Ecological opportunity as a driving force of radiation events and time-dependent evolutionary rates in papillomaviruses

Anouk Willemsen[1] **and Ignacio G. Bravo**[1]

[1]Centre National de la Recherche Scientique (CNRS), Laboratory MIVEGEC (CNRS IRD Univ. Montpellier), Montpellier, France
[*]Corresponding author: willemsen@microbial-ecology.net

ABSTRACT

Papillomaviruses (PVs) have a wide host range, infecting mammals, birds, turtles, and snakes. The recent discovery of PVs in different fish species allows for a more complete reconstruction of the evolutionary history of the viral family. In this study we combine fossil record estimates and molecular dating to analyze evolutionary events that occurred during PV evolution, as well as to estimate speciation and evolutionary rates.
We have used four different data sets to explore and correct for potential biases that particular taxa combinations may introduce in phylogenetic inference. When considering the evolution of speciation rates, we show that these are not constant through time, suggesting the occurrence of evolutionary events like adaptive radiations. We identified four periods along the evolutionary timeline of PVs: (i) a slow increase of PV lineages with an old ancestry that goes parallel to the basal diversification of amniotes; (ii) an initial radiation event of PVs infecting fish, birds, turtles and ancestral mammals; (iii) a secondary radiation event of PVs in parallel to the radiation of placental mammals, that generated a large number of viral lineages; and (iv) a final virus-host co-divergence in parallel to the diversification of PVs within their hosts. When considering the evolution of substitution rates, we show that there are strikingly different trends when stratifying by PV genus, revealing clade-specific events during the evolution of a highly diverse primate PV lineage: the clinically relevant *AlphaPVs*.
We have further explored the effect of heterogeneous taxa representation and the choice of calibration points on the molecular dating of the PV tree. When simultaneously using all calibration points, similar ancestral node ages and overall rates of evolution were obtained with all databases. On the contrary, in our inferences with each calibration point separately, calibrating with younger nodes tended to render higher substitution rates compared to older nodes, also known as the time-dependent rate phenomenon. We discuss that in our case this behavior may be a function of the available host niches and show that a power-law rate decay model performs well to correct for this phenomenon within PVs.
Our results provide new insights into the evolutionary history of PVs, where ecological opportunity seems to be the main driving force for the different radiation and key-innovation events we observe and showcases both the potential and limitations of using host-derived calibrations to infer viral speciation times.

Keywords: phylogenetics, molecular dating, virus evolution, time-dependent rate phenomenon

Introduction

Papillomaviruses (PVs) are small non-enveloped circular dsDNA viruses with genome size between 6 and 8 kb. The minimal PV genome consists of an upstream regulatory region (URR), an early gene region encoding for the E1 and E2 proteins, with in most cases an *E4* gene nested within *E2*, and a late gene region encoding for the L2 and L1 capsid proteins (García-Vallvé *et al.*, 2005). Proteins in the early region are involved in viral replication and cell transformation, while the capsid proteins self assemble to yield virions and encapsidate the genome. During PV evolution, the genome has acquired, and in cases subsequently lost, the *E5*, *E6* and *E7* oncogenes (Van Doorslaer and McBride, 2016; Willemsen and Bravo, 2019; Willemsen *et al.*, 2019). In some hosts, these genes are directly involved in the onset of cancer.

PVs have been mainly described to infect mammals, but have also been found in reptiles and birds. More recently, PVs have also been discovered in fish, where the first described fish PV was isolated from a gilt-head sea bream (López-Bueno *et al.*, 2016), and later other PVs isolated from a rainbow trout, two haddocks, and a red snapper were made available at GenBank (accessions: MH510267, MH616908, MH617143, MH617579). This discovery challenged our perspective on the origin of these viruses. Phylogenetic dating studies including this single fish PV genome date back the time to the most recent common ancestor (tmrca) of PVs to 481 (95% HPD 326 - 656, Van Doorslaer *et al.*, 2017b) and 424 million years ago (95% HPD: 402 - 446, Willemsen and Bravo, 2019), and thus more ancient than previously thought.

The continuous discovery of novel PVs allows to add pieces to the puzzle of their evolutionary history. Notwithstanding, an enormous bias towards human taxon sampling remains because certain human papillomaviruses (HPVs) are a major public health concern. Indeed, while the majority of PVs cause asymptomatic infections in skin and mucosa, a small number of oncogenic HPVs are associated to malignant lesions that can develop into cancer (IARC-Monographs, 2012). For other animals, PVs associated to malignant lesions have been found in a few lineages that mainly infect horses (Scase *et al.*, 2010), cows (Campo, 1997), rabbits (Kreider and Bartlett, 1981) and chamois (Mengual-Chuliá *et al.*, 2014).

Phylogenetic studies have revealed that virus-host codivergence is one of the main driving forces of PVs evolution: one third of the viruses' divergence pattern can be explained by the hosts' divergence patterns (Gottschling *et al.*, 2011b). This match allows to identify hosts fossil records of the host that can be used to date the PV tree (Pimenoff *et al.*, 2017; Rector *et al.*, 2007; Shah *et al.*, 2010). Although virus-host codivergence plays an important evolutionary role in PVs, it is not the only event that has shaped PV diversification. Other evolutionary processes such as recombination (Rector *et al.*, 2008), lineage sorting and host switches have also played a fundamental role in PV evolution, even in recent times (Pimenoff *et al.*, 2017). The best supported scenario proposes a biphasic evolution (Bravo and Félez-Sánchez, 2015; Gottschling *et al.*, 2007, 2011b), where an early PV radiation would have generated the main viral crown groups, followed by independent co-divergence between PVs and their hosts. Consequently, inconsistencies between the virus and host tree are often detected, which challenges the inference of ancestral node ages of the PV tree.

Besides node ages, evolutionary rate is one of the key parameters used to characterise the evolutionary history of viruses. Being dsDNA viruses, PVs belong within the group of slow evolving viruses with rates in the order of 10^{-7} to 10^{-9} nucleotide substitutions per site per year (Pimenoff *et al.*, 2017; Rector *et al.*, 2007; Shah *et al.*, 2010). These values are several orders of magnitudes lower than the rates of their fast evolving counterparts, the RNA viruses (Duffy *et al.*, 2008). However, it is becoming widely accepted that the division between viral types is not so strict as stipulated above. Over the last decade it has become more evident that viral evolutionary rates vary over time (Aiewsakun and Katzourakis, 2016, 2017; Ho *et al.*, 2011), depending on the timescale used to infer them (Aiewsakun and Katzourakis, 2015; Duchene *et al.*, 2014). As a consequence erroneous evolutionary rates might be inferred, where long-term substitution rates tend to be underestimated and short-term rates are prone to overestimation. This time-dependent rate phenomenon (TDRP) is present across multiple levels of viral taxonomy (Aiewsakun and Katzourakis, 2016), and can severely bias divergence time estimates. Despite of the differences in substitution rates between viral types (*e.g.* between DNA and RNA viruses), the rate decay speed of the TDRP is independent of viral taxonomy (Aiewsakun and Katzourakis, 2016). Therefore, corrections proposed to compensate for this TDRP have proven to be useful in providing better estimation of the evolutionary time scale of any virus (Aiewsakun and Katzourakis, 2015, 2017).

The biological significance of TDRP is not clear. As reviewed in Simmonds *et al.*, 2019, the most widely accepted interpretation of TDRP is that short-term rate measurements capture population-level processes (*e.g.* transient deleterious mutations and transient short-sighted adaptations for the current host), while long-term rates represent the fixation rate of mutations over macroevolutionary timescales. However, this model does not explain the paradox of why viral genomes are conserved over the long-term while having an apparent unlimited evolutionary potential to evolve and adapt rapidly. Simmonds *et al.*, 2019 propose an alternative explanatory model where viral genome conservation is best explained by a niche-filling model in which fitness optimization is rapidly achieved in the viral hosts, and therefore viral long-term rates increasingly resemble those of their hosts.

In this study we revisit the evolutionary history of PVs with newly available ancestral fish PV genomes. We selected our calibration points and dated the PV tree using four different data sets based on the consistencies in virus-host co-divergence. Two of these data sets contain the almost complete PV collection that we know up to date, while the other two are different versions of reduced data sets to correct for the over-representation of humans and other economically important hosts. We analyze these data sets by looking at the viral and host diversification rates and viral substitution rates. Moreover, we explore the effect of using either multiple or single calibration points, and correct for the outcome. Based on our observations we propose a novel evolutionary scenario and point out important events that occurred during the evolution of PVs.

Materials and Methods

Data collection and alignments

At the start of this study, we downloaded 354 full length PV genomes (154 animal and 200 human) from the PaVE (pave.niaid.nih.gov, Van Doorslaer *et al.*, 2017a) and GenBank (https://www.ncbi.nlm.nih.gov/genbank/) databases (table S1). During the study, new PV genomes were made available in the databases. Those that were relevant to the study were added to certain analyses, totalling the number of full length PV genomes to 359 (159 animal and 200 human). The *E1*, *E2*, *L2* and *L1* genes were extracted and 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). For tree construction, the *E1*, *E2*, *L2* and *L1* were concatenated using a custom perl script.

Prior phylogenetic analyses

To detect recombinant taxa as well as incongruent taxa between the early gene tree and the late gene tree, we used the filtered concatenated early genes (E1-E2) and late genes (L2-L1). For these two alignments, the LG+I+Γ protein substitution model was identified as the best suited model using ProtTest v.3.4.2 (Darriba *et al.*, 2011). ML-based phylogenetic analyses were conducted using RAxML v.8.2.9 (Stamatakis, 2014) under the GTR+Γ4 model for the nucleotide alignment using six partitions (three for each gene corresponding to each codon position), under the LG+I+Γ model for the amino acid alignment using two partitions (one for each gene), and using 1000 bootstrap replicates. The trees were rooted using the SaPV1 (*Sparus aurata papillomavirus 1*) sequence (López-Bueno *et al.*, 2016) (fig. S1 and S2). Rogue taxa were identified using the algorithm implemented in RAxML (Pattengale *et al.*, 2011). The topologies of the E1-E2 and L2-L1 trees were compared using the disagree method in TOPD-fMtS v.3.3 (Puigbò *et al.*, 2007), which allowed us to identify the disagreeing taxa between the two tree topologies.

