Evidence for a billion-years arms race between viruses, virophages and eukaryotes

Authors: Jose Gabriel Nino Barreat[1], Aris Katzourakis[1]*

[1]Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3SY, United Kingdom.

*Corresponding author. Email: aris.katzourakis@zoo.ox.ac.uk

Abstract: The PRD1-adenovirus lineage is one of the oldest and most diverse lineages of viruses. In eukaryotes, they have diversified to an unprecedented extent giving rise to adenoviruses, virophages, *Mavericks*, *Polinton*-like viruses and Nucleocytoplasmic Large DNA viruses (NCLDVs) which include the poxviruses, asfarviruses and iridoviruses, among others. Two major hypotheses for their origins have been proposed: the 'virophage first' and 'nuclear escape' hypotheses, but their plausibility until now has remained unexplored. Here, we use maximum-likelihood and Bayesian hypothesis-testing to compare the two scenarios based on the shared proteins forming the virus particle and a comprehensive genomic character matrix. We also compare the phylogenetic origin of the transcriptional proteins shared by NCLDVs and cytoplasmic linear plasmids. Our analyses overwhelmingly favour the virophage first model. These findings shed light on one of the earliest diversifications seen in the virosphere, supporting a billion-years arms race between viruses, virophages and eukaryotes.

Introduction

The PRD1-adenovirus lineage of viruses and related mobile genetic elements is a remarkable example of adaptive radiation in both terms of ecology and genomic complexity. In eukaryotes, they have diversified giving rise to many unique viral lineages (Koonin & Krupovic, 2017; Krupovič & Bamford, 2008; Krupovic & Koonin, 2015). These include 1) the Nucleocytoplasmic Large DNA viruses (NCLDVs), which have evolved the largest viruses known (Barthélémy et al., 2019; Legendre et al., 2013, 2014), 2) the virophages, which are viral parasites of giant viruses (Fischer & Suttle, 2011; la Scola et al., 2008), 3) the *Maverick*/*Polinton* genome-integrating viruses, which colonise the genomes of all major eukaryotic lineages (Kapitonov & Jurka, 2006; Pritham et al., 2007), 4) the adenoviruses, which infect vertebrates (Harrach et al., 2019), 5) the marine *Polinton*-like viruses (Yutin et al., 2015) and 6) the cytoplasmic linear plasmids of fungi, which have lost the capacity to form viral particles (Meinhardt et al., 1997). The common ancestry of elements in this lineage is supported by a shared dsDNA genome flanked by terminal inverted repeats (present in most forms), and an ancestral gene module consisting of double and single jelly-roll capsid proteins, an adenoviral-like protease and a HerA/FtsK superfamily DNA-packaging ATPase (Krupovic & Koonin, 2016; Yutin et al., 2015). Members of the eukaryotic PRD1-adenovirus lineage have been recently classified into the Kingdom *Bamfordvirae* (Walker et al., 2020). Understanding the origin and evolution of this viral lineage is challenging given the genetic, ecological and morphological diversity seen in these viruses as well as the long timescales involved.

One of the fundamental questions about the evolutionary biology of this lineage relates to the origin of virophages and NCLDVs. Virophages (Family *Lavidaviridae*) are small viruses which

parasitise the transcriptional machinery of giant viruses, by sharing the same promoter and polyadenylation sequences (Fischer & Suttle, 2011; la Scola et al., 2008; Mougari et al., 2019). These virophages have been shown to reduce the viable progeny of their host viruses, and both *Sputnik* and *Mavirus* virophages have the ability of genomic integration (Berjón-Otero et al., 2019; Blanc et al., 2015; Fischer & Hackl, 2016). Upon coinfection with a giant virus, the integrated virophages are reactivated and function as an altruistic immune response in the wider host population (Fischer & Hackl, 2016). These observations have led to the 'virophage first' hypothesis, where an ancestral large DNA virus coevolved with a primitive virophage which was the immediate ancestor of other elements in the lineage (Fischer & Suttle, 2011) (Figure 1, left). An alternative origins scenario is the 'nuclear escape' hypothesis, where NCLDVs evolved from a lineage of genomic transposon-like viruses, which then escaped from the nucleus and adapted to the cytoplasm (Koonin & Krupovic, 2017; Krupovic & Koonin, 2015) (Figure 1, right). Under this view, virophages would have evolved their promoter and poly-A sequences *de novo*, in order to parasitise the transcriptional machinery of NCLDVs (once these had escaped from the nucleus).

