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
[+]these authors contributed equally to this work
[*]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. These *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 four events lead to the presence of a long non-coding DNA stretch between the *E2* and the *L2* genes. In three of these events, the novel regions evolved coding capacity, becoming the extant *E5* ORFs. We then focused on the evolution of the *E5* genes in *AlphaPVs* infecting humans. The sharp match between the type of E5 protein encoded in *AlphaPVs* and the infection phenotype (cutaneous warts, genital warts or anogenital cancers) supports the role of E5 in the differential oncogenic potential of these PVs. In our analyses, the best-supported scenario is that the five types of extant E5 proteins within the *AlphaPV* genomes may not have a common ancestor. However, the chemical similarities between E5s regarding amino acid composition prevent us from confidently rejecting the model of a common origin. 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 that display 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 phenotypic diversity of the viral infections.

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 bony 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 (Gottschling *et al.*, 2011b) 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 (Moody and Laimins, 2010; Münger *et al.*, 1992; Tomaić, 2016). However, there is also a minor oncoprotein termed E5, whose functions and origin remain to be fully elucidated (DiMaio and Petti, 2013).

The *E5* ORFs are located in the intergenic region between the *E2* and the *L2* genes. This inter-E2–L2 region is highly variable between 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 (Bravo and Félez-Sánchez, 2015). PVs within the *Alpha-* and *DeltaPV* genera encode different E5 proteins in the inter-E2–L2 region (Bravo and Alonso, 2004). 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 and/or function (García-Pérez *et al.*, 2014).

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 (Bravo and Alonso, 2004). 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 (Bravo and Alonso, 2004). 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 primates other than humans 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 (Bravo and Félez-Sánchez, 2015).

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, and the differences in immediate interaction partners, 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 (Conrad *et al.*, 1993; Pim *et al.*, 1992; Straight *et al.*, 1993), as well as to immune evasion by modifying the membrane chemistry (Bra, 2005; Suprynowicz *et al.*, 2008) and decreasing the presentation of viral epitopes (Ashrafi *et al.*, 2005). BPV1 E5 is a very short protein (half the size of HPV16 E5) that also localizes in the membranes. It displays a strong transforming activity, largely by activating the platelet-derived growth factor receptor (DiMaio and Mattoon, 2001; Petti *et al.*, 1997), and it downregulates as well the presentation of viral epitopes in the context of the MHC-I molecules (Ashrafi *et al.*, 2002).

In this study, we have explored the evolutionary history of the *E5* ORFs found within the inter-E2–L2 region in PVs. First, we identified the PV clades that contain a long intergenic region between *E2* and *L2*, and therewith putative *E5* ORFs. Then, we assessed whether the E5 ORFs in the identified clades 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

We collected 354 full length PV genomes from the PaVE (`pave.niaid.nih.gov`) and GenBank (`https://www.ncbi.nlm.nih.gov/genbank/`) databases (table S1). The corresponding E5 sequences were retrieved from these genomes as well as the intergenic region between the *E2* and *L2* genes (inter-E2–L2). Based on the size of the inter-E2–L2 region in which E5s are present, we selected those with a minimum lenght of 250 nucleotides (fig.1 and fig. S1). For comparison in the tree figures, we extended our analysis and also indicated inter-E2–L2 regions with a minimum length of 125 nucleotides. The URR, *E6*, *E7*, *E1*, *E2*, *L2* and *L1* were also extracted from the collected genomes and analyzed in parallel to the *E5* sequences. We excluded the *E4* ORFs from our analyses 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 (Félez-Sánchez *et al.*, 2015; Hughes and Hughes, 2005). Genes were aligned individually at the amino acid level using MAFFT v.7.271 (Katoh and Standley, 2013), corrected manually, and backtranslated to nucleotides using PAL2NAL v.14 (Suyama *et al.*, 2006) The alignment was filtered using Gblocks v.0.91b (Castresana, 2000). The URR and the inter-E2–L2 region (non-coding regions) were aligned at the nucleotide level.

Phylogenetic Analyses

For tree construction of the concatenated *E1*, *E2*, *L2* and *L1* genes, the previously identified recombinant PVs isolated from Cetaceans (PphPV1-2, TtPV1-7, DdPV1, PsPV1) (Gottschling *et al.*, 2011b; Rector *et al.*, 2008; Robles-Sikisaka *et al.*, 2012) were removed before alignment, leaving us with a data set of 343 PVs. The concatenated E1-E2-L2-L1 alignment was used to construct Maximum Likelihood (ML) trees with RAxML v.8.2.9 (Stamatakis, 2014) under the GTR+Γ4 model for the nucleotide alignment, using 12 partitions (three for each gene corresponding to each codon position), or under the LG+I+Γ model for the amino acid alignment using 4 partitions (one for each gene), and using 1000 bootstrap replicates.

To measure the distances between the URR, *E6*, *E7*, *E1*, *E2*, *E5*, inter-E2–L2, *L2* and *L1* trees, we reduced the dataset to 69 PVs so that the taxa are in common among all trees. We reconstructed a phylogenetic tree for each gene separately, as well as for the URR and the inter-E2–L2 region. ML trees were constructed at the nucleotide level using RAxML v.8.2.9 under the GTR+Γ4 model. The weighted and unweighted Robinson-Foulds (RF) distances between trees were calculated (Robinson and Foulds, 1981). The unweighted RF distance only depends on the topology of the trees, while the weighted RF distance considers edge weights. A correspondence analysis was performed to identify similarities between the topologies of the trees reconstructed for each gene.

Testing for Common Ancestry using BAli-Phy

In order to evaluate the common ancestry of the E5 ORFs, we used the BAli-Phy algorithm (Suchard and Redelings, 2006). Under this bayesian 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 (Redelings and Suchard, 2005). 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) (de Oliveira Martins and Posada, 2014). Therefore, we have ΔBF = log[Prob(CA)]-log[Prob(IO)], such that positive values support CA and negative values support IO. The likelihood for the CA model was obtained running the software for all the E5 sequences together. For the IO scenarios, we ran one analysis for each group independently. We started with the different PV clades that contain an E5 ORF in the inter-E2–L2 region, located within the Alpha-Omikron (red) and Delta-Zeta (blue) crown groups (fig. 1 and fig. S1). 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. Then we ran the analyses on the E5 ORFs within *AlphaPVs* stratifying by the different E5 types that are associated to three distinct clinical presentations; mucosal lesions, cutaneous warts, and genital warts. The values for the independent groups of E5α1, E5α2, E5β, E5$\gamma\delta$, E5δ, and E5$\varepsilon\zeta$, and the sum of combinations of these, rendered the likelihood for the IO models. For instance, (α1-α2-$\varepsilon\zeta$) + ($\gamma\delta$-δ) + β denotes a hypothesis of three independent ancestries, one tree for the E5 types associated to mucosal lesions (E5α1, E5α2, and E5$\varepsilon\zeta$ together), another separate tree for the E5 types associated to genital warts (E5$\gamma\delta$ and E5δ together), and another tree for the E5 type associated to cutaneous warts (E5β). The likelihood of this example was obtained running BAli-Phy three times: one run for E5α1, E5α2, and E5$\varepsilon\zeta$, one for E5$\gamma\delta$ and E5δ, and one for E5β. The sum of these three analyses corresponded to the likelihood of the model. We only considered the IO scenarios that were biologically plausible based on the phylogeny of PVs (fig. 1 and fig. S1). The same procedure was applied to the E5 sequences belonging to both the Alpha-Omikron and Delta-Zeta crown groups. This analysis was performed at the amino acid level using the LG substitution model. For each model, three independent MCMC chains were run for at least 100000 iterations. The three runs were combined and checked for convergence.