The previously identified recombinant PVs isolated from Cetaceans (PphPV1-2, TtPV1-7, DdPV1, PsPV1) (Gottschling *et al.*, 2011a; Rector *et al.*, 2008; Robles-Sikisaka *et al.*, 2012) also showed to confidently disagree in position between the early and late gene trees in our analyses. Therefore, these recombinant PVs were removed for generating the Full Data set (FD), leaving us with a data set of 343 PV genomes. Based on the taxa that agree between the E1-E2 and L2-L1 trees, we generated two different versions of reduced data sets containing representatives for each PV species and PV type: Representative Data set 1 (RD1) and Representative Data set 2 (RD2). RD1 and RD2 both contain 130 PVs (82 animal and 48 human). Our FD, RD1 and RD2 contained one fish PV genome (López-Bueno *et al.*, 2016), but during the study four other fish PV genomes were made available in GenBank (accessions: MH510267, MH616908, MH617143, MH617579). For dating purposes, these genomes were added to the FD (FD+fish), and this new data set containing 347 PV genomes was named FDf.

For all four data sets (FDf, FD, RD1, and RD2), the concatenated E1-E2-L2-L1 alignments were used to construct ML trees with RAxML v.8.2.9 under the GTR+Γ4 model for the nucleotide alignments, using twelve partitions (three for each gene corresponding to each codon position), under the LG+I+Γ model for the amino acid alignments using four partitions (one for each gene), and in both cases using 1000 bootstrap replicates (fig. S3-S6).

For RD1 and RD2, Bayesian phylogenetic inference at the nucleotide level was performed with BEAST v.1.8.3 (Drummond *et al.*, 2012), under the GTR+Γ4 model, using twelve partitions and the uncorrelated relaxed clock model (Drummond *et al.*, 2006). Three different tree population growth models were tested; Bayesian Skyline, Constant Size and Yule Speciation Process, based on a pure birth process. For each population growth model, three independent MCMC chains were run for $2*10^8$ generations and the trees were compared by marginal likelihood (S.E. estimated from 1000 bootstrap replicates) (tables S2 and S3). Based on these results we chose the Yule Process (Yule, 1925) for phylogenetic dating.

For RD1 and RD2, Bayesian phylogenetic reconstruction at the amino acid level was performed with PhyloBayes MPI v.1.7 (Lartillot *et al.*, 2013), under the CAT-GTR+Γ8 model (Lartillot and Philippe, 2004). For each data set, three independent MCMC chains were run for 10000 cycles. Convergence diagnostics showed that the discrepancy observed across all bipartitions (meandiff; RD1: 0.0174, RD2: 0.0070) was acceptable, however the largest discrepancy (maxdiff; RD1: 0.9351, RD2: 0.3814) suggests that the chains had not converged. The same analysis was done on the FD, and also here we found that the three independent chains did not converge (meandiff: 0.0172, maxdiff: 1). Therefore, phylogenetic analyses and subsequent dating at the amino acid level with Phylobayes was excluded from this study.

Phylogenetic dating

Based on the ML constructed trees, 18 calibration points (table 1) were selected on subtrees where the E1-E2 and L2-L1 trees did not show discrepancies (fig. S1 and S2) and where the host tree matched the PV tree. Calibration point estimates were based on host fossil records (TimeTree: http://www.timetree.org/). The effect of the calibration points, and therewith forced clades, on the topology of the tree was validated by constructing ML trees constrained to the calibrations used and subsequent comparison to the corresponding unconstrained tree (table S4) using an Shimodaira-Hasegawa test (Shimodaira and Hasegawa, 1999), as implemented in RAxML (Stamatakis, 2014).

Bayesian time inference was performed at the nucleotide level using BEAST v.1.8.3 (Drummond *et al.*, 2012), under the GTR+Γ4 model, using twelve partitions, the uncorrelated relaxed clock model (Drummond *et al.*, 2006) with a lognormal distribution and a continuous quantile parametrisation (Li and Drummond, 2012), and the Yule Speciation Process tree prior (Gernhard, 2008; Yule, 1925). The constrained ML tree was used as starting tree for time inference. For each data set two independent MCMC chains were run for a maximum of 10^9 generations, sampling every 10^5, and combined when convergence was reached. Before combining the chains, 10% of the total amount sampled in each of them was discarded as a burn-in.

For RD1, two chains consisting each of 10^9 total states were combined. The combined Effective Sample Sizes (ESS) were 545, 196, and 4636 for the posterior, prior, and likelihood respectively. For RD2, chain #1 consisted of $6.470*10^8$ states and

chain #2 consisted of $6.532*10^8$ states. The combined ESS were 4411, 2006, and 9069 for the posterior, prior, and likelihood respectively. For the FD, chain #1 consisted of $5.249*10^8$ states and chain #2 consisted of $5.265*10^8$ states. The combined ESS were 1682, 792, and 6049 for the posterior, prior, and likelihood respectively. For the FDf, a total of 3 chains were combined. Chain #1 consisted of $8.083*10^8$ states, chain #2 consisted of $4.336*10^8$ states, and chain #3 consisted of $3.694*10^8$ states. The combined ESS were 2923, 1242, and 11160 for the posterior, prior, and likelihood respectively. Un-collapsed versions of the dated tree figures can be found in the Supplementary Material (fig. S7-S10).

number	Calibration point	Estimate (Ma)	CI (Ma)
1	Pan troglodytes/Pan paniscus	2.82	2.40 - 3.24
2	Homo/Pan	6.65	6.23 - 7.07
3	Bovidae1/Cervidae1	27.31	23.13 - 31.49
4	Bovidae2/Cervidae2	27.31	23.13 - 31.49
5	Bovidae3/Cervidae3	27.31	23.13 - 31.49
6	Homo/Colobus	29.44	27.61 - 31.28
7	Cricetidae/Muridae	32.70	27.90 - 37.40
8	Hyaenidae/Felidae	40	33 - 46
9	Platyrrhini/Catarrhini	43.20	40.60 - 45.70
10	Feliformia/Caniformia	54	52 - 57
11	Perissodactyla/Cetartiodactyla	78	74 - 81
12	Insectivora/Carnivora	89	83 - 96
13	Rodentia/Primates	90	85 - 94
14	Primates/Glires	90	85 - 94
15	Laurasiatheria/Primates	96	91 - 102
16	Laurasiatheria/Euarchontoglires	96	91 - 102
17	Aves/Testudines	254	240 - 268
18	Root	435	425 - 446

Table 1. Calibration points used in the molecular dating study. The estimates and the corresponding confidence interval (CI) are based on host fossil records from (TimeTree: http://www.timetree.org/). The numbers correspond to those used in subsequent graphs in this article to refer to the calibration points.

Time-Dependent Rate Phenomenon analysis

To verify the effect of the node age on the substitution rate, phylogenetic inference was performed separately with each calibration point on both RD1 and RD2. This implies that per data set, 18 independent Bayesian time inferences using a relaxed molecular clock were performed. For each time inference two independent MCMC chains were run for a maximum of 20^8 generations, sampling every 20^5, and combined when convergence was reached. The convergence statistics can be found in table S5.

From the 18 Bayesian time inferences with single calibration points we obtained, for each of the calibration points used, the posterior distributions of the tmrca and the substitution rate (substitutions per site per million years). The expected (input) and inferred (output) calibration times were compared (fig. S11) as well as the inferred calibration times versus the inferred substitution rate. Subsequently we focused on the relationship between the number of substitutions per site (S) and the evolutionary timescale in million years ago (T) as a function of the age of the node used for calibration. We fitted a power-law model to these data and use this model to correct for the TDRP (Aiewsakun and Katzourakis, 2015, 2017). This correction however, could only be done for clades where we sampled the posterior distribution of the tmrca, which allowed us to calculate the corresponding S values.

For TDRP correction random sub-samples of 500 values per calibration point (thus a total of 9000) were taken. The power-law model was fitted to the sampled S and T values. Model coefficients and summary statistics on the goodness of fit were obtained. The model was then used to predict the T values of selected nodes using their S estimates. For each node to be corrected, this analysis was repeated 1000 times. Summary statistics were obtained for each iteration, stratifying by calibration point, allowing to calculate the Grand Means and 95% CIs for the corrected T (fig. S13).

