<tr>
<th>Metapopulation</th>
<th>n</th>
<th>bp</th>
<th>S</th>
<th>H</th>
<th>Segregating sites (S)</th>
<th>MPD</th>
<th>Tajima's D</th>
<th>P-value</th>
<th>WIMP</th>
<th>Fu's-Li's D*</th>
<th>P-value</th>
<th>R2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Saharan</td>
<td>213</td>
<td>456</td>
<td>34</td>
<td>32</td>
<td>0.754 (0.672–0.837)</td>
<td>3.049</td>
<td>1.459</td>
<td>&lt;0.1</td>
<td>0.047</td>
<td>&lt;0.1</td>
<td>&lt;0.05</td>
<td>0.004</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Africa</td>
<td>221</td>
<td>456</td>
<td>30</td>
<td>26</td>
<td>0.861 (0.839–0.883)</td>
<td>4.985</td>
<td>0.969</td>
<td>&lt;0.1</td>
<td>0.019</td>
<td>&lt;0.1</td>
<td>&lt;0.05</td>
<td>0.019</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>North Africa</td>
<td>161</td>
<td>456</td>
<td>27</td>
<td>25</td>
<td>0.734 (0.638–0.808)</td>
<td>1.62</td>
<td>1.187</td>
<td>&lt;0.05</td>
<td>0.001</td>
<td>&lt;0.1</td>
<td>&lt;0.05</td>
<td>0.001</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Europe</td>
<td>139</td>
<td>458</td>
<td>26</td>
<td>23</td>
<td>0.523 (0.449–0.623)</td>
<td>0.998</td>
<td>0.464</td>
<td>&gt;0.1</td>
<td>0.253</td>
<td>&lt;0.1</td>
<td>&lt;0.05</td>
<td>0.026</td>
<td>&lt;0.02</td>
</tr>
<tr>
<td>South Asia</td>
<td>207</td>
<td>456</td>
<td>30</td>
<td>27</td>
<td>0.817 (0.504–0.799)</td>
<td>3.425</td>
<td>1.712</td>
<td>&gt;0.1</td>
<td>0.112</td>
<td>&lt;0.01</td>
<td>&lt;0.05</td>
<td>0.017</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>East Asia</td>
<td>213</td>
<td>458</td>
<td>30</td>
<td>27</td>
<td>0.803 (0.779–0.825)</td>
<td>4.551</td>
<td>1.803</td>
<td>&lt;0.1</td>
<td>0.627</td>
<td>&lt;0.05</td>
<td>&lt;0.05</td>
<td>0.03</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Central America</td>
<td>132</td>
<td>402</td>
<td>17</td>
<td>18</td>
<td>0.589 (0.507–0.671)</td>
<td>4.451</td>
<td>1.673</td>
<td>&lt;0.1</td>
<td>0.272</td>
<td>&lt;0.05</td>
<td>&lt;0.05</td>
<td>0.019</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>South America</td>
<td>207</td>
<td>456</td>
<td>19</td>
<td>18</td>
<td>0.803 (0.817–0.901)</td>
<td>5.671</td>
<td>1.703</td>
<td>&lt;0.1</td>
<td>0.216</td>
<td>&lt;0.05</td>
<td>&lt;0.05</td>
<td>0.016</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

Note.—Number of samples (n), base pair (bp), Segregating sites (S), haplotypes (h), haplotype diversity (Hd), mean pairwise difference (MPD), weighted intralineage mean pairwise difference (WIMP) (Hurles et al. 2002), Tajima’s D, Fu’s and Li’s D* and R2 statistics were calculated using DnaSP (Librado and Rozas 2009).

aCalculated from empirical distribution.

bCalculated from a coalescent simulation with 1,000 replicates and no recombination.

