Papillomaviruses infecting cetaceans exhibit signs of genome adaptation following a recombination event

Fanni Borvető, Ignacio G. Bravo, Anouk Willemsen

Virus Evolution, 2020, 6(1): veaa038

doi: 10.1093/ve/veaa038
Research Article

Papillomaviruses infecting cetaceans exhibit signs of genome adaptation following a recombination event

Fanni Borvető,[‡] Ignacio G. Bravo[§], and Anouk Willemsen[*,†,**]

Centre National de la Recherche Scientifique (CNRS), Laboratory MIVEGEC (CNRS IRD Univ, Montpellier), 911 Avenue Agropolis, BP 64501, 34394 Montpellier, France

Corresponding author: E-mail: anouk.willemsen@univie.ac.at

[†]Present address: University of Vienna, Centre for Microbiology and Environmental Systems Science, Division of Microbial Ecology, Vienna, Austria.

[‡]https://orcid.org/0000-0002-2532-7160

[**]https://orcid.org/0000-0002-8511-3244

[§]https://orcid.org/0000-0003-3389-3389

Abstract

Papillomaviruses (PVs) have evolved through a complex evolutionary scenario where virus–host co-evolution alone is not enough to explain the phenotypic and genotypic PV diversity observed today. Other evolutionary processes, such as host switch and recombination, also appear to play an important role in PV evolution. In this study, we have examined the genomic impact of a recombination event between distantly related PVs infecting Cetartiodactyla (even-toed ungulates and cetaceans). Our phylogenetic analyses suggest that one single recombination was responsible for the generation of extant 'chimeric' PV genomes infecting cetaceans. By correlating the phylogenetic relationships to the genomic content, we observed important differences between the recombinant and non-recombinant cetartiodactyle PV genomes. Notably, recombinant PVs contain a unique set of conserved motifs in the upstream regulatory region (URR). We interpret these regulatory changes as an adaptive response to drastic changes in the PV genome. In terms of codon usage preferences (CUPrefs), we did not detect any particular differences between orthologous open reading frames in recombinant and non-recombinant PVs. Instead, our results are in line with previous observations suggesting that CUPrefs in PVs are rather linked to gene expression patterns as well as to gene function. We show that the non-coding URR of PVs infecting cetaceans, the central regulatory element in these viruses, exhibits signs of adaptation following a recombination event. Our results suggest that also in PVs, the evolution of gene regulation can play an important role in speciation and adaptation to novel environments.

Key words: virus evolution; recombination; gene regulation; papillomavirus.

1. Introduction

Papillomaviruses (PVs) are small, non-enveloped viruses with double-stranded DNA genomes varying between 5.7 and 8.6 kb in size. The minimal PV genome consists of an upstream regulatory region (URR), an early gene region encoding the *E1* and *E2* genes, and a late gene region encoding the *L2* and *L1* genes. Other genes that are not strictly conserved in all PV genomes

are *E4* (nested within *E2*), *E5*, *E6*, *E7*, and *E10*. As the names suggest, the early genes are expressed during the early stages of PV infection, while the capsid proteins L2 and L1 are expressed during later stages.

According to the International Committee on Taxonomy of Viruses (https://talk.ictvonline.org/taxonomy/), the *Papillomaviridae* family currently consists of >50 genera and >130 species (Van Doorslaer et al. 2018). Based on the phylogenetic relationships of the concatenated early and late core genes (*E1-E2-L2-L1*) PVs have been classified into a limited number of crown groups: Alpha-Omikron, Beta-Xi, Lambda-Mu, and Delta-Zeta (Gottschling et al. 2011b; Bravo and Felez-Sanchez 2015). PVs have a wide host range, infecting bony fishes, birds, reptiles, and virtually all mammals (Antonsson and Hansson 2002; Rector and Van Ranst 2013; López-Bueno et al. 2016). However, the best-known members of the *Papillomaviridae* are PVs infecting humans, because of the clinical importance of some of these infections.

Although PVs have evolved in close relationship with their hosts, virus–host co-evolution alone is not enough to explain the phenotypic and genotypic viral diversity observed today (Bravo and Félez-Sánchez 2015; Gottschling et al. 2011b). Other processes such as host switch and recombination also play an important role in PV evolution (Rector et al. 2008; Gottschling et al. 2011b). Recombination remains a rare event for PVs, because even if individual mammals are very often infected by several different PVs at any given time, recombination requires the simultaneous presence of two different PV genomes within the same infected cell. Nevertheless, the result of a recombination event is most often conspicuous, rendering a chimeric daughter genome easily identifiable because of their differential similarities with the parental ones along the sequence. Evidence of recombination has been described within the group of PVs infecting Primates that includes the most oncogenic PVs to humans (Bravo and Alonso 2004; Narechania et al. 2005; Angulo and Carvajal-Rodríguez 2007). Another compelling example of recombination between distant viral sequences are two viruses isolated from bandicoots, where the early gene region resembles those of Polyomaviruses and the late gene region resembles those of PVs (Woolford et al. 2007; Bennett et al. 2008). However, the most noticeable lineage of recombinant PVs is a group of viral genomes isolated from different cetacean species (whales, dolphins, and porpoises), with the early gene region resembling that of PVs in the Alpha-Omikron crown group and the late gene region resembling that of PVs in the Beta-Xi crown group (Rector et al. 2008; Gottschling et al. 2011a; Robles-Sikisaka et al. 2012).

Recombination events between distantly related viruses can lead to drastic genomic changes. For example, a recombination event may change the repertoire of genes present in the genome, or modify the match between the codon usage preferences (CUPrefs) of virus and host. As a consequence, upon recombination adaptive changes may occur in both coding and non-coding regions of the viral genome. For the non-coding regions, sequence changes may occur in regulatory sites. For PVs, regulatory elements are mainly found in the URR, which contains transcription-factor binding sites (TFBSs) and other regulatory motifs that are necessary to regulate replication and transcription of the virus, with viral E1 and E2 as the central interaction partners (Bernard 2013). As an ATP-dependent DNA helicase, the PV E1 protein is essential for replication and amplification of the viral episome. Viral DNA replication is initiated by E1 binding to specific sequence motifs, such as the palindromic AT-rich E1-binding site (E1BS) and other versions of E1BSs, located within the URR (Bergvall, Melendy, and Archambault 2013). These E1BSs are often regarded as the 'origin of DNA replication'. The E2 viral protein is an essential transcription regulator that binds specifically to 12 bp motifs—E2-binding sites (E2BSs)—located mostly within the URR (McBride 2013). In addition, E2 modulates the shift from early to late transcript production, acting independently on E2BS outside the URR (Johansson et al. 2012).

In the coding regions, the CUPrefs of a virus may shift after drastic genomic changes. Since PVs depend on the host translation machinery and on the available host tRNAs, one can expect that viruses would evolve to match their CUPrefs to those of the host. Therefore, proteins required in large amounts are usually encoded by genes optimized to the host's CUPrefs while a poor match of CUPrefs generally results in lower protein production (Bahir et al. 2009). Despite this observation, it has been shown that CUPrefs of human PVs do not match those of their host, but can instead be associated to different clinical presentations of the infections; viruses causing productive lesions display CUPrefs closer to those of the host than viruses causing more oncogenic lesions (Félez-Sánchez et al. 2015). In addition, the timing of expression—early gene expression in basal epithelium versus late gene expression in differentiating epithelium—largely determines the differential CUPrefs (Zhou et al. 1999; Félez-Sánchez et al. 2015).

In this study, we have examined the recombinant cetacean PVs as well as closely related PVs infecting Cetartiodactyla (even-toed ungulates and cetaceans). To better understand what drives the evolution of these viruses, we have correlated their genomic content to their phylogenetic relationships. In particular, we have investigated whether recombinant PVs contain unique regulatory motifs and whether the recombinant and non-recombinant PVs are different in their CUPrefs. In addition, we have analysed whether viral CUPrefs are similar to those of the hosts they infect and whether macroscopic traits of the corresponding infection (e.g. clinical presentation or anatomical site of the infection) correlate with CUPrefs, motif distribution, and phylogenetic clustering. These tests allowed us to investigate the impact of recombination on the genomes of PVs infecting Cetartiodactyla.