Random permutations to test for Common Ancestry

To support the results of the BAli-Phy analyses, we performed a random permutation test as described in de Oliveira Martins and Posada, 2016. In this test the sequences for one of the groups are randomly shuffled and statistics are recalculated after realignment with MUSCLE (Edgar, 2004), which tells us how much the results using the original data departs from those with phylogenetic structure partially removed. The statistics used in this test are ML tree length and Log Likelihood calculated with PhyMLv3.0 (Guindon *et al.*, 2010). As for the BAli-Phy test, these analyses were performed at the amino acid level using the LG substitution model. We obtained a distribution by reshuffling one of the groups (for example the E5$\varepsilon\zeta$ sequences) 100 times, each time realigning against the other groups from the dataset, and comparing the resulting phylogeny with those if we separate again the groups. For each iteration, the alignment is always optimised and the statistics are calculated. To make the statistics comparable, the same alignment is used for both the IO and CA hypotheses. We compare the distrubution fot the CA and IO hypotheses with a Kruskal-Wallis rank sum test and a multiple comparison test after Kruskal-Wallis. The results were confirmed by performing Wilcoxon rank sum tests with continuity correction. Lower ML tree length and superior Log likelihood values are expected to support the best model.

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. 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 (Doron-Faigenboim *et al.*, 2005), using the MEC model (Doron-Faigenboim and Pupko, 2006). The likelihood of MEC model was tested against the model M8a (Yang *et al.*, 2000), 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 *AlphaPV* 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 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 AlphaPV* 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 *AlphaPV* ORFs, as described in (Félez-Sánchez *et al.*, 2015). 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 (Kyte and Doolittle, 1982).

Statistics and Graphics

Statistical analyses and graphics were done using R (R Core Team, 2014), with the aid of the packages "ape", "ade4", and "phangorn". The final display of the graphics was designed using Inkscape v.0.92 (https://inkscape.org/en/).

Results

Do the E5 ORFs Present in the Genomes of PVs Belonging to Different Crown Groups Have a Common Ancestor?

We collected 354 full length PV genomes from the PaVE (pave.niaid.nih.gov) and GenBank (https://www.ncbi.nlm.nih.gov/genbank/) databases (table S1). After removing eleven recombinant sequences we constructed a maximum likelihood phylogenetic tree of the concatenated *E1E2L2L1* sequences at the nucleotide and amino acid levels. Out of the 354 PV genomes, we identified 339 with an intergenic region (of at least 1 nucleotide) between the *E2* and *L2* genes. Of these, 83 contain an E5 ORF in the inter-E2–L2 region (fig. 1 and fig. S1). The E5 ORFs have a median size of 144 nucleotides (min: 126, max: 306). Based on the size of inter-E2–L2 region in which E5s are present (min: 289, median: 517, max: 938), we identified four PV clades containing an intergenic region selecting for a minimum size of 250 (min: 262, median: 512, max: 1579). This threshold is below the minimum of 289 nucleotides to allow for inclusion of possible unidentified PV lineages containing unknown E5-like ORFs in the inter-E2–L2 region. The identified clades are indicated with a star in fig. 1 and fig. S1, and are located in the four PV crown groups: Alpha-Omikron (coloured red), Delta-Zeta (coloured blue), Lambda-Mu (coloured yellow), and Beta-Xi (coloured green). Additionally, three recombinant bottlenose dolphin PVs (TtPV1-3) belonging to the *UpsilonPV* genus, also present an inter-E2–L2 region. Only the clades identified in the Alpha-Omikron and Delta-Zeta crown groups, have an E5 ORF present within the inter-E2–L2 region. The two other clades that locate within the Lambda-Mu and Beta-Xi crown groups also contain this relatively long intergenic region. Although, for these clades the inter-E2–L2 region does not contain any apparent ORFs. Interestingly, an ORF named *E5* is present in the Lambda-Mu clade in two rabbit PV genomes (SfPV1 and OcPV1), where no intergenic non-coding region is present and E5 largely overlaps with both the *E2* and *L2* genes in the case of SfPV1 and with *L2* in the case of OcPV1. There are other cases, like HPV16, where E5 partially overlaps with the *E2* gene. Nonetheless this overlap is small (4 nucleotides) compared to the almost complete overlap of E5