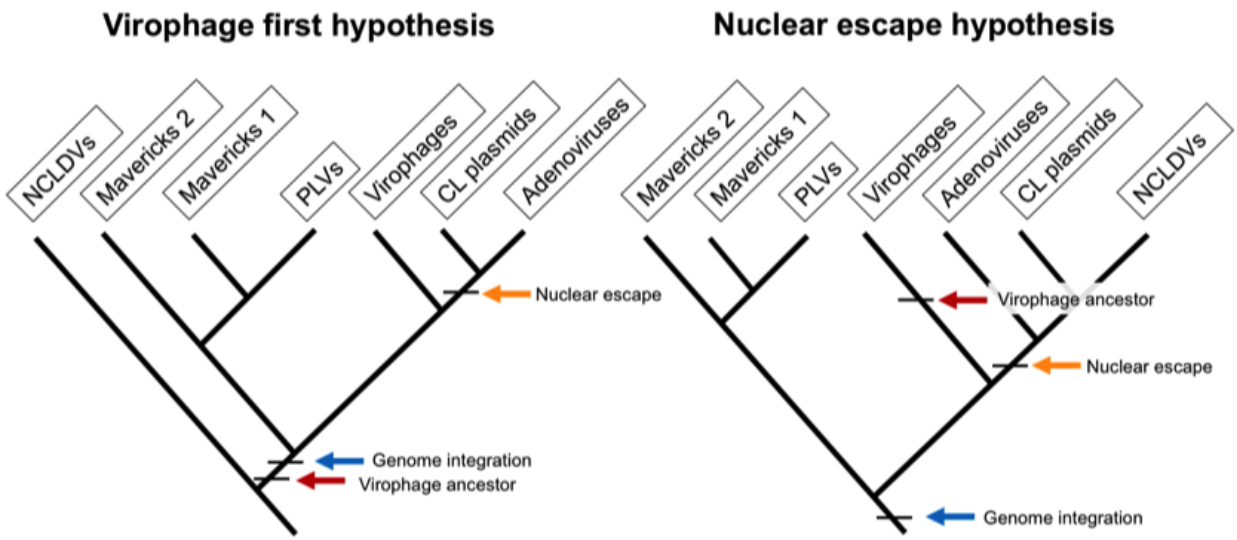

Figure 1. The two main hypotheses for the origin of NCLDVs in the eukaryotic PRD1-adenovirus lineage. In the virophage first hypothesis, NCLDVs diverge early with its sister lineage evolving into primitive virophages. In the nuclear escape hypothesis NCLDVs descend from endogenous elements that became exogenous and virophages then evolved to parasitise them. The trees are based on our analysis of the four core virion proteins, with the inclusion of cytoplasmic linear plasmids as the sister to adenoviruses or NCLDVs, respectively.

The two models make different evolutionary predictions. In the 'virophage first' scenario, NCLDVs diverged early and their sister lineage evolved into virophages with the ability to integrate into the eukaryotic host genome (Fischer & Suttle, 2011). These primitive virophages would have further diversified into *Mavericks*, PLVs, modern virophages (lavidaviruses), cytoplasmic linear plasmids and adenoviruses. Importantly, NCLDVs would belong to a

different clade to the group of cytoplasmic linear plasmids + adenoviruses. In addition, this model predicts that the 3 transcriptional gene homologues shared by cytoplasmic linear plasmids and NCLDVs (DNA-directed RNA polymerase II subunit Rpb2, helicase and mRNA-capping enzyme), were acquired independently during evolution. In contrast, the 'nuclear escape' model predicts that NCLDVs should form a clade with adenoviruses and cytoplasmic linear plasmids, as well as a single origin for the transcriptional homologues of cytoplasmic linear plasmids and NCLDVs (Koonin & Krupovic, 2017; Krupovic & Koonin, 2015). This model also suggests a much more recent origin for NCLDVs and virophages. Here we explore the plausibility of each hypothesis by using a combination of maximum-likelihood and Bayesian hypothesis testing, as well as by examining the origin of the transcriptional gene homologues in cytoplasmic linear plasmids and NCLDVs to ensure the robustness of our conclusions.

Results

To compare the two origins hypotheses between virophages and NCLDVs, we built a concatenated multiple sequence alignment of the four core virion proteins (major and minor capsid proteins, adenoviral-like protease and FtsK/HerA DNA-packaging ATPase), together with a matrix of genomic characters (Figure S1). The topology of the Bayesian majority-rule consensus tree calculated from the amino acid characters (Figure 2), is consistent with the virophage first model while it rules out the sister grouping of adenoviruses and NCLDVs (as expected by the nuclear escape hypothesis).

