Title

Evolution of the PRD1-adenovirus lineage: a viral tree of life incongruent with the cellular universal tree of life

Authors

Anthony C. Woo^{1,2*}, Morgan Gaia^{1,2}, Julien Guglielmini³, Violette Da Cunha^{1,2} and Patrick Forterre^{1,2*}

Affiliations

¹Unité de Biologie Moléculaire du Gène chez les Extrêmophiles (BMGE), Department of Microbiology, Institut Pasteur, 25-28 Rue du Docteur Roux, 75015 Paris, France.

²Department of Microbiology, CEA, CNRS, Université Paris-Sud, Université Paris-Saclay, Institute for Integrative Biology of the Cell (I2BC), Bâtiment 21, Avenue de la Terrasse, 91190 Gif-sur-Yvette cedex, France.

³HUB Bioinformatique et Biostatistique, C3BI, USR 3756 IP CNRS, Institut Pasteur, 25-28 Rue du Docteur Roux, 75015 Paris, France

*corresponding authors:

Anthony Woo: anthony.woo@pasteur.fr

Patrick Forterre: patrick.forterre@pasteur.fr

Abstract

Double-stranded DNA viruses of the PRD1-adenovirus lineage are characterized by homologous major capsid proteins containing one or two β -barrel domains known as the jelly roll folds. Most of them also share homologous packaging ATPases of the FtsK/HerA superfamily P-loop ATPases. Remarkably, members of this lineage infect hosts from the three domains of life, suggesting that viruses from this lineage could be very ancient and share a common ancestor. Here we analyzed the evolutionary history of these cosmopolitan viruses by inferring phylogenies based on single or concatenated genes. These viruses can be divided into two supergroups infecting either eukaryotes or prokaryotes. The latter can be further divided into two groups of bacteriophages and one group of archaeoviruses. This viral tree is thus incongruent with the cellular tree of life in which Archaea are closer to Eukarya and more divergent from Bacteria. We discuss various evolutionary scenarios that could explain this paradox.

34

35 **Introduction**

36 Studying virus origin and evolution is a challenging exercise, especially when
 37 addressing early co-evolution with cellular domains. While cellular domains (Archaea,
 38 Bacteria, and Eukarya) have been established based on ribosome sequences and more recently
 39 single-copy genes, viral lineages (or supergroups) have been suggested based on the
 40 conservation of their major capsid proteins (MCPs) structure (1–4). It has been proposed that
 41 MCP can be considered as the “viral self” or the “viral hallmark protein” (1, 5, 6). Viruses can
 42 be broadly divided into two types of lineages: domain-specific (identified in hosts from a
 43 single domain) and cosmopolitan lineages (identified in hosts from more than one domain) (7,
 44 8). To date, only viruses from the HK97 and the PRD1-adenovirus lineages, both
 45 corresponding mostly to double-stranded (ds) DNA viruses, infect host from the three
 46 domains of life (1, 3, 9, 10). The HK97 lineage mostly consists of archaeal and bacterial
 47 viruses, whereas the viruses of the PRD1-adenovirus lineage are well represented in the
 48 virosphere associated with all three domains (archaeoviruses, bacteriophages, and
 49 eukaryoviruses) and this lineage thus becomes an ideal subject to study the evolution of this
 50 viral lineage in the context of the universal tree of life.

51 The PRD1-adenovirus lineage was initially defined by the conservation of a double-
 52 jelly roll (DJR) fold in the MCP (11). This lineage is now characterized by either a single
 53 MCP with the DJR fold (1, 3, 12) or two MCPs, each with a single-jelly roll fold (SJR) (13–
 54 16). In both cases, the β -strands of their folds are oriented vertically relative to the capsid
 55 surface. These viruses will be called hereinafter DJR and SJR DNA viruses, respectively. The
 56 first attempt to determine the evolutionary relationships between DJR viruses was
 57 consequently based on structural alignments and confirmed that all viruses with DJR fold
 58 MCPs are indeed evolutionarily related (11).

59 The PRD1-adenovirus lineage encompasses different families as listed in Table 1.
 60 Among eukaryoviruses, all members of the PRD1-adenovirus lineage are DJR DNA viruses.
 61 Beside *Adenoviridae*, this lineage includes all members of the Nucleo-Cytoplasmic Large
 62 DNA Viruses (NCLDVs) assemblage (members of the families *Mimiviridae*,
 63 *Phycodnaviridae*, *Marseilleviridae*, *Iridoviridae*, *Ascoviridae*, *Poxviridae*, *Asfarviridae*, and
 64 related viruses), the *Lavidaviridae* (virophages related to viruses from the family of
 65 *Mimiviridae*) and the Polintoviruses (12, 17). In prokaryotes, most members of the PRD1-
 66 adenovirus lineage are also DJR DNA viruses, except *Sphaerolipoviridae*, which consists of
 67 SJR DNA viruses grouping two subfamilies of archaeal viruses and one subfamily of bacterial