cCalculated assuming beta distribution.
analyses including positions under selection still preferred the HHS model but could not differentiate the two alternative scenarios, namely, the ROOA and the HHS model, tested in this study (table 2). More importantly, our MCMC inference showed that irrespective of the evolutionary rate prior used, divergence times for extant HPV16 lineages largely predate the estimated recent out-of-Africa migration of modern human ancestors 60–120 kya. Furthermore, although HPV16A is the oldest lineage, the tmrca for the HPV16BCD variants (197 kya, 95% HPD 121–291 kya) was older than the tmrca for the HPV16A lineage (88 kya, 95% HPD 50–134 kya), and the HPV16A lineage encompasses less genetic diversity than the sister HPV16BCD lineages. Taken together, the lower diversity within HPV16A, the good match for the tmrca of this lineage with the out-of-Africa migration of modern humans and the delayed expansion of HPV16A reinforce the HHS scenario: the exceptional interbreeding events between archaic and modern human populations resulted in a strong bottleneck for the transmission of HPV16A, so that only a small fraction of the HPV16A diversity available in the Neandertal/Denisovan populations underwent effectively a horizontal transfer toward populations of ancestral modern humans.

Mounting evidence suggests that virus-host codivergence is not the only essential driving force for PV evolution. At higher taxonomic levels virus-host co-phylogeny events explain only around 30% of the PV evolutionary history, while additional events of lineage sorting, lineage duplication and host switch have played major roles during the evolution of PVs (Gottschling et al. 2011). In the present study, we have quantified for the first time that codivergence between modern humans and HPV16 explains <30% of the current geographical distribution of the viral extant diversity, largely based on the virtual absence of HPV16B and HPV16C lineages outside Africa and on the enrichment of HPV16A and HPV16D outside Africa (fig. 2). The contribution of virus-host codivergence to HPV16 evolution, regardless of excluding the viral coding region sites under selection, showed to be consistent only for mitochondrial genome variability and less consistent using European ascertained autosomal single nucleotide polymorphisms (SNPs) compared with sub-Saharan African ascertained autosomal SNPs. It is tempting to speculate that this difference in consistency might be partly due to the SNPs introgressed from archaic populations into the genomes of non-African modern humans. However, the underlying biological complexities and limitations of the available data to compare the human genome and of HPV16 genetic diversity restrained us from further estimations of local adaptation between virus and host genomes. Nevertheless, under the ROOA scenario, one would have expected a consistently high correlation between human and viral phylogeography. Instead, several lines of evidence argue further against this ROOA scenario for HPV16. First, the HHS scenario is preferred over ROOA in terms of HPV16 lineage phylogeny (supplementary fig. S6, Supplementary Material online). Second, the hypothesis of exclusive codivergence of all HPV16 lineages with the ancestral dispersals of modern humans does not explain the largely predominant role of HPV16A (fig. 2), the most basal HPV16 lineage, in all continents and in all indigenous populations, except in sub-Saharan Africa (Picconi et al., 2003; Cornet et al. 2012; Mendoza et al. 2013; Qmichou et al. 2013; Tan et al. 2013; Lopera et al. 2014). Moreover, for indigenous populations in South America, for instance, the increased presence of HPV16A lineage variants has been proposed to reflect the influence of recent European occupation (Picconi et al. 2003). However, such a rapid selective sweep of the putative pre-Columbian HPV16A genetic diversity would require strong selection forces for the viral dynamics in very short time scale, which are not compatible with our current understanding of PV evolution (Bravo and Félez-Sánchez 2015). Third, the geographical dichotomy for the HPV16 E6 gene haplotype backgrounds consistently supported our HHS model over ROOA: most common HPV16A lineage haplotypes were observed mostly in Eurasian populations while most common HPV16B, C, and D lineage haplotypes were observed mainly in African populations (table 5). And fourth, while human genetic diversity is highest in sub-Saharan Africa (1000 Genomes Project Consortium et al. 2015), our results of HPV16 E6 variability showed consistently higher genetic diversity estimates outside sub-Saharan Africa (table 4).

