Genome plasticity in Papillomaviruses and *de novo* emergence of *E5* oncogenes

Marta Félez-Sánchez[1], Anouk Willemsen[2,*], and Ignacio G. Bravo[2]

[1]Infections and Cancer Laboratory, Catalan Institute of Oncology (ICO), Barcelona, Spain
[2]Laboratory MIVEGEC (UMR CNRS IRD UM), Centre National de la Recherche Scientifique (CNRS), Montpellier, France
[*]Corresponding author: anouk.willemsen@ird.fr

ABSTRACT

The clinical presentations of papillomavirus (PV) infections come in many different flavors. While most PVs are part of a healthy skin microbiota and are not associated to physical lesions, other PVs cause benign lesions, and only a handful of PVs are associated to malignant transformations linked to the specific activities of the *E5*, *E6* and *E7* oncogenes. The functions and origin of *E5* remain to be elucidated. The E5 ORFs are present in the genomes of a few polyphyletic PV lineages, located between the early and the late viral gene cassettes. We have computationally assessed whether these E5 ORFs have a common origin and whether they display the properties of a genuine gene. Our results suggest that during the evolution of *Papillomaviridae*, at least five independent events resulted in the insertion of a non-coding DNA stretch between the *E2* and the *L2* genes. In three of these events, the novel regions evolved independently coding capacity, becoming the extant non-orthologous E5 ORFs. We then focused on the evolution of the *E5* genes in *AlphaPVs* infecting humans. Interestingly, while the nucleotide sequences in the intergenic E2–L2 region in *AlphaPVs* have a common ancestor, the four types of E5 that evolved within this region do not. The sharp match between the type of E5 protein encoded and the infection phenotype (cutaneous warts, genital warts or anogenital cancers) supports the role of E5 in the differential oncogenic potential of these PVs. Our evolutionary interpretation is that an originally non-coding region entered the genome of the ancestral *AlphaPVs*. This genetic novelty allowed to explore novel transcription potential, triggering an adaptive radiation that yielded three main viral lineages encoding for different E5 proteins, and displaying distinct infection phenotypes. Overall, our results provide an evolutionary scenario for the *de novo* emergence of viral genes and illustrate the impact of such genotypic novelty in the viral phenotypic diversity.

Keywords: oncogenes, virus evolution, papillomavirus, genome evolution

Introduction

Papillomaviruses (PVs) constitute a numerous family of small, non-encapsulated viruses infecting virtually all mammals, and possibly amniotes and fishes. According to the International Committee on Taxonomy of Viruses (ICTV: `https://talk.ictvonline.org/taxonomy/`), the *Papillomaviridae* family currently consists of 53 genera, which can be organized into a few crown groups according to their phylogenetic relationships[1] The PV genome consists of a double stranded circular DNA genome, roughly organized into three parts: an early region coding for six open reading frames (ORFs: *E1*, *E2*, *E4*, *E5*, *E6* and *E7*) involved in multiple functions including viral replication and cell transformation; a late region coding for structural proteins (L1 and L2); and a non-coding regulatory region (URR) that contains the *cis*-elements necessary for replication and transcription of the viral genome. The major oncoproteins encoded by PVs are E6 and E7, which have been extensively studied[2–4]. However, there is also a minor oncoprotein termed E5, whose functions and origin remain to be fully elucidated[5].

The E5 ORFs are located in the intergenic E2–L2 region. The inter-E2–L2 region is variable among PV genomes. In most PV lineages the early and late gene cassettes are located in direct apposition. In a few, non-monophyletic PV lineages, this region accommodates both coding and non-coding genomic segments, which may have gained access to the PV genomes through recombination events with hitherto non-identified donors[6]. PVs within the *Alpha*-, *Delta*- and *TauPVs* genera encode different E5 proteins in the inter-E2–L2 region[7]. Additionally members of the Lambda-MuPV and Beta-XiPV crown groups present in the inter-E2-L2 region large non-coding stretches of unknown significance[8].

The largest wealth of scientific literature about PVs deals with *AlphaPVs*. These are a clinically important group of PVs that infect primates, and are associated to largely different clinical manifestations: non-oncogenic PVs causing anogenital warts, oncogenic and non-oncogenic PVs causing mucosal lesions, and non-oncogenic PVs causing cutaneous warts. The E5 proteins in *AlphaPVs* can be classified into four different groups according to their hydrophobic profiles and phylogeny[7]. The presence of a given E5 type sharply correlates with the clinical presentation of the corresponding PV infection: viruses that contain E5α (*e.g.* HPV16) are associated with malignant mucosal lesions such as cervical cancer; viruses coding for E5β (*e.g.* HPV2) are associated with benign cutaneous lesions, commonly warts on fingers and face; and viruses that contain two putative E5 proteins, termed E5γ and E5δ (*e.g.* HPV6) are associated with benign mucosal lesions such as anogenital warts[7]. Two additional putative E5 proteins, E5ε and E5ζ (PaVE; `https://pave.niaid.nih.gov`), have been identified in *AlphaPVs* infecting *Cercopithecinae* (macaques and baboons). Contrary to the other E5 proteins, the E5ε and E5ζ are not associated with a specific clinical presentation, although our knowledge about the epidemiology of the infections in other primates is still very limited. It has been suggested that the integration of an *E5* proto-oncogene in the ancestor of *AlphaPVs* supplied the viruses with genotypic novelty, which triggered an adaptive radiation through exploration of phenotypic space, and eventually generated the extant three clades of PVs[6].

The only feature that all E5 proteins have in common is their highly hydrophobic nature and their location in the inter-E2–L2 region of the PV genome. It remains unclear whether all E5 proteins are evolutionary related. The E5 proteins of HPV16 and of BPV1 are the only E5s for which the biology is partially known. Despite the absence of sequence similarity, the cellular roles during infection are comparable. HPV16 E5 is a membrane protein that localizes in the Golgi apparatus and in the early endosomes. It has been associated to different oncogenic mechanisms related to the induction of cell replication through manipulation of the epidermal growth receptor response[9–11], as well as to immune evasion by modifying the membrane chemistry[12,13] and decreasing the presentation of viral epitopes[14]. BPV1 E5 is a very short protein that also localizes in the membranes. It displays a strong transforming activity, largely by activating the platelet-derived growth factor receptor[15,16], and it downregulates as well the presentation of viral epitopes in the context of the MHC-I molecules[17].

In this study, we describe the evolutionary history of the *E5* ORFs found within the inter-E2–L2 region in PVs. First, we identified the PV clades that contain an intergenic region between *E2* and *L2*, and therewith putative *E5* ORFs. Then, we assessed whether the inter-E2–L2 region in the identified clades had originated from a single common ancestor. Next, we verified whether the evolutionary history of the inter-E2–L2 region and of the *E5* ORFs therein encoded is similar to that of the other PVs genes, by comparing their sequences and phylogenies. Finally, we examined whether the different *E5* ORFs exhibited the characteristics of a *bona fide* gene to exclude the conjecture that these are simply spurious translations.

Materials and Methods

DNA and Protein Sequences

The inter-E2–L2 sequences were retrieved from the Papillomavirus Episteme Database (PaVE: `https://pave.niaid.nih.gov`). We also obtained all *E5* sequences belonging to *AlphaPVs*, including 17 E5α, 28 E5β, 6 E5γ, 10 E5δ, 11 E5ε, and 11 E5ζ sequences. The corresponding URR, *E6*, *E7*, *E1*, *E2*, *L1* and *L2* sequences from these viruses were also retrieved and analyzed in parallel to the *E5* sequences. We excluded the *E4* ORFs from our analysis as most of its coding sequence overlaps the *E2* gene in a different reading frame and it is supposed to be under different evolutionary pressures[18,19].