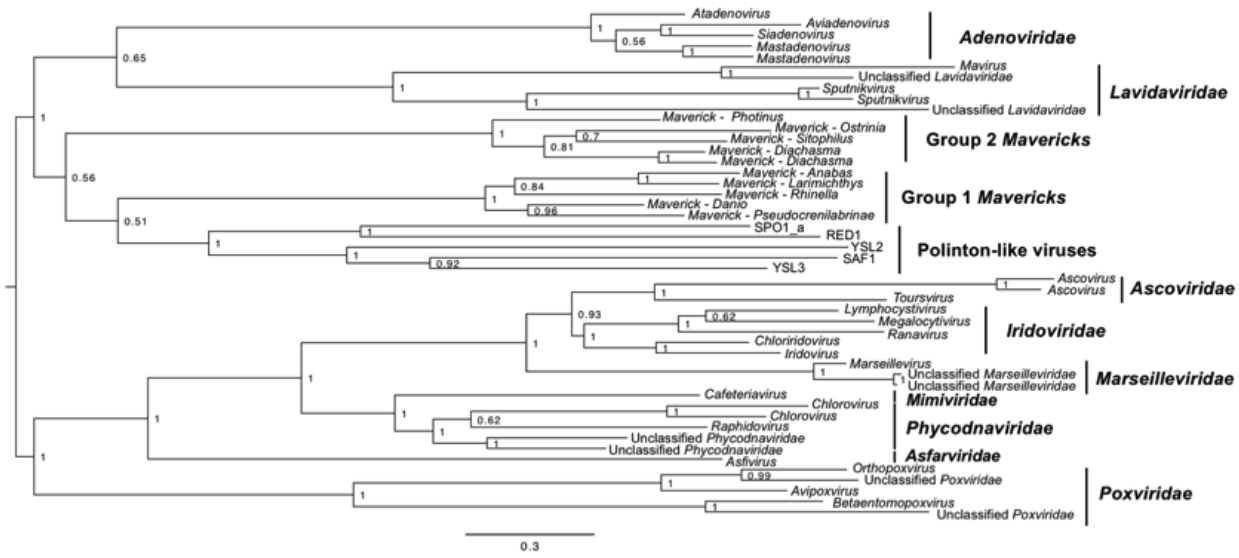

Figure 2. Bayesian majority-rule consensus tree based on the four core virion proteins (major and minor capsid proteins, adenoviral-like protease and DNA-packaging ATPase). The favoured topology is consistent with the virophage first hypothesis (compare with figure 1). The MCMC was run on MrBayes 3.2.7a with a length of 5,000,000 generations, with sampling every 10,000th generation. The concatenated alignment was unlinked for analysis.

A test of the Bayes factors calculated through stepping-stone analysis and contrasting the two model topologies, strongly favours the virophage first hypothesis both in the amino acid and combined (amino acid + genomic characters) datasets (Table 1). Alternatively, by using the posterior odds method, which is based on filtering the MCMC tree sample and counting the number of trees consistent with each hypothesis, we arrive at the same conclusion, the virophage first model is preferred (posterior odds >> 1) in both the amino acid and combined datasets (Table 2). Furthermore, by using a maximum-likelihood approach on the tree

topologies which is based exclusively on the amino acid data, the nuclear escape hypothesis is rejected at the level of $p < 0.01$ in all the 8 evaluated test statistics (Table 3).

Table 1. Bayes factor test for the two competing hypotheses estimated by stepping-stone analysis. The virophage first hypothesis is preferred for both the amino acid and combined datasets.

Hypothesis	**Characters**	**Marginal likelihood (ln)**	**Mean Marginal likelihood (ln)**	**ln Bayes Factor (M_0, M_1)**
Virophage first (M_0)	Combined	Run 1: -34,902.14	-34,902.01	+ 86.66***
		Run 2: -34,901.90		
Nuclear escape (M_1)	Combined	Run 1: -34,988.13	-34,988.67	
		Run 2: -34,989.88		
Virophage first (M_0)	Amino acid	Run 1: -33,021.80	-33,020.98	+ 354.23***
		Run 2: -33,020.54		
Nuclear escape (M_1)	Amino acid	Run 1: -33,374.75	-33,375.21	
		Run 2: -33,376.06		

*****Strong support for M_0**

Table 2. Bayesian topological hypothesis test following the method of Bergsten et al. (2013). A posterior sample of 750 trees (after 25% burn-in), was filtered to examine the frequency of topologies consistent with each hypothesis. The virophage first hypothesis is preferred.

Hypothesis	**Characters**	**MCMC tree frequency**	**Posterior model odds $P(M_0)/P(M_1)$**
Virophage first (M_0)	Combined	506/750 (0.67)	7.23
Nuclear escape (M_1)	Combined	70/750 (0.093)	
Virophage first (M_0)	Amino acid	595/750 (0.79)	31.32
Nuclear escape (M_1)	Amino acid	19/750 (0.025)	

Table 3. Probabilities of each hypothesis in Consel. The sister grouping of NCLDVs and adenoviruses (consistent with the nuclear escape hypothesis), is rejected with $p < 0.01$.

Hypothesis	**AU**[1]	**NP**[2]	**BP**[3]	**PP**[4]	**KH**[5]	**SH**[6]	**wKH**[7]	**wSH**[8]
Virophage first	0.995	0.995	0.996	1.000	0.996	0.996	0.996	0.996
Nuclear escape	0.005	0.005	0.004	5e-08	0.004	0.004	0.004	0.004

[1]Approximately-Unbiased test, [2]Bootstrap probability, [3]Bootstrap probability (calculated directly from replicates, $r_k = 1$), [4]Bayesian posterior probability (BIC approximation), [5]Kishino-Hasegawa test, [6]Shimodaira-Hasegawa test, [7]Weighted KH test, [8]Weighted SH test.