After rejecting the ROOA model as explanatory framework for HPV16 evolution, we have explored the HHS model as alternative scenario, implying a host switch of an ancestral HPV16 lineage between archaic and modern human ancestral populations (fig. 4). Indeed, recent genomic studies have confirmed the admixture of modern human ancestors with Neandertals and Denisovans through repeated interbreeding after migration to Europe and Asia, respectively (Reich et al. 2010; Lazaridis et al. 2014; Qin and Stoneking 2015; Vernot and Akey 2015). Such interbreeding events lead to...
adaptive introgression of genetic material from our closest evolutionary relatives into the genome of a subset of modern human ancestors. We propose here that in parallel to the introgression of genetic material, sexual intercourse between Neanderthals/Denisovans and modern human ancestors also prompted sexual transmission of PVs, chiefly of the HPV16A lineage to the ancestors of modern human populations in Eurasia. Viral lineages HPV16A, B, C, and D are labeled at the bottom.

Fig. 4. Cartoon timeline depicting interbreeding and subsequent gene flow of archaic alleles from Neanderthals and Denisovans into modern humans and the proposed sexual transmission of HPV16A lineage to the ancestors of modern human populations in Eurasia. Viral lineages HPV16A, B, C, and D are labeled at the bottom.

The virtual absence of HPV16B outside Sub-Saharan Africa is parsimoniously explained if in the last out-of-Africa expansion, the modern human ancestors that left Africa probably lost the ancestral HPV16B lineage by a lineage sorting event (Johnson et al. 2003). Similar pattern of virus extinction after host population bottleneck has been observed in non-human primates (Kapusinszky et al. 2015). After the modern human dispersal, the HPV16B/CD ancestor generated in allopatry the HPV16C lineage in the populations remaining in Africa, and the HPV16D lineage in the populations outside Africa (fig. 4). During their expansion in Europe and in Asia, modern human ancestors experienced limited admixture with Neanderthal and Denisovan populations and were exposed to the HPV16A lineage, most likely through sexual contact. After the adaptive introgression of Neanderthal-Denisovan alleles, anatomically modern humans of non-African ancestry carry nowadays in average between 2% and 4% of their genomes of archaic human origin (Meyer et al. 2012; Prüfer et al. 2014).Remarkably, the horizontal transfer of genetic loci upon interbreeding of Neanderthal/Denisovan ancestry into the genome of modern human ancestors did not occur at random (Vernot and Akkey 2014; Kühlwilm et al. 2016; Vernot et al. 2016). Instead, genes involved in keratinocyte differentiation and innate immunity, which are loci directly involved in host–PV interaction, are particularly enriched in introgressed genomic loci, displaying >60% of Neanderthal/Denisovan ancestry in present day Eurasians (table 6) (Abi-Rached et al. 2011; Sankaraman et al. 2014; Vernot and Akkey 2014; Vernot et al. 2016). Such genetic changes possibly altered the HPV16-human interplay, as this virus exclusively infects keratinocytes in particular epithelia and requires keratinocyte differentiation to accomplish the virus life cycle and virion production (Doorbar et al. 2012). Hence, we speculate that the adaptive introgression in modern humans of archaic alleles involved in keratinocyte differentiation and innate immunity may have elicited changes in the ecological niche of HPV16 that significantly enhanced the adaptive value of HPV16A and lead to increased prevalence of this viral lineage in modern human populations with Eurasian origin (figs. 2, 3B and C, and 4).