Testing for Common Ancestry using Bali-Phy

In order to evaluate the common ancestry of the inter-E2–L2 sequences, we used the software Bali-Phy[20]. Under this maximum-likelihood framework, the input data are the unaligned sequences, as the alignment itself is one of the parameters of the model to be treated as an unknown random variable[21]. We ran our analysis under the null hypothesis of common ancestry of the intergenic regions. We used the marginal likelihood calculated as the harmonic mean of the sample likelihood to estimate the Bayes Factor between the null hypothesis *Common Ancestry* (CA) and the alternative hypothesis *Independent Origin* (IO)[22]. Therefore, we have ΔBF = log[Prob(IO)]-log[Prob(CA)], such that negative values support Common Ancestry. The likelihood for the *Common Ancestry* model was obtained running the software for all the inter-E2–L2 sequences together. For the *Independent Origin* scenarios, we ran one analysis for each group independently. We started with the different PV clades that contain an inter-E2–L2 sequence, arbitrarily named in this study as clades C1-C5 (fig. 1). Then we ran the analyses on the inter-E2–L2 region within *AlphaPVs* stratifying by the clinical presentation of each PV; mucosal lesions (MUC), cutaneous warts (CUT), and anogenital warts (GW). The values for the independent groups for MUC, CUT, and/or GW, and the sum of these, rendered the likelihood for the *Independent Origin* models. For instance, [MUC-GW]+CUT denotes a hypothesis of two independent ancestries, one tree for the inter-E2–L2 of MUC and GW PVs together, and another separate tree for the inter-E2–L2 of CUT PVs. The likelihood of this example was obtained running Bali-Phy two times: one for all inter-E2–L2 sequences of MUC and GW PVs combined; and another run with all inter-E2–L2 sequences of CUT PVs. The sum of these two analyses corresponded to the likelihood of the model. We only considered the *Independent Origin* scenarios that were biologically plausible based on the phylogeny of PVs (fig. 1). The same procedure was applied to E5 sequences (E5α, E5β, E5γ, E5δ, E5ε, and E5ζ), at both the nucleotide and amino acid level, in order to test the common ancestry of the putative coding sequences contained in the inter-E2–L2 region. For nucleotide analysis, we used GTR+Γ substitution model, whereas for the amino acid analysis we used the LG model. In the cases where two putative E5 ORFs were located in the same inter-E2–L2 fragment (for instance for E5γ and E5δ, and E5ε and E5ζ) sequences were concatenated. As the harmonic mean tends to overestimate the marginal likelihood[23], and thus favors the *Independent Origin* hypothesis, each analysis was performed three times, in order to ensure the validity of the results. Moreover, in order to test the validity of the procedure we also assessed the common ancestry for the *E6* ORF (at nucleotide and amino acid levels) and for the URR fragment. E6 is a highly divergent ORF[24], while the URR is a highly divergent and heterogeneous non coding region[25]. In both cases (*E6* and URR), the *Common Ancestry* scenario was confirmed (table S1, S2), confirming the soundness of the methodology.

Phylogenetic Analyses

For the sequences retrieved, we reconstructed a phylogenetic tree for each gene separately, as well as for the URR and the inter-E2–L2 region. Coding sequences were aligned at the amino acid level using MUSCLE[26] and back-translated into the corresponding codon-aligned nucleotide sequences. Informative positions were filtered with GBLOCKS under non-stringent conditions[27]. For the non-coding regions (URR and the inter-E2–L2 region) nucleotide sequences were aligned. Phylogenetic relationships were inferred in a Maximum Likelihood framework using RAxML v.8.2.3 (`http://www.exelixis-lab.org/`)[28]. The Robinson-Foulds (RF) distances between trees were calculated[29]. Multiple correspondence analysis (MCA) was performed to identify similarities between the topologies of the trees reconstructed for each gene. The statistical relationships between RF distances were displayed graphically.

Generation of Random ORFs

In order to assess whether the *E5* sequences were larger than expected by chance, we estimated first the median A/T/G/C composition of the inter-E2–L2 regions of *AlphaPVs* (A:0.22; T:0.41; G:0.20; C:0.17). Using in-house perl scripts, we created a set of 10,000 random DNA sequences with this median nucleotide composition and with a median length of 400 nt (which is the median length of the inter-E2–L2 region). Then, we computed the length of all putative ORFs that may have appeared in this set of randomly generated DNA sequences.

dN/dS Values

In order to assess whether the *E5* ORFs are protein-coding sequences, we computed the dN/dS values for all *E5* ORFs as well as for the other PV ORFs (*E1*, *E2*, *E6*, *E7*, *L1*, *L2*). The dN/dS values were computed with SELECTON (`http://selecton.tau.ac.il/overview.html`[30], using the MEC model[31]. The likelihood of MEC model was tested against the model M8a[32], which does not allow for positive selection. For all the sequence sets, the MEC model was preferred over the M8a model.

Pairwise Distances

In order to assess the diversity of the *AlphaPVs* genes, we calculated the pair-wise distances between aligned sequences within each group of the *E5* ORFs, the other PV ORFs (*E1*, *E2*, *E6*, *E7*, *L1*, *L2*), and the URR. These random intergenic CDS were generated by extracting the non-coding region of the E2–L2 fragments of all *AlphaPVs*. Then, for each non-coding region, we extracted a random subregion with the same length as the *E5* ORF of this PV. These random intergenic regions were truncated at the 5' to get a sequence length multiple of 3. All internal stop codons were replaced by N's. Pair-wise distances between aligned DNA sequences were calculated with the package *ape* in R (`https://www.r-project.org/`)[33] using the TN93 model. All distances were normalized with respect to the corresponding one obtained for *L1*.

Codon Usage Preferences

We calculated the codon usage preferences (CUPrefs) for the *E5 AlphaPVs* ORFs. The frequencies for the 59 codons with redundancy (i.e. excluding Met, Trp and stop codons) was retrieved using an in-house perl script. For each of the 18 families of synonymous codons, we calculated the relative frequencies of each codon. We performed the same analysis for all other ORFs in the same genomes (*E1*, *E2*, *E6*, *E7*, *L1* and *L2*) as well as to the randomly generated intergenic CDS. A matrix was created in which the rows corresponded to the ORFs on one PV genome and the columns to the 59 relative frequency values, such that each row had the codon usage information for a specific ORF. We performed a non-metric Multidimensional Scaling (MDS) analysis with Z-transformation of the variables in order to assess similarities in codon usage preferences of the *E5* ORFs with respect to the other *AlphaPVs* ORFs, as described in[19]. In parallel, we performed a two-step cluster analysis with the same relative frequency values. The optimal number of clusters was automatically determined using the Bayesian Information Criterion (BIC).

GRAVY Index

For all E5 proteins the grand average hydropathy (GRAVY) was calculated by adding the hydropathy value for each residue and dividing this value was by the length of the protein sequence[34].