Statistics and graphics

Statistical analyses and graphics were done using R (R Core Team, 2014), with the aid of the packages ”ape”, ”car”, ”dplyr”, ”ggfortify”, ”ggplot2”, ”ggtree”, ”lawstat”, ”stats”, ”strap”, ”pgirmess”, and ”reshape”. The final display of the

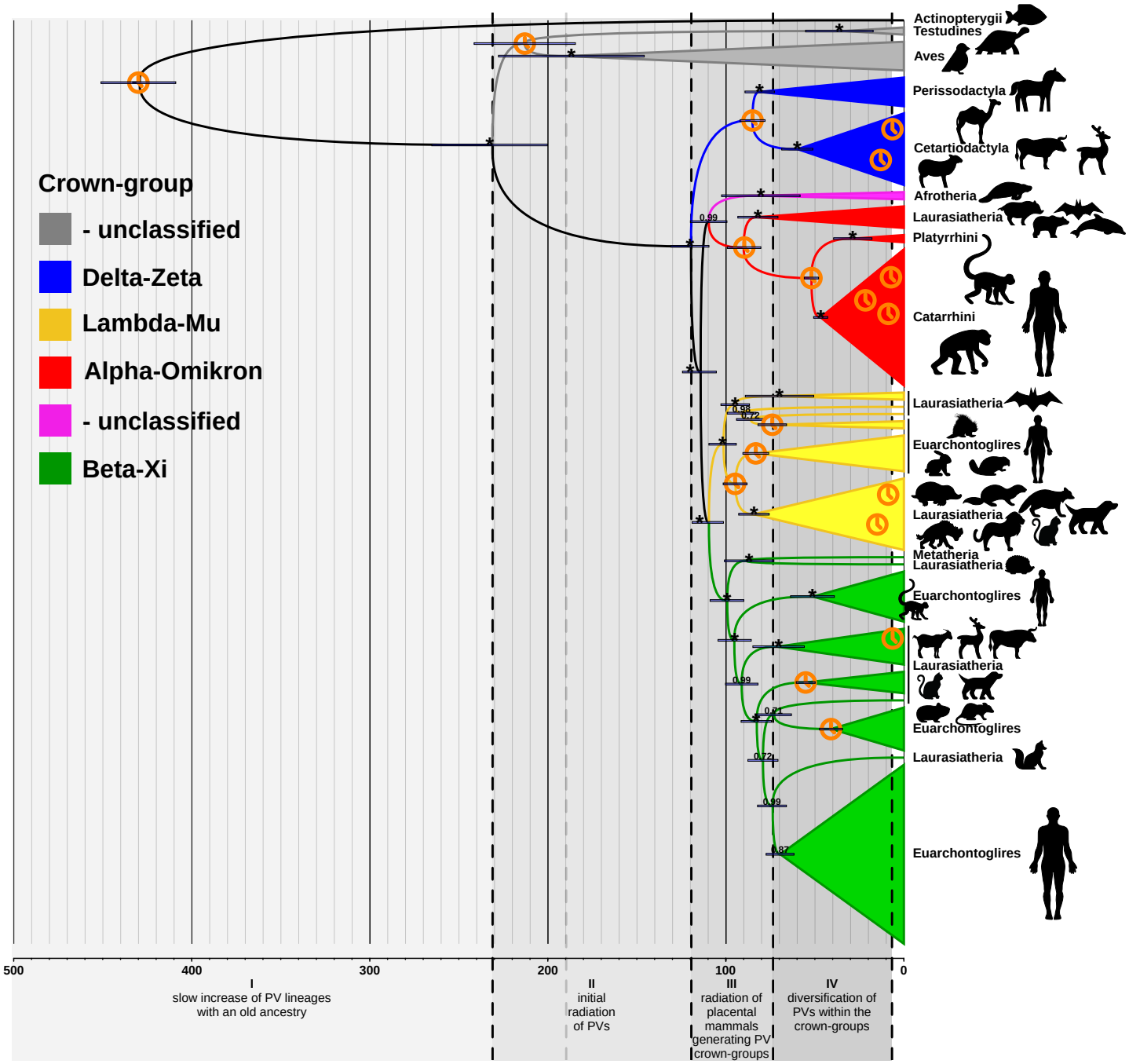

Figure 1. Dated Bayesian phylogenetic tree for representative data set 2 (RD2) containing 130 PVs. The tree was constructed at the nucleotide level based on the concatenated E1-E2-L2-L1 genes. The scale bar is given in million years ago (Ma). Values at the nodes correspond to posterior probabilities, where asterisks indicate full support. Error bars encompass 95% HPD for the age of the nodes. Clock symbols indicate the nodes used for calibration. The clades are colored according to the PV crown group classification, as indicated in the legend on the left. Next to the tree on the right, the taxonomic host group (superorder, superclass, class, order, parvorder, no rank) corresponds to the one in which the host clades could best be summarized. The grey shaded areas (delimited with black dashed lines) give an overall overview of four periods that we can distinguish along the evolutionary timeline of PVs. Event II is separated by a faded dashed line as the initial radiation event of PVs does not occur at similar time points when comparing the reduced data sets (RD1 and RD2) with the full data sets (FDf and FD), as further described in the results and shown in fig. 2

graphics was designed using Inkscape v.0.92 (https://inkscape.org/en/). The silhouettes in fig. 1 were obtained from Freepik (www.freepik.com).

Results

Recombination and papillomavirus evolution

From the databases PaVE (https://pave.niaid.nih.gov/) and GenBank (https://www.ncbi.nlm.nih.gov/genbank/) we collected 354 full length PV genomes, including human and animal PVs (table S1). The concatenated early genes (E1-E2) and the late genes (L2-L1) were used to perform ML-based phylogenetic analyses at both the nucleotide and amino acid levels (see Materials and Methods). We observed phylogenetic inconsistencies when comparing the E1-E2 and the L2-L1 trees at both the nucleotide level (fig. S1) and the amino acid level (fig. S2). At the nucleotide level 83 taxa showed disagreement between the E1-E2 and the L2-L1 trees, while at the amino acid level this was the case for 84 taxa. One suspect of recombination is *Equus asinus papillomavirus 1*(EaPV1). When phylogenetic reconstruction is done on the E1-E2 genes, this PV clusters within the Delta-Zeta crown group (colored blue in fig. 1 and fig. S1-S2), and is closely related to, but not monophyletic with, other PVs infecting Perissodactyla. When phylogenetic reconstruction is done on the L2-L1 genes, EaPV1 clusters within the Lambda-Mu crown group (colored yellow), close to other PVs infecting Laurasiatheria. The evidence for this recombination event is not strong, as the bootstrap support of EaPV1 in the late gene tree is low. The most obvious disagreement is seen with eleven Cetacean (porpoise and dolphin) PVs (PphPV1-2, TtPV1-7, DdPV1, PsPV1), for which previous studies already found evidence of recombination (Gottschling *et al.*, 2011a; Rector *et al.*, 2008; Robles-Sikisaka *et al.*, 2012). This recombination event occurred between the early and the late gene region, where certain PVs infecting Cetaceans appear to have an early gene region that resembles a non-recombinant cetacean PV in the Alpha-Omikron crown group (colored red in fig. 1, and fig. S1-S2) and a late gene region that resembles *XiPVs* infecting Bovidae and Cervidae in the Beta-Xi crown group (colored green). This is to our knowledge the only strong evidence for recombination found within PVs.

PV crown groups form well-supported clades, albeit with an unclear relative positions among them

After removing the Cetacean PVs from the data set, the Full Data set (FD) contained 343 genomes. During this study, four other fish PV genomes were made available in Genbank (accessions: MH510267, MH616908, MH617143, MH617579). As the sequences of these new PVs add relevant information to the basal clade of the tree, these genomes were added to the FD. This new data set containing 347 PV sequences was named FDf. For the FDf and FD, the early genes and the late genes were concatenated (E1-E2-L2-L1) and we compared trees constructed at the nucleotide and amino acid level under maximum likelihood framework. The trees were rooted using the PV isolated from fish. Overall, we observed well supported clades of the different PV crown groups (fig. S3-S4): Alpha-OmikronPVs (colored in red), Beta-XiPVs (green), Lambda-MuPVs (yellow), and Delta-ZetaPVs (blue). The relative positions of these well supported clades, however, change between the two trees. The grey clade consists of PVs isolated from Aves and Testudines, and conforms a yet unclassified crown group. The position of FgPV1 isolated from *Fulmarus glacialis* (Northern fulmar) is not clear within this clade. In some cases it clusters together with the rest of Aves PVs, and otherwise, it is basal to the Aves/Testudines clade.

As some PVs –mainly PVs infecting humans and some economically important species such as cattle, dogs and horses– are over-represented in the FDf and FD, we chose to work with two different reduced data sets (RD1 and RD2) to eventually corroborate the results on the FDf and FD. Both RD1 and RD2 contain 130 PVs (82 animal and 48 human), where we selected at least one representative for each PV species and PV type. Maximum likelihood-based phylogenetic analyses were performed on RD1 and RD2 using the concatenated E1-E2-L2-L1 genes (fig. S5-S6). Again, the position of the well-supported clades representing the different PV crown groups changes between trees constructed at the nucleotide and amino acid levels, as well as between the two reduced data sets.

Data set	Number of taxa	γ-statistic	*P* value
FDf	347	-3.7763	0.0002
FD	343	-3.7492	0.0002
RD1	130	-4.1391	< 0.0001
RD2	130	-4.0950	< 0.0001
hosts	81	3.0041	0.0013

Table 2. Values for the γ-statistic calculated for the dated phylogenetic trees of the FDf, FD, RD1 and RD2 data sets. The γ-statistic of the host tree is given for comparative purposes. The negative γ-values of the PV trees indicate that the internal nodes are closer to the root than expected under the pure birth model, indicating that the null hypothesis of constant diversification rates is rejected.

A total of 18 calibration points were selected on local topologies (table 1) where the E1-E2 and L2-L1 tree did not show discrepancies (fig. S1-S2), where the E1-E2-L2-L1 trees of the reduced data sets were congruent, and where the host tree matched the PV tree. Therefore, the FgPV1 sequence was forced to cluster within the Aves host group, and the EaPV1 sequence was forced to cluster within the Perissodactyla host group. The effect of the calibration points, and therewith forced clades, on the topology of the trees was validated by constructing constrained ML trees and comparing these to the unconstrained trees using an Shimodaira-Hasegawa test. At the nucleotide level, the constrained trees were not significantly worse than the unconstrained trees of the FD, RD1 and RD2 (table. S4). At the amino acid level, the same results were obtained for the constrained trees of the FD and RD1, however for the FDf and RD2 the constrained tree tested significantly worse. Based on these results, and the fact that Bayesian phylogenetic reconstruction at the amino acid level did not reach convergence (Materials and Methods), we only performed phylogenetic dating at the nucleotide level.