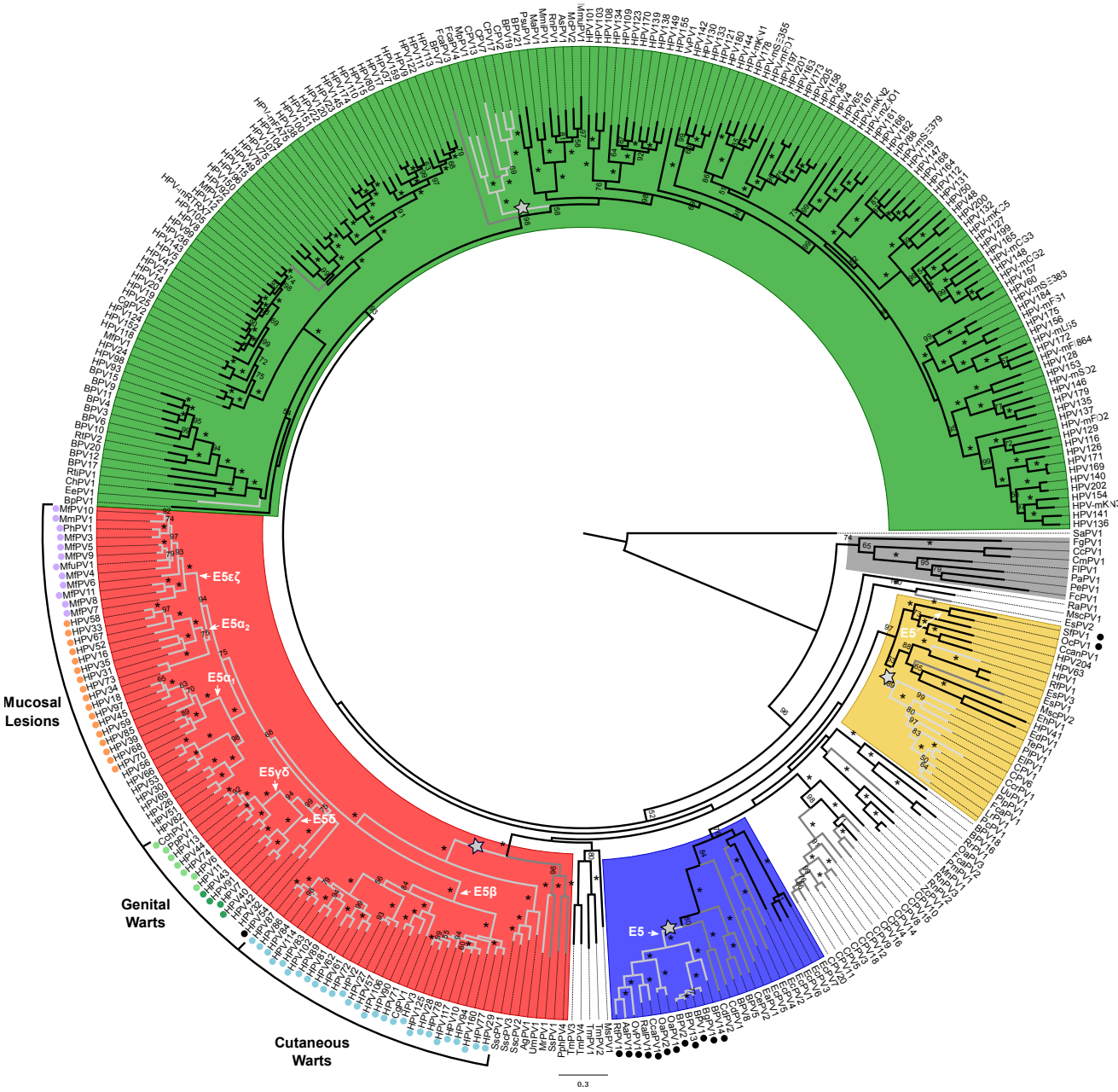

Figure 1. PV phylogenetic reconstruction and identification of clades with an intergenic E2–L2 region. Best-known maximum likelihood phylogenetic tree of the concatenated *E1E2L2L1* amino acid sequences of 343 PVs. Color code highlights the four PV crown groups: red, Alpha-OmikronPVs; green, Beta-XiPVs; yellow, Lambda-MuPVs; blue, Delta-ZetaPVs; gray, a yet unclassified crown group consisting of PVs infecting birds and turtles; and white, PVs without well-supported phylogenetic relationships. Outer labels, *Mucosal Lesions*, *Genital Warts* and *Cutaneous Warts*, indicate the most common tropism for the *AlphaPVs*. Values on branches correspond to ML bootstrap support values. Asterisks indicate a maximal support of 100, and values under 50 are not shown. Branches in light-gray correspond to PV genomes containing an inter-E2–L2 region longer than 250 nt; branches in dark-gray correspond to PV genomes with an inter-E2–L2 region longer than 125 nt. The basal nodes of the four clades containing a relatively long intergenic region between the E2 and the L2 ORFs are labelled with a star. The basal node of the lineages containing an E5 coding sequence is indicated with an arrow, and the corresponding terminal taxa are labelled with a color-coded dot indicating the E5 type. Purple dots indicate: E5$\varepsilon\zeta$, orange dots: E5α, light green dots: E5$\gamma\delta$, dark green dots: E5δ, blue dots: E5β, and black dots are lineages containing unclassified E5 types.

ORFs with *L2* in the rabbit PV genomes. All things being equal, the E5 ORFs in the rabbit PV genomes seem unique in a way that no inter-E2–L2 region is present at all.

In order to determine whether the E5 ORFs in the different PV crown groups share a single common ancestor, we tested for common ancestry using BAli-Phy (as described in de Oliveira Martins and Posada, 2014). We named the clades according to their coloured crown groups, therefore we have the red clade (including 69 E5 sequences), the blue clade (12 E5 sequences), and the yellow clade (2 E5 sequences). For the common ancestry test, trees are inferred for all groups combined as well as separately (see Materials an Methods). Therefore, we could not include the yellow clade in this test, as this clade contains only two sequences and no trees can be inferred. We performed the analysis on the full data set (excluding the two yellow clade sequences) containing 81 sequences and on a reduced dat set containing 24 sequences; twelve representative E5 sequences from the red clade and the twelve E5 sequences from the blue clade. We made the choice between the alternative hypotheses *Common Ancestry* (CA) and *Independent Origin* (IO) by computing the marginal likelihoods using the stabilized harmonic mean estimator. We ran our analysis under the null hypothesis of CA of the E5 ORF. Therefore, we have ΔBF = log[Prob(CA)] - log[Prob(IO)], such that positive values support CA and negative values support IO. Other statistics that we take into account are the alignment length and the Bayesian tree length, calculated as the sum of the branch lengths. For both the alignment length and tree length, lower values support the best model.

	full data set				reduced data set			
Model	P(data|M)	ΔBF	ali length	tree length	P(data|M)	ΔBF	ali length	tree length
H0: (red-blue)	-7129.167	0	402	26.946	-2608.666	0	344	10.893
H1: red + blue	-7139.029	9.862	373	24.301	-2617.114	8.448	335	10.706

Table 1. Hypothesis testing on the origin of the E5 ORFs in the Alpha-Omikron (red) and Delta-Zeta (blue) PV crown groups. For each hypothesis tested, common ancestry (H0) and independent origin (H1), we show the marginal likelihood (P(data|M)) value, the ΔBF, the alignment (ali) length and tree length. Cells highlighted in gray indicate the best-supported scenario for the respective statistic.

The results are contradictory between the different statistics tested. On the one hand, based on the likelihood the best supported model is CA for the E5 ORFs in the Alpha-Omikron and Delta-Zeta PV crown groups (table 1). Nonetheless, the difference in Log likelihood (ΔBF) between the CA and IO hypotheses is very small for both the full and reduced data sets. On the other hand, the alignment length and tree length statistics support the IO hypothesis. Previous approaches of other type of CA tests have shown to give misleading conclusions on alignments without any phylogenetic structure (Koonin and Wolf, 2010), as well as on unrelated families of protein coding sequences (Yonezawa and Hasegawa, 2012). As these approaches all started from a fixed alignment there could be an initial bias towards CA (de Oliveira Martins and Posada, 2014; Theobald, 2011; Yonezawa and Hasegawa, 2010). The BAli-Phy approach used here partly reduces this bias, as it starts from unaligned sequences and estimates simultaneously the alignment and the phylogeny. Given the inconclusive results, we performed a random permutation test as described in de Oliveira Martins and Posada, 2016. In this test the columns of the alignment for one of the groups are randomly shuffled and statistics are recalculated after realignment. Contrary to the BAli-Phy test, all trees are produced within a maximum likelihood (ML) framework (see Materials and Methods). We performed this test on both the full and the reduced data sets, using 100 iterations. For each iteration we recovered the ML tree length and Log likelihood, and estimated the empirical distribution of these value. If the E5 ORFs have an IO, we expect lower ML tree length and superior Log likelihood values for this hypothesis (H1). Our results show that for both the full and reduced data sets we obtained significant differences between the ML tree length distributions of CA and IO, where the IO hypothesis is favoured fig. S2A-B. However, for the Log likelihood distributions there is no significant difference between CA and IO for the full data set, and for the reduced data set the CA hypothesis is slightly favoured fig. S2C-D.