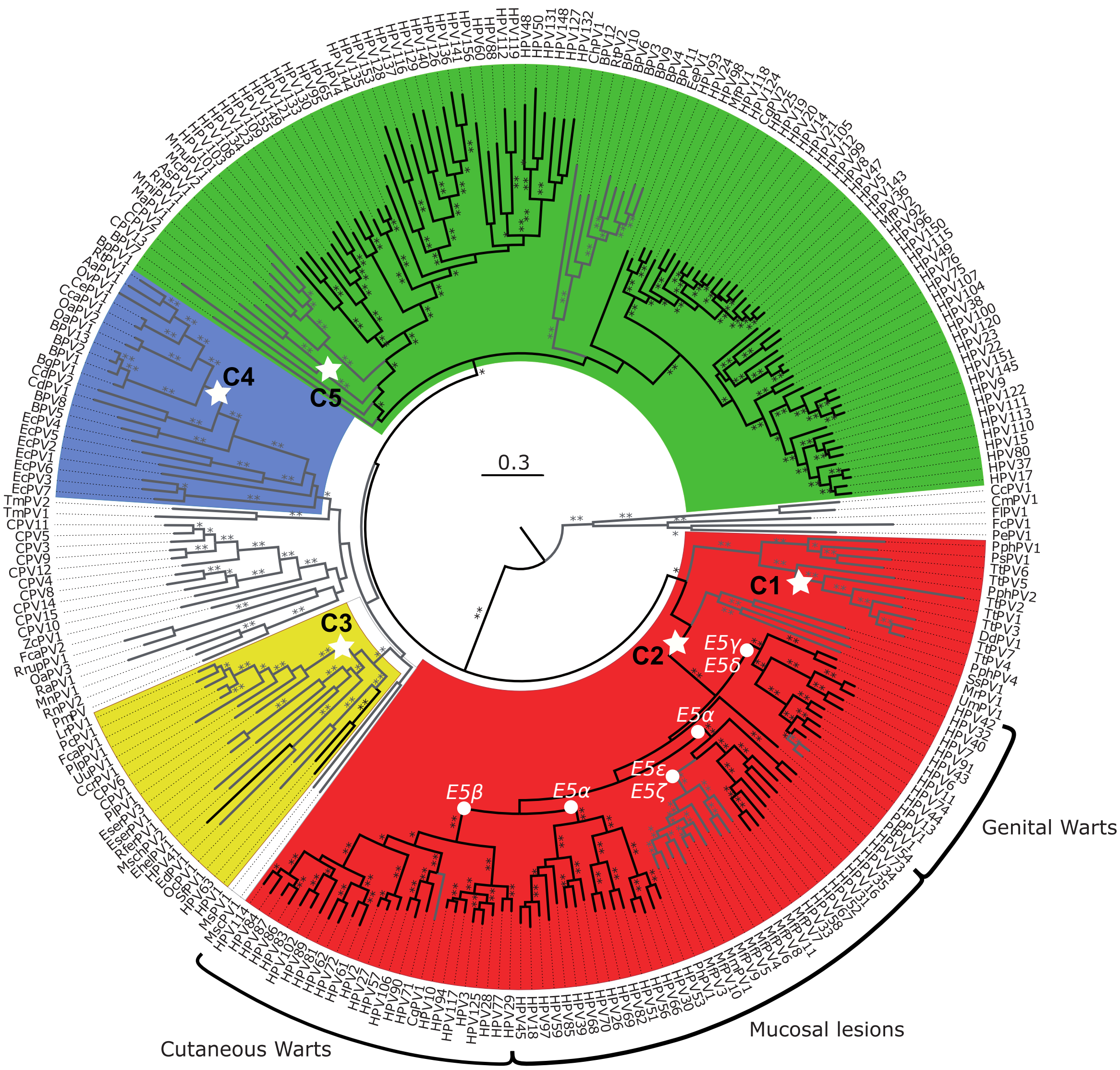

Figure 1. PV phylogenetic reconstruction and identification of clades with an intergenic E2–L2 region. Best-known maximum likelihood nucleotide phylogenetic tree of the concatenated *E1E2L2L1* gene sequences of 263 PVs, modified from[6]. Color code highlights the four PVs crown groups: red, Alpha-Omikron-PVs; green, Beta-Xi-PVs; ochre, Lambda-Mu-PVs; blue, Delta-Zeta-PVs and white, PVs without well-supported phylogenetic relationships to be assigned into a different crown group. Branches in black correspond to HPVs (human PVs) and branches in gray to animal HPVs. Outer labels indicate the most common tropism for the *AlphaPVs*. Asterisks on branches correspond to ML bootstrap support values. Two asterisks indicate maximal support values; one indicates support values between 90 and 50; values under 50 are not shown. The basal node of the five clades containing an intergenic region between the E2 and the L2 ORFs are labeled with a star. Clade C1 includes *UpsilonPVs*, C2 includes *AlphaPVs*, C3 includes *LambdaPVs*, C4 includes *DeltaPVs*, and C5 includes *TauPVs*. The basal nodes of the *AlphaPVs* are labeled with a circle indicating which clade contains which E5 type

Results

The inter-E2–L2 Regions Present in Different PV Genomes Are Not Monophyletic

From all PV sequences available at the Papillomavirus Episteme Database (PaVE; https://pave.niaid.nih.gov), we retrieved 316 intergenic E2–L2 segment sequences. Based on the best-known maximum likelihood phylogenetic tree of the concatenated *E1E2L2L1* gene sequences of the available full-length PV genomes[6], we identified five PV clades containing a intergenic region between *E2* and *L2*, labeled respectively C1 to C5 in (fig. 1). These clades are located in the four PV crown groups: Alpha-Omikron (clade C1, including *UpsilonPVs*; clade C2, including *AlphaPVs*), Lambda-Mu (clade C3, including *LambdaPVs*), Delta-Zeta (clade C4, including *DeltaPVs*), and Beta-Xi (clade C5, including *TauPVs*).

In order to determine whether the genome fragments comprised between the *E2* and *L2* genes of the different PV clades (C1-C5) share a single common ancestor, we tested for common ancestry using Bali-Phy (as described in de Oliveira Martins and Posada 2014[22]). We made the choice between the alternative hypotheses *Common Ancestry* (CO) and *Independent Origin* (IO) by estimating marginal likelihoods calculated as the harmonic mean of the sample likelihoods. We ran our analysis under the null hypothesis of Common Ancestry of the fragment. Therefore, we have ΔBF = log[Prob(IO)] - log[Prob(CA)], such that negative values support *Common Ancestry*. We considered different plausible *Independent Origin* scenarios based on the phylogeny and we found that the *Independent Origin* hypothesis is the best-supported scenario (table 1). Thus, our results suggest that the inter-E2–L2 segments present in the different PV crown groups did not originate from a single common ancestor, but rather from multiple ancestors.

Model	LogLik	ΔBF
(C1-C2-C3-C4-C5)	-51548.9	0
(C1-C2)+C3+C4+C5	-48324.8	3224.1
C1+C2+(C3-C4)+C5	-48364.4	3184.5
C1+C2+C3+C4+C5	-47907.7	3641.2

Table 1. Hypothesis testing on the origin of the inter-E2–L2 region. For each hypothesis tested, common ancestry (CA) and independent origin (IO), we show the LogLikelihood value and the ΔBF (which equals log[Prob(IO)] - log[Prob(CA)]). Clade C1 includes *UpsilonPVs*, C2 includes *AlphaPVs*, C3 includes *LambdaPVs*, C4 includes *DeltaPVs*, and C5 includes *TauPVs*. The row highlighted in gray is the best-supported scenario.

DNA Sequences in The inter-E2–L2 Region in *AlphaPVs* are Monophyletic but The *E5* ORFs Therein Encoded are Not

The C1 and the C2 lineages of inter-E2–L2 sequences are present in viruses within the Alpha-Omikron PV crown group (fig. 1), and our results in (table 1) show that sequences in these C1 and C2 clades do not have a single common ancestor. We addressed then the question of the evolutionary history of the C2 lineage of inter-E2–L2 region and of the E5 ORFs therein encoded, present in the genomes of the *AlphaPVs*. We first checked whether this entire region present in extant *AlphaPVs* originated from the same ancestor, at the nucleotide level. We considered different plausible *Independent Origin* scenarios based on the phylogeny of the *AlphaPVs*. Specifically, we splitted the inter-E2–L2 regions in the *AlphaPVs* in three clusters that correspond to three different lineages: PVs causing cutaneous warts (CUT), mucosal lesions (MUC), and anogenital warts (GW) (fig. 1). The results showed that the *Common Ancestry* hypothesis was the best-supported model, while the *Independent Origin* hypothesis had the lowest support (table 2). We propose thus that in *AlphaPVs*, the region comprised between the *E2* and the *L2* ORFs has a single ancestor, and originated from the same recombination donor and/or gained access to the ancestral genome through a single integration event.