PVs have not diversified at a constant rate

Bayesian time inference was performed at the nucleotide level using BEAST v.1.8.3 (Drummond *et al.*, 2012) for the FDf, FD, RD1 and RD2. The constrained ML trees were used as starting trees for time inference. For each data set, two independent MCMC chains were run for a maximum of $1*10^9$ generations and combined when convergence was reached (see Materials and Methods). The dated tree for RD1 is depicted in fig. S7 and that of RD2 in fig. 1 and fig. S8.

To test whether the per-lineage speciation rates have remained constant through time, we calculated the γ-statistic (Pybus and Harvey, 2000) on the resulting phylogenetic trees (table 2). Since for all PV trees $\gamma < 0$, the internal nodes are closer to the root than expected under the pure birth model, compatible with an early burst of speciation. These results suggests that we can reject the null hypothesis of constant diversification rates on the FDf, FD, RD1, and RD2, indicating that evolutionary events such as adaptive radiations or key adaptations may have taken place. The negative values of γ relative to that of the hosts indicate decelerating diversification through time for PVs.

To visualize the timing of PV diversification, we constructed lineages-through-time (LTT) plots and generalized skyline plots based on the dated trees. fig. 2. For viral data in all data sets we can distinguish four evolutionary periods (fig. 2A): (I) a slow increase of PV lineages with an old ancestry, (II) an initial radiation event of PVs that generated the basal PV clades, (III) the radiation of mammals generating the mammalian PV crown-groups, paralleled by a secondary viral radiation, and (IV) the diversification of placental mammals within the host crown-groups and the parallel diversification of PVs with their hosts (fig.1 and fig. 2). Only in period II we observe clear differences between the data sets, where the nodes are generally dated older in the reduced data sets (RD1 and RD2) compared to the full data sets (FDf and FD). This data set based difference is also detected when we infer a generalized skyline plot on the data (fig. 2B). An overall similar pattern is observed in the analysis of changes in effective population size with time. We focused on the LTT plot to calculate the increase in taxa across time, comparing the result for the FDf tree (fig. 2C) and and the result for the host phylogenetic tree constructed using TimeTree (http://www.timetree.org/). Our analysis of the LTT slope shows that for the viruses the rate of emergence of novel viral lineages increased $\sim$7 times from evolutionary period I to period II , while it remained constant for the hosts (table 3). We interpret that this evolutionary period II reflects a radiation event limited to the viruses, which would have generated a reduced number of lineages that correspond to the crown groups of PVs infecting mammals. For both viruses and hosts, the slopes for the number of taxa through time increased from around 3 (PVs) to 5 (hosts) times from period II to period III, while the slopes slow down by almost half in from period III to IV (table 3). It is interesting to note that the ratio of LTT slopes between viruses and hosts remains constant in events III and IV, with viruses diversifying at twice the pace of their hosts. We interpret that evolutionary period III corresponds to the crown radiation of placental mammals $\sim$100 Ma, while evolutionary period IV corresponds to the subsequent diversification of placental mammals within the large clades. Our results for PVs suggest that, in their turn, the viruses paralleled the increase in host lineage number during both periods. In addition to virus host co-evolution, diversification of PVs is observed within their hosts. It is important to note that our estimates lose in certainty as we travel close to the root of our tree, essentially due to the limited number of viral taxa retrieved from non-mammalian hosts (table 3). We subsequently focused on the LTT patterns in recent times, for different PV groups within evolutionary period IV (fig. 2D), by stratifying the data into eight viral clades corresponding to the six PV genera *Alpha*, *Beta*, *Xi*, *Gamma*, *Delta* and *Lambda*) plus two unclassified PV clades infecting Aves-Testudines and Fish. Although we have a limited sample size in these basal unclassified clades, both clades appear to have originated in period II and follow a similar LTT pattern. After the radiation event that led to the generation of the PV crown groups (period III), *GammaPVs* appeared first followed by the other PV genera. For the six PV genera we observe an important increase in LTT values in period IV, that is steeper for viral lineages infecting primates (*i.e. AlphaPVs*: slope = 0.0315 , SE = 0.0006;, *BetaPVs*: slope = 0.0322, SE = 0.0008; and *GammaPVs*: slope = 0.0242, SE = 0.0007), as compared to smaller clades containing viral lineages infecting mainly other mammals (*i.e. XiPVs*: slope = 0.0185, SE = 0.0014; *DeltaPVs*: slope = 0.0205, SE = 0.0010; and *LambdaPVs*: slope = 0.0158, SE = 0.0020).

For each of the 8 sub-clades of the dated FDf tree, the mean age between the internal nodes was calculated and displayed with a density graph fig. 3A. The densities were compared with a permutation test of equality using 100 bootstraps as well

as with a Kruskal-Wallis rank sum test. Significant differences were found between the clades (χ^2 = 60.202, *df* = 7, *p* = 1.376e-10), where the distribution of the *AlphaPVs* is significantly different from that of the *GammaPVs* (χ^2 = 27.552, *df* = 1, *p* = 1.529e-07), and all PV genera are different from the basal Aves-Testudines and Fish PV clades. Subsequently, the substitution rate was obtained for the internal nodes of the dated FDf tree fig. 3B. As a general trend, we observe that younger

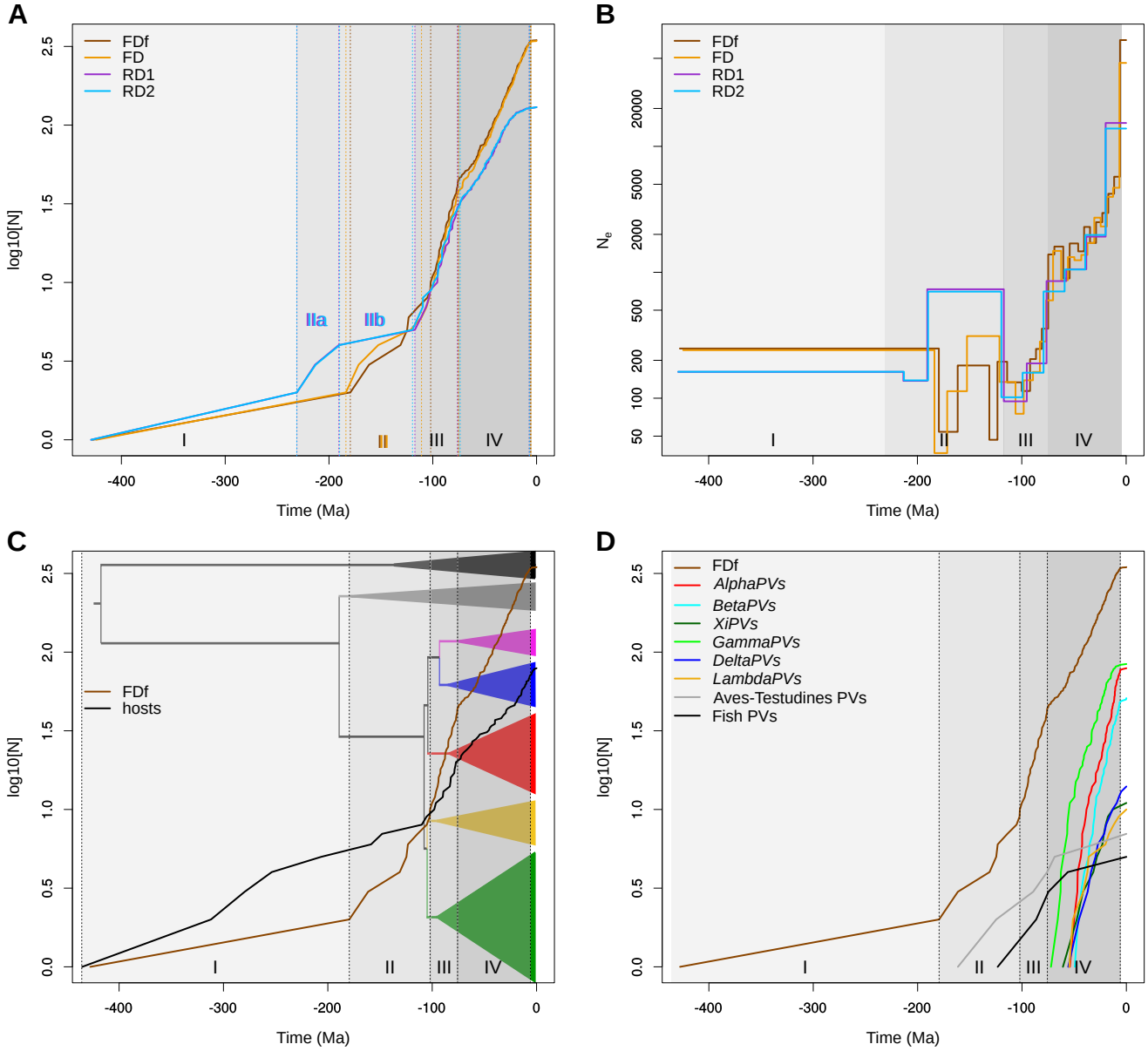

Figure 2. Plots showing the overall pattern of diversification for the dated phylogenetic trees. In each plot the grey shaded areas give an overview of four evolutionary periods that we can distinguish in the four data sets. **A**: Lineages through time (LTT) plot for the FDf, FD, RD1 and RD2 data sets. The dotted lines indicate the separate periods identified on the different data sets. Event II is separated in IIa and IIb as this initial radiation event occurs earlier in the reduced data sets (RD1 and RD2), as compared to the full data sets (FDf and FD). **B**: Generalized skyline plot showing an estimate of the effective population size over time for FDf, FD, RD1 and RD2. **C**: LTT plot for the FDf tree and the host tree. A sketch of the dated bayesian FDf tree is depicted behind the LTT lines, where the node ages correspond to those inferred on the FDf. For the LTT values of the FDf and the hosts, the corresponding slope in number of new clades per million of years is calculated for each of the four evolutionary periods (see table 3). **D**: LTT plot for FDf and this same tree stratified in eight PV clades. The PV genera are written in italics and unclassified clades are named according to the corresponding host species. In panels C and D, the dotted lines indicate the separate events identified on the FDf.