viruses (18, 19). The best-known archaeal and bacterial DJR viruses of the PRD1-adenovirus lineage are small viruses, such as *Turriviridae*, *Tectiviridae*, and *Corticoviridae*, exemplified by the *Sulfolobus* Turreted Icosahedral Virus (STIV), the bacteriophage PRD1, and the *Pseudoalteromonas* virus PM2 infecting *Sulfolobus*, *Escherichia coli*, and *Pseudoalteromonas*, respectively (12, 20). Most of these prokaryotic viruses are also known to integrate into hosts genomes (20, 21) or exist as free plasmids corresponding to defective viruses (22). In terms of representation, the prokaryotic portion of the lineage was considered scarce compared to the eukaryotic counterpart, but until recently Koonin and co-workers highlighted the unexpected diversity of these viruses in archaeal and bacterial genomes and metagenomes (20). The authors identified five groups of prokaryotic dsDNA viruses related to the PRD1-adenovirus lineage based on detecting sequence similarities between their MCP: Odin (named after an integrated element present in the metagenome-assembled genome of an Odinararchaeon), PM2 (with some members, the Autolykoviruses, abundant in marine microbial metagenomes(21)), STIV (all the viruses infecting Archaea belong to this group, with the exception of the one presumably infecting the Odinararchaeon), Bam35/Toil/Cellulomonas, and PRD1. An additional group, FLiP (*Flavobacterium*-infecting, lipid-containing phage), composed of rolling-circle ssDNA viruses with a homologous, yet very divergent MCP, has also been identified (23).

In a phylogenetic tree based on the common core of the MCP published in 2013, the viruses of the DJR PRD1-adenovirus were divided into three groups: one corresponding to Adenoviruses, another grouping NCLDVs (represented by two members) with *Lavidaviridae*, and the third one including viruses infecting Archaea and Bacteria (11). Furthermore, the evolutionary relationships among the PRD1-adenovirus lineage were recently investigated using pairwise sequence similarities between DJR viruses (20, 24). These analyses demonstrated that there are evolutionary links between archaeoviruses and bacteriophages as well as eukaryoviruses and bacteriophages. With the advent of metagenomics approaches, many new viruses related to the PRD1-adenovirus lineage have been identified recently, providing important insights into the evolution of viruses from this lineage. This strengthens the call for a more thorough study to examine the evolutionary relationships among all MCPs from the DJR viruses.

In addition to the MCP, most members of the PRD1-adenovirus lineage, with a few exceptions, share a packaging ATPase (pATPase) of the FtsK/HerA superfamily P-loop ATPases (Table 1) (3, 20). The MCP and pATPase were also part of the core genes defined in our recent study on the evolution of the NCLDV assemblage (25), and concatenating the two

proteins proved useful in rooting the phylogenetic tree of NCLDV with the Polintoviruses as an outgroup, suggesting that such an approach could be extended to the whole PRD1-adenovirus lineage.

Here we show that the concatenated MCP and pATPases genes of the viruses of the PRD1-adenovirus lineage produces a robust viral phylogenetic tree characterized by a major division between eukaryotic and prokaryotic viruses, in contradiction with the topology of the universal cellular tree, which is characterized by the grouping of Archaea with eukaryotes. Accordingly, the viral phylogenetic tree is not in congruence with the cellular phylogenetic trees inferred using universal markers shared between the three cellular domains. We explore various scenarios that could explain this discrepancy between the viral tree and the cellular universal tree of life.

Results

The MCP and the pATPase genes were used as the hallmark proteins to identify viruses from the PRD1-adenovirus lineage

We retrieved MCP and pATPase sequences using PSI-BLAST searches against the GenBank non-redundant protein sequence database (nr), including sequences recovered from proviruses (Materials and Methods). To validate the selected MCP and pATPase sequences in our dataset, we generated protein models for all selected sequences and compared these predicted structures to the PDB database. The DJR MCPs associated to groups and families previously described indeed matched their corresponding structures in the public databases, except for the putative MCPs identified from the Odin group, confirming the structural conservation of the motif in the MCP of most of the viruses from the PRD1-adenovirus lineage (fig. S1 and data file S1). The results also showed that MCPs from the SJR groups were very divergent from those of the DJR groups. The MCP from *Adenoviridae* was unique in exhibiting several additional structural elements (fig. S1). Our analysis also showed that all pATPases share similar predicted structures including this time those from the SJR groups. The only exception was the pATPase of the *Adenoviridae*, which had again, like in the MCP analysis, the most distant structure compared to the other groups of DJR viruses (fig. S2 and data file S1).

A recent study has used the MCP genes of the PRD1-adenovirus lineage to create amino acid sequence similarity networks to analyze evolutionary relationships between viruses from this lineage (20). We thus performed a similar analysis based on the pATPases (fig. S3). Almost all viruses of the PRD1-adenovirus lineage were clustered together when

using the pATPases for the analysis, except for the *Adenoviridae*. This corroborates previous observations based on amino-acid signatures and secondary structure prediction suggesting that pATPases of the *Adenoviridae* were not specifically related to other pATPases of viruses of the PRD1-adenovirus lineage but were derived from ATPases of the ABC superfamily (26), indicating an exchange of pATPases during the evolution of this viral family. Surprisingly, we found that the pATPases of the SJR DNA viruses did not form a distinct cluster, as expected for an outgroup, but strongly clustered together with the pATPases of DJR archaeoviruses and bacteriophages. It was possible to distinguish two large assemblages in the pATPase network, one corresponding to a core NCLDV loosely associated to the *Asfarviridae* on one side and the *Poxviridae* and the Polintoviruses on the other, and a second one corresponding to archaeoviruses and bacteriophages (including SJR viruses), loosely associated to the *Lavidaviridae*.