2. Materials and methods

2.1 PV genome sequences and their characteristics

We collected the complete genomes of PVs infecting Cetartiodactyla from the Papillomavirus Episteme database (PaVE: https://pave.niaid.nih.gov/) and GenBank (https://www.ncbi.nlm.nih.gov/genbank/) between March and May 2018. The ORFs (*E10*, *E6*, *E7*, *E1*, *E2*, *E4*, *E5*, *L2*, and *L1*) and the URR of 58 reference PV genomes were extracted for subsequent analyses. For each PV genome, we collected information on the corresponding host species, the clinical presentation of the infection, the anatomical location, and viral taxonomy, as reported by the authors in the corresponding PaVE and GenBank entries or publications (Supplementary Table S1). The PVs in this study were sampled from twenty different host species that belong to seven distinct host families (Bovidae, $n = 30$; Camelidae, $n = 2$; Cervidae, $n = 11$; Delphinidae, $n = 8$; Giraffidae, $n = 1$; Phocoenidae, $n = 4$; Suidae, $n = 2$). They represent three viral crown groups: Alpha-Omikron ($n = 13$), Beta-Xi ($n = 18$), and Delta-Zeta ($n = 20$), along with several unclassified viral genomes ($n = 7$). Most of the viral genomes have been retrieved from benign epithelial lesions ($n = 47$), albeit a number of

samples correspond to malignant lesions ($n = 3$), asymptomatic infections ($n = 7$), and one eye fluid sample. The data set contains eleven recombinant PV genotypes infecting members of the Delphinidae and Phocoenidae host families. These genomes have already been identified as being recombinant by previous studies (Rector et al. 2008; Gottschling et al. 2011a; Robles-Sikisaka et al. 2012).

2.2 Phylogenetic inference

For the construction of phylogenetic trees, we first used the concatenated *E1-E2* genes and the concatenated *L2-L1* genes. Two different data sets were used, one including all PVs collected for this study, and a second one removing the recombinant PVs. The individual gene sequences were aligned at the amino acid level using MUSCLE in Geneious v8.0.5 (https://www.geneious.com/), and subsequently back-translated to nucleotides. The nucleotide alignments were filtered with Gblocks v.0.91b (Castresana 2000) to exclude the non-informative positions. The Gblocks parameters used were as follows: type of sequence: codons; minimum number of sequences for a conserved position: thirty; minimum number of sequences for a flank position: thirty; maximum number of contiguous non-conserved positions: twelve; minimum length of a block: six; allowed gap positions: all; use similarity matrices: yes. The phylogenies of the concatenated *E1-E2* and *L2-L1* alignments were used to construct maximum likelihood (ML) trees. ML phylogenetic inference was done at the nucleotide and amino acid level with RAxML 8.2.9 (Stamatakis 2014), under the GTR + Γ4 model, using 5,000 bootstrap cycles and three partitions (one for each codon position). Additional ML trees were constructed (GTR + Γ4 model, 10,000 bootstrap cycles, one partition per codon position) for each of the individual *E1*, *E2*, *L1*, and *L2* genes that were used for comparing the phylogenetic signal with the CUPrefs.

2.3 Comparison of early gene and late gene phylogenetic trees

To measure topological distances between the early (*E1-E2*) and late (*L2-L1*) gene trees, we compared pairwise distances, the Robinson–Foulds (RF) (Robinson and Foulds, 1981) distance, and the K-tree score, using Ktreedist v.1.0 (Soria-Carrasco et al. 2007). The calculated pairwise distances in the two corresponding trees were compared by a Mantel test, to evaluate whether correlation between the two matrices was higher than expected by chance. The RF distance evaluates the differences between two trees by counting the number of partitions that are not present in both trees. The maximum RF distance is thus the total number of nodes in both trees and would correspond to two trees that do not share any partition. The K-tree score is the minimum branch length distance one can get from one tree to another after scaling one of them. The higher the RF distance and K-tree score, the bigger the topological dissimilarity between the two trees. The tree distance measures were calculated between nucleotide-based trees, amino acid-based trees, and between trees with and without recombinant taxa.

2.4 Distribution of conserved motifs in the upstream regulatory region

We used the MEME Suite v.4.11.0 (Bailey et al., 2009) to identify conserved motifs in the URR. Some of the PV genomes studied here contain a very short URR that is followed by the *E10* ORF. For these PVs, we concatenated the URR and *E10* for the analysis, as we suspect that *E10* may be functionally linked to the short URRs. We scanned for motifs on both strands of the URR, with a length between six and fifty nucleotides, and with a minimum of four occurrences in total per motif. To determine the *E*-value cut-off ($E = 3.63 \times 10^6$) for the discovered motifs, we shuffled each of the sequences from the same data set (conserving the sequence length and nucleotide composition), and repeated the analyses. We constructed a matrix containing the absolute counts of the detected motifs and analysed this matrix by a centred principal component analysis (PCA), and a correspondence analysis (COA). The detected motifs were also compared with the known regulatory motifs in PVs (Bergvall, Melendy, and Archambault 2013; Bernard 2013; McBride 2013), as well as with those in the online databases TOMTOM (Gupta et al. 2007) and TRANSFAC (Wingender 2000). For certain PV genomes for which important motifs were not detected in the URR, we used FIMO implemented in the MEME Suite to scan for the presence of these motifs elsewhere in the genome.

2.5 Codon usage preferences

We calculated the CUPrefs for all ORFs of the fifty-eight PV genomes included in this study. The relative frequencies for each of the eighteen families of synonymous codons were calculated using COUSIN v.1.0 (Bourret et al. 2019). We only considered the frequencies of the fifty-nine codons with redundancy (i.e. excluding Met, Trp, and stop codons). A matrix was created in which the rows correspond to the ORFs and the columns to the fifty-nine relative frequency values, such that each row contains the codon usage information for a specific ORF. We performed a PCA to display the variance distribution and dispersion of CUPrefs for orthologous ORFs as well as for all ORFs present within the same genome.

In addition, we used COUSIN to compare the viral CUPrefs to those of the corresponding host species. The algorithm in this program allows us to compare the CUPrefs of a query (ORFs of PV genomes) to those of a reference data set (ORFs of host genomes) and outputs a normalized value. The COUSIN score can be interpreted as follows: COUSIN = 1, the CUPrefs of the PV ORFs are similar to those of the corresponding host; COUSIN = 0, the CUPrefs of the PV ORFs are similar to a random usage of synonymous codons; COUSIN < 0, the CUPrefs of the PV ORFs are opposite to those of the corresponding host (i.e. the less used codons in the host reference are used more often in the query than in the null hypothesis of equal frequency), and COUSIN > 1, the CUPrefs of the PV ORFs are superior to those in the reference (i.e. the more frequent codons in the host reference are even more frequently used in the query) (Bourret et al. 2019). To calculate the CUPrefs of the hosts, a representative genome for each host family was chosen and the respective codon usage tables were calculated. The representatives used are: Bovidae—*Bos Taurus* (accession: AC_000158), Camelidae—*Camelus dromedarius* (accession: NW_011590949), Cervidae—*Odocoileus virginianus* (accession: NW_018326927), and Suidae—*Sus scrofa* (accession: NC_010443). For PVs infecting Delphinidae and Phocoenidae, we chose a common representative, *Tursiops truncatus* (accession: NW_017842062), as both host families are closely related. We did not calculate the CUPrefs for the Giraffidae family as the available giraffe genomes (GenBank accessions: LVKQ00000000.1 and SJXV00000000.1) are not annotated, hence GcPV1 was removed from the COUSIN analysis.

2.6 Statistical analyses

Statistical analyses and graphics were done using R v.3.4.3 (R Core Team 2018), with the aid of the packages 'ape' and 'vegan'. To compare the phylogenetic trees, we calculated pairwise distances between the concatenated *E1-E2* and *L2-L1* trees and between all single gene trees (*E1*, *E2*, *L2*, and *L1*). Jaccard distances were calculated for the distribution of motifs, and Euclidian distances for the CUPrefs of the different genes. Correlation between distance matrices were then evaluated with a Mantel test. To investigate whether the viral taxonomy, host taxonomy, sampling location, and clinical presentation correlate with the CUPrefs of all PV ORFs, the phylogenetic signal of the early gene and late gene trees, and the distribution of motifs, we used a permutational multivariate analysis of variance (PERMANOVA).

3. Results

3.1 Phylogenetic reconstruction of PVs infecting Cetartiodactyla

We collected 58 PVs infecting Cetartiodactyla from the PaVE and GenBank databases (Supplementary Table S1). ML phylogenetic trees of the concatenated early genes (*E1-E2*) and the concatenated late genes (*L2-L1*) were constructed at the nucleotide (Fig. 1) and amino acid (Supplementary Fig. S1) levels. The constructed trees are well supported with high bootstrap values, although few inner branches have low (>30 and <50) bootstrap values.