Data set	Period	Slope	SE	*df*	*F*-stat.	adj. R^2	*p*-value
FDf	I	0.0012	NA	0	NA	NA	NA
hosts	I	0.0037	0.0006	2	33.35	0.915	0.0287
FDf	II	0.0088	0.0010	6	76.60	0.915	0.0001
hosts	II	0.0027	0.0006	2	20.59	0.867	0.0453
FDf	I + II	0.0028	0.0006	8	25.97	0.735	0.0009
hosts	I + II	0.0028	0.0002	6	160.20	0.958	<0.0001
FDf	III	0.0243	0.0003	32	6426	0.995	<2.2e-16
hosts	III	0.0130	0.0010	9	153.50	0.939	<0.0001
FDf	IV	0.0142	0.0001	296	50230	0.994	<2.2e-16
hosts	IV	0.0073	0.0001	49	2487	0.980	<2.2e-16

Table 3. The slopes (measured in log10 of the number of new clades per million of years) of the LTT lines in fig. 2C. The slope is given for each of the four periods we observe in the LTT plot of the FDf. This allows us to compare the slopes of the host and PV trees. Since periods I and II do not have a clear breakpoint in the host LTT plot, we also give the combined slope of periods I and II. In addition to the slope, the standard error (SE), the degrees of freedom (*df*, N−2), the *F*-statistic, adjusted R^2 and the *p*-value are given, to indicate the significance of the calculated slope.

nodes have a higher substitution rate as compared to older nodes (Pearson's product-moment correlation FDf: -0.157). This relationship is significant when fitted to a linear model, however this model does not explain well the data (adj. R^2: 0.022, F = 8.668, df = 343, p = 0.003). When stratifying the data in the 8 PV sub-clades, we observe that these PV groups have striking different trends, where the *Alpha-* and *GammaPVs* actually have a lower substitution rate in the younger nodes as compared to the older nodes. The point with the highest substitution rate is present within *AlphaPVs* fig. 3B. This sudden increase in substitution rate coincides with the appearance of the *E5* oncogene within the *AlphaPV* clade (Willemsen and Bravo, 2019). Although this high rate might contribute to the significant linear relationship between the node age and substitution rate within *AlphaPVs*, this relationship remains significant upon eliminating this point (adj. R^2 = 0.124, F = 11.74, df = 75, $p < 0.001$).

Dating results on different data sets

For the dated trees constructed on our full and reduced data sets, we obtained the tmrca for different clades along the PV tree (fig. 4; for the raw data see table S6). Overall, the median tmrca estimates obtained by using different data sets are similar. However, when comparing the posterior distributions by multiple comparison tests after Kruskal-Wallis and Wilcoxon rank sum tests, significant differences are found comparing all data sets (fig. S12). For the inferred node ages, significant differences are found between the FD and FDf as well as between these two full data sets and the two reduced data sets (RD1 and RD2). No significant differences were found between RD1 and RD2 comparing the posterior distributions of the *Root*, *AvesTestudines/Mammals*, and *Aves/Testudines* nodes. A simple explanation is that a few number of PVs are described in the hosts corresponding to the underlying PV clades, and therefore the representative extant taxa are the same in these data sets. However, this hypothesis is discarded by the FD that is significantly different from both reduced data sets and also contains the same set of taxa in these clades. These results suggest that by using representative data sets, compensation occurs for the over-represented taxa in the PV clades infecting mammals.

For all four data sets, we obtained an overall substitution rate of around $8*10^{-9}$ substitutions per site per year (table 4). This result is in line with rates estimated for dsDNA viruses (10^{-7} to 10^{-9}) and previous overall rates obtained for the *E1* ($7.1*10^{-9}$) and *L1* ($9.6*10^{-9}$) genes of PVs (Shah *et al.*, 2010). However, our estimates are more than 2 times lower than those obtained in Pimenoff *et al.*, 2017 ($18.4*10^{-9}$) and Rector *et al.*, 2007 ($19.5*10^{-9}$). This difference can be explained by the inference of only short-term evolutionary rates in these studies (within the HPV16 clade, based on a data set spanning 0.7 Ma, and within the feline PV clade, using a data set spanning 12 Ma, respectively), advocating that a time dependency of the evolutionary rate is present within PVs.

Time-Dependent Rate Phenomenon in PVs

To verify the effect of the node age on the substitution rate, each of the 18 calibration points were inferred separately on both reduced data sets using BEAST. As shown in fig. 5A and 5B there is a negative correlation (RD1: *Spearman's rho* = -0.8225, S = 1766, p = 2.419e-05; RD2; *Spearman's rho* = -0.5170, S = 1470, p = 0.0299) between the substitution rate (substitutions/site/million years) and the inferred time. When performing a multiple comparison test after Kruskal-Wallis, most calibration points tested to have a significantly different substitution rate table S7. The data set used also has an effect on the substitution rate, where in RD1 the rates range from $4.2*10^{-9}$ to $2.4*10^{-8}$ substitutions per site per year, while in RD2 the upper value of this range is more than two times lower, ranging from $4.5*10^{-9}$ to $1.0*10^{-8}$. Nonetheless, when using single

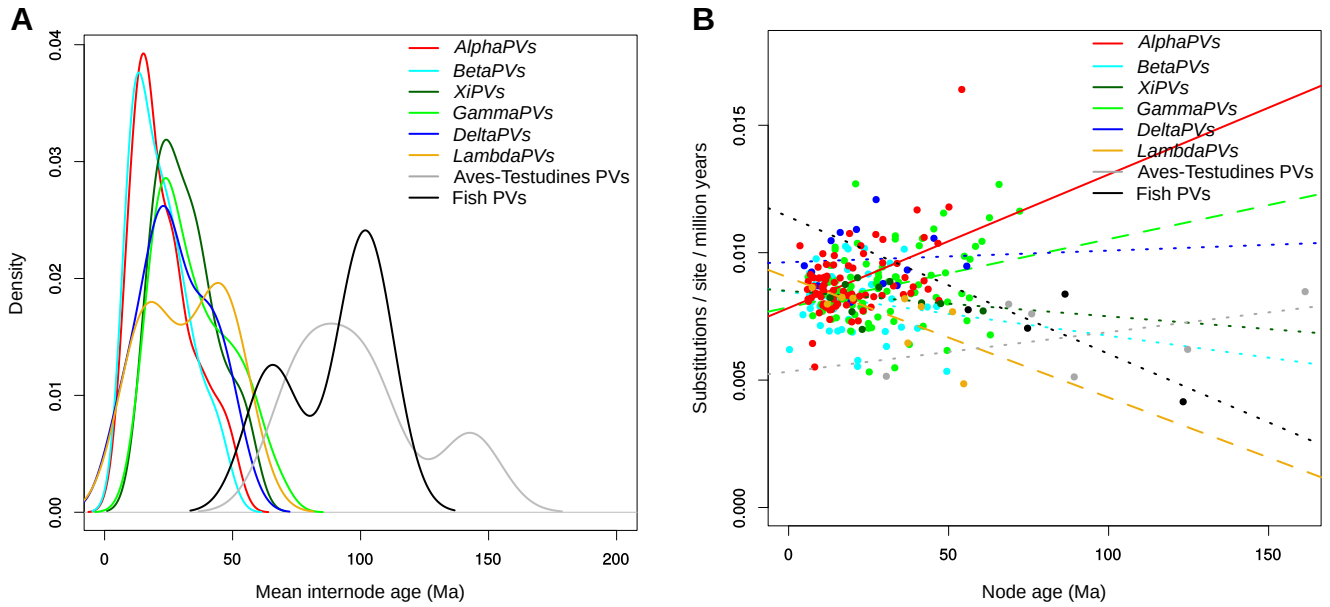

Figure 3. Internode age and substitution rate for eight PV clades in the FDf dataset. The PV clades are indicated in the legend, where PV genera are written in italics and unclassified clades are named according to the corresponding host species. **A**: Density graph of the mean age between internal nodes of the PV clades. **B**: Relationship of the node age and the substitution rate of the PV clades. A straight line was fitted to the data point for each of the clades. The continuous line indicates a convincing significant trend (*AlphaPVs*: adj. $R^2 = 0.215$, $F = 22.070$, $df = 76$, $p = 1.147e\text{-}05$), dashed lines indicate a near significant trend (*GammaPVs* and *LambdaPVs*), and dotted lines indicate no significant trend.

calibration points, the median substitution rates are similar for RD1 and RD2, and approximate those obtained when using all calibration points together (table 4). All together, these results show that younger nodes tend to have a higher substitution rate compared to older nodes, a typical signature of a time-dependent rate phenomenon (TDRP) (Aiewsakun and Katzourakis, 2015; Duchene *et al.*, 2014; Ho *et al.*, 2007). Therefore, these results show that using a single or few calibration points for dating the PV phylogenetic tree, can severely bias the outcome of the inferred the substitution rates.