The initial idea of the permutation test was to resort only to simple summary statistics such as the ML tree length, rather than to rely on Log likelihood values (de Oliveira Martins and Posada, 2016). If we only regard the ML tree length values, IO for the red and blue clades is suggested to be the best supported model. Nevertheless, we can not ignore the Log likelihood values of the permutation test nor the results of the BAli-Phy test, and therefore, we cannot make a conclusive choice between the alternative hypotheses CA and IO. Finally, when we look at the final trees produced by BAli-Phy fig. S3, we observe that the branch lengths leading to each group are long compared to the other branches, suggesting that IO is the preferred model. In addition, these trees suggest that the E5 ORFs within the *AlphaPV* (red clade) do not originate from a single CA, which may have introduced a bias in our CA test.

Do the E5 ORFs Present in the Genomes within the AlphaPV Clade Have a Common Ancestor?

In the *AlphaPV* clade within the Alpha-Omikron crown group (red), the six E5 types are present in five different clades (fig. 1 and fig. S1). E5α exists in two different clades of PVs associated to mucosal lesions, hereafter named E5α1 and E5α2,

	full data set				reduced data set			
Model	P(data|M)	ΔBF	ali length	tree length	P(data|M)	ΔBF	ali length	tree length
H0: (α1-α2-β-$\gamma\delta$-δ-$\varepsilon\zeta$)	-6400.049	0	305	1.059	-3288.708	0	216	1.207
H1: (α1-α2-$\gamma\delta$-δ-$\varepsilon\zeta$) + β	-6415.579	15.530	328	2.238	-3300.208	11.500	275	2.533
H2: (α1-α2-$\gamma\delta$-δ) + β + $\varepsilon\zeta$	-6460.830	60.781	370	3.288	-3336.950	48.242	344	3.581
H3: (α1-α2-$\varepsilon\zeta$) + ($\gamma\delta$-δ)+ β	-6460.851	60.802	401	3.329	-3333.322	44.614	384	3.689
H4: (α1-α2-$\varepsilon\zeta$) + β + $\gamma\delta$ + δ	-6515.861	115.812	444	4.306	-3388.247	99.539	431	4.641
H5: (α1-α2) + ($\gamma\delta$-δ)+ β + $\varepsilon\zeta$	-6491.504	91.455	438	4.431	-3362.185	73.477	414	4.767
H6: α1 + α2 + $\gamma\delta$ + δ + β + $\varepsilon\zeta$	-6609.832	209.783	551	6.489	-3472.122	183.414	535	6.879

Table 2. Hypothesis testing on the origin of the E5 ORFs within the *AlphaPV* clade (red). For each hypothesis tested, common ancestry (H0) and independent origins (H1-H6), we show the marginal likelihood (P(data|M)) value, the ΔBF, the alignment (ali) length and tree length. Cells highlighted in gray indicate the best-supported scenario for the respective statistic.

consisting of eight and nine sequences respectively. E5β is present in all PVs associated to cutaneuous warts, consisting of 28 sequences. E5δ exists in all PVs associated to anogenital warts. Of these, only four PV genomes contain E5δ in isolation. The other seven PV genomes contain two E5 types; E5γ and E5δ, hereafter named E5$\gamma\delta$. Finally, E5$\varepsilon\zeta$ is present in twelve non-human *AlphaPV* genomes that infect *Cercopithecinae* and that are associated to mucosal lesions.

The BAli-Phy trees obtained in the CA test above, suggest that the E5 ORFs within *AlphaPVs* may have an IO fig. S3. These trees, that are based on the E5 amino acid sequences, show a clear separation depending on the clinical presentation of the infections: mucosal lesions (E5α1, E5α2, and E5$\varepsilon\zeta$), genital warts (E5γ, and E5$\gamma\delta$), and cutaneous warts (E5β). One exception is HPV54, which has an unclassified E5 type and is associated to genital warts. This PV clusters with the E5α1 type of mucosal lesions. To address whether the E5 ORFs present in the genomes of the *AlphaPVs* have a CA, we applied the same procedures as described above. We considered different plausible IO scenarios based on the E5 types and the phylogeny of the *AlphaPVs* (fig. 1 and fig. S1). The BAli-Phy analysis showed that the CA hypothesis was the best-supported model for all statistics, while the hypothesis of each clade having an IO (H6) had the lowest support (table 2). The second best-supported IO model (H1) –where E5β has an IO– has a small difference in Log likelihood with the CA model (H0). As in the results described above, the random permutation tests disagree with the results of the BAli-Phy approach. The results of the random permutation test suggest that based on ML tree lenght the IO H6 is the best supported model, while based on Log likelihood the CA model (H0) and IO H1 model are equally probable (fig. S4). Although the CA tests performed here give inconclusive results, the IO H1 model is also supported by the trees produced, where long branches separate E5β and the other E5 types. In this scenario E5α1, E5α2, $\gamma\delta$, E5δ, and E5$\varepsilon\zeta$ (encoded in PVs with mucosal and anogenital tropism) have a CA, but E5β (encoded in PVs with cutaneous tropism) has an IO. We therefore propose that at least E5α1, E5α2, $\gamma\delta$, E5δ, and E5$\varepsilon\zeta$ have a single ancestor, and originated from the same recombination donor and/or gained access to the ancestral genome through a single integration event. Further tests are needed to conclude whether E5β originated from the same ancestor as the other E5 types or whether it has an independent origin.

In *AlphaPVs*, The Evolutionary History of The inter-E2–L2 Region is Different from That of E5

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 and the *E5* ORF with the topologies obtained for each of the PV ORFs (*E6*, *E7*, *E1*, *E2*, *L2* and *L1*) as well as for the non-coding URR. We calculated the weighted and unweighted Robinson-Foulds (RF) distances between paired trees and we performed a correspondence analysis in order to identify similarities among the topologies of the PV gene trees (fig. 2). The first axis captured a large fraction of the variance (more than 50% in the weighted RF distance) and splitted the *E5* and the inter-E2–L2 reconstructions from those of core PV genes. The second axis contained more than 15% of the overall variance and splitted the topologies of the early genes *E6*, *E7*, *E1* and *E2*, from those of the late genes *L2* and *L1*, and the URR. Interestingly, in this second axis the inter-E2–L2 clustered together with the late genes, while the *E5* genes clustered together with the early genes. These results suggest that the inter-E2–L2 region and E5 may have different evolutionary histories.

0.1 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 appearing by chance in these randomly generated DNA sequences and we computed their nucleotide length. (fig. 3) shows the cumulative frequency of the