We also inferred the maximum-likelihood phylogenetic trees of the transcriptional proteins in cytoplasmic linear plasmids which are also present in eukaryotes and NCLDVs. As shown independently for the Rpb2 subunit, mRNA-capping enzyme and helicase (Fig. 3A-C), the homologues present in cytoplasmic linear plasmids have a distinct origin from those of NCLDVs. These observations are in line with the virophage first hypothesis, indicating that these genes were acquired independently by NCLDVs and cytoplasmic linear plasmids. Cytoplasmic linear plasmids have probably acquired these genes more recently from their eukaryotic hosts. In contrast, none of the phylogenies showed a topology consistent with a sister grouping of the NCLDV and cytoplasmic linear plasmid homologues, which is incompatible with the prediction of the nuclear escape model.

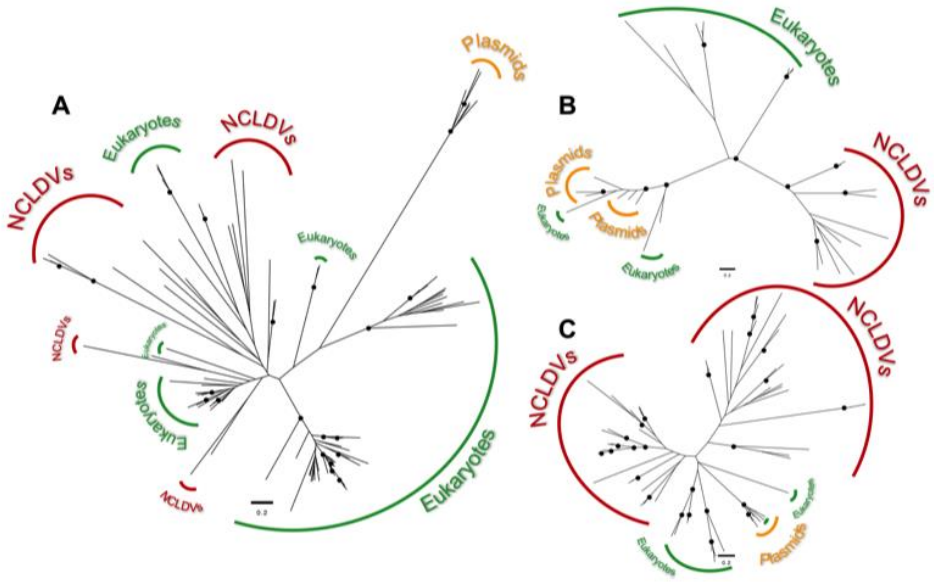

Figure 3. Maximum-likelihood phylogenetic trees (unrooted) of the transcriptional homologues encoded in cytoplasmic linear plasmids. (**A**) Trees for the DNA-directed RNA polymerase II, subunit Rpb2, (B) mRNA capping enzyme, and (C) helicase. In all cases, the topologies rule out a sister grouping of the NCLDV and cytoplasmic linear plasmid homologues, suggesting they were acquired independently. Black circles indicate bootstrap support ≥ 0.94 after 1,000 replicates.

Discussion

The analyses presented here build a strong case in favour of the virophage first model. These results highlight the ancient origins of NCLDVs and their coevolutionary arms race with virophages and eukaryotes. This virophage first model is able to explain a number of biological observations which are harder to reconcile with the nuclear escape scenario. The broad distribution of NCLDVs and *Mavericks* in all major eukaryotic lineages suggests that these groups had an early association with eukaryotes (Fischer & Suttle, 2011; Schulz et al., 2020). Recent evidence indicates that the NCLDV ancestor had already infected proto-eukaryotes 1-2 billion years ago, based on horizontal exchanges of the major subunits of the DNA-dependent RNA polymerase (Guglielmini et al., 2019). In addition, several previous works which separately analysed the major capsid protein, DNA-packaging ATPase and adenoviral-like protease also suggest this basal positioning of NCLDVs in the phylogenetic tree (Blanc et al., 2015; Krupovic et al., 2014; Yutin et al., 2013, 2015), consistent with our findings using the concatenated protein and combined character datasets.

A key feature of this model is the early divergence between NCLDVs and primitive virophages, which would then go on to diversify in eukaryotes. Once the ancestral large DNA virus started to exploit the host cytoplasm, this would have provided a window of opportunity for the emergence of parasitic virophages. At this early stage, the large DNA virus and the virophage would share the same promoter and poly-A sequences, since they have an immediate common ancestor. Several authors have emphasised it is unlikely that a transposon-like element would independently evolve the necessary control sequences *de novo* to successfully parasitise a coinfecting giant virus (Campbell et al., 2017; Fischer & Suttle, 2011). In the virophage first scenario, these sequences would have kept coevolving throughout evolutionary time, ensuring

the specificity between the virophage and its host virus. Furthermore, the parasitism of primitive virophages on the large DNA virus would have proved beneficial for the eukaryotic host, such that natural selection could favour the retention of endogenised virophages in the eukaryotic host's genome (Berjón-Otero et al., 2019). One possibility is that replacement of the protein-primed DNA polymerase in NCLDVs by an RNA-primed DNA polymerase, was an early strategy to reduce parasitism of virophages on NCLDV replication. These coevolutionary dynamics would explain the origin of *Mavericks*, their broad distribution across eukaryotes as well as the ability of virophages to integrate into the genome of their hosts. *Mavericks* could potentially represent a lineage of endogenised virophages which require coinfection by a host virus for reactivation (Barreat & Katzourakis, 2021).