Model	LogLik	ΔBF
(MUC-CUT-GW)	-28776.6	0
(MUC-GW)+Cut	-28840.9	-64.3
(MUC-CUT)+GW	-29498.9	-722.3
MUC+(CUT-GW)	-29542.5	-765.9
MUC+CUT+GW	-29568.3	-791.7

Table 2. Hypothesis testing on the origin of the inter-E2–L2 region within *AlphaPVs*. PVs were stratified according their clinical presentation: MUC, *AlphaPVs* causing mucosal lesions; CUT, *AlphaPVs* causing cutaneous lesions; GW, *AlphaPVs* causing anogenital warts. The row highlighted in gray is the best-supported scenario.

Once common ancestry for the inter-E2–L2 region within the *AlphaPVs* was confirmed as the best-supported model, we asked whether the *E5* ORFs therein encoded also had a single common ancestor. We applied the same procedure and calculated the likelihood for different plausible scenarios (*Common Ancestry* and the *Independent Origin*) for the *E5* ORFs, at both the nucleotide and amino acid levels. Our results supported an *Independent Origin* scenario (table 3), where E5α, E5γ - E5δ, and E5ε - E5ζ (encoded in PVs with mucosal, anogenital tropism) have a common ancestor, but where E5β (encoded in PVs with cutaneous tropism) has an independent origin.

Model	LogLik(nt)	ΔBF(nt)	LogLik(aa)	ΔBF(aa)
(α-β-γ_δ-ε_ζ)	-11951.9	0	-6839.8	0
(α-γ_δ-ε_ζ)+β	-11936.8	15.1	-6816.9	22.9
(α-γ_δ)+β+ε_ζ	-11999.8	-47.9	-6853.9	-14.1
(α-ε_ζ)+β+γ_δ	-11972.4	-20.5	-6879.9	-40.1
α+β+γ_δ+ε_ζ	-12029.5	-77.6	-6909.6	-69.8

Table 3. Hypothesis testing on the origin of *E5* within *AlphaPVs*, at the nt and aa level. In the cases where two putative E5 ORFs are located in the same inter-E2–L2 fragment, as for E5γ and E5δ (E5γ_δ), and E5ε and E5ζ (E5ε_ζ) sequences were concatenated. The row highlighted in gray is the best-supported scenario.

In *AlphaPVs*, The Evolutionary History of The inter-E2–L2 Region is Similar to That of The Early Genes

In order to look deeper into the evolutionary history of the inter-E2–L2 region within *AlphaPVs*, we performed phylogenetic analyses and compared the tree topology for the inter-E2–L2 fragment sequences with the topologies obtained for each of the PV ORFs (*E1*, *E2*, *E6*, *E7*, *L1* and *L2*) as well as for the non-coding URR. The E5 tree was not included in this analysis because we could not reconstruct a single tree, as the E5β did not share a common ancestor with the other *E5* ORFs. We calculated the Robinson-Foulds (RF) distances between paired trees and we performed a multiple correspondence analysis using a distance matrix in order to identify similarities among the topologies of the PV gene trees. We found that the topology of the tree reconstructed from the inter-E2–L2 fragment was close to the topology of the early genes (*E1* and *E2*) in the PV genome (fig. 2). The late genes (*L1* and *L2*) clustered together but separated from the early genes. Finally, the non-coding URR appears separated from all the PV genes and the E2–L2 fragment.

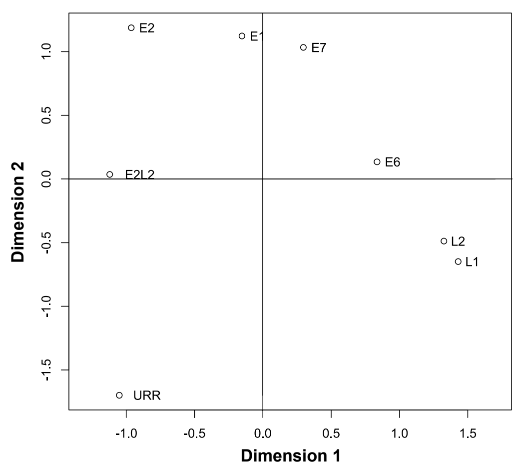

Figure 2. Multiple correspondence analysis of the Robinson-Foulds tree distance comparing tree topologies for each of the PV ORFs, the inter-E2–L2 region, and the URR.

The *E5* ORFs in *AlphaPVs* Display the Characteristics of a Genuine Gene

Since it is often discussed whether the *E5* ORFs in *AlphaPVs* are actual coding sequences, we performed a number of analyses in order to assess whether the different *E5* ORFs exhibit the characteristics of a *bona fide* gene. In order to determine whether the *E5* ORFs are larger than expected by chance, we constructed first 1000 random DNA sequences with the same median nucleotide composition as the inter-E2–L2 region of *AlphaPVs*, we identified all putative ORFs in these randomly generated DNA sequences and we computed their nucleotide length. (fig. 3) shows the cumulative frequency of the *E5* genes length and of the random ORFs. A one-way ANOVA followed by a post-hoc Tukey HSD test was performed, with *gene* as a factor (table S3) shows that ORFs in randomly generated sequences are shorter than any of the E5 ORFs (Tukey HSD: $p < 0.0001$).

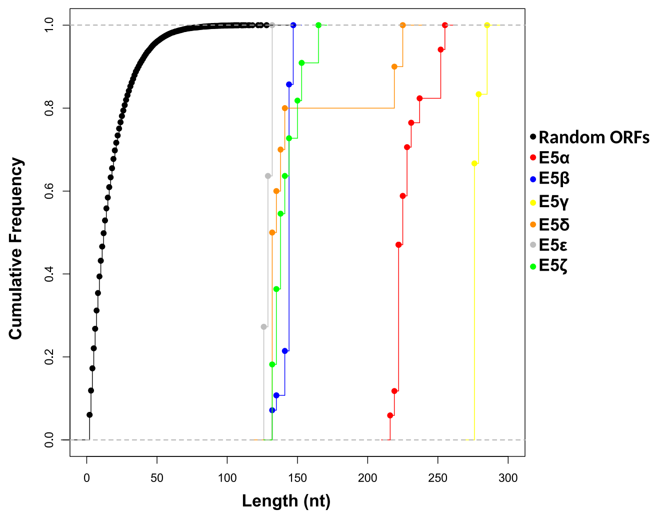

Figure 3. Cumulative frequency of the length for each group of the *E5* genes and random ORFs. Color-codes are indicated in the legend.

Besides length, evidence of selective pressure is another signature of *bona fide* genes. We calculated the dN/dS values for all *E5* sequences (fig. 4). Our results showed that the *E5* genes display a dN/dS distribution that is significantly lower than 1 (Wilcoxon-Mann-Whitney one side test: $p < 0.001$), with median values ranging from 0.13 to 0.40. All other PV genes presented median dN/dS values lower than the *E5* sequences (Tukey HSD: $p < 0.001$) (fig. 4).

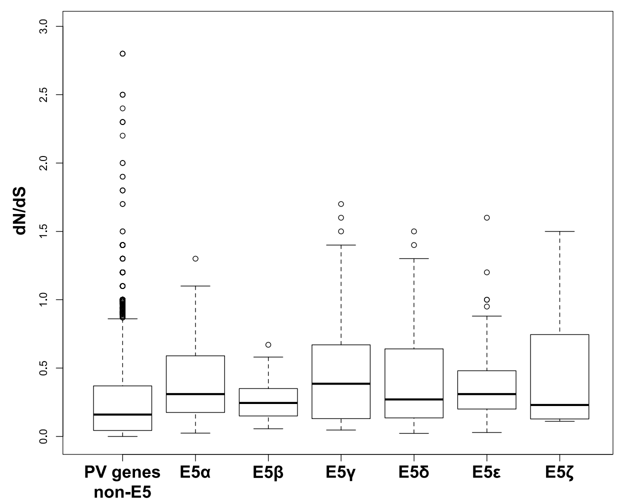

Figure 4. dN/dS values for each group of the *E5* genes and the other PV genes (E1, E2, E6, E7, L1 and L2).