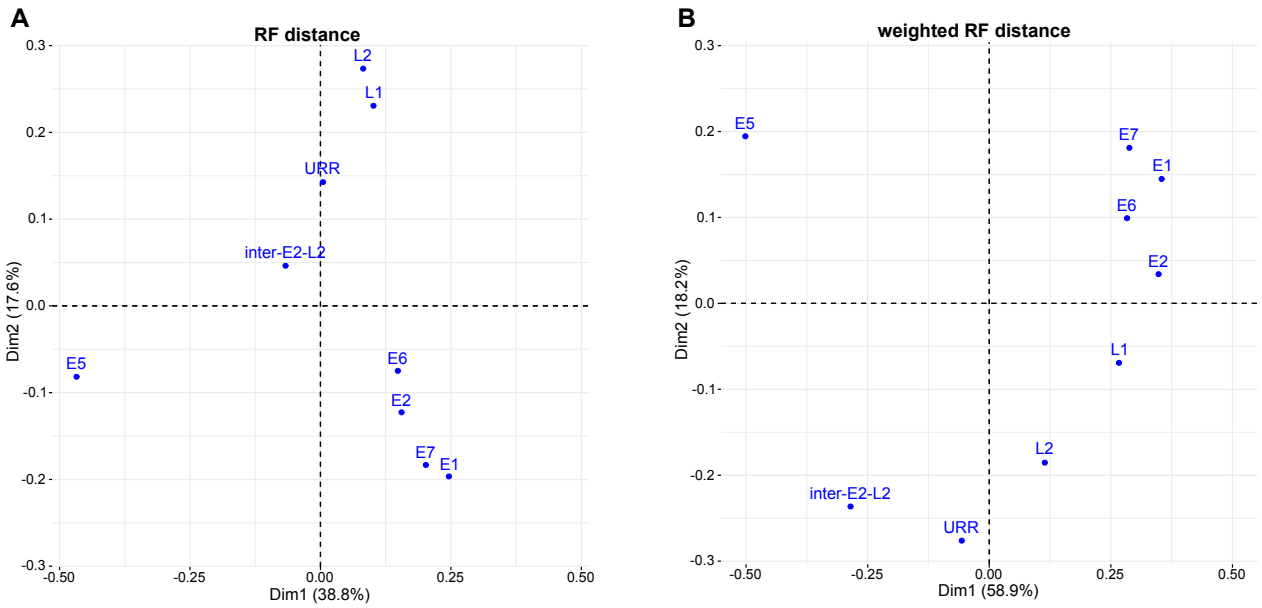

Figure 2. Correspondence analysis of the unweighted (A) and weighted (B) Robinson-Foulds tree distance comparing maximum likelihood trees construced for each of the PV ORFs, the inter-E2–L2 region, and the URR.

E5 genes length and of the random ORFs. A one-way ANOVA followed by a post-hoc Tukey honestly significant difference test was performed, with *gene* as a factor (table S2) shows that ORFs in randomly generated sequences are shorter than any of the E5 ORFs (Tukey HSD: $p < 0.0001$).

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).

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

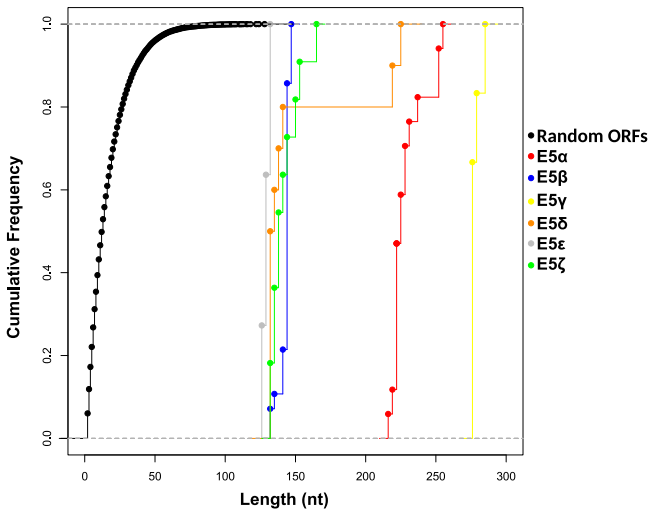

Figure 3. Cumulative frequency of the nucleotide length for each group of the *E5* genes and random ORFs. The different types of E5 are colour-coded as indicated in the legend.

were found as expected 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, $p < 0.001$).

In PVs, codon usage preferences (CUPrefs) are different from those of their hosts, and viral genes with similar expression patterns display similar CUPrefs (Félez-Sánchez *et al.*, 2015). To corroborate whether the 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 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 the best-studied E5 proteins are transmembrane proteins, we hypothesized that a *bona fide* E5 protein should be more hydrophobic than expected by chance. We calculated the GRAVY index for the E5 proteins as well as for the ORFs enconded in 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).

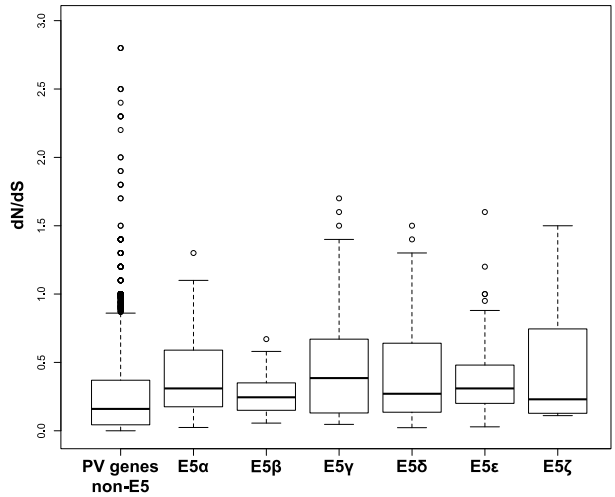

Figure 4. dN/dS values for each group of the *E5* genes and the other PV genes (E1, E2, E6, E7, L1 and L2).

Discussion

Reconstructing 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, if we eventually aim to understand why certain PVs are oncogenic while their close relative cause anodyne infections. 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 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 (Doorbar *et al.*, 2012; Forman *et al.*, 2012). All *AlphaPVs* present a region between the *E2* and *L2* genes, potentially encoding in all cases for conserved ORFs. With few exceptions (Cartin and Alonso, 2003), 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 (Bravo and Alonso, 2004). These E5α behave as oncoproteins, promoting cell division and allowing the infected cells to avoid immune recognition (Bra, 2005; Ashrafi *et al.*, 2005; Suprynowicz *et al.*, 2008).

Since all 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

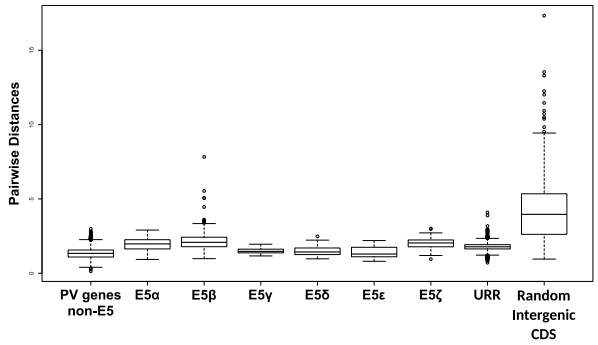

Figure 5. Pairwise distances between *AlphaPVs* for the all genes, the URR, and a set of randomly generated intergenic CDS. All values have been normalized to the corresponding *L1* pairwise distances.

displaying an intergenic region between *E2* and *L2* are not monophyletic, and belong instead to four polyphyletic clades in the PV tree (fig. 1 and fig. S1). It can 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 (*E6* and *E7*) in certain PVs (Van Doorslaer and McBride, 2016). Alternatively, the different inter-E2–L2 regions present in extant PV genomes could derive from one or from several genetic events in which an ancestral sequence could have gained access to one ancestral PV genome.

We can formulate two main non-exclusive mechanisms to explain the origin of the four 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 (Flores and Lambert, 1997). During the PV bidirectional replication, the replication forks start at the URR and converge opposite to the origin of replication, which 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 (Roerink *et al.*, 2014), which may lead to the emergence of an ancestral inter-E2–L2 region in one or in several instances during the evolutionary history of PVs.