Taken together, our HHS evolutionary model provides explanatory potential for the three central flaws of the alternative ROOA scenario for the evolution of HPV16, namely, 1) the distant phylogenetic relationship between HPV16A and HPV16BCD clades (fig. 1A and B), which predates the most recent out-of-Africa migration of modern human ancestors (table 3); 2) the global increased prevalence of HPV16A outside Africa (fig. 2), and 3) the increased viral genetic diversity outside sub-Saharan Africa (table 4). In agreement with our HHS model, recent studies of parasites evolution have concluded that the divergence of at least five present day human parasites (lice, tapeworm, follicle mites, a protozoan, and bedbug) predates modern human origin. The most likely explanation for the presence of these closely related ancient pairs of parasite taxa involves host switch from an archaic to modern human ancestors (Ashford 2000; Reed et al. 2004).
Furthermore, evolutionary studies of sexually transmitted herpes simplex-virus suggest ancient codivergence and later cross-species transmission of this pathogen between ancestors of chimpanzees and extinct Hominin groups (Wertheim et al. 2014).

Despite the explanatory power of our results, our study suffers from a number of limitations. Ideally, the global host and pathogen sequence data should be analyzed at the level of full genomes from the same individuals, but such data set is not available. Furthermore, there is hitherto no evidence of the presence of any PV sequences from ancient human samples. Indeed, we analyzed the currently available Neanderthal and Denisovan preassembly sequence data, and we could not find any significant traces of any known HPVs in these data sets. Unfortunately, the epithelial tropism of these viruses likely prevents retrieving viral sequences from fossil bones, the common reservoir for ancient DNA studies. We have also tried to identify sequence traces of PV DNA from the metagenomic sequence data extracted from tissue remains of the European Copper Age glacier mummy (Maixner et al. 2014) but could not find any traces of PV DNA in any of these sample data sets. However, future studies encompassing larger global sample of individuals with full human and HPV genome sequence data and/or well-preserved ancestral human samples containing PVs DNA will allow to test and refute or validate our hypothesis.

We have presented here the most comprehensive study of the evolution and diversity of HPV16, the most oncogenic infectious agent for humans. Our phylogeographic analyses confirm the limited contribution of virus-host codivergence to the evolution of the HPV16 lineage; most of the geographic diversity within HPV16 cannot be explained alone by co-speciation with modern humans. Instead, our results suggest that ancestral HPV16 already infected the ancestor of H. sapiens and H. neanderthalensis half a million years ago, and that two main HPV16 lineages co-diverged with either human lineage. When a population of modern humans migrated out of Africa some 60–120 kya, they carried with them a subset of the HPV16 diversity evolved in that continent. When modern humans encountered and interbred with Neanderthal/Denisovan populations in Europe and in Asia, a transfer of sexually transmitted pathogens occurred, in parallel with the genomic introgression. Gene alleles involved in keratinocyte differentiation and in innate immunity evolved among Neanderthals/Denisovans and transferred to modern humans, may have facilitated niche colonization and expansion of HPV16A in modern human populations with archaic admixture. This evolutionary advantage led to the enhanced prevalence of HPV16A lineage among modern humans with introgressed archaic genetic background. Indeed, our results show that sexually transmitted pathogens may have been effectively transferred between different Hominin populations in recent events. We propose that the repertoire of the HPV16 variants coevolved with archaic humans, the diversity of keratinocyte differentiation and innate immune genes that have undergone adaptive introgression, and eventually the interaction between viral and host genotypes may be largely responsible for the differential association of certain HPV16 variants with cancer risk in certain modern human populations.

### Materials and Methods

**HPV16 Complete Genomes and Partial Sequences**

The full-genome sequences for 118 HPV16 isolates used for the analysis were retrieved from GenBank. Sequences were aligned with MUSCLE (Edgar 2004), at amino acid level for the coding region and at nucleotide level for the UTR, and tested for putative recombination break points using GARD algorithm (Kosakovsky Pond et al. 2006). The final full-genome alignment encompassed 118 full-length HPV16 genomes, 7,926 bp and 638 distinct alignment patterns (alignment available as supplementary material S1, Supplementary Material online). Best substitution model for the HPV16 full-genome data, manually curated and filtered using Gblocks (Castresana 2000), was inferred using jModelTest 2.0 (Darriba et al. 2012). Phylogenetic inference was performed under ML framework with RAxML v8.0.16 using the general time reversible (GTR) model of nucleotide substitution with $\Gamma$-rate heterogeneity parameter ($\Gamma + \Gamma_i$) and 500 bootstrap replicates (Stamatakis 2014). Besides the unpartitioned analysis, four partition schemes were also explored: 1) partitions of noncoding and coding regions; 2) the same as before, but further partitioning the coding region into three codon