An additional line of evidence in favour of the virophage first model are the distinct placements of the cytoplasmic linear plasmid and NCLDV transcriptional homologues in all the protein phylogenies. These results are a strong indication that these genes were captured independently during evolution. Convergent gene captures in viruses are a common occurrence, they have been reported for the retroviral superantigens in rhadinoviruses (*Herpesviridae*) (Aswad & Katzourakis, 2015), 32 homologous genes shared between insect entomopoxviruses and baculoviruses (Thézé et al., 2015), and in the translational machinery of NCLDVs (Koonin & Yutin, 2018). A separate origin for the transcriptional genes of cytoplasmic linear plasmids suggest these captures occurred more recently, after the divergence of NCLDVs, which would be consistent with their apparently restricted distribution in members of the Order Saccharomycetales (Fungi, Ascomycota) (Meinhardt et al., 1997).

Our results shed light on some of the earliest and least understood periods of viral evolution, giving clear evidence of the timescales involved and supporting the viral origins of one of the

major groups of eukaryotic mobile genetic elements. By using a combination of maximum-likelihood and Bayesian analyses, we have shown that the best supported model for the origin of virophages and NCLDVs is the 'virophage first' hypothesis. This model is strongly supported by our hypothesis tests of topology, and it is also able to explain the broad host range of *Mavericks* and NCLDVs in eukaryotes, the convergent gene captures between cytoplasmic linear plasmids and NCLDVs as well as the selective pressures which may have driven endogenisation of early virophages after their divergence from NCLDVs. These findings illuminate our understanding of eukaryotic virus evolution and support the ancient origins of virophages and NCLDVs, which have engaged in a coevolutionary arms race for at least a billion years.

Materials and Methods

Search for protein homologues

We selected a representative set of taxa in the families *Adenoviridae*, *Ascoviridae*, *Asfarviridae*, *Iridoviridae*, *Lavidaviridae*, *Mimiviridae*, *Marseilleviridae*, *Phycodnaviridae* and *Poxviridae* as reported in the ICTV Master Species List 2018b.v2. We obtained sequences for Group 1 *Mavericks* from Barreat & Katzourakis (Barreat & Katzourakis, 2021), and Group 2 elements from insect genomes following the same procedure described by these authors. Finally, we used sequences for the *Polinton*-like viruses reported by Yutin et al. (Yutin et al., 2015). A total of 48 taxa representing these groups were selected for our analysis (Table S1).

Elements in the PRD1-adenovirus lineage share an ancestral gene module encoding four core proteins: the major and minor jelly-roll fold capsid proteins, the adenoviral-like protease and

the HerA/FtsK DNA packaging ATPase. Although not all elements have the four genetic markers, gene sets are overlapping. In order to build a data set of these four proteins, we used existing gene annotations, open reading frame (ORF) predictions and blastp similarity searches (evalue < $1e^{-5}$) (Altschul et al., 1990). Homology of candidate proteins was further assessed using HHpred (Söding et al., 2005).

Multiple sequence alignment

Each set of homologous proteins was aligned in MAFFT v7.407 (Katoh et al., 2002). We then recovered conserved blocks for phylogenetic analysis by trimming the alignments in trimAl using the -automated1 option (which uses a heuristic to select the best trimming method) (Capella-Gutiérrez et al., 2009). Next, we concatenated the alignments for each taxon, and in case a protein marker was absent, we introduced the corresponding number of missing characters "?". The concatenated alignment had a total length of 486 amino acid characters (Figure S1).

Morphological matrix

To complement the concatenated protein alignments, we constructed a morphological matrix of genomic characters. This has been useful in recovering the phylogenetic relationships of retroviruses (Renoux-Elbé et al., 2002). We coded a total of 136 characters for presence/absence of genomic and virologic features (Additional Excel file 1). Terminal inverted repeats were found by blasting each genome sequence against itself and observing the resulting dot-plots (blastn), genome lengths were counted directly and characteristics of virus replication were obtained from the literature (Arantes et al., 2016; Asgari, 2006; Huang et al.,

2009; Jouvenet et al., 2004; Mutsafi et al., 2010; Risco et al., 2002; J. van Etten, 2009; J. L. van Etten et al., 2020). To find homologous gene sets, we predicted all the ORFs > 300 nucleotides in each genome and clustered them using vmatch (Kurtz, 2003). The identity of proteins in each cluster was confirmed with HHpred. Finally, we combined both data sets into a single alignment of amino acid and binary features.