Recombination can also be invoked as a mechanism that could 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 (Flores and Lambert, 1997; McBride, 2017), generating large linear molecules of concatenated viral genomes (Dasgupta *et al.*, 1992). Unidirectional replication relies on homologous recombination, as this mechanisms is required for resolving, excising and recircularizing the concatenated genomes into individual plasmid genomes (Gillespie *et al.*, 2012; Mehta and Laimins, 2018; Sakakibara *et al.*, 2013). Additionally, productive replication concurs with a virus-mediated impairment of the cellular DNA damage repair mechanisms (Chappell *et al.*, 2016; Wallace *et al.*, 2017), 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) (Bravo and Alonso, 2004; Narechania *et al.*, 2005). 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) (Gottschling *et al.*, 2011a; Rector *et al.*, 2008; Robles-Sikisaka *et al.*, 2012). 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 (Bennett *et al.*, 2008; Woolford *et al.*, 2007).

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 (García-Pérez *et al.*, 2014; Rector *et al.*, 2007; Schulz *et al.*, 2009). Despite the

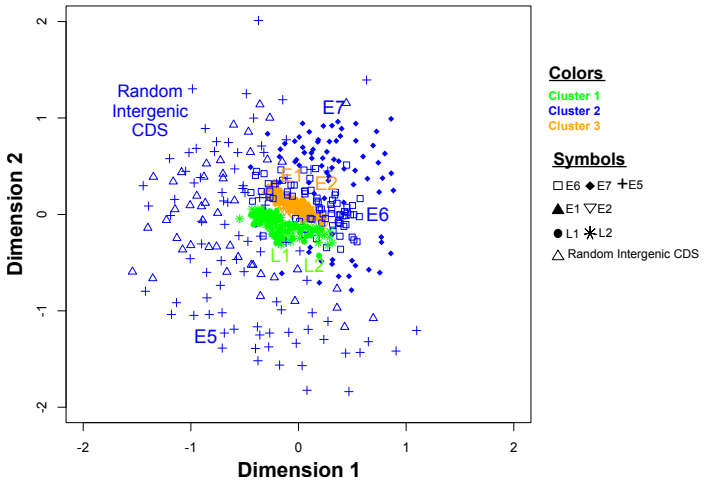

Figure 6. Multidimensional Scaling (MDS) plot of codon usage preferences for the *AlphaPV* ORFs. The ORFs were independently clustered by an unsupervised two-step clustering algorithm. The best assembly included three clusters, displayed onto the MDS plot as with a color code, composed respectively by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

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 (Rector *et al.*, 2007). 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 and fig. S1) 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.

The putative ORFs that emerged in the inter-E2–L2 region are often named *E5*. Notwithstanding, our results suggest the E5 proteins encoded in the different clades may not be monophyletic. Specifically, this would 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 (Ashby *et al.*, 2001; Venuti *et al.*, 2011). Yet, the E5 sequences are short and display similar amino acid composition because of their transmembrane nature, and these two facts combined reduce the power of the algorithms used to pinpoint common ancestry between genes. Further tests are needed to resolve the riddle on the origin of E5s: either *in silico* by improving the CA test or experimentally by evolving a predicted ancestor(s) of E5 or by performing *de novo* gene evolution on the inter–E2-L2 region.

When restricting our analysis to the E5 ORFs within the *AlphaPVs*, we found support for monophyly (table 2), indicating that a single event on the backbone of the ancestral *AlphaPV* genome could have led to its emergence. Nevertheless, the alternative hypothesis of E5β having an independent origin was not significanlty worse. This hypothesis is supported by the different tropism of lineages within *AlphaPVs*: those containing an E5β display an essentially cutaneous tropism, while all other lineages encoding for E5α, E5γ, E5δ, E5ε, and E5ζ, display a mucosal tropism. 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% (Bravo and Alonso, 2004). Phylogenetic reconstruction based on the *E5* ORFs showed a star-like pattern with the main branches emerging close to a putative central point (Bravo and Alonso, 2004). 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 *AlphaPV* genomes. Our interpretation based on the evidence here provided is as follows. Under the hypothesis of recombination, within *AlphaPVs*, a non-coding sequence was integrated in a single event between the early and the late genes in the genome of an ancestral PV lineage, which infected the ancestors of Old World monkeys and apes. Mutations in this originally non-coding region gave birth to the different *E5* ORFs. Such *de novo* birth of new protein-coding sequences from non-coding genomic regions is not unfamiliar and has been reported in for example *Drosophila* (Lev, 2006; Zhou *et al.*, 2008), yeast (Cai *et al.*, 2008) and mammals (Toll-Riera *et al.*, 2009). Experimentally, it has been shown that random, E5-like short peptide sequences can indeed insert in the cellular membranes and display a biological activity (Chacón *et al.*, 2014). Using genetic selection, these small artificial transmembrane 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 (Chacón *et al.*, 2014). Therefore we consider *de novo* birth of the *E5* genes

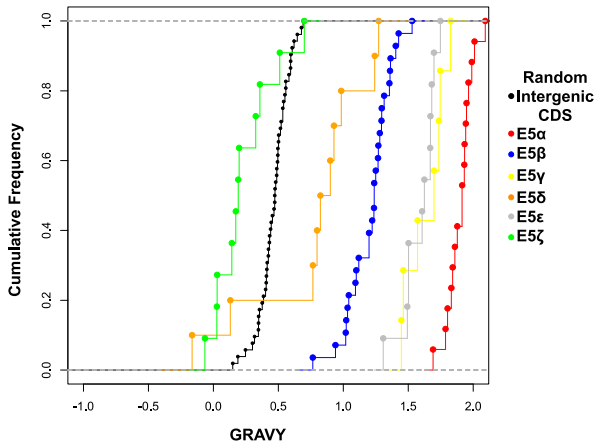

Figure 7. Cumulative frequency of the GRAVY index for the *E5* ORFs and the randomly generated intergenic CDS.

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, for most of these ORFs we do not have evidence of their expression *in vivo*. Moreover, the possible independent origins of *E5*, raise 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 orthologs of the *E5* genes are not found in other viruses or in their hosts, we have studied the *E5* ORFs in the context of orphan genes. In agreement with studies of orphan genes in other species (Carvunis *et al.*, 2012; Toll-Riera *et al.*, 2009; Wolf *et al.*, 2009), 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 (Albà and Castresana, 2005; Palmieri *et al.*, 2014; Toll-Riera *et al.*, 2009). However, a *bona fide* gene should be longer than expected by chance (Schlötterer, 2015), and this is what we actually find 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.* found evidence to support this hypothesis (Ruiz-Orera *et al.*, 2018). 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 extant E5s may be functionally relevant. Gene CUPrefs have a strong 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 may remain untranslated (Ruiz-Orera *et al.*, 2018). We have thus evaluated whether CUPrefs in *E5* resemble those in other *AlphaPV* genes. 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 (Félez-Sánchez *et al.*, 2015). 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 life cycle of oncogenic human *AlphaPVs*.