### Table 6. Global Prevalence of HPV16A Lineages (Upper Panel) and Proportion of Nonrandomly Introgressed Archaic Alleles into Modern Human Ancestors (Lower Panel).

<table>
<thead>
<tr>
<th>Human papillomavirus 16A lineages</th>
<th>Sub-Saharan Africa</th>
<th>West Eurasia</th>
<th>East Eurasia</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPV16A1-3</td>
<td>&lt;5%</td>
<td>93%</td>
<td>25-91%</td>
<td>This study</td>
</tr>
<tr>
<td>HPV16A4</td>
<td>Absent</td>
<td>~1%</td>
<td>60%</td>
<td>This study</td>
</tr>
</tbody>
</table>

**Host’s genome loci directly involved in PVs life-cycle**

- Neanderthal ancestry loci:
  - Absent
  - 64% 62% (Sankararaman et al. 2014)

- Loci involved in keratinocyte differentiation:
  - Absent 40–70% 40–66% (Vernot and Akey 2014)

- HLA I loci (innate immunity):
  - <7% 52–59% 72–82% (Abi-Rached et al. 2011)

- Toll-like receptor loci (innate immunity):
  - Absent 15–39% 17–51% (Dannemann et al. 2015)

- AP0BEC3A deletion (innate immunity):
  - <1% 7% 14-93% (Kidd et al. 2007)

**NOTE.—** A selection of host’s genome loci is listed, directly involved in the Papillomavirus life-cycle through keratinocyte differentiation and innate immunity.

The full-genome data, manually curated and filtered using Gblocks (Castresana 2000), was inferred using jModelTest 2.0 (Darriba et al. 2012). Phylogenetic inference was performed under ML framework with RAxML v8.0.16 using the general time reversible (GTR) model of nucleotide substitution with $\Gamma$-rate heterogeneity parameter ($\Gamma + \Gamma_i$) and 500 bootstrap replicates (Stamatakis 2014). Besides the unpartitioned analysis, four partition schemes were also explored: 1) partitions of noncoding and coding regions; 2) the same as before, but further partitioning the coding region into three codon...
of 1,338 bp partial sequences spanning the LCR and E6 loci and 519 isolates of 665 bp partial sequences spanning the LCR, E6 and L2 loci, see available alignments in supplementary materials S3–S4, Supplementary Material online). For each partial sequence, we integrated the likelihood weights for all nodes using 0.7 as a likelihood cutoff value to confidently assign each sample into one of the five main HPV16 variant lineages (supplementary table S2, Supplementary Material online). Finally, we stratified the HPV16 variant isolate data into 10 geographical population groups: North/East/West Southern Africa, South Asia, East Asia, Europe, Central and South America using the United Nations standard geographical criteria (supplementary table S3, Supplementary Material online). To compare the genetic variability patterns observed for HPV16 isolates within each geographic population group we calculated the summary statistics of the available 1,123 HPV16 E6 sequences (table 4). Further to visualize the genetic variability of the 1,123 HPV16 E6 456bp nearly complete gene sequences within each geographic background a haplotype-specific median joining network was constructed using NETWORK 4.6.1.3 (www.fluxus-technology.com) with equal weight for variable positions.