Approximately-Unbiased test

A Bayesian phylogeny was inferred from the concatenated alignment in MrBayes version 3.2.7a (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003). Models for protein evolution were selected in ModelTest (Darriba et al., 2020) and applied to each partition. The families of viruses as well as NCLDVs were then set to be monophyletic with hard topological constraints. Markov-Chain Monte-Carlo was performed for 5,000,000 generations, sampling every 10,000 generations. The Potential Scale Reduction Factor (PSRF), was examined at the end to check all values were close to 1. The majority rule- consensus tree was consistent with the virophage first hypothesis, although it had low posterior probabilities for internal nodes (Figure 2). To obtain a topology for the alternative hypothesis, we manually introduced the branch leading to NCLDVs as the sister to adenoviruses. We optimised branch lengths for both topologies in RaxML (Kozlov et al., 2019) and then calculated sitewise likelihoods in PAML (Yang, 2007). The resulting sitewise likelihood files were used to contrast each hypothesis in Consel, under the AU and additional likelihood-based statistical tests (Shimodaira, 2002).

Marginal likelihood and Bayes factors

We estimated the marginal likelihood of two topological models using Bayesian stepping-stone analysis in MrBayes. First, we conducted a partitioned analysis using the combined protein and morphological data, using the best model of evolution for each protein partition and a discrete Dirichlet distribution for the morphological characters. We also conducted the same analysis only with the protein data, to assess the robustness of our conclusions to the exclusion of the morphological data matrix. To test the virophage-first hypothesis, we imposed monophyletic constraints on each of the main groups of elements and we explicitly prohibited a sister grouping between NCLDVs and adenoviruses. Conversely, for the nuclear-escape hypothesis, we imposed phylogenetic constraints for each of the main groups and enforced a sister grouping of adenoviruses with NCLDVs. Bayesian stepping-stone was conducted for 5,000,000 generations to obtain the marginal likelihood estimates under each model. Significance was measured as e to the power of the difference in log-marginal likelihoods, and compared to the table of Kass and Raftery (Kass & Raftery, 1995).

Bergsten et al. posterior odds method

We used a posterior tree sample from a Bayesian analysis conducted in MrBayes for calculation of the posterior model odds as described by Bergsten et al. (Bergsten et al., 2013). Monophyletic constraints were introduced for each of the major groups of elements but no specific sister groupings were enforced. We ran the analysis for 5,000,000 generations sampling every 10,000 generations, and checked the PSRF was close to 1 for all model parameters. The frequency of tree topologies consistent with each hypothesis was obtained by filtering the trees using topological constraints in PAUP version 4.0a (Swofford, 1998). Since

the trees are unrooted, we considered a topology to be consistent with the nuclear-escape hypothesis if a sister relationship between adenoviruses and NCLDVs could be observed. Alternatively, we considered trees to be consistent with the virophage-first hypothesis if a sister relationship between adenoviruses and any other group, except for NCLDVs, was possible. Posterior model odds were calculated as the ratio of the frequencies of trees consistent with each hypothesis.

Cytoplasmic linear plasmids

Cytoplasmic linear plasmids have lost the ancestral module of genes involved in formation of the capsid, but they have gained three genes encoding a mRNA capping-enzyme, a helicase and the Rpb2 subunit of the DNA-dependent RNA-polymerase II. These are homologous to the ones found in eukaryotes and NCLDVs. We used the proteins in cytoplasmic linear plasmids (Table S2), to search for homologous sequences in the non-redundant protein database using blastp and restricting searches to ‘Eukaryota (taxid:2759)’ and ‘Viruses (taxid:10239)’. The identity of significant matches (evalue < 1e-10) was confirmed with HHpred, and in the case of eukaryotes, only sequences mapping to chromosome assemblies were used. Virus, eukaryotic and plasmid homologues were aligned in Mafft and conserved blocks recovered from trimAl. We chose the best models for protein evolution in ModelTest and ran a maximum-likelihood tree search in RAxML-ng version 0.9.0 with 1,000 non-parametric bootstrap replicates for each protein separately.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Arantes, T. S., Lima, A., Silva, S., Oliveira, P., Souza, H. L. de, Khalil, J. Y. B., Oliveira, B. de, Torres, A., Colson, P., Kroon, G., Bonjardim, A., & Scola, L. (2016). The Large Marseillevirus Explores Different Entry Pathways by Forming Giant Infectiuos Vesicles. *Journal of Virology*, *90*(11), 5246–5255. https://doi.org/10.1128/JVI.00177-16.Editor

Asgari, S. (2006). Replication of Heliothis virescens ascovirus in insect cell lines. *Archives of Virology*, *151*(9), 1689–1699. https://doi.org/10.1007/s00705-006-0762-7

Aswad, A., & Katzourakis, A. (2015). Convergent capture of retroviral superantigens by mammalian herpesviruses. *Nature Communications*. https://doi.org/10.1038/ncomms9299

Barreat, J. G. N., & Katzourakis, A. (2021). Phylogenomics of the Maverick Virus-Like Mobile Genetic Elements of Vertebrates . *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msaa291

Barthélémy, R., Faure, E., & Goto, T. (2019). Serendipitous Discovery in a Marine Invertebrate (Phylum Chaetognatha) of the Longest Giant Viruses Reported till Date Virology : Current Research. *Virology: Current Research*, *3*(1), 1–13.