The TDRP in viruses can be empirically described well by a power-law curve (Aiewsakun and Katzourakis, 2016). We fitted a power-law model to the relationship between the number of substitutions per site (S) and the evolutionary time scale (T) in million years ago (fig. 5C and 5D). We used this model to correct for TDRP by fitting the power-law model on random sub-samples of corresponding S and T estimates. Subsequently, we extrapolated the model by predicting the T values of a node to be corrected using its S estimates. This correction could only be done for clades where we sampled the posterior distribution of the tmrca. For each node to be corrected, 1000 iterations of sub-sampling and model fitting were performed using the 18 single calibration points. We corrected a total of 23 ancestral nodes ($RD1_{corr}$ and $RD2_{corr}$ in fig. 4), corresponding to the tmrca of all 18 calibration points, the tmrca of each crown-group, the tmrca of all *Mammals* (with one exception), the tmrca of the

	Data set	Mean	Median	Geometric mean	95%HPD
multiple calibrations	FDf	$7.675*10^{-9}$	$7.672*10^{-9}$	$7.672*10^{-9}$	$7.313*10^{-9} – 8.020*10^{-9}$
	FD	$7.989*10^{-9}$	$7.983*10^{-9}$	$7.987*10^{-9}$	$7.615*10^{-9} – 8.357*10^{-9}$
	RD1	$7.896*10^{-9}$	$7.894*10^{-9}$	$7.893*10^{-9}$	$7.452*10^{-9} – 8.345*10^{-9}$
	RD2	$8.013*10^{-9}$	$8.008*10^{-9}$	$8.009*10^{-9}$	$7.548*10^{-9} – 8.468*10^{-9}$
single calibrations	RD1	$9.560*10^{-9}$	$7.392*10^{-9}$	NA	NA
	RD2	$6.930*10^{-9}$	$6.710*10^{-9}$	NA	NA

Table 4. Overall substitution rates of evolution for PVs obtained by using different data sets. The substitution rates were inferred by (i) constructing one dated Bayesian phylogenetic tree for each data set (FDf, FD, RD1 and RD2) using all 18 calibrations points together (multiple calibrations), and by (ii) constructing 18 dated Bayesian phylogenetic trees for RD1 and RD2 using each calibration point separately (single calibrations). For the single calibrations the mean and median values were calculated on the 18 combined posterior distributions. Values are in substitutions per site per year.

Aves/Testudines clade, and the tmrca of the split between the *Aves/Testudines* and *Mammals*.

Our results show an overall good fit of the power-law model on the PV data (RD1: adj. R^2 = 0.912 [0.904 - 0.920]; RD2: adj. R^2 = 0.930 [0.924 - 0.935]). Using the 18 single calibration points to correct for TDRP, we observe that by using these calibration points –for which we initially obtained very dissimilar time estimates– we obtain similar time estimates after correction for each of the selected nodes (fig. S13). An exception is the root of the tree for which we also obtain variable time estimates after correction. This is not unexpected, since the time estimates of this node are situated at the edge of the model, and no good extrapolation can be made. Overall, our results confirm that the power-law model performs well when correcting for TDRP within PVs. Interestingly, the data set used for correction seems to have an impact, as in RD2 the corrected node

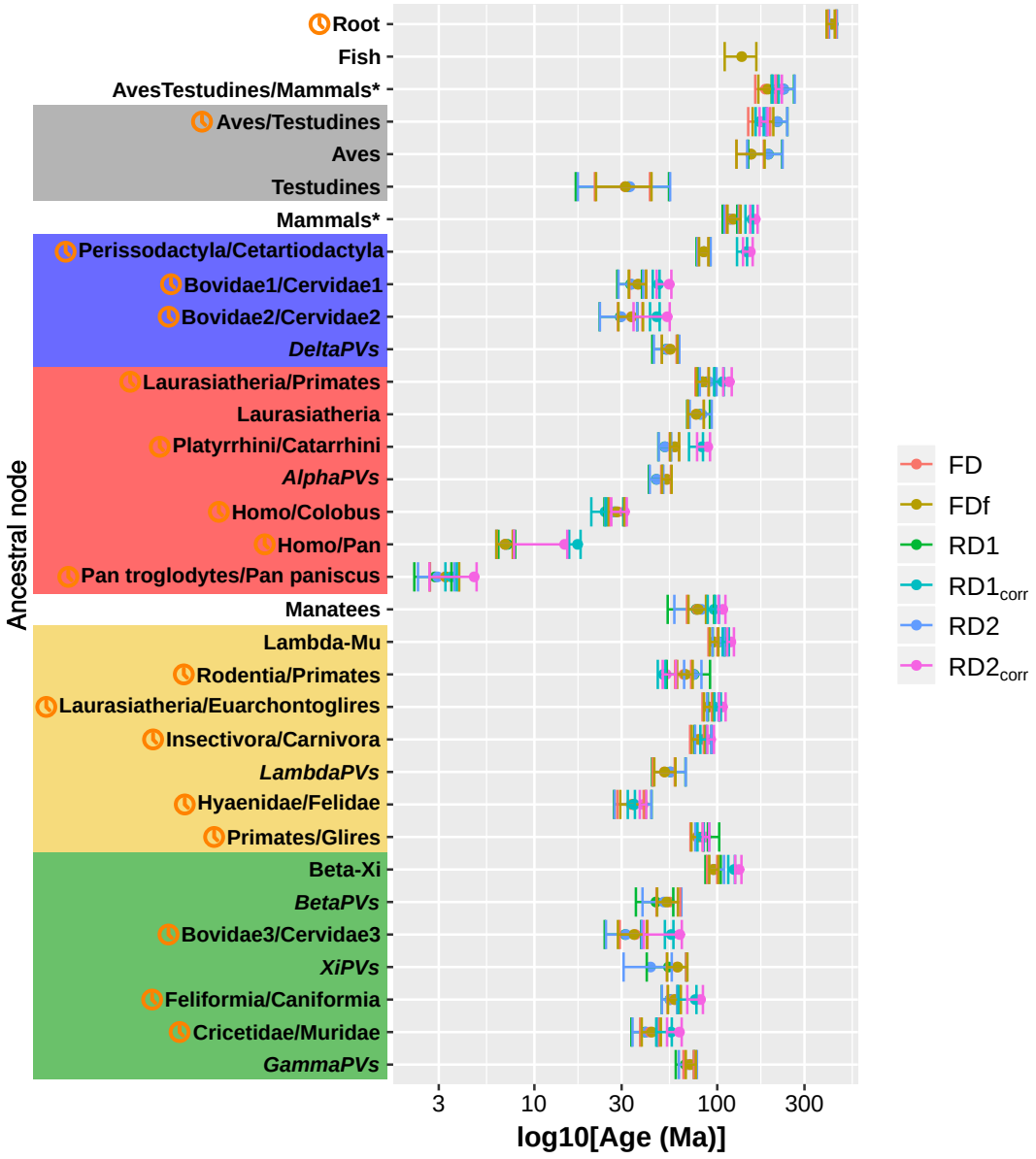

Figure 4. Inferred age for ancestral nodes on dated Bayesian phylogenetic trees. This figure allows to compare the inferred age on the different data sets: FD, FDf, RD1, and RD2. The nodes are ancestors of clades within the PV crowngroups and unclassified clades, that are in the same colour code as in fig. 1. The nodes used for calibration are indicated with a clock symbol and the corresponding number. The node age is in million years ago (Ma) and on a log scale. For certain nodes, a correction for the age has been made using the TDRP: $RD1_{corr}$ and $RD2_{corr}$. For FD, FDf, RD1, and RD2, error bars encompass 95% HPD for the age of the nodes. For $RD1_{corr}$ and $RD2_{corr}$, error bars encompass the lowest and highest inferred 95% CI for the 18 different nodes used for error correction.

age is always older than the corresponding correction in RD1 (fig. 4, table S6, and fig. S14). This might be due to the lower substitution rates obtained in RD2, as a consequence of differences in representative taxa in the two reduced data sets. When comparing the corrected times using single calibration points ($RD1_{corr}$ and $RD2_{corr}$ in fig. 4) versus the inferred time using multiple calibration points on RD1 and RD2 (RD1 and RD2 in fig. 4), we observe that older nodes are corrected to have a younger age, while younger nodes are corrected to older ages (fig. S14).

To obtain a better overview of the differences between using 18 multiple calibration points, 18 single calibration points, and the 18 corrections made for these, we calculated the relative difference as $T_{multiple} - T_{single}/T_{single}$ and $T_{corrected} - T_{single}/T_{single}$ (fig. S15). We observe that for both RD1 and RD2, the differences between multiple and single calibrations are smaller than between corrected and single calibrations. The differences for the corrected calibrations are more important for younger nodes: the younger the node used for calibration, the faster the molecular clock ticks (as also described above). When fitting a linear model to the relative differences of the multiple an corrected calibrations, we observe that the negative correlation between the corrected calibrations and the host divergence time is significant for nodes between 0 and 100 Ma, in both RD1 and RD2 (fig. S15).

Discussion

We provide here a fine-grained overview of the evolution of PVs. Our results show that the initial PV evolution seems to be marked by virus-host codivergence, where during the basal diversification of amniotes, the number of PV lineages slowly increases. This increase is probably much faster than the one we communicate here, because a proper sampling of PVs in basal amniotes and in fishes is still wanting. Subsequently we observe two radiation events, one radiation that appears to be exclusive to the viruses, and a second radiation that appears to be driven by the radiation of the placental mammals. The ecological opportunity for PVs to explore new hosts (vacant niches), can be considered a trigger of the second adaptive radiation event we observe. Previous studies have already reported this link between ecological opportunity and adaptive radiation (Glor, 2010; Yoder *et al.*, 2010). In the case of PVs, adaptive radiation increases the chances that the descendants that are better able to exploit new hosts and results in species that possess different types of adaptations. When all available niche space becomes filled, we expect rates of lineage accumulation to decrease (Rabosky and Lovette, 2008), and this is exactly what we observe for PVs. After the secondary radiation event, the slope of the LTT plot decreases, and besides virus-host codivergence, independent diversification of PVs occurred. Within the major PV crown groups, the evolution of key innovation allowed PVs to exploit new resources. The most compelling example is that of *AlphaPVs* infecting humans, where the appearance of the *E5* oncogene triggered an adaptive radiation that generated three viral lineages with different tissue tropisms. Within these three clades, PVs further diversified some evolved the potential to degrade tumor suppressor proteins (Fu *et al.*, 2010; Mesplede *et al.*, 2012; Willemsen and Bravo, 2019).