We next calculated the pair-wise distances between terminal taxa for all ORFs and for the URR in *AlphaPVs*, as well as for a set of randomly generated intergenic CDS (fig. 5). These random CDS were generated using the average nucleotide composition from the inter-E2–L2 region of *AlphaPVs*, selecting for the same length distribution as the E5 ORFs (see Materials and Methods). Pairwise distances were normalized with respect to the corresponding *L1* distance. The highest rates of variation were found in the random intergenic CDS region and the lowest rates in the PV genes that are not *E5* (Tukey HSD, $p < 0.001$). Our results also showed that all *E5* genes presented lower rates of variation than the random intergenic CDS but higher rates than the other PV genes. The E5α, E5β and E5ζ showed higher rates of variation compared to the URR (Tukey HSD, $p < 0.001$). Contrary, the E5γ, E5δ, and E5ε showed lower rates of divergence in comparison to the URR (Tukey HSD, $p < 0.001$).

To corroborate whether the codon usage preferences (CUPrefs) of the *E5* genes are similar to those of the other PV genes, we calculated the relative frequencies of the 59 codons in synonymous families in the *E5* genes and in the rest of PV genes and

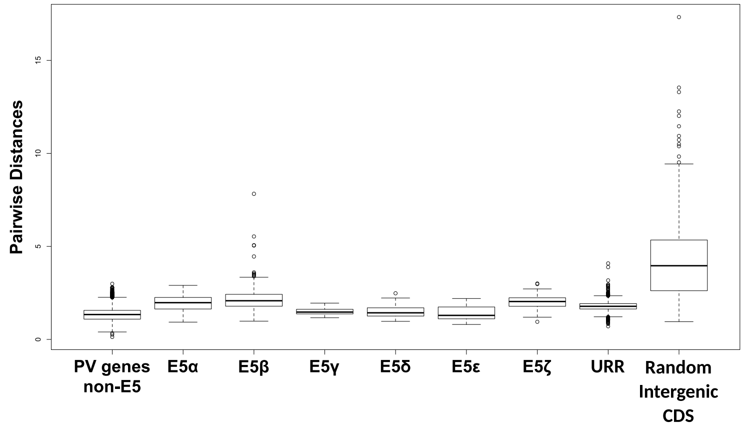

Figure 5. Pairwise distances between *AlphaPVs* for the all genes, the URR, and a set of randomly generated intergenic CDS. All values have been normalized to the corresponding *L1* pairwise distances.

the randomly generated intergenic CDS. Then we performed a multidimensional scaling (MDS) analysis on the 59-dimensional codon usage vectors, and in parallel, an unsupervised two-step cluster analysis (fig. 6). The optimal number of clusters was three: one cluster containing the early *E1* and *E2* genes; a second cluster containing late *L2* and *L1* genes; and a third cluster containing the *E5*, *E6*, *E7* oncogenes.

As E5 is a transmembrane protein, we hypothesized that a real *E5* genes should be more hydrophobic than expected by chance. We calculated the GRAVY index for the *E5* genes as well as for the randomly generated intergenic CDS (fig. 7). We found that E5α, E5β, E5γ, E5δ, and E5ε are more hydrophobic than the random intergenic CDS (Wilcoxon-Mann-Whitney test, $p < 0.0001$). The E5ζ is the only E5 protein that did not tested significantly more hydrophobic than the random intergenic CDS (Wilcoxon-Mann-Whitney, p =0.125).

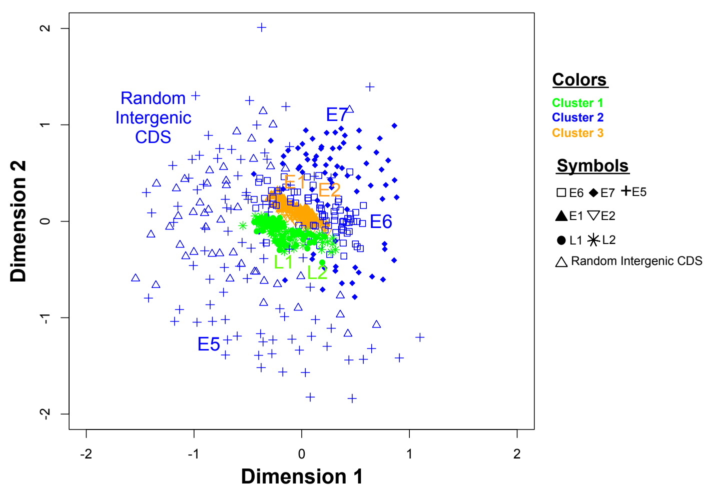

Figure 6. Multidimensional Scaling (MDS) plot of codon usage preferences for the *AlphaPV* ORFs. The ORFs were independently clustered by an unsupervised two-step clustering algorithm. The best assembly included three clusters, displayed onto the MDS plot as with a color code, composed respectively by the oncogenes *E5, E6* and *E7*; the early genes *E1* and *E2*; and the capsid genes *L1* and *L2*.

Discussion

Understanding how PV genes have originated and evolved is crucial for explaining the genetic basis of the origin and evolution of phenotypic diversity found in PVs. In this work our first aim was to study the origin of the *E5* oncogenes in *AlphaPVs*. This viral genus hosts around fifty viral genotypes with a relative narrow host distribution (they seem to be restricted to Primates), but with very diverse phenotypic presentations of the infections: many of them are associated to asymptomatic infections of

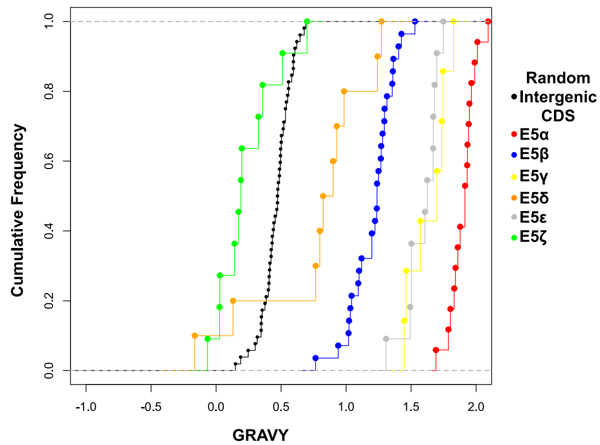

Figure 7. Cumulative frequency of the GRAVY index for the *E5* ORFs and the randomly generated intergenic CDS.

the skin, but also of the oral, nasal, or anogenital mucosas; some of them cause productive infections that result in common skin warts, or in genital warts; and a number of them cause chronic infections that may result in anogenital or oropharyngeal cancers[35,36]. All *AlphaPVs* present a region between the *E2* and *L2* genes, potentially encoding in all cases for conserved ORFs. With few exceptions[37], actual gene expression and protein function for *E5* oncogenes have only been characterized for the more oncogenic HPVs, which carry E5 proteins of type E5α[7]. These E5α behave as oncoproteins, promoting cell division and allowing the infected cells to avoid immune recognition[12–14].

Since the *E5* ORFs in *AlphaPVs* map between the *E2* and *L2* genes we extended our analysis to the evolution of this intergenic region in the Alpha-Omikron crown group. Finally, since a number of non-monophyletic PVs also contain a sometimes long non-coding region between the *E2* and *L2* genes in their genomes that may also encode for genes named *E5*, we expanded our analyses to the full set of PV sequences containing a long non-coding region at this genomic location. PVs displaying an intergenic region between *E2* and *L2* are not monophyletic, and belong instead to five clades in the PV tree (fig. 1). It could be argued that the ancestral PV genomes could have already presented an inter-E2–L2 region, which may have undergone several loss events. Such repeated losses have been invoked as a mechanism to explain the repeated absence of early genes in certain PVs[38]. However, our results clearly show that all extant nucleotide sequences present in the inter-E2–L2 region of PVs do not share a common ancestor (table 1). Instead, the occurrence of the ancestral inter-E2–L2 regions most likely occurred as five independent events, where each event took place in a separate PV clade.