Bayesian MCMC Inference

Evolutionary analysis for HPV16 lineages was estimated using both the complete 7,926 bp genome alignment and the selection-filtered 6,093 bp coding region alignment of 118 HPV16 variants (supplementary materials S1–S2, Supplementary Material online), and using Bayesian MCMC analysis as implemented in BEAST v.1.8.1 (Drummond et al. 2012). Bayesian MCMC inferences were performed with the GTR + Γ4 model. Genealogies were estimated using both strict and relaxed molecular clocks with an uncorrelated log-normal distribution of rates. Two credible PVs substitution rate estimates from the literature were used as uniform priors: 9.6 × 10^{-9} [5.5 × 10^{-9}–13.6 × 10^{-9}] subs/site/yr (Shah et al. 2010) and 19.5 × 10^{-9} [13.2 × 10^{-9}–24.7 × 10^{-9}] subs/site/yr (Rector et al. 2007). We also assessed the suitability of two additional plausible substitution rates, namely, the values inferred for mammalian genomes 2.2 × 10^{-9} [2.0 × 10^{-9}–2.4 × 10^{-9}] subs/site/yr (Kumar and Subramanian 2002) and for human mitochondrial genomes 25.3 × 10^{-9} [17.6 × 10^{-9}–32.3 × 10^{-9}] subs/site/yr (Fu et al. 2014). In addition, a noninformative uniform prior between 4.5 × 10^{-11} and 4.5 subs/site/yr covering all plausible PVs rate for both molecular clock models was used (table 2). Two recently revised evolutionary time point estimates for human populations were used as prior normal distribution for calibration of the HPV16 variant phylogenetic tree (table 2 and fig. 3C): first, the archaic divergence of modern humans and Neanderthals/Denisovans was set at 500 kya (95% CI, 400–600 kya) (Reich et al. 2010; Scally and Durbin 2012; Buck and Stringer 2014; Mendez et al. 2016); second, modern humans migration out-of-Africa was set at 90 kya (95% CI, 60–120 kya) (Scally and Durbin 2012; Liu et al. 2015). All evolutionary related parameters were estimated with the following demographic models: 1) constant population size, 2) exponential growth, and 3) Bayesian skyline coalescent models (Pybus and Stringer 2014; Mendez et al. 2016).
Two independent runs were performed for each parameter combination with 100 million generations and subsampling every 10,000 generations. The resulting tree and log-files were combined for further analysis after removing a 10% burn-in from each run. Statistical confidence in parameter estimates was assessed by reporting marginal posterior means and their associated 95% HPD intervals (95% HPD). Efficient mixing of the chains in the Bayesian MCMC analysis was assessed with effective sample size values $>200$ for all parameters. Convergence of each of the simulations was visually confirmed before data were merged for further analysis. The combined Bayesian phylogenetic inferences were analyzed using Tracer v.1.6 and TreeAnnotator v.1.8.1 and visualized using Figtree v.1.4.2. Best model estimates to explain the HPV16 variant diversity were selected using two independent path sampling runs with at least 1 million iterations per path step and until no significant chances were observed in log marginal likelihoods for Bayes Factor (2lnBF) estimates (Baele et al. 2012, 2013) as implemented in BEAST v.1.8.1 (Drummond et al. 2012). The support of a particular model was assessed with $2\lnBF < 2$ indicating no support, $2\lnBF = 2 - 6$ indicating positive model support, $2\lnBF = 6 - 10$ indicating strong positive support and $2\lnBF > 10$ indicating definite distinction between competing models.

To determine the best topology for the HPV16 viral lineages, the maximum clade credibility trees were also inferred for both complete and selection-filtered coding region HPV16 genome alignments, without imposing any prior on tree topology, using BEAST v.1.8.1 (Drummond et al. 2012), and further assessing the distance between the obtained trees using K tree score (Soria-Carrasco et al. 2007). Wilcoxon–Mann–Whitney $U$ test was used to determine the significance of intertaxa pairwise differences between the complete and selection filtered coding region HPV16 ML trees. Summary statistics, including Tajima’s $D$, Fu’s and Li’s $D^*$ and $R^2$ were calculated using DnaSP v.5.10.1. (Librado and Rozas 2009).