In both *E1-E2* and *L2-L1* phylogenetic trees at nucleotide and amino acid levels, the Delta-Zeta crown group (blue in Fig. 1 and Supplementary Fig. S1) forms a monophyletic clade. The other crown groups, Alpha-Omikron (coloured orange/red) and Beta-Xi (coloured green), and unclassified PVs (coloured purple), form monophyletic clades in the early gene trees. However, these do not appear to be monophyletic in the late gene trees (Fig. 1 and Supplementary Fig. S1). This incongruence is due to the 'chimeric' genomic composition of the recombinant cetacean PVs, and thereby, a position in the phylogenetic trees that varies depending on the genome region considered. In the early gene tree, these recombinant PVs (in red) cluster with non-recombinant Alpha-OmikronPVs (PphPV4, SsPV1, in orange), while in the late gene tree, the recombinant PVs cluster with Beta-XiPVs (in green) infecting Bovidae and Cervidae. Despite this displacement and several internal changes (as shown with the tanglegram in Fig. 1 and Supplementary Fig. S1), recombinant PVs remain monophyletic in both trees, suggesting that only one main recombination event occurred in the ancestral genome of these PVs.

To measure topological distances between the constructed phylogenetic trees, we calculated the pairwise distances, the K-tree scores, and the RF distances (Table 1 and Supplementary Table S2). The pairwise distances were compared with a Mantel test, a statistical test indicating correlation between the two matrices. We first compared the distances between all amino acid and nucleotide-based *E1-E2* trees and did the same for the *L2-L1* trees. None of the three distance measures indicate a significant difference between the amino acid and their corresponding nucleotide-based phylogenetic trees (early vs. early and late vs. late in Supplementary Table S2). Upon comparing the early and late gene trees without the recombinant strains, we also observe a high correlation (>0.95) between trees (Table 1 and Supplementary Table S2). However, upon

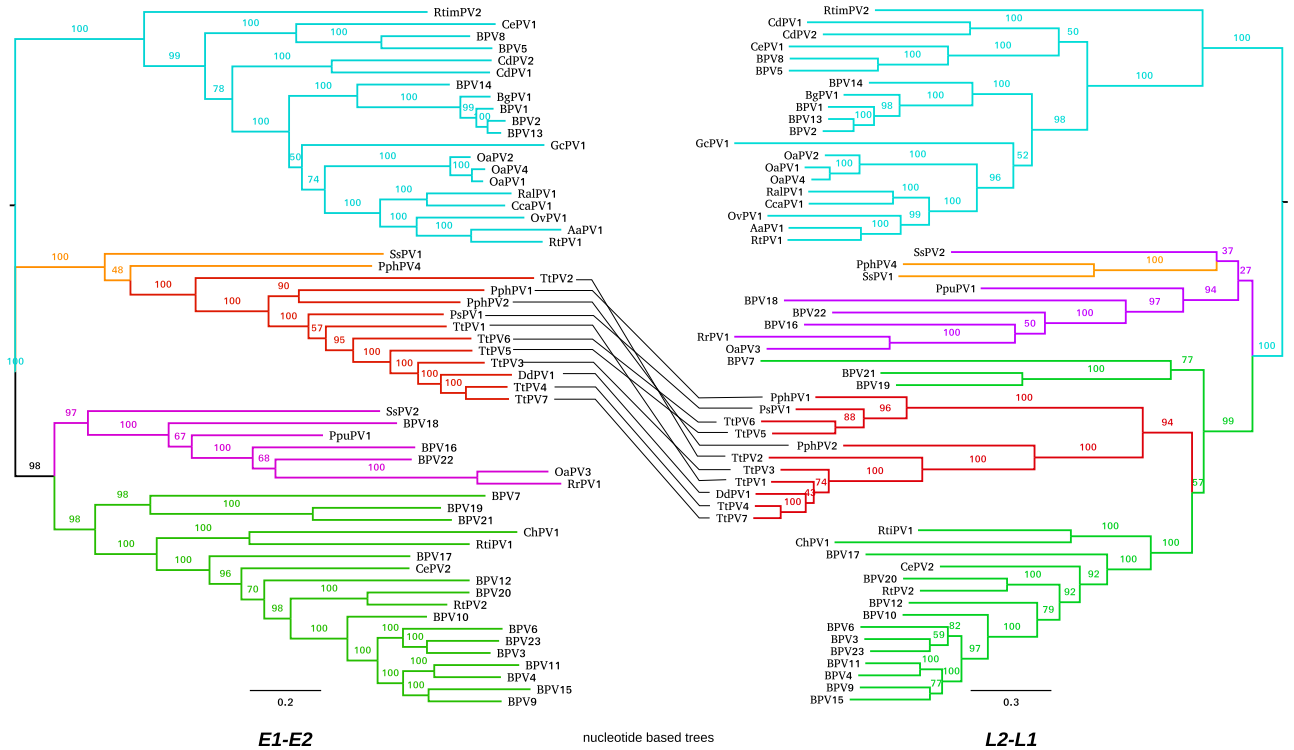

Figure 1. ML nucleotide-based phylogenetic trees of the concatenated *E1-E2* (early) and *L2-L1* (late) gene alignments. Both trees comprise fifty-eight PVs infecting Cetartiodactyla. The colour code highlights the different PV clades based on the PV crown groups: orange, Alpha-OmikronPVs; red, recombinant PVs clustering with the Alpha-OmikronPVs in the *E1-E2* tree; green, Beta-XiPVs; blue, Delta-ZetaPVs; and purple, yet unclassified PVs. Values at the branches correspond to bootstrap support values. A tanglegram connects the recombinant cetacean PVs between the early and late gene trees, emphasizing the differences in positioning of these PVs.

Table 1. Distances between phylogenetic trees based on the early and late gene regions.

Trees compared	Mantel test correlation	P-value	K-tree score	RF distance
E1-E2 nt–L2-L1 nt	0.9577	<0.001	0.9330	12
E1-E2[a] nt–L2-L1[a] nt	0.8824	<0.001	1.4766	32
E1-E2 aa–L2-L1 aa	0.9506	<0.001	1.1680	10
E1-E2[a] aa–L2-L1[a] aa	0.8745	<0.001	1.9529	28

The nucleotide and amino acid-based E1-E2 and L2-L1 trees are compared by using pairwise distances and a subsequent Mantel test with the corresponding *P* values, by the K-tree score, and by the RF distances. The introduction of the recombinant taxa in the phylogenetic inference is accompanied by a loss in concordance between the phylogenetic reconstructions for early and late genes.
nt, nucleotide-based tree; aa, amino acid-based tree.
aTree includes recombinant taxa.

introducing the recombinant taxa, this correlation is lower (~0.88). In concordance with the Mantel test, the K-tree scores and RF distances are higher for comparisons of trees that include the recombinant PVs, indicating that the number of topological incongruences is higher.

3.2 The distribution of conserved motifs in the URR reflects the phylogenetic relationships

The URR in PV genomes harbours TFBSs and other conserved motifs that regulate viral replication and transcription. The number and occurrence of these conserved motifs are more important than their order of appearance. To investigate whether the recombination event led to changes in the presence/absence of regulatory motifs and therewith possible changes in PV replication, we scanned for conserved motifs in the URR of the PV genomes. The MEME algorithm detected twenty-two conserved motifs throughout the URR of the fifty-eight query sequences (Fig. 2 and Supplementary Fig. S2). The most recurrent ones were identified as E2BSs (M1 in Fig. 2 and Supplementary Fig. S2) and the preferred E1BS (M2 in Fig. 2 and Supplementary Fig. S2). The E2BS was detected 298 times in 56 out of 58 sequences, as we could not detect this motif in the URR of BPV19 and BPV21. The E1BS was detected sixty times in all fifty-eight sequences. Since the E1- and E2BSs are pivotal for the PV life cycle, we suspect that the E2BS is located elsewhere in the genome of PVs lacking these in the URR. Indeed, for both BPV19 and BPV21, we detected an E2BS within the *L2* gene. Moreover, it is likely that additional E1- and/or E2BSs were not detected due to sequence divergence from the consensus motif sequence of PVs included in this study. Apart from the E2BS and E1BS, we detected twenty other motifs. However, we were not able to match these motifs with other known PV regulatory motifs or with those known in the online databases (TOMTOM and TRANSFAC). Certain of the URR motifs seemed to be exclusive to specific PVs; motifs M8, M9, M10, and M15 are present only in recombinant PVs, while motif M6 is solely found in a smaller group of Beta-XiPVs (Fig. 2 and Supplementary Fig. S2). The conservation of these motifs in these phylogenetic clades indicates a genuine role for the life cycle of these PVs.