The putative ORFs that emerged in the inter-E2–L2 region are often named *E5*. Notwithstanding, our results show that the E5 proteins encoded in the different clades are not monophyletic. Specifically, these results imply that the *E5* ORFs in *AlphaPVs* (*e.g.* HPV16 *E5*) are not evolutionarily related to the *E5* ORFs in *DeltaPVs* (*e.g.* BPV1 *E5*). This is an important change in perspective, because these two proteins are often referred to and their cellular activities compared as if they were orthologs[39,40].

We can formulate two main non-exclusive mechanisms to explain the origin of the five extant groups of inter-E2–L2 regions in the PVs genomes: random nucleotide addition and recombination. Random nucleotide addition is a plausible mechanism, based on the way the PV genome replicates. The replication of the PV genome occurs bidirectionally during the non-productive stages of the infection, yielding episomes[41]. During bidirectional replication, the replication forks converge opposite to the origin of replication, which in the case of PVs is located in the URR. The opposite region to the URR happens to lay between the *E2* and *L2* genes. At this point, concerted DNA breaks are required for decatenation, which eventually generates two separate circular dsDNA molecules. The end joining of these DNA breaks is error prone. Indeed, the DNA close to the break site can be used as a template for *de novo* synthesis before the DNA ends are joined, resulting in the non-templated introduction of a stretch of additional nucleotides[42].

Recombination can also be invoked as a mechanism that may result in the integration of novel DNA sequences into the PV genome. In parallel to the host keratinocyte differentiation, replication of the viral genome switches from bidirectional to unidirectional[41,43], generating large linear molecules of concatenated viral genomes[44]. Unidirectional replication relies on homologous recombination, as this mechanisms is required for resolving, excising and recircularizing the concatenated genomes into individual plasmid genomes[45–47]. Additionally, productive replication concurs with a virus-mediated impairment of the cellular DNA damage repair mechanisms[48,49], thus rendering the overall viral replication process error-prone by increasing the probability of integrating exogenous DNA during recircularization. Phylogenetic evidence for the existence and fixation of such recombination events is provided by the incongruence in the reconstruction of the evolutionary history for different regions of the PV genome. In all cases, such inconsistencies appear when comparing the phylogenetic inference for the early and for the late genes of the genome, respectively upstream and downstream the recombination-prone genomic region. Evidence for recombination has been described at several nodes in the PV tree. The first example occurs at the root of *AlphaPVs*, with the species containing oncogenic PVs being monophyletic according to the early genes (involved in oncogenesis and genome replication), and paraphyletic according to the late genes (involved in capsid formation)[7,50]. The second example is provided by certain PVs infecting cetaceans, which display the early genes related to those in other cetacean PVs in the Alpha-Omikron crown group (in red in fig. 1) and the late genes related to those in bovine PVs in the Beta-Xi crown group (in green in fig. 1)[51–53]. Finally, the most cogent examples of recombination between distant viral sequences are two viruses isolated from bandicoots and displaying the early genes related to Polyomaviruses and the late genes related to PVs[54,55].

The inter-E2–L2 sequences may occasionally be very long and span more than 1 Kbp, a considerable size for an average genome length of around 8 Kbp. Additionally, for many viral genomes, the sequences in the inter-E2–L2 region do not resemble other sequences in the databases, and do not seem to contain any functional elements, neither ORFs nor transcription factor binding sites or conserved regulatory regions[8,56,57]. Despite the lack of obvious function and of their length, these sequences seem to belong *bona fide* in the viral genome in which they are found, as they are fixed and conserved in viral lineages[57]. Although the two hypothesis referred above to explain the origin of the inter-E2–L2 regions (random nucleotide addition and recombination) are plausible, we interpret that the presence of long and conserved sequences in certain monophyletic clades (labeled with a star in (fig. 1) suggests that the respective insertions of each of these long sequences in the ancestral genomes occurred during single episodes, pointing thus towards a recombination event.

When restricting our analysis to the inter-E2–L2 region within the *AlphaPVs*, we found support for monophyly (table 2), indicating that a single event on the backbone of the ancestral *AlphaPV* genome led to its emergence. On the contrary the different *E5* ORFs that arose from this region in *AlphaPVs*, were found to be not monophyletic (table 3). In our analysis, the E5β, which is present in *AlphaPVs* with a cutaneous tropism, presents a different origin than the rest of the E5 proteins, which are present in *AlphaPVs* with a mucosal tropism (E5α, E5γ, E5δ, E5ε, and E5ζ). Indeed, there is no evident sequence similarity between the E5 proteins, inasmuch as the evolutionary divergence between E5β and the other *E5* ORFs rises to 80%[7]. Phylogenetic reconstruction based on the *E5* ORFs showed a star-like pattern with the main branches emerging close to a putative central point[7]. These features could be related to the multiple ancestries of the different *E5* ORFs.

It remains unclear how the different *E5* genes emerged in the viral genome. Our interpretation of the evidence here provided is as follows. Under the hypothesis of recombination, within the *AlphaPVs*, a non-coding sequence was integrated between the early and the late genes in the genome of a PV lineage infecting the ancestor of Old World monkeys and apes. After several mutations in this non-coding region the different *E5* ORFs were generated. *De novo* birth of new protein-coding sequences from non-coding genomic regions is not unfamiliar and has been reported in for example *Drosophila*[58,59], yeast[60] and mammals[61]. Experimentally, protein structures that have not been observed in nature have also been isolated, more specifically Chacón *et al.* 2014[62] replaced the BPV E5 oncoprotein with randomized hydrophobic segments and used genetic selection to isolate artificial transmembrane proteins lacking any preexisting sequences. These amino acid sequences that do not occur in nature were able to bind and activate the platelet derived growth factor (PDGF) β receptor (just like BPV E5 does), resulting in cell transformation and tumorogenicity[62]. Therefore we consider *de novo* birth of the *E5* genes in the inter-E2–L2 region a plausible hypothesis. The randomly appeared *E5* genes, short and enriched in hydrophobic amino acids, could thus have provided with a rudimentary function by binding to membrane receptors or by modifying membrane environment. Such activities may have lead to an increase in viral fitness and could have been selected and enhanced, resulting in the different *E5* genes lineages observed today.

The location within the inter-E2–L2 region and the hydrophobic nature of the protein have up to date been the criteria to classify the *E5* ORFs as putative genes. This is probably the reason for which we found all *E5* ORFs, with the only exception of E5ζ, more hydrophobic than expected by chance (fig. 7). However, we do not have evidence of the expression of these ORFs *in vivo*. Moreover, the possible independent origins of *E5*, rise the concern of whether all *E5* ORFs are actually coding sequences. In this study, we have used several approaches in order to distinguish true *E5* genes from spurious ORFs that are not functional. As *E5* genes are not found in other related species, we studied the *E5* ORFs in the context of orphan genes. In agreement with studies of orphan genes in other species[61,63,64], the *E5* genes are shorter than the other PV genes. It has previously been proposed that there is a direct relationship between the length of a gene and its age[61,65,66]. However, a real gene must be longer than expected by chance[67], and this is what we found for the different *E5* ORFs (fig. 3).