In this study, we used four different data sets to explore potential biases of the chosen taxa to infer the evolutionary time scale of PVs. This comprises mainly the over represented human taxa. The two full data sets only differ in four fish PV genomes that were made available during the study. Compared to these, both reduced data sets show that there is an effect in the ancestral nodes of the PV tree between roughly 120 and 230 Ma. The LTT plots and dating results show that these are generally inferred to be older when using representative sets of taxa, which leads to higher early estimates of the number of lineages and the effective population size. Interestingly, we observe an effect of species choice in our reduced data sets, as we observe differences in the inferred node times as well as in substitution rates. RD1 contains representative species with initial shorter branches as compared to those chosen for RD2. Meaning that more diverse PV genomes are being compared, and thus RD2 represents a more greedy strategy as compared to RD1. Although it has been shown that greedy algorithms are rarely the best option for statistical power (McAuliffe *et al.*, 2005), it has also been shown that for genomic comparisons the greedy strategy of maximizing the evolutionary divergence among species chosen from a known phylogeny can provide optimal solutions (Pardi and Goldman, 2005). The latter is in line with the analyses done on our "greedy" data set (RD2), where both maximum likelihood and bayesian phylogenetic trees were better resolved (Fig. S5-S8) and where the chains of the dated tree reached convergence faster with higher ESSs (see Materials and Methods), as compared to the "non-greedy" data set (RD1).

Previous studies have shown that different substitution rates can be obtained for the same species depending on whether they are inferred on a long-term or short-term evolutionary scale. Here we show that this time-dependent rate phenomenon is also present within PVs. By using 18 single calibration points, we observe a vast effect on the substitution rate, where long-term rates are inferred to be lower and short-term rates to be higher. Moreover, when using single calibration points the outcome depends on the data set used. Nevertheless, when using all calibration points together there is no effect on the inferred substitution rate as the result was similar in all data sets (table 4), indicating that the uncorrelated relaxed clock model accommodates well the TDRP when using all possible calibration points.

The TDRP has been previously detected in animals (García-Moreno, 2004; Wayne *et al.*, 1991), bacteria (Feng *et al.*, 2008; Rocha *et al.*, 2006), and across multiple levels of viral taxonomy (Aiewsakun and Katzourakis, 2016; Gibbs *et al.*, 2010). The TDRP in PVs follows the general rule of viruses as it can be well fitted by a power-law model (Aiewsakun and Katzourakis,

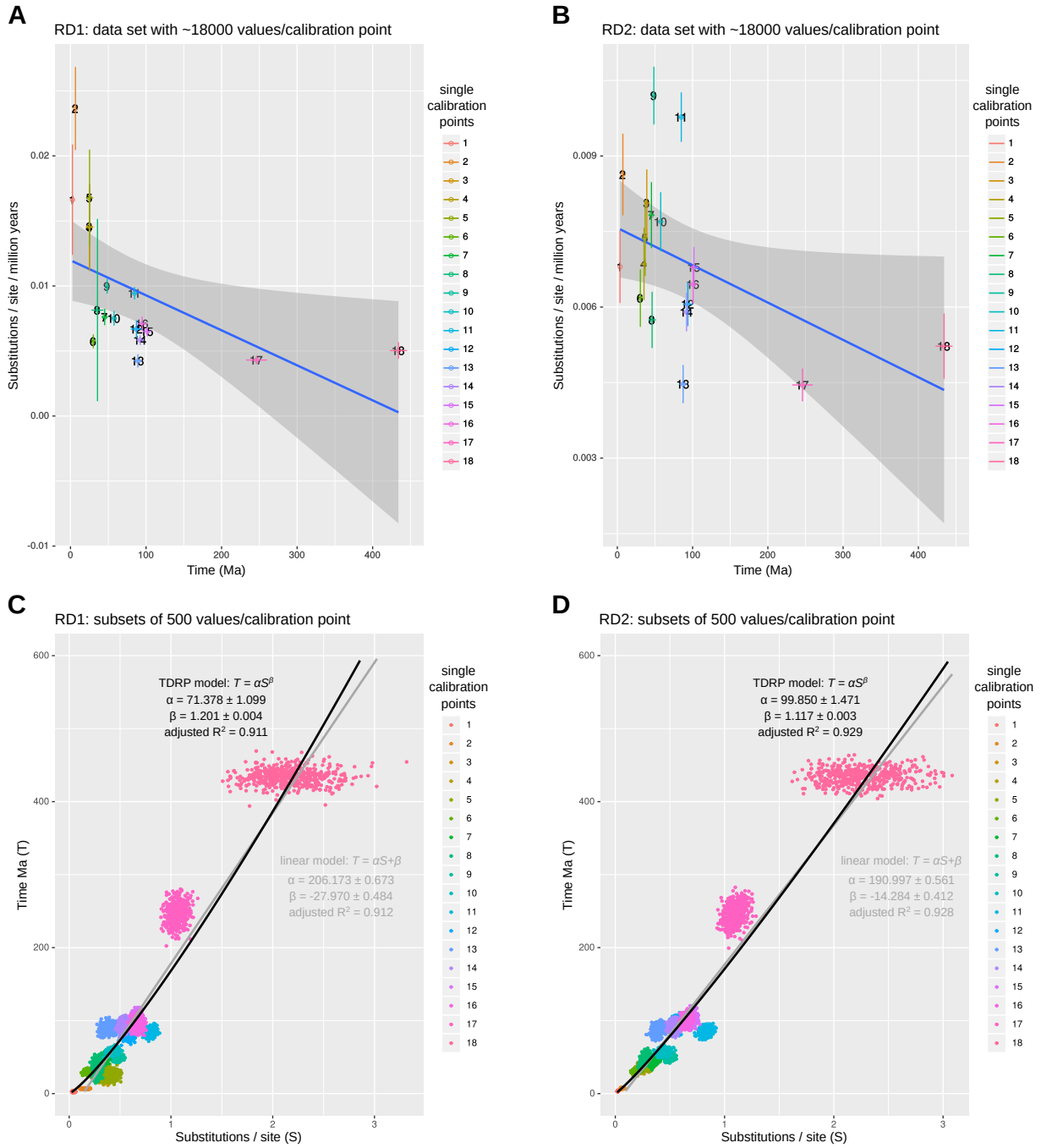

Figure 5. Time-Dependent Rate Phenomenon for PVs. **A**: The substitution rate and time were inferred independently on 18 calibration points using RD1. Therefore, 18 dated Bayesian phylogenetic trees were constructed, each with a single calibration point. For each calibration point, the mean and sd of the inferred time in million years ago (Ma) and the substitution rate in substitutions per site per million years is displayed. **B**: The same analysis as in panel A was performed on RD2. Note that the y-axes in panels A and B are not on the same scale. **C**: The relationship between the number of substitutions per site (S) and the corresponding time points (T) using RD1. A power-law TDRP model was fitted to the data. Values for the goodness and significance of fit are provided. The fit of a linear model is given for comparative purposes. This is an example of one model fit on a subset of 500 values per calibration point. For TDRP correction, the power-law model fitting was repeated 1000 times per ancestral node to be corrected. During each iteration the S values of the ancestral were used to predict the corresponding corrected T values. **D**: The same analysis as in panel C was performed on RD2.

2016). This model has been used in previous studies for TDRP correction, although the strategies that were used for inferring *T* and *S* estimates are different from the one we present here. Aiewsakun and Katzourakis, 2015 inferred the *T* and *S* estimates from one single tree constructed under the strict molecular clock assumption with a fixed rate of 1, while Aiewsakun and Katzourakis, 2017 used no clock at all. In this study, the *T* and *S* estimates were inferred from 18 different trees using 18 different calibration points under the uncorrelated relaxed clock model. Even though this is computationally more costly, we show that TDRP correction is also attainable on our data. Time dependent rates appear to be ubiquitous in viruses, and although corrections can be made, the biological significance of TDRP is not clear and we currently lack biological models that can account for TDRP together with the variation in virus genome evolution. Nonetheless, also for PVs, the substitution rate appears to be a function of the host niches a virus can exploit (Simmonds *et al.*, 2019). Ancient ancestral PV genomes appear to have lower substitution rates, and as the niche availability increased at the time where mammals started to evolve, the substitution rates increased accordingly. However, it does not seem that PV evolution has reached an evolutionary stasis, as the PV substitution rates still vary after more than 400 million years of virus-host co-evolution (fig. 3b). The frequent host jumps and more recent jumps of tissue tropisms might prevent a reduction in the substitution rate. As a consequence of this niche variation, specific events such as genome expansion, might account for the variation in substitution rates we observe.

With the current set of available genomes we can distinguish four periods, along the evolutionary time scale of PVs. Our results suggest that there is at least a triphasic evolution of PVs, where two separate radiation events allowed viruses to adapt to new niches, and a final period where independent diversification between viruses and their hosts occurred within the major virus-host clades. Our results improve and refine the previous evolutionary scenario for PVs, that had suggested a biphasic evolution of PVs, with a primary radiation was directly followed by a secondary diversification event (Bravo and Félez-Sánchez, 2015). The discovery of novel PVs infecting ancestral hosts, like in this case fish, allowed for the detection of these events. Further increase in taxon sampling of both mammalian and non-mammalian hosts will lead to a more detailed overview of the evolutionary history of this successful viral family.