Phylogenetic positions of the PRD1-adenovirus lineage viruses inferred from the single protein genes

The result of the sequence similarity network suggested that the pATPase gene could be a good marker for delineating the phylogeny of the viruses from the PRD1-adenovirus lineage, in addition to the MCP that was used in recent studies (11, 20). Sequences of the MCPs (excluding and both SJR viruses and the *Adenoviridae*) and pATPases were aligned and trimmed (Materials and Methods). Phylogenetic analyses were performed separately on the MCP and the pATPase using the resulting datasets within the maximum likelihood (ML) framework. DJR eukaryoviruses were clearly separated from archaeoviruses and bacteriophages in both trees, in agreement with previous observations made using pairwise sequence similarities or comparative structural analyses (11, 24). In particular, unlike the result obtained with the sequence similarity network, members of the *Lavidaviridae* were grouped with other eukaryoviruses in the pATPase tree. We thus decided to root the trees between “prokaryotic” and “eukaryotic” DJR viruses. As a result, DJR eukaryoviruses and viruses infecting Archaea or Bacteria formed two monophyletic clades. Moreover, we recovered the monophyly of most previously defined assemblages at various taxonomic ranks (groups, families or superfamilies). The few exceptions were the STIV, the *Phycodnaviridae* and the *Iridoviridae* that were paraphyletic in both the MCP and pATPase trees, and the PM2 group which was paraphyletic in the pATPase tree (fig. S4 and S5). Mollivirus, which is related to pandoraviruses, branched within the *Phycodnaviridae* and this corroborates previous observations (25, 27).

In the MCP tree, the NCLDV formed a monophyletic sister group to the *Lavidaviridae*, whereas in the pATPase tree, the *Lavidaviridae* branched at the base of the eukaryotic clade and the NCLDV were not monophyletic as the *Poxviridae* were sister group to the Polintoviruses. On the prokaryotic side, viruses of the STIV group were the sister group to viruses of the PM2 group (except for the two *Sulfolobus* STIV), whereas the five other groups of DJR bacteriophages (the Bam35, the PRD1, the Toil, the Cellulomonas and the FLiP) emerged at the base of this part of the tree (fig. S4).

In the pATPase tree, the Bam35, the PRD1, the Toil and the Cellulomonas groups formed a monophyletic cluster that was sister group to all STIV (except two members of the STIV group) whereas the PM2 group and the two STIV members branched at the base of the tree, within SJR viruses of the *Sphaerolipoviridae* family (fig. S5). Remarkably, the SJR pATPases did not form a monophyletic group well separated from DJR viruses, as expected if DJR viruses originated from SJR viruses (28), but grouped with DJR viruses infecting prokaryotes in agreement with the network analysis (fig. S3 and S5). To better address the evolutionary relationship between the pATPases of SJR and DJR viruses, we performed another phylogenetic analysis using only sequences of bacterial and archaeal viruses (fig. S6). Interestingly, the pATPases of archaeal SJR DNA viruses branched within the clade of archaeal DJR DNA viruses, whereas the pATPases of bacterial SJR viruses formed a separate clade. It has been previously suggested that DJR viruses originated from the simpler SJR type by ancestral gene duplication of one of the two SJR fold MCPs that possibly occurred before the emergence of the Last Universal Common Ancestor (LUCA) (29, 30). If one assumes that SJR viruses, defined by their very divergent MCPs, predated DJR viruses as previously proposed (28), this suggests that they replaced later on their pATPases with those of DJR viruses infecting the same hosts. Alternatively, SJR viruses might have emerged twice independently from DJR viruses and the MCP genes of SJR viruses might have diverged rapidly following their structural rearrangements.

To better compare the MCP and pATPase trees, we removed the sequences of the SJR viruses whose MCP cannot be included in the MCP tree from the pATPase dataset and we removed viruses of the FLiP group that have no detectable pATPases from the MCP dataset (Fig. 1 and fig. S7 and S8). The removal of SJR did not change the topology of the pATPase tree, except for some modifications in the relative branching of DJR viruses from the PM2 and the STIV groups. Surprisingly, the removal of the FLiP viruses from the MCP tree also introduced some modifications in the relationships between the different groups of viruses infecting eukaryotes, especially the *Poxviridae* and the *Asfarviridae*. The *Poxviridae* and the

Asfarviridae are however known to be unstable taxa, possibly due to their frequent long branches. Nevertheless, most of the clades were still monophyletic and the bipartition corresponding to eukaryoviruses and prokaryoviruses was still present.

Concatenation of the MCP and the pATPase genes highlights an unexpected evolutionary scenario

The small differences observed between the MCP and the pATPase trees, the uncertainties in the topologies due to the removal of some sequences, and the low support values of most branches in these single trees were anticipated considering the divergence of the analyzed proteins. However, these trees also exhibited some clear-cut congruence, such as the eukaryotic/prokaryotic split, the monophyly of the NCLDV other than the *Poxviridae*, the monophyly of several previously defined groups of viruses infecting Archaea and Bacteria, the grouping of the STIV and the PM2 on one side and the grouping of the PRD1, the Bam35, the PRD1, the Toil, and the Cellulomonas on the other (Fig.1 and fig. S7 and S8). This suggested that we could concatenate the MCP and the pATPase sequences to obtain a more reliable phylogeny. We adopted the new version of transfer bootstrap in which the presence of inferred branches in replications is measured using a gradual ‘transfer’ distance and the resulting supports do not induce falsely supported branches (31). In the concatenated ML tree, we recovered the division between DJR viruses infecting eukaryotes and those infecting prokaryotes, as well as the monophyly of all previously defined groups of DJR viruses, except for the STIV group. Interestingly, the STIV group was now divided into two subgroups corresponding to DJR viruses infecting either Archaea or Bacteria and most of the branch supports were above 0.9 (in the range of 0 to 1), indicating that the tree is very robust (Fig. 2).