For a new functional protein to evolve from randomly occurring ORFs, it needs to be produced in significant amounts. These proteins are expected to evolve under neutral selection, as these are unlikely to be functional at first. By combining ribosome profiling RNA sequencing with proteomics and SNP information Ruiz-Orera *et al.* 2018 found evidence to support this hypothesis[68]. By analyzing mouse tissue they found hundreds of small proteins that evolve under no purifying selection. Regarding the *E5* ORFs, we obtained dN/dS ratios below 1 (fig. 4), indicating negative or purifying selection, reinforcing the idea that E5 is functionally relevant. Apparently, the codon composition has an effect on ORF translation, where a favorable codon composition may facilitate the translation of certain ORFs, while other ORFs with a less favorable codon composition remain untranslated[68]. To measure whether E5 has a favorable codon composition that resembles the other PV genes, we compared their codon usage preferences (CUPrefs). The *E5* genes exhibited CUPrefs similar to those in the early (*E6* and *E7*) genes (fig. 6), which are both implicated in oncogenesis. This is in line with previous work reporting that genes expressed at similar stages during viral infection have similar CUPrefs[19]. The observation that the E5 ORFs are under purifying selection and the clustering of the CPUrefs of *E5* together with the two other oncogenes, reinforces the oncogenic role of the different *E5* proteins in the PV life cycle.

Our results strongly suggest that *E5* in *AlphaPVs* are *bona fide* genes and not merely spurious translations. This is supported by previous studies that already assigned different properties to *E5*, such as the alteration of membrane composition and dynamics[12,13] and the down-regulation of surface MHC class I molecules[37,69] for immune evasion. However, many questions about *E5* remain to be elucidated. Further experimental studies should be performed to provide evidence of the expression of the different *E5* ORFs *in vivo* and to elucidate whether *E5* originated through recombination, random nucleotide addition or another unknown mechanism.

Supplementary Material

Supplementary tables S1-S3 are available online.

Acknowledgements

This work was supported by the European Research Council Consolidator Grant CODOVIREVOL (Contract Number 647916) to IGB and by the European Union Horizon 2020 Marie Sklodowska-Curie research and innovation programme grant ONCOGENEVOL (Contract Number 750180) to AW.

References

1. Gottschling, M. *et al.* Quantifying the phylodynamic forces driving papillomavirus evolution. *Molecular biology and evolution* **28**, 2101–13 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21285031.

2. Münger, K., Scheffner, M., Huibregtse, J. M. & Howley, P. M. Interactions of HPV E6 and E7 oncoproteins with tumour suppressor gene products. *Cancer surveys* **12**, 197–217 (1992). URL http://www.ncbi.nlm.nih.gov/pubmed/1322242.

3. Moody, C. A. & Laimins, L. A. Human papillomavirus oncoproteins: pathways to transformation. *Nature Reviews Cancer* **10**, 550–560 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20592731.

4. Tomaić, V. Functional Roles of E6 and E7 Oncoproteins in HPV-Induced Malignancies at Diverse Anatomical Sites. *Cancers* **8**, 95 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27775564.

5. DiMaio, D. & Petti, L. M. The E5 proteins. *Virology* **445**, 99–114 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23731971.

6. Bravo, I. G. & Félez-Sánchez, M. Papillomaviruses. *Evolution, Medicine, and Public Health* **2015**, 32–51 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25634317.

7. Bravo, I. G. & Alonso, A. Mucosal Human Papillomaviruses Encode Four Different E5 Proteins Whose Chemistry and Phylogeny Correlate with Malignant or Benign Growth. *Journal of Virology* **78**, 13613–13626 (2004). URL https://www.ncbi.nlm.nih.gov/pubmed/15564472.

8. García-Pérez, R. *et al.* Novel Papillomaviruses in Free-Ranging Iberian Bats: No Virus–Host Co-evolution, No Strict Host Specificity, and Hints for Recombination. *Genome Biology and Evolution* **6**, 94–104 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24391150.

9. Pim, D., Collins, M. & Banks, L. Human papillomavirus type 16 *E5* gene stimulates the transforming activity of the epidermal growth factor receptor. *Oncogene* **7**, 27–32 (1992). URL http://www.ncbi.nlm.nih.gov/pubmed/1311063.

10. Conrad, M., Bubb, V. J. & Schlegel, R. The human papillomavirus type 6 and 16 E5 proteins are membrane-associated proteins which associate with the 16-kilodalton pore-forming protein. *Journal of virology* **67**, 6170–8 (1993). URL http://www.ncbi.nlm.nih.gov/pubmed/7690419.

11. Straight, S. W., Hinkle, P. M., Jewers, R. J. & McCance, D. J. The E5 oncoprotein of human papillomavirus type 16 transforms fibroblasts and effects the downregulation of the epidermal growth factor receptor in keratinocytes. *Journal of virology* **67**, 4521–4532 (1993). URL http://www.ncbi.nlm.nih.gov/pubmed/8392596.

12. Bravo, I. G., Crusius, K. & Alonso, A. The E5 protein of the human papillomavirus type 16 modulates composition and dynamics of membrane lipids in keratinocytes. *Archives of Virology* **150**, 231–246 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/15503216.

13. Suprynowicz, F. A. *et al.* HPV-16 E5 oncoprotein upregulates lipid raft components caveolin-1 and ganglioside GM1 at the plasma membrane of cervical cells. *Oncogene* **27**, 1071–1078 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/17704805.

14. Ashrafi, G. H., Haghshenas, M. R., Marchetti, B., O'Brien, P. M. & Campo, M. S. E5 protein of human papillomavirus type 16 selectively downregulates surface HLA class I. *International Journal of Cancer* **113**, 276–283 (2005). URL https://www.ncbi.nlm.nih.gov/pubmed/15386416.

15. Petti, L. M., Reddy, V., Smith, S. O. & DiMaio, D. Identification of amino acids in the transmembrane and juxtamembrane domains of the platelet-derived growth factor receptor required for productive interaction with the bovine papillomavirus E5 protein. *Journal of virology* **71**, 7318–7327 (1997). URL http://www.ncbi.nlm.nih.gov/pubmed/9311809.

16. DiMaio, D. & Mattoon, D. Mechanisms of cell transformation by papillomavirus E5 proteins (2001). URL http://www.ncbi.nlm.nih.gov/pubmed/11753669.

17. Ashrafi, G. H. *et al.* Down-regulation of MHC class I by bovine papillomavirus E5 oncoproteins. *Oncogene* **21**, 248–259 (2002). URL http://www.ncbi.nlm.nih.gov/pubmed/11803468.

18. Hughes, A. L. & Hughes, M. A. K. Patterns of nucleotide difference in overlapping and non-overlapping reading frames of papillomavirus genomes. *Virus research* **113**, 81–8 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/15913825.

19. Félez-Sánchez, M. *et al.* Cancer, Warts, or Asymptomatic Infections: Clinical Presentation Matches Codon Usage Preferences in Human Papillomaviruses. *Genome Biology and Evolution* **7**, 2117–2135 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26139833.

20. Suchard, M. A. & Redelings, B. D. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* **22**, 2047–2048 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/16679334.

21. Redelings, B. D. & Suchard, M. A. Joint bayesian estimation of alignment and phylogeny. *Systematic Biology* **54**, 401–418 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/16012107.

22. de Oliveira Martins, L. & Posada, D. Testing for Universal Common Ancestry. *Systematic Biology* **63**, 838–842 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24958930.

23. Xie, W., Lewis, P. O., Fan, Y., Kuo, L. & Chen, M. H. Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic Biology* **60**, 150–160 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21187451.

24. Mengual-Chuliá, B., Bedhomme, S., Lafforgue, G., Elena, S. F. & Bravo, I. G. Assessing parallel gene histories in viral genomes. *BMC Evolutionary Biology* **16**, 32 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26847371.

25. García-Vallvé, S., Iglesias-Rozas, J. R., Alonso, Á. & Bravo, I. G. Different papillomaviruses have different repertoires of transcription factor binding sites: convergence and divergence in the upstream regulatory region. *BMC Evolutionary Biology* **6**, 20 (2006). URL https://www.ncbi.nlm.nih.gov/pubmed/16526953.

26. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797 (2004). URL http://www.ncbi.nlm.nih.gov/pubmed/15034147.

27. Castresana, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution* **17**, 540–552 (2000). URL http://www.ncbi.nlm.nih.gov/pubmed/10742046.

28. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24451623.