To evaluate the match between the phylogenetic signal and the distribution of detected motifs, we calculated Jaccard distances on the presence/absence matrix of motifs, and compared these to the pairwise distances calculated on the early and the late gene trees. The results show that there is a correlation of 47.8 per cent ($P < 0.001$) between the distribution of motifs and the early gene phylogeny, and a 35.6 per cent ($P < 0.001$) correlation between motifs and the late gene phylogeny.

To analyse the motif distribution in the URR of the different PV genomes, we performed a centred PCA (Fig. 3a). The first axis explains 26 per cent of the observed variance and clearly separates the recombinant, Alpha-Omikron, Beta-Xi, and unclassified PVs from most PVs in the Delta-Zeta crown group. The second axis, explaining 17 per cent of the variance, separates the recombinant PVs (except one) and certain Delta-ZetaPVs from the Beta-Xi and unclassified PVs. More importantly, ten out of the eleven recombinant PVs (in red) are clearly separated from the non-recombinant Alpha-OmikronPVs (in orange). The one exception is a recombinant PV isolated from a bottlenose dolphin (TtPV2), that surprisingly does not cluster with the other recombinant PVs, including six other TtPVs. We relate this observation to the lack of sequence motifs M8, M9, M10, and M15 in the URR of TtPV2 (Fig. 2), which are conserved in and exclusive to all other recombinant PV genomes. In addition to a centred PCA, we also performed a COA to analyse the proportions between the motifs detected (Fig. 3b). The results are highly similar as those obtained for the PCA, where the recombinant PVs are separated from the non-recombinant PVs. The non-recombinant Alpha-OmikronPV (PphPV4) that is positioned closest to the recombinant PVs is also the PV with the closest phylogenetic relationship in the early gene tree (Fig. 1). The main difference between the PCA and the COA results is that certain Beta-Xi PVs, that contain motif M6, are separated from all other PVs (including other Beta-Xi PVs), that do not contain motif M6.

3.3 Orthologous Cetartiodactyla PV genes have similar codon usage preferences

To test whether the CUPrefs of the genes in the recombinant PV genomes are similar to those in the other Cetartiodactyla PV genomes, we calculated the relative frequencies of the fifty-nine codons in synonymous families and displayed this multidimensional information using a PCA. When including all ORFs in the analysis (*E1*, *E2*, *E4*, *E5*, *E6*, *E7*, *E10*, *L2*, and *L1*), we observe that the first axis (explaining 14% of the variance) separates the *E4* ORFs from the rest (Supplementary Fig. S3). This PCA also separates *E10* of BPV4, BPV9, BPV12, BPV15, and BPV23 from the rest. The centre of the PCA contains the *E1*, *E2*, *L2* and *L1* 'core' genes, indicating that these display similar CUPrefs. Although the CUPrefs of *E6* and *E7* do not display a clear pattern, these ORFs cluster closer to the core genes than *E4*, *E5*, or *E10* do. Subsequently, we performed a PCA on the CUPrefs of only the core genes (*E1*, *E2*, *L2*, and *L1*; Fig. 4). The first axis captured 16 per cent of the variance and separates the *E1* ORFS from the *E2* ORFs. The second axis contained 8 per cent of the overall variance and roughly separates the early genes (*E1* and *E2*) from the late genes (*L2* and *L1*). Although the CUPrefs of the late genes partially overlap, the recombinant PVs separate clearly from the other PVs and the first axis splits recombinant *L1* from recombinant *L2*. The relatively low median absolute deviation for each of the studied groups indicates that PVs belonging to the same clade tend to have similar CUPrefs. Unexpectedly, we observed that the CUPrefs of SsPV1 (a non-recombinant Alpha-OmikronPV, recovered from pigs) are very different from those of other PVs and SsPV2 (a non-recombinant unclassified PV, also recovered from pigs) (Fig. 4 and Supplementary Fig. S3).

Subsequently, we investigated whether the differential gene CUPrefs are related to phylogenetic clustering and/or to the

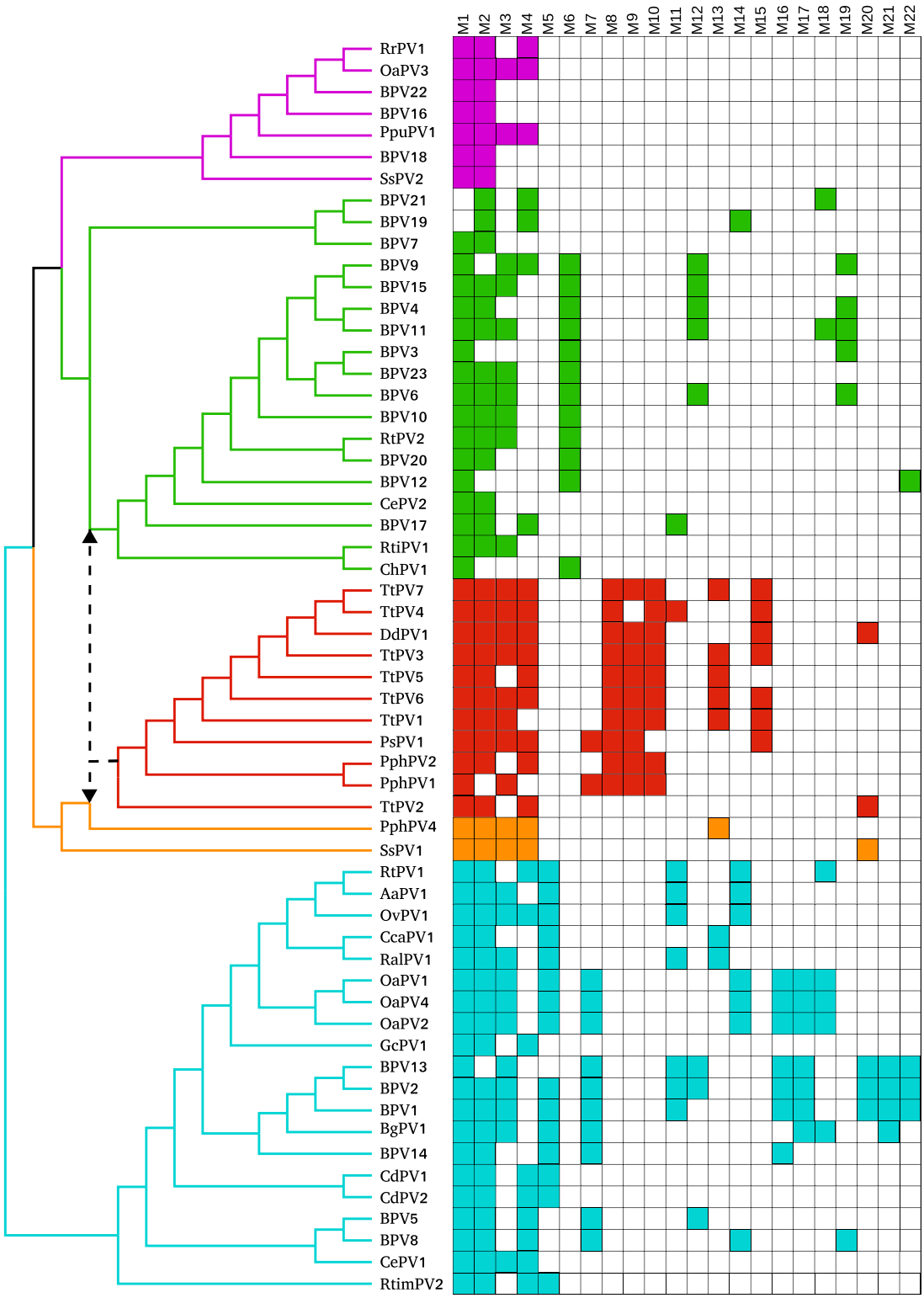

Figure 2. Presence–absence matrix of conserved motifs detected in the URR of fifty-eight PVs infecting Cetartiodactyla. In total, twenty-two motifs were identified by the MEME algorithm, as indicated with M1 to M22 on top of the columns of the matrix. Left to the rows of the matrix, the names of the studied PVs are given and a schematic representation of their phylogenetic relationships is shown. The dashed lines with arrows indicate the different phylogenetic positions of the recombinant PV clade (in red) in the early (placed with PVs in orange) and late gene trees (placed with PVs in green). Colour code corresponds to the different PV clades based on the PV crown groups: orange, Alpha-OmikronPVs; red, recombinant PVs clustering with the Alpha-OmikronPVs in the *E1-E2* tree; green, Beta-XiPVs; blue, Delta-ZetaPVs; and purple, yet unclassified PVs. A filled rectangle means that the given motif was detected in the URR of the given PV. Motifs are numbered and ordered by their abundance. M1 and M2 correspond respectively to the canonical E2BS and E1BS.