Among the eukaryoviruses of the tree, the Polintoviruses were basal to all eukaryotic groups whereas the NCLDV assemblage was sister group to the *Lavidaviridae*. The *Poxviridae* were the first NCLDVs branching family, followed by the *Asfarviridae*. Notably, we recovered the two major groups of NCLDVs that we previously identified based on 8 marker core genes (25), the MAPI (*Marseilleviridae*, *Ascoviridae*, Pitho-like viruses, and *Iridoviridae*) and the PAM (*Phycodnaviridae*, *Asfarviridae* and the recently proposed order “Megavirales”), except that the *Asfarviridae* were sister group to these two superclades instead to be part of the PAM and the *Phycodnaviridae* were paraphyletic. Remarkably, we also previously obtained *Asfarviridae* as sister group to other NCLDVs in a tree based on the two large RNA polymerase subunits, using cellular sequences as an outgroup (25). We previously noticed that the *Poxviridae* had exceptionally long branches and variable positions

in single-gene trees of NCLDV proteins. In particular, they tended to attract the long branches of the *Asfarviridae* in our previous analyses. We thus removed both the *Poxviridae* and the *Asfarviridae* from the dataset and we obtained a tree in which NCLDVs are no more sister group to the *Lavidaviridae* (fig. S9), but the Polintoviruses, as previously proposed based on phenotypic analyses (32).

Noticeably, two strongly supported clades of viruses infecting bacteria and one clade of viruses infecting Archaea were recovered on the prokaryotic side of the tree (Fig. 2). The bacterial members of the STIV formed a strongly supported monophyletic clade with bacterial viruses/proviruses of the PM2 group that branched at the base of this clade and will be hereinafter called the DJR bacterial cluster I. The second bacterial clade with the *Cellulomonas*, the Bam, the Toil and the PRD1 (hereinafter called the DJR bacterial cluster II) was similar to a clade first observed with the pATPase tree. The third clade included all archaeal members of the STIV group (hereinafter called the DJR archaeal cluster). The DJR archaeal cluster became sister group to the DJR bacterial cluster I when both the *Poxviridae* and the *Asfarviridae* were removed from the analysis, with members the DJR bacterial cluster II becoming paraphyletic at the base of the “prokaryotic clade” (fig. S9).

The monophyly of DJR archaeoviruses, previously observed with the complete MCP tree was only weakly supported in the concatenated tree with or without the *Poxviridae* and the *Asfarviridae*. To improve the resolution, we removed eukaryoviruses from the concatenated tree. We again obtained the monophyly of the DJR bacterial clusters I, II, and the archaeal cluster if we rooted the tree between Archaea and Bacteria (Fig. 3). On this occasion, we obtained good support for the monophyly of Archaea. Interestingly, viruses infecting Euryarchaea and those infecting Crenarchaea or Thaumarchaea were clearly separated in the tree, suggesting that these viruses might have been present at the time of the last common ancestor of Archaea and later on co-evolved with cellular hosts from this domain.

Discussion

The PRD1-adenovirus lineage is one of the two major viral lineages with members infecting hosts from all three domains of life. It has long been thought that inferring phylogeny of this lineage was out of reach because of the absence of significant sequence similarities between their MCPs (11). This could be because the first viruses identified in this lineage, PRD1 and Adenoviruses, are the most divergent viruses of this lineage. However, many additional large groups of DNA viruses from the three domains of life were later on progressively incorporated in this lineage, revealing previously undetectable connection at the

primary sequence level between their MCPs and pATPases. Here, we show that most members of this lineage can be placed in a robust “universal tree” based on the concatenation of these two proteins. A similar strategy has been recently adopted to produce a global evolutionary history of RNA viruses based on the phylogeny of their RNA-dependent RNA polymerases (33). Our results are supported by the fact that we recovered most groups previously defined on different criteria, with the only exception of the STIV group (20). In particular, the internal phylogeny of the NCLDV is comparable to the one that we previously obtained with eight-core genes (25).

The tree presented here will certainly change in the future with the discovery of new viral groups belonging to this lineage and the discovery of new members of the existing groups. Nevertheless, our findings suggest that the phylogenetic tree of the concatenated MCP and pATPase can be used as a backbone to review current hypotheses about the evolution of this lineage and propose new ones.

Interestingly, our in-depth phylogenetic analyses reveal a great divide between DJR viruses infecting prokaryotes (Archaea and Bacteria) and those infecting eukaryotes. This produces a “viral tree of life” strikingly different from the cellular tree based on universal proteins in which Archaea and eukaryotes are closely related (34–37). Interestingly, the topology of this viral tree of life corroborates phenotypic observations showing that the mobilomes of Archaea and Bacteria are both very similar and strikingly different from the eukaryotic mobilome (38). For instance, Caudovirales (head and tailed “phages”), conjugative plasmids, casposons, or several families of IS elements are only present in Archaea and Bacteria (39–44), whereas several families of RNA viruses, retroviruses, and certain families of IS elements are specific to eukaryotes (38, 45). Similarly, defense systems are also distributed according to the “eukaryote/prokaryote” divide, with the CRISPR-Cas system or the toxin-antitoxin systems only present in Archaea and Bacteria whereas the RNAi system only exists in eukaryotes (46, 47).