In summary, 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 (Bra, 2005; Suprynowicz *et al.*, 2008) and the down-regulation of surface MHC class I molecules (Campo *et al.*, 2010; Cartin and Alonso, 2003) 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 and S2 are available online. The data from the Bali-Phy analyses and the random permutation test are available at https://github.com/anoukwillemsen/ONCOGENEVOL.

Acknowledgements

We are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing and storage resources. The authors acknowledge the IRD itrop HPC (South Green Platform) at IRD montpellier for providing HPC resources that have contributed to the research results reported within this paper. 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

2005. The E5 protein of the human papillomavirus type 16 modulates composition and dynamics of membrane lipids in keratinocytes. *Archives of Virology*, 150(2): 231–246.

2006. 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(26): 9935–9939.

Albà, M. M. and Castresana, J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. *Molecular Biology and Evolution*, 22(3): 598–606.

Ashby, A. D., Meagher, L., Campo, M. S., and Finbow, M. E. 2001. E5 transforming proteins of papillomaviruses do not disturb the activity of the vacuolar H+-ATPase. *Journal of General Virology*, 82(10): 2353–2362.

Ashrafi, G. H., Tsirimonaki, E., Marchetti, B., O'Brien, P. M., Sibbet, G. J., Andrew, L., and Campo, M. S. 2002. Down-regulation of MHC class I by bovine papillomavirus E5 oncoproteins. *Oncogene*, 21(2): 248–259.

Ashrafi, G. H., Haghshenas, M. R., Marchetti, B., O'Brien, P. M., and Campo, M. S. 2005. E5 protein of human papillomavirus type 16 selectively downregulates surface HLA class I. *International Journal of Cancer*, 113(2): 276–283.

Bennett, M. D., Woolford, L., Stevens, H., Van Ranst, M., Oldfield, T., Slaven, M., O'Hara, A. J., Warren, K. S., and Nicholls, P. K. 2008. Genomic characterization of a novel virus found in papillomatous lesions from a southern brown bandicoot (Isoodon obesulus) in Western Australia. *Virology*, 376(1): 173–182.

Bravo, I. G. and Alonso, A. 2004. Mucosal Human Papillomaviruses Encode Four Different E5 Proteins Whose Chemistry and Phylogeny Correlate with Malignant or Benign Growth. *Journal of Virology*, 78(24): 13613–13626.

Bravo, I. G. and Félez-Sánchez, M. 2015. Papillomaviruses. *Evolution, Medicine, and Public Health*, 2015(1): 32–51.

Cai, J., Zhao, R., Jiang, H., and Wang, W. 2008. De Novo Origination of a New Protein-Coding Gene in Saccharomyces cerevisiae. *Genetics*, 179(1): 487–496.

Campo, M., Graham, S., Cortese, M., Ashrafi, G., Araibi, E., Dornan, E., Miners, K., Nunes, C., and Man, S. 2010. HPV-16 E5 down-regulates expression of surface HLA class I and reduces recognition by CD8 T cells. *Virology*, 407(1): 137–142.

Cartin, W. and Alonso, A. 2003. The human papillomavirus HPV2a E5 protein localizes to the Golgi apparatus and modulates signal transduction. *Virology*, 314(2): 572–579.

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E., and Vidal, M. 2012. Proto-genes and de novo gene birth. *Nature*, 487(7407): 370–374.

Castresana, J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, 17(4): 540–552.

Chacón, K. M., Petti, L. M., Scheideman, E. H., Pirazzoli, V., Politi, K., and DiMaio, D. 2014. De novo selection of oncogenes. *Proceedings of the National Academy of Sciences*, 111(1): E6–E14.

Chappell, W. H., Gautam, D., Ok, S. T., Johnson, B. A., Anacker, D. C., and Moody, C. A. 2016. Homologous Recombination Repair Factors Rad51 and BRCA1 Are Necessary for Productive Replication of Human Papillomavirus 31. *Journal of Virology*, 90(5): 2639–2652.

Conrad, M., Bubb, V. J., and Schlegel, R. 1993. 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(10): 6170–8.

Dasgupta, S., Zabielski, J., Simonsson, M., and Burnett, S. 1992. Rolling-circle replication of a high-copy BPV-1 plasmid. *Journal of Molecular Biology*, 228(1): 1–6.

de Oliveira Martins, L. and Posada, D. 2014. Testing for Universal Common Ancestry. *Systematic Biology*, 63(5): 838–842.

de Oliveira Martins, L. and Posada, D. 2016. Infinitely long branches and an informal test of common ancestry. *Biology Direct*, 11(1): 19.

DiMaio, D. and Mattoon, D. 2001. Mechanisms of cell transformation by papillomavirus E5 proteins.

DiMaio, D. and Petti, L. M. 2013. The E5 proteins. *Virology*, 445(1-2): 99–114.

Doorbar, J., Quint, W., Banks, L., Bravo, I. G., Stoler, M., Broker, T. R., and Stanley, M. A. 2012. The biology and life-cycle of human papillomaviruses.

Doron-Faigenboim, A. and Pupko, T. 2006. A Combined Empirical and Mechanistic Codon Model. *Molecular Biology and Evolution*, 24(2): 388–397.

Doron-Faigenboim, A., Stern, A., Mayrose, I., Bacharach, E., and Pupko, T. 2005. Selection: A server for detecting evolutionary forces at a single amino-acid site. *Bioinformatics*, 21(9): 2101–2103.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5): 1792–1797.

Félez-Sánchez, M., Trösemeier, J.-H., Bedhomme, S., González-Bravo, M. I., Kamp, C., and Bravo, I. G. 2015. Cancer, Warts, or Asymptomatic Infections: Clinical Presentation Matches Codon Usage Preferences in Human Papillomaviruses. *Genome Biology and Evolution*, 7(8): 2117–2135.

Flores, E. R. and Lambert, P. F. 1997. Evidence for a switch in the mode of human papillomavirus type 16 DNA replication during the viral life cycle. *Journal of virology*, 71(10): 7167–79.

Forman, D., de Martel, C., Lacey, C. J., Soerjomataram, I., Lortet-Tieulent, J., Bruni, L., Vignat, J., Ferlay, J., Bray, F., Plummer, M., and Franceschi, S. 2012. Global Burden of Human Papillomavirus and Related Diseases. *Vaccine*, 30: F12–F23.

García-Pérez, R., Ibáñez, C., Godínez, J. M., Aréchiga, N., Garin, I., Pérez-Suárez, G., de Paz, O., Juste, J., Echevarría, J. E., and Bravo, I. G. 2014. 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(1): 94–104.

Gillespie, K. A., Mehta, K. P., Laimins, L. A., and Moody, C. A. 2012. Human Papillomaviruses Recruit Cellular DNA Repair and Homologous Recombination Factors to Viral Replication Centers. *Journal of Virology*, 86(17): 9520–9526.

Gottschling, M., Bravo, I. G., Schulz, E., Bracho, M. A., Deaville, R., Jepson, P. D., Bressem, M.-F. V., Stockfleth, E., and Nindl, I. 2011a. Modular organizations of novel cetacean papillomaviruses. *Molecular Phylogenetics and Evolution*, 59(1): 34–42.