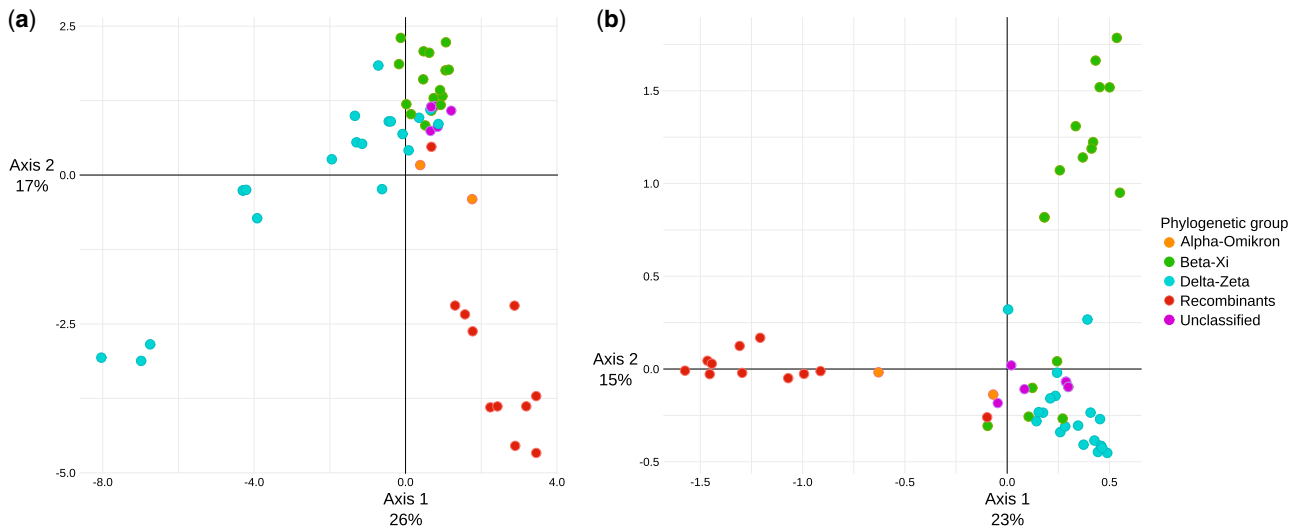

Figure 3. Centred PCA (a) and COA (b) on the distribution of motifs detected in the URR of fifty-eight PVs infecting Cetartiodactyla. As indicated in the legend on the right, colour code corresponds to the different PV clades based on the PV crown groups: orange, Alpha-OmikronPVs; red, recombinant PVs clustering with the Alpha-OmikronPVs in the *E1-E2* tree; green, Beta-XiPVs; blue, Delta-ZetaPVs; and purple, yet unclassified PVs. Values next to the axes represent the percentage of total variance explained by the corresponding axis. For the PCA, the first and second axes represent 43 per cent of the total information. For the COA, the first and second axes represent 38 per cent of the total information.

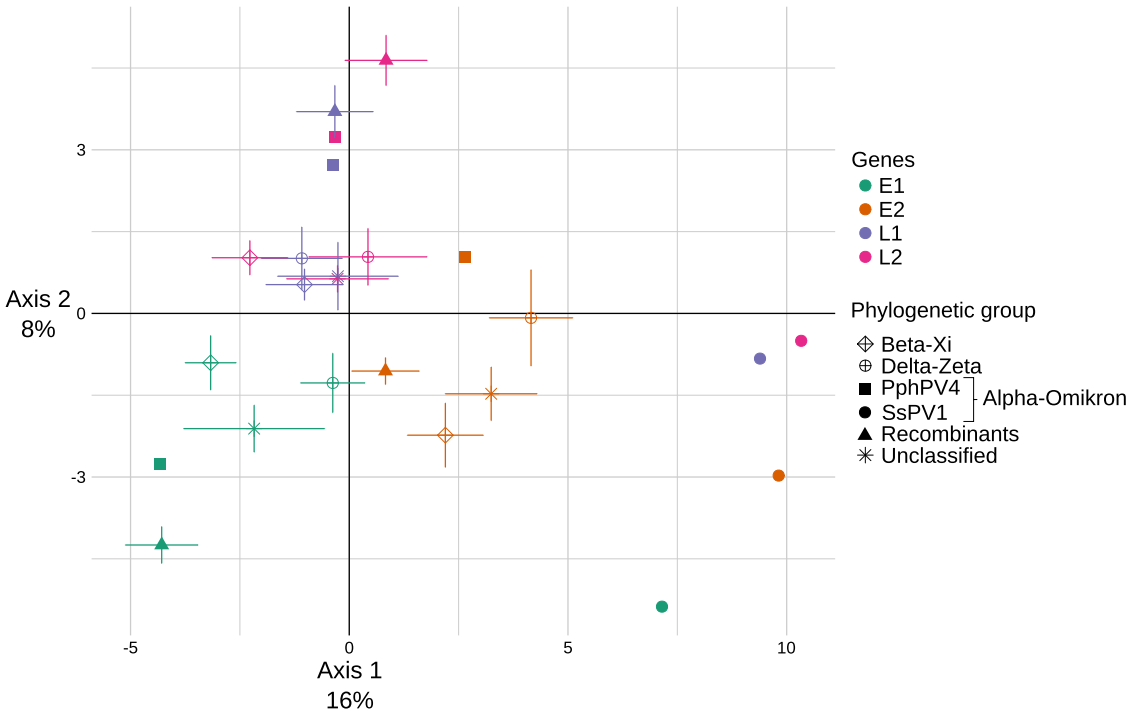

Figure 4. PCA on the CUPrefs of the PV core genes (*E1*, *E2*, *L2* and *L1*) of fifty-eight PVs infecting Cetartiodactyla. The data points are Huber M-estimator values, and the error bars correspond to the median absolute deviation. Colour code corresponds to data stratification by gene. Shapes for data points correspond to data stratification by PV crown group. Values next to the axes represent the percentage of total variance explained by the corresponding axis. Combined, the first and second axes represent 24 per cent of the total information. The main explanatory factor seems to be driven by all genes in SsPV1, infecting pigs. Secondarily, axis 1 splits the early genes *E1* and *E2*, while axis 2 splits the late and the early genes.

presence/absence of motifs in the URR. Therefore, we compared the CUPrefs of the *E1*, *E2*, *L1*, and *L2* genes to the respective gene phylogenetic trees and the URR motif distribution. We observe a higher correlation between CUPrefs and phylogenetic signal than between CUPrefs and motif distribution (Table 2). This is not surprising, as the regulatory motifs analysed in this study are located in a non-coding region, and codon usage is thus not expected to be an important factor in the evolution of this region. Even so, the CUPrefs of the early genes are better correlated to both phylogenetic signal and motif distribution than the CUPrefs of the late genes that show no correlation at all with motif distribution.

Table 2. Comparison of the CUPrefs with phylogenetic pairwise distances and URR motif distribution.

	Codon usage versus pairwise phylogenetic distances		Codon usage versus URR motif distribution	
ORF	Mantel test	*P*-value	Mantel test	*P*-value
E1	0.4606	<0.001	0.2246	0.001
E2	0.3424	<0.001	0.2111	0.001
L2	0.1386	<0.001	−0.0443	0.722
L1	0.2555	<0.001	0.0666	0.098

A Mantel test was used to compare the pairwise Euclidian distances of CUPrefs with the corresponding pairwise phylogenetic distances. Similarly, a Mantel test was used to compare the pairwise Euclidian distances of CUPrefs with the corresponding pairwise Jaccard distances of the presence/absence matrix of conserved motifs detected in the URR. This comparison was done for each of the PV core ORFs (*E1*, *E2*, *L2*, and *L1*). Phylogenetic relatedness correlates stronger with CUPrefs for early than for late genes, Similarly, a significant correlation between CUPrefs and the repertoire of motifs in the URR is only observed for the early genes.