It has been suggested that all eukaryoviruses, including those from the PRD1-adenovirus lineage, originated from a melting pot of bacteriophages (bacteriophages) infecting the bacterium at the origin of mitochondria (48). Regarding the origin of DJR viruses, it was suggested that Polintoviruses evolved from a *Tectiviridae* and the Polintoviruses subsequently became the ancestor of *Lavidaviridae* and NCLDVs (32). However, in that scenario, DJR eukaryoviruses, especially Polintoviruses, should have been more closely related to *Tectiviridae* than to other DJR bacteriophages in the MCP/pATPase DJR tree, which is not the case.

Phylogenetic analyses of cellular and NCLDV RNA polymerases revealed that NCLDVs had already diverged before the Last Eukaryotic Common Ancestor (LECA) (25). Our present study indicates that the divergence between NCLDVs, Polintoviruses and *Lavidaviridae* took place even earlier. This renders unlikely that the transformation of an ancestral prokaryotic-like DJR virus into these three groups of eukaryoviruses had enough time to occur during the evolutionary period between the first eukaryote with a proto-mitochondrion and LECA. If they all originated from the mitochondrion, it would leave a rather short lapse of time before LECA for so many various families to emerge. Notably, among these three groups of eukaryoviruses, one group encompasses diverse large and giant viruses infecting a range of host from unicellular algae to human beings. A more plausible related hypothesis could be that DJR eukaryoviruses evolved from DJR viruses that infected prokaryotes that entered into symbiosis with proto-eukaryotes much before the final endosymbiosis that led to mitochondria, leaving more time for the evolution and diversification of DJR eukaryoviruses.

If DJR eukaryoviruses originated indeed from bacteriophages or archaeoviruses (hereinafter called the “prokaryotic” ancestor hypothesis), the DJR universal tree could be rooted with the DJR archaeal cluster (fig. S10), the DJR bacterial clusters (fig. S11 and S12) or between the DJR archaeal cluster and the DJR bacterial cluster I (Fig. 4A and fig. S13). In this scenario, according to our MCP/pATPase tree, DJR eukaryoviruses could have originated from a third unknown group of DJR viruses, sister group to DJR bacterial cluster I. In any case, it remains unclear in the “prokaryotic” ancestor hypothesis how a virus infecting a prokaryotic symbiont could have acquired the capacity to infect the proto-eukaryotic host.

The presence of DJR viruses in the three domains has been often used as an argument suggesting that these viruses were already present at the time of the Last Universal Common Ancestor (LUCA) (1)(49)(50). In the “prokaryotic” ancestor hypothesis, the presence of DJR viruses in LUCA favors rooting the whole DJR tree between archaeoviruses and bacteriophages, as in fig. S13. However, the branch between archaeoviruses and bacteriophages was very short in fig. S13, whereas the branch between archaeal and bacterial proteins is always very long in the trees of universal proteins that were present in LUCA (34)(51). In the “prokaryotic” ancestor hypothesis, it thus seems more likely that DJR viruses originated after LUCA, either in proto-archaea or proto-bacteria and were later on transferred from one domain to the other.

Alternatively, the presence of DJR viruses in LUCA would explain the clear-cut separation of prokaryotic and eukaryotic DJR viruses in the whole DJR tree. However,

considering that viruses generally co-evolve with their hosts, it remains obscure why DJR archaeoviruses are much more similar to DJR bacteriophages than to DJR eukaryoviruses, since Archaea and Eukarya are closely related in the universal tree of life. A plausible explanation for this paradox could be that eukaryotic-like DJR viruses were first lost in the stem lineage of Archaea and reintroduced later on in this domain from Bacteria (or protobacteria) before the diversification of Archaea (Fig. 4B). The present absence of eukaryotic-like DJR viruses in Archaea could be also due to a sampling bias, some of these archaeal viruses being still hidden in some unexplored places around the globe.

The presence of DJR viruses in the three cellular domains thus raises challenging questions about their origin and evolution. Here, we have shown that phylogenies based on the concatenation of their MCP and pATPase not only can help to formulate more precisely previous evolutionary hypotheses but also to propose new ones, such as the monophyly of DJR archaeoviruses and to define new assemblages, such as DJR clusters I and II in Bacteria. Future identification and isolation of new viral families in the PRD1-adenovirus lineage will be essential to provide insights into various scenarios for the history of this lineage.

Materials and Methods

Selection of MCP/pATPase sequences

Representative MCP/pATPase sequences from different groups of the DJR viruses were used as queries for PSI-BLAST (52) searches against the GenBank non-redundant protein sequence database (nr). The query sequences are listed below:

Group	Name	MCP	pATPase
<i>Lavidaviridae</i>	Sputnik virophage	YP_002122381	YP_002122364
<i>Adenoviridae</i>	Frog adenovirus 1	NP_062443	NP_062434
STIV	Sulfolobus turreted icosahedral virus 1	YP_025022	YP_025021
Bam35	Bacillus phage Bam35c	NP_943764	NP_943760
PRD1	Enterobacteria phage PRD1	NP_040692	NP_040689
Toil	Rhodococcus phage Toil	ARK07697	ARK07695
PM2	Pseudoalteromonas phage PM2	NP_049903	NP_049900
FLiP	Flavobacterium phage FLiP	ASQ41214	-

The MCP/pATPase sequences of the NCLDVs were retrieved from a previous study that we conducted (25). The Polintoviruses sequences were gathered from the Repbase collection (53) (http://www.girinst.org/Repbase_Update.html): Polinton-2_NV, Polinton-1_DY, Polinton-

1_TC, Polinton-1_SP, Polinton-2_SP, Polinton-2_DR, Polinton-1_DR. Finally, the sequences from the SJR group were recovered based on previously identified sequences (19).