Gottschling, M., Göker, M., Stamatakis, A., Bininda-Emonds, O. R. P., Nindl, I., and Bravo, I. G. 2011b. Quantifying the phylodynamic

forces driving papillomavirus evolution. *Molecular biology and evolution*, 28(7): 2101–13.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3): 307–321.

Hughes, A. L. and Hughes, M. A. K. 2005. Patterns of nucleotide difference in overlapping and non-overlapping reading frames of papillomavirus genomes. *Virus research*, 113(2): 81–8.

Katoh, K. and Standley, D. M. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4): 772–780.

Koonin, E. V. and Wolf, Y. I. 2010. The common ancestry of life. *Biology direct*, 5: 64.

Kyte, J. and Doolittle, R. F. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1): 105–32.

McBride, A. A. 2017. Mechanisms and strategies of papillomavirus replication. *Biological Chemistry*, 398(8): 919–927.

Mehta, K. and Laimins, L. 2018. Human Papillomaviruses Preferentially Recruit DNA Repair Factors to Viral Genomes for Rapid Repair and Amplification. *mBio*, 9(1): e00064–18.

Moody, C. A. and Laimins, L. A. 2010. Human papillomavirus oncoproteins: pathways to transformation. *Nature Reviews Cancer*, 10(8): 550–560.

Münger, K., Scheffner, M., Huibregtse, J. M., and Howley, P. M. 1992. Interactions of HPV E6 and E7 oncoproteins with tumour suppressor gene products. *Cancer surveys*, 12: 197–217.

Narechania, A., Chen, Z., DeSalle, R., and Burk, R. D. 2005. Phylogenetic incongruence among oncogenic genital alpha human papillomaviruses. *Journal of virology*, 79(24): 15503–10.

Palmieri, N., Kosiol, C., and Schlötterer, C. 2014. The life cycle of Drosophila orphan genes. *eLife*, 3: e01311.

Petti, L. M., Reddy, V., Smith, S. O., and DiMaio, D. 1997. 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(10): 7318–7327.

Pim, D., Collins, M., and Banks, L. 1992. Human papillomavirus type 16 *E5* gene stimulates the transforming activity of the epidermal growth factor receptor. *Oncogene*, 7(1): 27–32.

R Core Team 2014. R: A language and environment for statistical computing.

Rector, A., Lemey, P., Tachezy, R., Mostmans, S., Ghim, S.-J., Van Doorslaer, K., Roelke, M., Bush, M., Montali, R. J., Joslin, J., Burk, R. D., Jenson, A. B., Sundberg, J. P., Shapiro, B., and Van Ranst, M. 2007. Ancient papillomavirus-host co-speciation in Felidae. *Genome Biology*, 8(4): R57.

Rector, A., Stevens, H., Lacave, G., Lemey, P., Mostmans, S., Salbany, A., Vos, M., Van Doorslaer, K., Ghim, S.-J., Rehtanz, M., Bossart, G. D., Jenson, A. B., and Van Ranst, M. 2008. Genomic characterization of novel dolphin papillomaviruses provides indications for recombination within the Papillomaviridae. *Virology*, 378(1): 151–161.

Redelings, B. D. and Suchard, M. A. 2005. Joint bayesian estimation of alignment and phylogeny. *Systematic Biology*, 54(3): 401–418.

Robinson, D. and Foulds, L. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2): 131–147.

Robles-Sikisaka, R., Rivera, R., Nollens, H. H., St. Leger, J., Durden, W. N., Stolen, M., Burchell, J., and Wellehan, J. F. 2012. Evidence of recombination and positive selection in cetacean papillomaviruses. *Virology*, 427(2): 189–197.

Roerink, S. F., Schendel, R., and Tijsterman, M. 2014. Polymerase theta-mediated end joining of replication-associated DNA breaks in C.

elegans. *Genome Research*, 24(6): 954–962.

Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Cañas, J. L., Messeguer, X., and Albà, M. M. 2018. Translation of neutrally evolving peptides provides a basis for de novo gene evolution.

Sakakibara, N., Chen, D., and McBride, A. A. 2013. Papillomaviruses Use Recombination-Dependent Replication to Vegetatively Amplify Their Genomes in Differentiated Cells. *PLoS Pathogens*, 9(7): e1003321.

Schlötterer, C. 2015. Genes from scratch–the evolutionary fate of de novo genes. *Trends in genetics : TIG*, 31(4): 215–9.

Schulz, E., Gottschling, M., Bravo, I. G., Wittstatt, U., Stockfleth, E., and Nindl, I. 2009. 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(3): 626–633.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312–1313.

Straight, S. W., Hinkle, P. M., Jewers, R. J., and McCance, D. J. 1993. 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(8): 4521–4532.

Suchard, M. A. and Redelings, B. D. 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, 22(16): 2047–2048.

Suprynowicz, F. A., Disbrow, G. L., Krawczyk, E., Simic, V., Lantzky, K., and Schlegel, R. 2008. HPV-16 E5 oncoprotein upregulates lipid raft components caveolin-1 and ganglioside GM1 at the plasma membrane of cervical cells. *Oncogene*, 27(8): 1071–1078.

Suyama, M., Torrents, D., and Bork, P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34(Web Server): W609–W612.

Theobald, D. L. 2011. On universal common ancestry, sequence similarity, and phylogenetic structure: the sins of P-values and the virtues of Bayesian evidence. *Biology Direct*, 6(1): 60.

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., and Mar Albà, M. 2009. Origin of primate orphan genes: A comparative genomics approach. *Molecular Biology and Evolution*, 26(3): 603–612.

Tomaić, V. 2016. Functional Roles of E6 and E7 Oncoproteins in HPV-Induced Malignancies at Diverse Anatomical Sites. *Cancers*, 8(10): 95.

Van Doorslaer, K. and McBride, A. A. 2016. Molecular archeological evidence in support of the repeated loss of a papillomavirus gene. *Scientific Reports*, 6(1): 33028.

Venuti, A., Paolini, F., Nasir, L., Corteggio, A., Roperto, S., Campo, M. S., and Borzacchiello, G. 2011. Papillomavirus E5: The smallest oncoprotein with many functions.

Wallace, N. A., Khanal, S., Robinson, K. L., Wendel, S. O., Messer, J. J., and Galloway, D. A. 2017. High-Risk Alphapapillomavirus Oncogenes Impair the Homologous Recombination Pathway. *Journal of Virology*, 91(20): e01084–17.

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V., and Lipman, D. J. 2009. 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(18): 7273–7280.

Woolford, L., Rector, A., Van Ranst, M., Ducki, A., Bennett, M. D., Nicholls, P. K., Warren, K. S., Swan, R. A., Wilcox, G. E., and O'Hara, A. J. 2007. 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(24): 13280–13290.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.-M. K. 2000. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. *Genetics*, 155(1): 431–49.

Yonezawa, T. and Hasegawa, M. 2010. Was the universal common ancestry proved? *Nature*, 468(7326): E9–E9.

Yonezawa, T. and Hasegawa, M. 2012. Some problems in proving the existence of the universal common ancestor of life on Earth. *TheScientificWorldJournal*, 2012: 479824.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S., and Wang, W. 2008. On the origin of new genes in Drosophila. *Genome Research*, 18(9): 1446–1455.