3.4 Cetartiodactyla PV codon usage preferences do not follow those of their respective hosts

To investigate whether the Certartiodactyla PVs, and in particular the recombinant cetacean PVs, have similar CUPrefs to the hosts they infect, we calculated the COUSIN score for each of the PV ORFs (see Section 2). As a general observation, for *E6*, *E7*, *E2*, and *E5*, we obtained a COUSIN score close to 0 (Supplementary Fig. S4), indicating that for these ORFs the CUPrefs are not different from a random usage of synonymous codons. The *E4* ORF has a COUSIN score of around 1, significantly higher than all other PV ORFs (Wilcoxon–Mann–Whitney two-sided test: $W = 1{,}7762$, $P < 2.2e\text{-}16$), indicating that the CUPrefs of *E4* are closer to those of the corresponding hosts. The *E10*, *E1*, *L2*, *and L1* genes display COUSIN scores lower than 0 (Wilcoxon–Mann–Whitney one sided test: $V = 1{,}406$, $P < 2.2e\text{-}16$), indicating that the less used codons in the host reference are used more often in the PV ORFs, going towards 'opposite' CUPrefs.

Most of the Certartiodactyla PVs follow the pattern described above, however, after stratifying the COUSIN data per gene and per taxa, we observe individual exceptions (Fig. 5). Contrary to the general observation, the CUPrefs of the *E6*, *L2*, and *L1* ORFs for most recombinant- and closely related non-recombinant PVs in the Alpha-Omikron crown group are closer to those of the hosts as compared to the other PVs. Also, in this phylogenetic group, the CUPrefs of *E4* for three taxa (DdPV1, TtPV4, and TtPV5) display high COUSIN scores (Fig. 5), meaning that the most frequent codons used in the host are even more often used in this ORF. With the PCA in Fig. 4 we already showed that the CUPrefs of SsPV1 are different from those of other PV taxa. With the COUSIN score (Fig. 5), SsPV1 also distinguishes itself from the other Cetartiodactyla PVs. For all ORFs, SsPV1 has CUPrefs close to those of the host (*S.scrofa*).

3.5 Viral taxonomy and host phylogeny explain most of the observed differences in clinical presentation of the infection

PERMANOVA tests were performed to investigate whether qualitative traits—viral taxonomy, host taxonomy, sampling location, and clinical presentation—correlate with CUPrefs, motif distribution, and phylogenetic clustering of PVs. The best correlation was found between the concatenated early (*E1-E2*) and late (*L2-L1*) gene trees and viral taxonomy (60% and 52%, respectively), while for the other traits this correlation was much lower (Table 3). We also observe that the distribution of conserved motifs in the URR correlates best with viral taxonomy (33%), followed by the host taxonomy (26%). For CUPrefs, we observed that all ORFs correlate better with host taxonomy than with viral taxonomy (Table 3). This is an unexpected result as we have shown that the CUPrefs of *E6*, *E7*, *E2*, and *E5* are similar to a random usage of synonymous codons and that the CUPrefs of *E10*, *E1*, *L2*, and *L1* are going towards and 'opposite' direction as compared to the CUPrefs of the hosts they infect (Supplementary Fig. S4). These results suggest that even though the PV CUPrefs do not necessarily match those of the hosts, the viral CUPrefs do seem to be partially modulated by interaction with the different host species. The CUPrefs of *E1* correlate best with host taxonomy (35%), followed by *E5* (29%, Table 3). As *E5* was not included in the CUPrefs analysis in Fig. 4, here we investigated this ORF separately. *E5* is only present in the genomes of PVs belonging to the Delta-ZetaPV crown group, consisting of PVs infecting bovids, cervids, and one giraffid. When performing a PCA on the E5 CUPrefs (Supplementary Fig. S5), the PVs infecting Bovidae are separated by the first axis (explaining 21% of the observed variance) into two clusters, one cluster with PVs infecting members of the Bos genus, and a second cluster with PVs infecting members of the Ovis genus. The giraffid PV (GcPV1) clusters with the Ovis group. Both the first and the second axis (explaining together 36% of the variance), separate PVs infecting Cervidae from the rest. Overall, these results suggest that also for *E5* the CUPrefs are host genus specific.

4. Discussion

Here we analysed PVs infecting Certartiodactyla with the main aim to better understand the evolution of recombinant PVs infecting cetaceans. Discrepancies between the early and late gene trees are compatible with a recombination event between ancestral PVs belonging to two distant viral clades, with extant descents classified today into two different crown groups (Alpha-Omikron and Beta-Xi PVs) (Gottschling et al. 2011a; Robles-Sikisaka et al. 2012). Our phylogenetic analyses suggest that one single recombination event occurred between the genomes of these distantly related PVs. Our results for the phylogenetic inference are consistent with those communicated for the complete viral family, with recombinant cetacean PVs clustering with non-recombinant cetacean PVs in the Alpha-Omikron crown group in the early gene tree and as a sister clade to the XiPVs in the Beta-Xi crown group in the late gene tree (Supplementary material in Willemsen and Bravo 2020). The ancestral genomes of these two clades were dated back to around 60 and 70 million years ago (Ma), respectively (Willemsen and Bravo 2020), suggesting that the recombination event occurred between 60 Ma and the present.

Our analyses here presented show that the recombinant cetacean PVs contain a unique set of motifs in the regulatory region, indicating that upon recombination these PVs have followed a particular evolutionary path. Presumably, these motifs evolved as an adaptive response to the need of additional/modified regulation for effective gene expression/replication/packaging of these chimeric genomes. Nonetheless, in one of the recombinant PVs, TtPV2, we did not identify any of these specific motifs. TtPV2 is indeed basal to all other recombinant cetacean PVs in the early gene tree (Fig. 1), suggesting that the