Putative MCP/pATPase sequences were aligned with the query sequences for the examination of the conserved structural elements using MAFFT (54). Prediction of the secondary structure was performed using Phyre2 (55) and the predicted protein structures were visualized using UCSF Chimera (56). The sequences used in this study are shown in the supplementary file. After removing sequences with no significant matches or low confidence levels, we obtained two different datasets of 145 and 128 sequences for the MCP and pATPase respectively.

Network analysis

After performing the structural protein prediction analysis, all-against-all blastp analyses were performed on the refined MCP and pATPase datasets. The all-against-all integrases BlastP results were grouped using the SiLiX (for *Single Linkage Clustering of Sequences*) package v1.2.8 (<http://lbbe.univ-lyon1.fr/SiLiX>) (57). This approach for the clustering of homologous sequences is based on single transitive links with alignment coverage constraints. Several different criteria can be used separately or in combination to infer homology (percentage of identity, alignment score or E-value, alignment coverage). The MCP and pATPase sequences were clustered independently by similarity using SiLiX with the expect threshold of 0.001 as previously used for MCP analysis (20). The clustering results were analyzed and visualized using the igraph package of the R programming language (<https://igraph.org/>).

Phylogenetic analysis

The alignments of the MCP sequences were performed using MAFFT v7.392 with the E-INS-i algorithm (54), which can align sequences with several conserved motifs embedded in long unalignable regions, whereas pATPase sequences were aligned using MAFFT with the L-INS-i algorithm (54), which can align a set of sequences containing sequences flanking around one alignable domain. Positions containing more than 30% of gaps were trimmed using goalign v0.2.8 (<https://github.com/evolbioinfo/goalign>).

Maximum likelihood phylogenies

Single protein and concatenated protein phylogenies were conducted within the Maximum Likelihood (ML) framework using IQ-TREE v1.6.3 (58). We first performed a model test with the Bayesian Information Criterion (BIC) by including protein mixture models (59). For

mixture model analyses, we used the PMSF models (60). The support values were either transfer bootstrap expectation (TBE) support (31) computed from 1000 bootstrap replicates in the case of nonparametric bootstrap using gotree v0.3.0 (<https://github.com/evolbioinfo/gotree>), or from 1,000 replicates for SH-like approximation likelihood ratio test (aLRT) (61) and ultrafast bootstrap approximation (UFBoot) (62).

Visualization

The phylogenetic trees were visualized with FigTree v1.4.3 (<http://tree.bio.ed.ac.uk/software/figtree/>) and iTOL (63).

Supplementary Materials

Fig. S1. Major capsid proteins of the PRD1-adenovirus lineage.
 Fig. S2. Packaging ATPases of the PRD1-adenovirus lineage.
 Fig. S3. Sequence similarity networks of the PRD1-adenovirus packaging ATPase proteins.
 Fig. S4. Maximum likelihood (ML) phylogenetic tree of the major capsid protein gene of the viruses from the PRD1-adenovirus lineage.
 Fig. S5. Maximum likelihood (ML) phylogenetic tree of the packaging ATPase gene of the viruses from the PRD1-adenovirus lineage.
 Fig. S6. Maximum likelihood (ML) phylogenetic tree of the packaging ATPase gene of the archaeoviruses and bacteriophages from the PRD1-adenovirus lineage.
 Fig. S7. Maximum likelihood (ML) phylogenetic tree of the packaging ATPase gene of the double jelly-roll viruses from the PRD1-adenovirus lineage, excluding single jelly-roll viruses.
 Fig. S8. Maximum likelihood (ML) phylogenetic tree of the major capsid protein gene of the double jelly-roll viruses from the PRD1-adenovirus lineage, excluding the FLiP members.
 Fig. S9. Maximum likelihood (ML) phylogenetic tree of the concatenated major capsid protein and packaging ATPase genes, excluding the *Poxviridae* and the *Asfarviridae*.
 Fig. S10. Maximum likelihood phylogenetic tree of the concatenated major capsid protein and packaging ATPase genes, rooting with the DJR archaeal cluster.
 Fig. S11. Maximum likelihood phylogenetic tree of the concatenated major capsid protein and packaging ATPase genes, rooting with the DJR bacterial cluster II.
 Fig. S12. Maximum likelihood phylogenetic tree of the concatenated major capsid protein and packaging ATPase genes, rooting with the DJR cluster I.
 Fig. S13. Maximum likelihood phylogenetic tree of the concatenated major capsid protein and packaging ATPase genes, rooting between the DJR archaeal cluster and the DJR bacterial cluster II.
 Data file S1. Results of the predicted structures of the major capsid protein and the packaging ATPase genes.

References

1. D. H. Bamford, Do viruses form lineages across different domains of life? *Res. Microbiol.* (2003), , doi:10.1016/S0923-2508(03)00065-2.