29. Robinson, G. A. & Wasnidge, D. C. Comparison of the accumulation of 125I and 144Ce in the growing oocytes of the Japanese quail. *Poultry science* **60**, 2195–9 (1981). URL http://www.ncbi.nlm.nih.gov/pubmed/7199144.

30. Doron-Faigenboim, A., Stern, A., Mayrose, I., Bacharach, E. & Pupko, T. Selection: A server for detecting evolutionary forces at a single amino-acid site. *Bioinformatics* **21**, 2101–2103 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/15647294.

31. Doron-Faigenboim, A. & Pupko, T. A combined empirical and mechanistic codon model. *Molecular Biology and Evolution* **24**, 388–397 (2007). URL http://www.ncbi.nlm.nih.gov/pubmed/17110464.

32. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A.-M. K. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. *Genetics* **155**, 431–49 (2000). URL https://www.ncbi.nlm.nih.gov/pubmed/10790415.

33. R Core Team. R: A language and environment for statistical computing. (2014). URL http://www.r-project.org/.

34. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology* **157**, 105–32 (1982). URL http://www.ncbi.nlm.nih.gov/pubmed/7108955.

35. Doorbar, J. *et al.* The biology and life-cycle of human papillomaviruses (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/23199966. NIHMS150003.

36. Forman, D. *et al.* Global Burden of Human Papillomavirus and Related Diseases. *Vaccine* **30**, F12–F23 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/23199955.

37. Cartin, W. & Alonso, A. The human papillomavirus HPV2a E5 protein localizes to the Golgi apparatus and modulates signal transduction. *Virology* **314**, 572–579 (2003). URL https://www.ncbi.nlm.nih.gov/pubmed/14554085.

38. Van Doorslaer, K. & McBride, A. A. Molecular archeological evidence in support of the repeated loss of a papillomavirus gene. *Scientific Reports* **6**, 33028 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27604338.

39. Ashby, A. D., Meagher, L., Campo, M. S. & Finbow, M. E. E5 transforming proteins of papillomaviruses do not disturb the activity of the vacuolar H+-ATPase. *Journal of General Virology* **82**, 2353–2362 (2001). URL http://www.ncbi.nlm.nih.gov/pubmed/11562529.

40. Venuti, A. *et al.* Papillomavirus E5: The smallest oncoprotein with many functions (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/22078316.

41. Flores, E. R. & Lambert, P. F. Evidence for a switch in the mode of human papillomavirus type 16 DNA replication during the viral life cycle. *Journal of virology* **71**, 7167–79 (1997). URL http://www.ncbi.nlm.nih.gov/pubmed/9311789.

42. Roerink, S. F., Schendel, R. & Tijsterman, M. Polymerase theta-mediated end joining of replication-associated DNA breaks in C. elegans. *Genome Research* **24**, 954–962 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24614976.

43. McBride, A. A. Mechanisms and strategies of papillomavirus replication. *Biological Chemistry* **398**, 919–927 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28315855.

44. Dasgupta, S., Zabielski, J., Simonsson, M. & Burnett, S. Rolling-circle replication of a high-copy BPV-1 plasmid. *Journal of Molecular Biology* **228**, 1–6 (1992). URL http://www.ncbi.nlm.nih.gov/pubmed/1333015.

45. Gillespie, K. A., Mehta, K. P., Laimins, L. A. & Moody, C. A. Human Papillomaviruses Recruit Cellular DNA Repair and Homologous Recombination Factors to Viral Replication Centers. *Journal of Virology* **86**, 9520–9526 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22740399.

46. Mehta, K. & Laimins, L. Human Papillomaviruses Preferentially Recruit DNA Repair Factors to Viral Genomes for Rapid Repair and Amplification. *mBio* **9**, e00064–18 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29440569.

47. Sakakibara, N., Chen, D. & McBride, A. A. Papillomaviruses Use Recombination-Dependent Replication to Vegetatively Amplify Their Genomes in Differentiated Cells. *PLoS Pathogens* **9**, e1003321 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23853576.

48. Chappell, W. H. *et al.* Homologous Recombination Repair Factors Rad51 and BRCA1 Are Necessary for Productive Replication of Human Papillomavirus 31. *Journal of Virology* **90**, 2639–2652 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26699641.

49. Wallace, N. A. *et al.* High-Risk Alphapapillomavirus Oncogenes Impair the Homologous Recombination Pathway. *Journal of Virology* **91**, e01084–17 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28768872.

50. Narechania, A., Chen, Z., DeSalle, R. & Burk, R. D. Phylogenetic incongruence among oncogenic genital alpha human papillomaviruses. *Journal of virology* **79**, 15503–10 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/16306621.

51. Gottschling, M. *et al.* Modular organizations of novel cetacean papillomaviruses. *Molecular Phylogenetics and Evolution* **59**, 34–42 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21195783.

52. Rector, A. *et al.* Genomic characterization of novel dolphin papillomaviruses provides indications for recombination within the Papillomaviridae. *Virology* **378**, 151–161 (2008). URL https://www.ncbi.nlm.nih.gov/pubmed/18579177.

53. Robles-Sikisaka, R. *et al.* Evidence of recombination and positive selection in cetacean papillomaviruses. *Virology* **427**, 189–197 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22386054.

54. Woolford, L. *et al.* A Novel Virus Detected in Papillomas and Carcinomas of the Endangered Western Barred Bandicoot (Perameles bougainville) Exhibits Genomic Features of both the Papillomaviridae and Polyomaviridae. *Journal of Virology* **81**, 13280–13290 (2007). URL https://www.ncbi.nlm.nih.gov/pubmed/17898069.

55. Bennett, M. D. *et al.* Genomic characterization of a novel virus found in papillomatous lesions from a southern brown bandicoot (Isoodon obesulus) in Western Australia. *Virology* **376**, 173–182 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18440042.

56. Schulz, E. *et al.* Genomic characterization of the first insectivoran papillomavirus reveals an unusually long, second non-coding region and indicates a close relationship to Betapapillomavirus. *Journal of General Virology* **90**, 626–633 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19218207.

57. Rector, A. *et al.* Ancient papillomavirus-host co-speciation in Felidae. *Genome Biology* **8**, R57 (2007). URL http://www.ncbi.nlm.nih.gov/pubmed/17430578.

58. Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences* **103**, 9935–9939 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/16777968.

59. Zhou, Q. *et al.* On the origin of new genes in Drosophila. *Genome Research* **18**, 1446–1455 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18550802.

60. Cai, J., Zhao, R., Jiang, H. & Wang, W. De Novo Origination of a New Protein-Coding Gene in Saccharomyces cerevisiae. *Genetics* **179**, 487–496 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18493065.

61. Toll-Riera, M. *et al.* Origin of primate orphan genes: A comparative genomics approach. *Molecular Biology and Evolution* **26**, 603–612 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19064677.

62. Chacón, K. M. *et al.* De novo selection of oncogenes. *Proceedings of the National Academy of Sciences* **111**, E6–E14 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24344264.

63. Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V. & Lipman, D. J. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences* **106**, 7273–7280 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19351897.

64. Carvunis, A.-R. *et al.* Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22722833.

65. Albà, M. M. & Castresana, J. Inverse relationship between evolutionary rate and age of mammalian genes. *Molecular Biology and Evolution* **22**, 598–606 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/15537804.

66. Palmieri, N., Kosiol, C. & Schlötterer, C. The life cycle of Drosophila orphan genes. *eLife* **3**, e01311 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24554240.

67. Schlötterer, C. Genes from scratch–the evolutionary fate of de novo genes. *Trends in genetics : TIG* **31**, 215–9 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25773713.

68. Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Cañas, J. L., Messeguer, X. & Albà, M. M. Translation of neutrally evolving peptides provides a basis for de novo gene evolution (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29556078.

69. Campo, M. *et al.* HPV-16 E5 down-regulates expression of surface HLA class I and reduces recognition by CD8 T cells. *Virology* **407**, 137–142 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20813390.