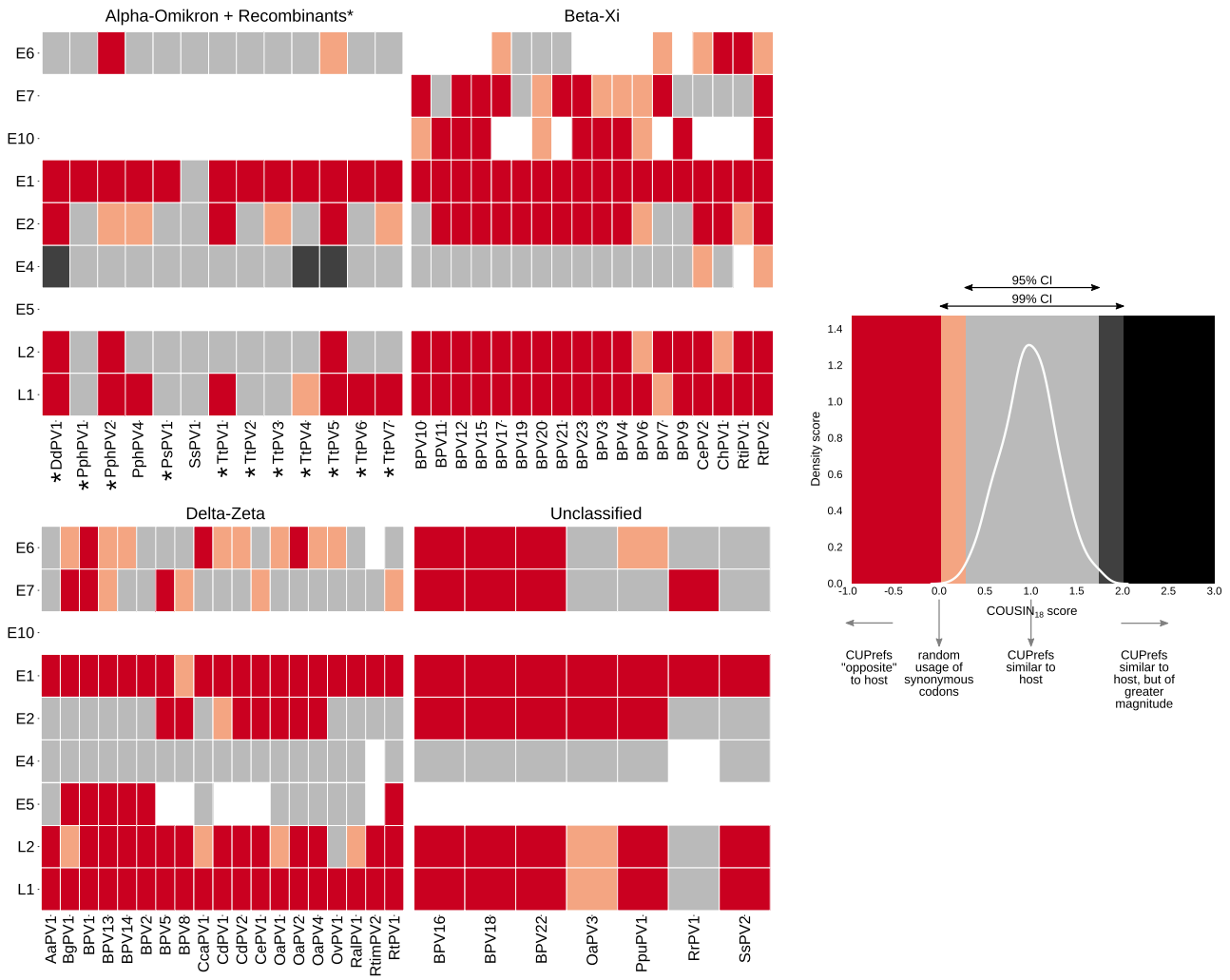

Figure 5. Heatmap of the COUSIN scores for all PV ORFs of fifty-eight PVs infecting Cetartiodactyla. COUSIN scores are stratified by PV ORFs (rows: *E6*, *E7*, *E10*—when present, *E1*, *E2*, *E4*, *E5*—when present, *L2*, and *L1*), listed in the order they are present in the PV genome, and by PV type (columns) that are grouped based on the PV crown groups: Alpha-OmikronPVs (including recombinant PVs), Beta-XiPVs, Delta-ZetaPVs, and unclassified PVs. Recombinant cetacean PVs are indicated with an asterisk. The COUSIN scores reflect the similarity between the CUPrefs in a given case gene (the corresponding viral gene) and those in a reference gene set (the full gene set in the corresponding host genome). Interpretation of the COUSIN score is given in the inset, and illustrated by colours that have been used as guideline for the heatmap. The curve in the inset corresponds to the COUSIN scores of a simulation of 500 random sequences composed of 100 codons, generated with the same CUPrefs as the different Cetartiodactyla hosts. The 95 per cent and 99 per cent confidence intervals (CIs) were calculated, and subsequently compared to the COUSIN score of the different PV ORFs. If the COUSIN score of a PV gene falls outside these intervals (coloured red, salmon, dark-grey, or black), it is considered significantly different from the reference, and when the score falls within the interval of the 95 per cent CI (coloured light grey), it is judged as matching the reference. Most viral genes display CUPrefs that are significantly different from those of the host, being systematically enriched in codons that are underrepresented in the host's genes.

appearance of the specific motifs in the URR occurred after the recombination event, as well as after the branching of TtPV2 from all other recombinant PVs. This observation supports our hypothesis of an adaptive response in the PV genome to drastic changes in the virus–host interactions associated to the recombination event.

When comparing the distribution of motifs with the evolutionary distances, we observe a better correspondence between motif composition and early genes phylogeny than with late genes phylogeny. We interpret that this agreement between early genes and motif repertoire reflects the fact that motifs in the URR are mostly involved in early gene expression regulation and genome replication, while control elements for late gene expression regulation are not located within the URR. In Alpha-Omikron PVs, the best characterized PVs, promoters for late gene expression are located within *E7* (Hummel, Hudson, and Laimins 1992; Ozbun and Meyers 1998; Bernard 2013). It is therefore not unexpected to observe such a correlation. On the contrary, it is surprising that the distribution of motifs in the URR correlates almost equally well with viral taxonomy and with host taxonomy. This suggests that besides the precise viral gene assembly, adaptation to the host species also play an important role in the evolution of regulatory PV motifs.

As the genomes of the recombinant cetacean PVs are composed of gene cassettes stemming from two distantly related viral lineages (Alpha-OmikronPVs and Beta-XiPVs), infecting distant hosts (Phocoenidae/Suidae and Bovidae/Cervidae), one could also expect to observe trends in extant gene CUPrefs, so that orthologous genes from viruses infecting closely related hosts would display closer CUPrefs than those infecting distantly related hosts. Our results do not show particular differences in CUPrefs between recombinant and non-recombinant PVs

Table 3. Comparison of the CUPrefs, motif distribution, and phylogenetic clustering with viral taxonomy, host taxonomy, sampling location, and clinical presentation.

	ORF or genomic region	Viral taxonomy (four categories)		Host taxonomy (seven categories)		Anatomical sampling location (five categories)		Clinical presentation (four categories)	
		Correlation	*P*-value	Correlation	*P*-value	Correlation	*P*-value	Correlation	*P*-value
Codon usage preferences	*E6*	0.1621	<0.001	0.2140	<0.001	0.1418	<0.001	0.0600	0.621
	E7	0.0700	0.005	0.1176	0.026	0.0938	0.420	0.0572	0.840
	E10	Only present in Beta-Xi	NA	0.1177	0.286	0.1982	0.500	0.2179	0.273
	E1	0.2822	<0.001	0.3528	<0.001	0.2354	<0.001	0.0635	0.206
	E2	0.2024	<0.001	0.2358	<0.001	0.1278	<0.001	0.0427	0.832
	E4	0.1610	<0.001	0.1726	<0.001	0.1594	<0.001	0.0630	0.231
	E5	Only present in Delta-Zeta	NA	0.2917	<0.001	Only one location	NA	Only one clinical pres.	NA
	L1	0.1428	<0.001	0.2060	<0.001	0.0941	0.054	0.0711	0.069
	L2	0.1419	<0.001	0.1648	0.007	0.1172	0.014	0.0732	0.085
Motif distribution	URR	0.3280	<0.001	0.2605	<0.001	0.1829	<0.001	0.0371	0.878
Phylogenetic clustering	*E1-E2*	0.5966	<0.001	0.3362	<0.001	0.2441	<0.001	0.0849	0.020
	L2-L1	0.5189	<0.001	0.2897	<0.001	0.1939	<0.001	0.0794	0.038

A PERMANOVA test was performed to test significance beyond null expectation for the respective correlation between qualitative traits (viral taxonomy: Alpha-Omikron, Beta-Xi, Delta-Zeta, and unclassified; host taxonomy: Bovidae, Camelidae, Cervidae, Delphinidae, Giraffidae, Phocoenidae, Suidae; anatomical sampling location: alimentary tract, anogenital, eye, hair follicles, and skin; clinical presentation: asymptomatic infection, benign (fibro)epithelial lesion, malignant lesion, and fluid running from eyes), and the Euclidian distances of CUPrefs of each PV ORF, the Jaccard distances of the presence/absence matrix of conserved motifs detected in the URR, and pairwise phylogenetic distances of the *E1-E2* and *L2-L1* trees. The good match between phylogenetic clustering and viral taxonomy is expected, as PV taxonomy boundaries are designed based on phylogenetic relatedness. The repertoire of motifs in the URR is more closely related to the viral taxonomy than to the host taxonomy. On the contrary, for all genes CUPrefs are better correlated with host taxonomy than with viral taxonomy.

infecting Certartiodactyla. Only for all genes in SsPV1, infecting pigs, the CUPrefs differ from those of all other PVs infecting cetartiodactyles. Otherwise, we observe that orthologous genes of PVs belonging to different crown groups display closer CUPrefs, than non-orthologous genes from the same virus, so that early and late genes tend to respectively display similar CUPrefs, independently of the viral genome. Such differences in CUPrefs between early and late genes have already been described for PVs infecting humans (Félez-Sánchez et al. 2015). Concordantly, CUPrefs in late genes are likely related to the cellular context in which they are expressed—differentiating epithelial cells—which provides with a particular tRNA pool for translation (Zhou et al. 1999).