2. D. H. Bamford, J. M. Grimes, D. I. Stuart, What does structure tell us about virus evolution? *Curr. Opin. Struct. Biol.* (2005), , doi:10.1016/j.sbi.2005.10.012.
3. N. G. A. Abrescia, D. H. Bamford, J. M. Grimes, D. I. Stuart, Structure Unifies the Viral Universe. *Annu. Rev. Biochem.* **81**, 795–822 (2012).
4. P. Forterre, M. Krupovic, D. Prangishvili, Cellular domains and viral lineages. *Trends Microbiol.* (2014), , doi:10.1016/j.tim.2014.07.004.
5. M. Krupovic, D. H. Bamford, Order to the Viral Universe. *J. Virol.* **84**, 12476–12479 (2010).
6. P. Forterre, The virocell concept and environmental microbiology. *ISME J.* (2013), , doi:10.1038/ismej.2012.110.
7. J. Iranzo, M. Krupovic, E. V. Koonin, The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *MBio.* **7**, 1–21 (2016).
8. D. Prangishvili, D. H. Bamford, P. Forterre, J. Iranzo, E. V. Koonin, M. Krupovic, The enigmatic archaeal virosphere. *Nat. Rev. Microbiol.* **15**, 724–739 (2017).
9. R. W. Hendrix, Evolution: The long evolutionary reach of viruses. *Curr. Biol.* (1999), , doi:10.1016/S0960-9822(00)80103-7.
10. M. L. Baker, W. Jiang, F. J. Rixon, W. Chiu, Common Ancestry of Herpesviruses and Tailed DNA Bacteriophages. *J. Virol.* **79**, 14967–14970 (2005).
11. J. Ravantti, D. Bamford, D. I. Stuart, Automatic comparison and classification of protein structures. *J. Struct. Biol.* **183**, 47–56 (2013).
12. C. San Martín, M. J. van Raaij, The so far farthest reaches of the double jelly roll capsid protein fold. *Virol. J.* (2018), doi:10.1186/s12985-018-1097-1.
13. H. T. Jaalinoja, E. Roine, P. Laurinmaki, H. M. Kivela, D. H. Bamford, S. J. Butcher, Structure and host-cell interaction of SH1, a membrane-containing, halophilic euryarchaeal virus. *Proc. Natl. Acad. Sci.* (2008), doi:10.1073/pnas.0801758105.
14. S. T. Jaatinen, L. J. Happonen, P. Laurinmäki, S. J. Butcher, D. H. Bamford, Biochemical and structural characterisation of membrane-containing icosahedral dsDNA bacteriophages infecting thermophilic *Thermus thermophilus*. *Virology* (2008), doi:10.1016/j.virol.2008.06.023.
15. S. T. Jaakkola, R. K. Penttinen, S. T. Vilen, M. Jalasvuori, G. Ronnholm, J. K. H. Bamford, D. H. Bamford, H. M. Oksanen, Closely Related Archaeal Haloarcula hispanica Icosahedral Viruses HHIV-2 and SH1 Have Nonhomologous Genes Encoding Host Recognition Functions. *J. Virol.* (2012), doi:10.1128/JVI.06666-11.
16. Z. Zhang, Y. Liu, S. Wang, D. Yang, Y. Cheng, J. Hu, J. Chen, Y. Mei, P. Shen, D. H.

- 476 Bamford, X. Chen, Temperate membrane-containing halophilic archaeal virus SNJ1
477 has a circular dsDNA genome identical to that of plasmid pHH205. *Virology*. **434**,
478 233–241 (2012).
- 479 17. M. Krupovic, J. H. Kuhn, M. G. Fischer, A classification system for virophages and
480 satellite viruses. *Arch. Virol.* **161**, 233–247 (2016).
- 481 18. D. Gil-Carton, S. T. Jaakkola, D. Charro, B. Peralta, D. Castaño-Díez, H. M. Oksanen,
482 D. H. Bamford, N. G. A. Abrescia, Insight into the assembly of viruses with vertical
483 single β -barrel major capsid proteins. *Structure*. **23**, 1866–1877 (2015).
- 484 19. T. A. Demina, M. K. Pietilä, J. Svirskaitė, J. J. Ravantti, N. S. Atanasova, D. H.
485 Bamford, H. M. Oksanen, HCIV-1 and other tailless icosahedral internal membrane-
486 containing viruses of the family Sphaerolipoviridae. *Viruses*. **9** (2017),
487 doi:10.3390/v9020032.
- 488 20. N. Yutin, D. Bäckström, T. J. G. Ettema, M. Krupovic, E. V. Koonin, Vast diversity of
489 prokaryotic virus genomes encoding double jelly-roll major capsid proteins uncovered
490 by genomic and metagenomic sequence analysis. *Virol. J.* **15**, 1–17 (2018).
- 491 21. K. M. Kauffman, F. A. Hussain, J. Yang, P. Arevalo, J. M. Brown, W. K. Chang, D.
492 Vaninsberghe, J. Elsherbini, R. S. Sharma, M. B. Cutler, L. Kelly, M. F. Polz, A major
493 lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature*
494 (2018), doi:10.1038/nature25474.
- 495 22. M. Gaudin, M. Krupovic, E. Marguet, E. Gauthier, V. Cvirkaite-Krupovic, E. Le Cam,
496 J. Oberto, P. Forterre, Extracellular membrane vesicles harbouring viral genomes.
497 *Environ. Microbiol.* **16**, 1167–1175 (2014).
- 498 23. E. Laanto, S. Mäntynen, L. De Colibus, J. Marjakangas, A. Gillum, D. I. Stuart, J. J.
499 Ravantti, J. T. Huiskonen, L.-R. Sundberg, Virus found in a boreal lake links ssDNA
500 and dsDNA viruses. *Proc. Natl. Acad. Sci.* **114**, 8378–8383 (2017).
- 501 24. R. Sinclair, J. Ravantti, D. H. Bamford, Nucleic and Amino Acid Sequences
502 Classification. *J. Virol.* **91**, 1–13 (2017).
- 503 25. J. Guglielmini, A. Woo, M. Krupovic, P. Forterre, M. Gaia, Diversification of giant
504 and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *bioRxiv*
505 (2018), doi:10.1101/455816.
- 506 26. A. Burroughs, L. Iyer, L. Aravind, Comparative genomics and evolutionary trajectories
507 of viral ATP dependent DNA-packaging systems. *Genome Dyn.* **3**, 48–65 (2007).
- 508 27. N. Yutin, E. V. Koonin, Pandoraviruses are highly derived phycodnaviruses. *Biol.*
509 *Direct* (2013), doi:10.1186/1745-6150-8-25.