Bergsten, J., Nilsson, A. N., & Ronquist, F. (2013). Bayesian tests of topology hypotheses with an example from diving beetles. *Systematic Biology*, *62*(5), 660–673. https://doi.org/10.1093/sysbio/syt029

Berjón-Otero, M., Koslová, A., & Fischer, M. G. (2019). The dual lifestyle of genome-integrating virophages in protists. In *Annals of the New York Academy of Sciences* (Vol. 1447, Issue 1, pp. 97–109). Blackwell Publishing Inc. https://doi.org/10.1111/nyas.14118

Blanc, G., Gallot-Lavallée, L., & Maumus, F. (2015). Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(38), E5318–E5326. https://doi.org/10.1073/pnas.1506469112

Campbell, S., Aswad, A., & Katzourakis, A. (2017). Disentangling the origins of virophages and polintons. In *Current Opinion in Virology*. https://doi.org/10.1016/j.coviro.2017.07.011

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Darriba, Di., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2020). ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution*, *37*(1), 291–294. https://doi.org/10.1093/molbev/msz189

Fischer, M. G., & Hackl, T. (2016). Host genome integration and giant virus-induced reactivation of the virophage mavirus. *Nature*, *540*(7632), 288–291. https://doi.org/10.1038/nature20593

Fischer, M. G., & Suttle, C. A. (2011). A virophage at the origin of large DNA transposons. *Science*, *April*, 231–234.

Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P., & Gaia, M. (2019). Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1912006116

Harrach, B., Tarján, Z. L., & Benkő, M. (2019). Adenoviruses across the animal kingdom: a walk in the zoo. In *FEBS Letters* (Vol. 593, Issue 24). https://doi.org/10.1002/1873-3468.13687

Huang, X., Huang, Y., Sun, J., Han, X., & Qin, Q. (2009). Characterization of two grouper Epinephelus akaara cell lines: Application to studies of Singapore grouper iridovirus (SGIV) propagation and virus-host interaction. *Aquaculture*, *292*(3–4), 172–179. https://doi.org/10.1016/j.aquaculture.2009.04.019

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754–755. https://doi.org/10.1093/bioinformatics/17.8.754

Jouvenet, N., Monaghan, P., Way, M., & Wileman, T. (2004). Transport of African Swine Fever Virus from Assembly Sites to the Plasma Membrane Is Dependent on Microtubules and

Conventional Kinesin. *Journal of Virology*, *78*(15), 7990–8001. https://doi.org/10.1128/jvi.78.15.7990-8001.2004

Kapitonov, V. v., & Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(12). https://doi.org/10.1073/pnas.0600833103

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 319–323. https://doi.org/10.1108/10775730610619007

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Koonin, E. v., & Krupovic, M. (2017). Polintons, virophages and transpovirons: a tangled web linking viruses, transposons and immunity. *Current Opinion in Virology*, *25*(June), 7–15. https://doi.org/10.1016/j.coviro.2017.06.008

Koonin, E. v., & Yutin, N. (2018). Multiple evolutionary origins of giant viruses. In *F1000Research*. https://doi.org/10.12688/f1000research.16248.1

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A., & Wren, J. (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*(21), 4453–4455. https://doi.org/10.1093/bioinformatics/btz305

Krupovič, M., & Bamford, D. H. (2008). Virus evolution: How far does the double β-barrel viral lineage extend? *Nature Reviews Microbiology*, *6*(12), 941–948. https://doi.org/10.1038/nrmicro2033

Krupovic, M., Bamford, D. H., & Koonin, E. v. (2014). Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. *Biology Direct*, *9*(1), 1–7. https://doi.org/10.1186/1745-6150-9-6

Krupovic, M., & Koonin, E. v. (2015). Polintons: A hotbed of eukaryotic virus, transposon and plasmid evolution. *Nature Reviews Microbiology*, *13*(2), 105–115. https://doi.org/10.1038/nrmicro3389

Krupovic, M., & Koonin, E. v. (2016). Self-synthesizing transposons: Unexpected key players in the evolution of viruses and defense systems. *Current Opinion in Microbiology*, *31*(February), 25–33. https://doi.org/10.1016/j.mib.2016.01.006

Kurtz, S. (2003). The Vmatch large scale sequence analysis software. In *Ref Type: Computer Program* (2.3.1; pp. 4–12). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.3262&rep=rep1&type=pdf

la Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E., & Raoult, D. (2008). The virophage as a unique parasite of the giant mimivirus. *Nature*, *455*(7209), 100–104. https://doi.org/10.1038/nature07218

Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., Lescot, M., Poirot, O., Bertaux, L., Bruley, C., Couté, Y., Rivkina, E., Abergel, C., & Claverie, J. M. (2014). Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(11), 4274–4279. https://doi.org/10.1073/pnas.1320670111

Legendre, M., Doutre, G., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., & Claverie, J. (2013). Pandoraviruses : Amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*, *341*(July), 281–286. https://doi.org/10.1126/science.1239181

Meinhardt, F., Schaffrath, R., & Larsen, M. (1997). Microbial linear plasmids. In *Applied Microbiology and Biotechnology*. https://doi.org/10.1007/s002530050936

Mougari, S., Bekliz, M., Abrahao, J., Pinto, F. di, Levasseur, A., & la Scola, B. (2019). Guarani virophage, a new sputnik-like isolate from a Brazilian lake. *Frontiers in Microbiology*. https://doi.org/10.3389/fmicb.2019.01003

Mutsafi, Y., Zauberman, N., Sabanay, I., & Minsky, A. (2010). Vaccinia-like cytoplasmic replication of the giant Mimivirus. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(13), 5978–5982. https://doi.org/10.1073/pnas.0912737107

Pritham, E. J., Putliwala, T., & Feschotte, C. (2007). Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene*, *390*(1–2). https://doi.org/10.1016/j.gene.2006.08.008

Renoux-Elbé, C., Cheynier, R., & Wain-Hobson, S. (2002). Phylogeny derived from coding retroviral genome organization. *Journal of Molecular Evolution*, *54*(3), 376–385. https://doi.org/10.1007/s00239-001-0028-7

Risco, C., Rodríguez, J. R., López-Iglesias, C., Carrascosa, J. L., Esteban, M., & Rodríguez, D. (2002). Endoplasmic Reticulum-Golgi Intermediate Compartment Membranes and Vimentin Filaments Participate in Vaccinia Virus Assembly. *Journal of Virology*, *76*(4), 1839–1855. https://doi.org/10.1128/jvi.76.4.1839-1855.2002

Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, *19*(12), 1572–1574. https://doi.org/10.1093/bioinformatics/btg180

Schulz, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D. A., Denef, V. J., McMahon, K. D., Konstantinidis, K. T., Eloe-Fadrosh, E. A., Kyrpides, N. C., & Woyke, T. (2020). Giant virus diversity and host interactions through global metagenomics. *Nature*, *578*(7795). https://doi.org/10.1038/s41586-020-1957-x

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, *51*(3), 492–508. https://doi.org/10.1080/10635150290069913

Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, *33*(SUPPL. 2), 244–248. https://doi.org/10.1093/nar/gki408

Swofford, D. L. (1998). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. (No. 4; p. 143). Sinauer Associates.

Thézé, J., Takatsuka, J., Nakai, M., Arif, B., & Herniou, E. A. (2015). Gene acquisition convergence between entomopoxviruses and baculoviruses. *Viruses*. https://doi.org/10.3390/v7041960

van Etten, J. (2009). Lesser Known Large dsDNA Viruses: Preface. In *Current Topics in Microbiology and Immunology* (Vol. 328).

van Etten, J. L., Dunigan, D. D., Nagasaki, K., Schroeder, D. C., Grimsley, N., Brussaard, C. P. D., & Nissimov, J. I. (2020). Phycodnaviruses (Phycodnaviridae). In *Reference Module in Life Sciences* (Issue April). Elsevier Ltd. https://doi.org/10.1016/b978-0-12-809633-8.21291-0

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Dempsey, D. M., Dutilh, B. E., Harrach, B., Harrison, R. L., Hendrickson, R. C., Junglen, S., Knowles, N. J., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Nibert, M., Orton, R. J., Rubino, L., Sabanadzovic, S., … Davison, A. J. (2020). Changes to virus taxonomy and the Statutes ratified by the International Committee on Taxonomy of Viruses (2020). *Archives of Virology*, *165*(11). https://doi.org/10.1007/s00705-020-04752-x

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yutin, N., Raoult, D., & Koonin, E. v. (2013). Virophages, polintons, and transpovirons: A complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. *Virology Journal*, *10*. https://doi.org/10.1186/1743-422X-10-158

Yutin, N., Shevchenko, S., Kapitonov, V., Krupovic, M., & Koonin, E. v. (2015). A novel group of diverse Polinton-like viruses discovered by metagenome analysis. *BMC Biology*, *13*(1), 1–14. https://doi.org/10.1186/s12915-015-0207-4

Acknowledgements

JGNB is funded by the Jose Gregorio Hernandez Award from the National Academy of Medicine of Venezuela and Pembroke College, Oxford.

Competing interests

Authors declare that they have no competing interests.

Additional data

The data that supports the findings of this study has been deposited in figshare with the DOI: 10.6084/m9.figshare.14178626. Code files and results for the topological hypothesis tests described in this study are available in https://github.com/josegabrielnb/prd1-adenovirus.