Human Genome Data

To compare the phylogeography of HPV16 with the global variability of the human genome, we compiled human mitochondrial, Y chromosome, and autosomal genotype data of the same 938 nonrelated individuals of the CEPH Human Genome Diversity Panel representing 51 human populations worldwide (Cann et al. 2002). In order to systematically compare the host and pathogen genome diversity distribution, we stratified the human population genetic data in nine subcontinental metapopulation groups (supplementary table S3, Supplementary Material online) using the same geographical classification presented for HPV16 isolate data in supplement table S6 (Supplementary Material online). This geographical grouping of human population genetic data reflected, as far as possible, the metapopulation groups from which the viral isolates were sampled. All mitochondrial complete genome sequences were manually aligned using MUSCLE (Edgar 2004) and 875 complete mitochondrial DNA genomes available from (Lippold et al. 2014) were used in this study. Y chromosome nonrecombining region sequences from 551 males were retrieved (Lippold et al. 2014) and all variable loci data were included in this study. Autosomal single nucleotide polymorphism (SNP) loci data ascertained using either a European or sub-Saharan African individual as previously reported (Reich et al. 2010) were used in this study. All autosomal SNP markers were previously genotyped in 934 CEPH diversity panel individuals (Green et al. 2010; Reich et al. 2010). The African and European ascertained data sets consisted of 12,162 and 111,970 SNPs, respectively, covering all autosomal chromosomes of the human genome. All autosomal SNP genotype quality control and data mining were performed using PLINK software (Purcell et al. 2007). Only SNPs with MAF $> 0.01$ were used in further analysis. Phasing for each genotype data sets per chromosome were performed using BEAGLE software (Browning and Browning 2007). Genetic $F_{ST}$ distances (WRIGHT 1951) between the nine geographical population groups defined above, and excluding North America, were estimated for 1,101 available HPV16 isolate sequences encompassing LCR (735bp) and E6 (457 bp) genome regions (supplementary fig. S3, Supplementary Material online) as well as for each of the human genome marker data sets (mitochondrial, Y chromosome, and autosomal SNPs). Mantel’s permutation test for matrix similarity was implemented for pairwise combinations of $F_{ST}$ distance matrices with 10,000 permutations. To search for possible traces of HPVs, including all known HPV16 variants, in archaic humans we retrieved all available Neanderthal and Denisovan pre-assembly sequence data from http://cnda.evampgd.de/neander
tal/altai/AltaiNeandertal and http://www.ebi.ac.uk/ena/data/
view/PRJEB3092, respectively. Reads were mapped against all known HPVs genomes using bowtie2 (version: 2.2.5) with –very-sensitive command for the alignment. Further details of data processing are available from the authors upon request.

Supplementary Material

Supplementary figures S1–S7, tables S1–S6 and materials S1–S4 are available at Molecular Biology and Evolution online.

Author Contributions

Designed and conceived the project: VNP, IGB. Retrieved data and performed the phylogenetic inference: VNP, CMO. Performed the modern and archaic human genome data analyses and Bayesian MCMC inference: VNP. Wrote the manuscript: VNP, CMO. All authors read and approved the final article.

Acknowledgments

This research was financially supported by The European Society of Clinical Microbiology and Infectious Diseases (ESCMID Research Grant 2015 VNP), Aaltonen Foundation (VNP), Otto A. Malmi Foundation (VNP), The S Paulo Research Foundation (FAPESP grant #2011/24035-2 and #2012/23290-1, CMO) and the Europen Research Council (Contract #647916, IGB). No agency had any role in the interpretation of the results or in the preparation of the article. Alexander Herbig is warmly thanked for the help in the analysis of the metagenomic data of the European Copper Age
glacier mummy. Xavier Bosch, Samuel Alizon and Yannis Michalakis are thanked for helpful comments and review of an initial version of the article.

References


Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575.