As viruses depend on the host machinery for translation, we also assessed whether the CUPrefs of the Certartiodactyla PVs match those of the hosts they infect. As already shown for PVs infecting humans (Félez-Sánchez et al. 2015), we observe that CUPrefs of the Certartiodactyla PVs do not match those of the hosts they infect, to the extent that viral genes are systematically enriched in codons that are rare in the host's genome, and this independent from their nature of early or late genes. Overall, the lack of match between PVs and host CUPrefs has been explained as a strategy to avoid overexposure to the immune system (Tindle 2002). Only in certain PVs, the *E4* gene displays CUPrefs closer to those of the host, whereas for *E5* the CUPrefs appear linked to those of the hosts. While little is known about the expression pattern of E5, the E4 protein is usually expressed at high levels, and interacts with cytoskeletal proteins facilitating virion release (Doorbar 2013). The differences in CUPrefs between PV ORFs relative to the CUPrefs of their hosts, suggests that also for Certartiodactyla PVs CUPrefs are linked to gene expression patterns as well as gene function, as proposed for human PVs (Félez-Sánchez et al. 2015).

SsPV1 is the only PV that clearly distinguishes itself from other Certartiodactyle PVs in terms of CUPrefs (Figs 4 and 5). This virus was isolated from different individual stabled pigs (Stevens et al. 2008), and has also been detected in pig slurry (Di Bonito et al. 2019). Differences in CUPrefs between SsPV1 and SsPV2 can be related to the presentation of the infection, as SsPV1 has been isolated from healthy skin (Stevens et al. 2008), while SsPV2 has been isolated from papillomatous lesions in wild boars (Link et al. 2017). This result matches again previous observations on human PVs showing that differences in CUPrefs correspond well to the different clinical presentations (Félez-Sánchez et al. 2015).

In summary, we have shown here that recombination in PVs infecting Certartiodactyla occurred most probably through one single recombination event. This event generated 'chimeric' genomes of distantly related PVs. As an adaptive response to this drastic change in genome composition and in cellular context for gene expression, new regulatory motifs evolved in the URR of recombinant PV genomes. A gene expression study among cetacean PVs could shed light on the adaptive phenotypes that were affected by the changes in regulatory motifs observed in this study.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

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 article.

Funding

This work was supported by the European Research Council Consolidator Grant CODOVIREVOL (contract number 647916 to I.G.B.) and by the European Union Horizon 2020 Marie Sklodowska-Curie research and innovation programme grant ONCOGENEVOL (contract number 750180 to A.W.).

References

Angulo, M., and Carvajal-Rodríguez, A. (2007) 'Evidence of Recombination within Human Alpha-Papillomavirus', *Virology Journal*, 4: 33.

Antonsson, A., and Hansson, B. G. (2002) 'Healthy Skin of Many Animal Species Harbors Papillomaviruses Which Are Closely Related to Their Human Counterparts', *Journal of Virology*, 76: 12537–42.

Bahir, I. et al (2009) 'Viral Adaptation to Host: A Proteome-Based Analysis of Codon Usage and Amino Acid Preferences', *Molecular Systems Biology*, 5: 311.

Bailey, T. L et al. (2009) 'MEME SUITE: tools for Motif Discovery and Searching', *Nucleic Acids Research*, 37: W202–8.

Bennett, M. D. et al (2008) 'Genomic Characterization of a Novel Virus Found in Papillomatous Lesions from a Southern Brown Bandicoot (*Isoodon obesulus*) in Western Australia', *Virology*, 376: 173–82.

Bergvall, M., Melendy, T., and Archambault, J. (2013) 'The E1 Proteins', *Virology*, 445: 35–56.

Bernard, H.-U. (2013) 'Regulatory Elements in the Viral Genome', *Virology*, 445: 197–204.

Di Bonito, P. et al (2019) 'Evidence for Swine and Human Papillomavirus in Pig Slurry in Italy', *Journal of Applied Microbiology*, 127: 1246–54.

Bourret, J., Alizon, S., and Bravo, I. G. (2019) 'COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences', *Genome Biology and Evolution*, 11: 3523–8.

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: 13613–26.

Bravo, I. G., and Felez-Sanchez, M. (2015) 'Papillomaviruses: Viral Evolution, Cancer and Evolutionary Medicine', *Evolution, Medicine, and Public Health*, 2015: 32–51.

Castresana, J. (2000) 'Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis', *Molecular Biology and Evolution*, 17: 540–52.

Doorbar, J. (2013) 'The E4 Protein; Structure, Function and Patterns of Expression', *Virology*, 445: 80–98.

Van Doorslaer, K. et al (2018) 'ICTV Virus Taxonomy Profile: Papillomaviridae', *Journal of General Virology*, 99: 989–90.

Félez-Sánchez, M. et al (2015) 'Cancer, Warts, or Asymptomatic Infections: Clinical Presentation Matches Codon Usage Preferences in Human Papillomaviruses', *Genome Biology and Evolution*, 7: 2117–35.

Gottschling, M. et al (2011a) 'Modular Organizations of Novel Cetacean Papillomaviruses', *Molecular Phylogenetics and Evolution*, 59: 34–42.

Gottschling, M. et al (2011b) 'Quantifying the Phylodynamic Forces Driving Papillomavirus Evolution', *Molecular Biology and Evolution*, 28: 2101–13.

Gupta, S. et al (2007) 'Quantifying Similarity between Motifs', *Genome Biology*, 8: R24.

Hummel, M., Hudson, J. B., and Laimins, L. A. (1992) 'Differentiation-Induced and Constitutive Transcription of Human Papillomavirus Type 31b in Cell Lines Containing Viral Episomes', *Journal of Virology*, 66: 6070–80.

Johansson, C. et al (2012) 'HPV-16 E2 Contributes to Induction of HPV-16 Late Gene Expression by Inhibiting Early Polyadenylation', *The EMBO Journal*, 31: 3212–27.

Link, E. K. et al (2017) '*Sus scrofa* Papillomavirus 2-Genetic Characterization of a Novel Suid Papillomavirus from Wild Boar in Germany', *Journal of General Virology*, 98: 2113–7.

López-Bueno, A. et al (2016) 'Concurrence of Iridovirus, Polyomavirus and a Unique Member of a New Group of Fish Papillomaviruses in Lymphocystis Disease Affected Gilthead Seabream', *Journal of Virology*, 90: 8768–79.

McBride, A. A. (2013) 'The Papillomavirus E2 Proteins', *Virology*, 445: 57–79.

Narechania, A. et al (2005) 'Phylogenetic Incongruence among Oncogenic Genital Alpha Human Papillomaviruses', *Journal of Virology*, 79: 15503–10.

Ozbun, M. A., and Meyers, C. (1998) 'Temporal Usage of Multiple Promoters during the Life Cycle of Human Papillomavirus Type 31b', *Journal of Virology*, 72: 2715–22.

R Core Team (2018) 'R: A language and environment for statistical computing', R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rector, A. et al (2008) 'Genomic Characterization of Novel Dolphin Papillomaviruses Provides Indications for Recombination within the Papillomaviridae', *Virology*, 378: 151–61.

Rector, A., and Van Ranst, M. (2013) 'Animal Papillomaviruses', *Virology*, 445: 213–23.

Robinson, D. F., and Foulds, L. R. (1981) 'Comparison of Phylogenetic Trees', Mathematical Biosciences, 53: 131–47.

Robles-Sikisaka, R. et al (2012) 'Evidence of Recombination and Positive Selection in Cetacean Papillomaviruses', *Virology*, 427: 189–97.

Soria-Carrasco, V. et al (2007) 'The K Tree Score: Quantification of Differences in the Relative Branch Length and Topology of Phylogenetic Trees', *Bioinformatics*, 23: 2954–6.

Stamatakis, A. (2014) 'RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies', *Bioinformatics*, 30: 1312–3.

Stevens, H. et al (2008) 'Isolation and Cloning of Two Variant Papillomaviruses from Domestic Pigs: *Sus scrofa* Papillomaviruses Type 1 Variants a and b', *Journal of General Virology*, 89: 2475–81.

Tindle, R. W. (2002) 'Immune Evasion in Human Papillomavirus-Associated Cervical Cancer', *Nature Reviews Cancer*, 2: 59–64.

Willemsen, A., and Bravo, I. G. (2020) 'Ecological Opportunity as a Driving Force of Radiation Events and Time-Dependent Evolutionary Rates in Papillomaviruses,' bioRxiv. 2020.03.08.982421. doi: 10.1101/2020.03.08.982421.

Wingender, E. (2000) 'TRANSFAC: An Integrated System for Gene Expression regulation', *Nucleic Acids Research*, 28: 316–9.

Woolford, L. et al (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: 13280–90.

Zhou, J. et al (1999) 'Papillomavirus Capsid Protein Expression Level Depends on the Match between Codon Usage and tRNA Availability', *Journal of Virology*, 73: 4972–82.