Supplementary Material

Supplementary tables S1-S7 and figures S1-S15 are available online.

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.

Abbreviations

FD: Full data set
FDf: Full data set + new fish PV genomes
ML: Maximum Likelihood
RD1: Representative data set version 1
RD2: Representative data set version 2
TDRP: time-dependent rate phenomenon
tmrca: time to most recent common ancestor

References

Aiewsakun, P. and Katzourakis, A. 2015. Time dependency of foamy virus evolutionary rate estimates. *BMC Evolutionary Biology*, 15(1): 119.

Aiewsakun, P. and Katzourakis, A. 2016. Time-Dependent Rate Phenomenon in Viruses. *Journal of Virology*, 90(16): 7184–7195.

Aiewsakun, P. and Katzourakis, A. 2017. Marine origin of retroviruses in the early Palaeozoic Era. *Nature Communications*, 8: 13954.

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

Campo, M. 1997. Bovine papillomavirus and cancer. *The Veterinary Journal*, 154(3): 175–188.

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

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8): 1164–1165.

Drummond, A. J., Ho, S. Y., Phillips, M. J., and Rambaut, A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5): 699–710.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8): 1969–1973.

Duchene, S., Holmes, E. C., and Ho, S. Y. W. 2014. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proceedings of the Royal Society B: Biological Sciences*, 281(1786): 20140732–20140732.

Duffy, S., Shackelton, L. A., and Holmes, E. C. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics*, 9(4): 267–276.

Feng, L., Reeves, P. R., Lan, R., Ren, Y., Gao, C., Zhou, Z., Ren, Y., Cheng, J., Wang, W., Wang, J., Qian, W., Li, D., and Wang, L. 2008. A Recalibrated Molecular Clock and Independent Origins for the Cholera Pandemic Clones. *PLoS ONE*, 3(12): e4053.

Fu, L., Van Doorslaer, K., Chen, Z., Ristriani, T., Masson, M., Travé, G., and Burk, R. D. 2010. Degradation of p53 by Human Alphapapillomavirus E6 Proteins Shows a Stronger Correlation with Phylogeny than Oncogenicity. *PLoS ONE*, 5(9): e12816.

García-Moreno, J. 2004. Is there a universal mtDNA clock for birds? *Journal of Avian Biology*, 35(6): 465–468.

García-Vallvé, S., Alonso, Á., and Bravo, I. G. 2005. Papillomaviruses: different genes have different histories. *Trends in Microbiology*, 13(11): 514–521.

Gernhard, T. 2008. The conditioned reconstructed process. *Journal of Theoretical Biology*, 253(4): 769–778.

Gibbs, A. J., Fargette, D., García-Arenal, F., and Gibbs, M. J. 2010. Time - The emerging dimension of plant virus studies. *Journal of General Virology*, 91(1): 13–22.

Glor, R. E. 2010. Phylogenetic Insights on Adaptive Radiation. *Annual Review of Ecology, Evolution, and Systematics*, 41(1): 251–270.

Gottschling, M., Stamatakis, A., Nindl, I., Stockfleth, E., Alonso, A., and Bravo, I. G. 2007. Multiple Evolutionary Mechanisms Drive Papillomavirus Diversification. *Molecular Biology and Evolution*, 24(5): 1242–1258.

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., Goker, 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): 2101–2113.

Ho, S. Y., Lanfear, R., Bromham, L., Phillips, M. J., Soubrier, J., Rodrigo, A. G., and Cooper, A. 2011. Time-dependent rates of molecular evolution. *Molecular Ecology*, 20(15): 3087–3101.

Ho, S. Y. W., Shapiro, B., Phillips, M. J., Cooper, A., and Drummond, A. J. 2007. Evidence for Time Dependency of Molecular Rate Estimates. *Systematic Biology*, 56(3): 515–522.

IARC-Monographs 2012. Biological Agents, Human Papillomaviruses. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, 100B.

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.

Kreider, J. W. and Bartlett, G. L. 1981. The Shope Papilloma-Carcinoma Complex of Rabbits: A Model System of Neoplastic Progression and Spontaneous Regression. In *Advances in Cancer Research*, volume 35, pages 81–110. Academic Press.

Lartillot, N. and Philippe, H. 2004. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Molecular Biology and Evolution*, 21(6): 1095–1109.

Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. 2013. PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Systematic Biology*, 62(4): 611–615.

Li, W. L. S. and Drummond, A. J. 2012. Model Averaging and Bayes Factor Calculation of Relaxed Molecular Clocks in Bayesian Phylogenetics. *Molecular Biology and Evolution*, 29(2): 751–761.

López-Bueno, A., Mavian, C., Labella, A. M., Castro, D., Borrego, J. J., Alcami, A., and Alejo, A. 2016. Concurrence of Iridovirus, Polyomavirus, and a Unique Member of a New Group of Fish Papillomaviruses in Lymphocystis Disease-Affected Gilthead Sea Bream. *Journal of Virology*, 90(19): 8768–8779.

McAuliffe, J. D., Jordan, M. I., and Pachter, L. 2005. Subtree power analysis and species selection for comparative genomics. *Proceedings of the National Academy of Sciences*, 102(22): 7900–7905.

Mengual-Chuliá, B., Domenis, L., Robetto, S., and Bravo, I. G. 2014. A novel papillomavirus isolated from a nasal neoplasia in an Italian free-ranging chamois (Rupicapra r. rupicapra). *Veterinary Microbiology*, 172(1-2): 108–119.

Mesplede, T., Gagnon, D., Bergeron-Labrecque, F., Azar, I., Senechal, H., Coutlee, F., and Archambault, J. 2012. p53 Degradation Activity, Expression, and Subcellular Localization of E6 Proteins from 29 Human Papillomavirus Genotypes. *Journal of Virology*, 86(1): 94–107.

Pardi, F. and Goldman, N. 2005. Species Choice for Comparative Genomics: Being Greedy Works. *PLoS Genetics*, 1(6): e71.

Pattengale, N. D., Aberer, A. J., Swenson, K. M., Stamatakis, A., and Moret, B. M. E. 2011. Uncovering Hidden Phylogenetic Consensus in Large Data Sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(4): 902–911.

Pimenoff, V. N., de Oliveira, C. M., and Bravo, I. G. 2017. Transmission between Archaic and Modern Human Ancestors during the Evolution of the Oncogenic Human Papillomavirus 16. *Molecular Biology and Evolution*, 34(1): 4–19.

Puigbò, P., Garcia-Vallvé, S., and McInerney, J. O. 2007. TOPD/FMTS: A new software to compare phylogenetic trees. *Bioinformatics*, 23(12): 1556–1558.

Pybus, O. G. and Harvey, P. H. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society B: Biological Sciences*, 267(1459): 2267–2272.

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

Rabosky, D. L. and Lovette, I. J. 2008. Density-dependent diversification in North American wood warblers. *Proceedings of the Royal Society B: Biological Sciences*, 275(1649): 2363–2371.

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.

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.

Rocha, E. P., Smith, J. M., Hurst, L. D., Holden, M. T., Cooper, J. E., Smith, N. H., and Feil, E. J. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology*, 239(2): 226–235.

Scase, T., Brandt, S., Kainzbauer, C., Sykora, S., Bijmholt, S., Hughes, K., Sharpe, S., and Foote, A. 2010. Equus caballus papillomavirus-2 (EcPV-2): An infectious cause for equine genital cancer? *Equine Veterinary Journal*, 42(8): 738–745.

Shah, S. D., Doorbar, J., and Goldstein, R. A. 2010. Analysis of host-parasite incongruence in papillomavirus evolution using importance sampling. *Molecular Biology and Evolution*, 27(6): 1301–1314.

Shimodaira, H. and Hasegawa, M. 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Molecular Biology and Evolution*, 16(8): 1114–1116.

Simmonds, P., Aiewsakun, P., and Katzourakis, A. 2019. Prisoners of war — host adaptation and its constraints on virus evolution. *Nature Reviews Microbiology*, 17(5): 321–328.

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

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

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.

Van Doorslaer, K., Li, Z., Xirasagar, S., Maes, P., Kaminsky, D., Liou, D., Sun, Q., Kaur, R., Huyen, Y., and McBride, A. A. 2017a. The Papillomavirus Episteme: A major update to the papillomavirus sequence database. *Nucleic Acids Research*, 45: D499–D506.

Van Doorslaer, K., Ruoppolo, V., Schmidt, A., Lescroël, A., Jongsomjit, D., Elrod, M., Kraberger, S., Stainton, D., Dugger, K. M., Ballard, G., Ainley, D. G., and Varsani, A. 2017b. Unique genome organization of non-mammalian papillomaviruses provides insights into the evolution of viral early proteins. *Virus Evolution*, 3(2).

Wayne, R. K., Van Valkenburgh, B., and O'Brien, S. J. 1991. Molecular distance and divergence time in carnivores and primates. *Molecular Biology and Evolution*, 8(3): 297–319.

Willemsen, A. and Bravo, I. G. 2019. Origin and evolution of papillomavirus (onco)genes and genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1773): 20180303.

Willemsen, A., Félez-Sánchez, M., and Bravo, I. G. 2019. Genome plasticity in Papillomaviruses and de novo emergence of E5 oncogenes. *Genome Biology and Evolution*, 11(6): 1602–1617.

Yoder, J. B., Clancey, E., Des Roches, S., Eastman, J. M., Gentry, L., Godsoe, W., Hagey, T. J., Jochimsen, D., Oswald, B. P., Robertson, J., Sarver, B. A. J., Schenk, J. J., Spear, S. F., and Harmon, L. J. 2010. Ecological opportunity and the origin of adaptive radiations. *Journal of Evolutionary Biology*, 23(8): 1581–1596.

Yule, G. U. 1925. A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 213(402-410): 21–87.