28. E. V. Koonin, M. Krupovic, Polintons, virophages and transpovirons: a tangled web linking viruses, transposons and immunity. *Curr. Opin. Virol.* (2017), , doi:10.1016/j.coviro.2017.06.008.
29. M. Krupović, D. H. Bamford, Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. *Virology*. **375**, 292–300 (2008).
30. N. G. A. Abrescia, J. M. Grimes, H. M. Kivelä, R. Assenberg, G. C. Sutton, S. J. Butcher, J. K. H. Bamford, D. H. Bamford, D. I. Stuart, Insights into Virus Evolution and Membrane Biogenesis from the Structure of the Marine Lipid-Containing Bacteriophage PM2. *Mol. Cell*. **31**, 749–761 (2008).
31. F. Lemoine, J. B. Domelevo Entfellner, E. Wilkinson, D. Correia, M. Dávila Felipe, T. De Oliveira, O. Gascuel, Renewing Felsenstein’s phylogenetic bootstrap in the era of big data. *Nature* (2018), doi:10.1038/s41586-018-0043-0.
32. M. Krupovic, E. V. Koonin, Polintons: A hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat. Rev. Microbiol.* **13**, 105–115 (2015).
33. Y. I. Wolf, D. Kazlauskas, J. Iranzo, A. Lucía-Sanz, J. H. Kuhn, M. Krupovic, V. V. Dolja, E. V. Koonin, Origins and Evolution of the Global RNA Virome. *MBio* (2018), doi:10.1128/mbio.02329-18.
34. V. Da Cunha, M. Gaia, D. Gadelle, A. Nasir, P. Forterre, Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.* **13** (2017), doi:10.1371/journal.pgen.1006810.
35. V. Da Cunha, M. Gaia, A. Nasir, P. Forterre, Asgard archaea do not close the debate about the universal tree of life topology. *PLOS Genet.* **14**, e1007215 (2018).
36. A. Spang, J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. Van Eijk, C. Schleper, L. Guy, T. J. G. Ettema, Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* (2015), doi:10.1038/nature14447.
37. A. Spang, L. Eme, J. H. Saw, E. F. Caceres, K. Zaremba-Niedzwiedzka, J. Lombard, L. Guy, T. J. G. Ettema, Asgard archaea are the closest prokaryotic relatives of eukaryotes. *PLOS Genet.* (2018), doi:10.1371/journal.pgen.1007080.
38. P. Forterre, The Common Ancestor of Archaea and Eukarya Was Not an Archaeon. *Archaea*. **2013**, 1–18 (2013).
39. M. Krupovic, D. Prangishvili, R. W. Hendrix, D. H. Bamford, Genomics of Bacterial and Archaeal Viruses: Dynamics within the Prokaryotic Virosphere. *Microbiol. Mol. Biol. Rev.* **75**, 610–635 (2011).
40. M. Krupović, P. Forterre, D. H. Bamford, Comparative Analysis of the Mosaic

Genomes of Tailed Archaeal Viruses and Proviruses Suggests Common Themes for
Virion Architecture and Assembly with Tailed Viruses of Bacteria. *J. Mol. Biol.* (2010),
doi:10.1016/j.jmb.2010.01.037.

41. M. K. Pietila, N. S. Atanasova, V. Manole, L. Liljeroos, S. J. Butcher, H. M. Oksanen,
D. H. Bamford, Virion Architecture Unifies Globally Distributed Pleolipoviruses
Infecting Halophilic Archaea. *J. Virol.* (2012), doi:10.1128/JVI.06915-11.

42. J. Filee, P. Siguier, M. Chandler, Insertion Sequence Diversity in Archaea. *Microbiol.
Mol. Biol. Rev.* **71**, 121–157 (2007).

43. N. Soler, M. Gaudin, E. Marguet, P. Forterre, Plasmids, viruses and virus-like
membrane vesicles from Thermococcales. *Biochem. Soc. Trans.* (2011),
doi:10.1042/bst0390036.

44. M. Krupovic, M. Gonnet, W. Ben Hania, P. Forterre, G. Erauso, Insights into
Dynamics of Mobile Genetic Elements in Hyperthermophilic Environments from Five
New Thermococcus Plasmids. *PLoS One* (2013), doi:10.1371/journal.pone.0049044.

45. P. Forterre, A new fusion hypothesis for the origin of Eukarya: Better than previous
ones, but probably also wrong. *Res. Microbiol.* (2011),
doi:10.1016/j.resmic.2010.10.005.

46. R. Sorek, C. M. Lawrence, B. Wiedenheft, CRISPR-Mediated Adaptive Immune
Systems in Bacteria and Archaea. *Annu. Rev. Biochem.* (2013), doi:10.1146/annurev-
biochem-072911-172315.

47. K. S. Makarova, Y. I. Wolf, E. V. Koonin, Comparative genomics of defense systems
in archaea and bacteria. *Nucleic Acids Res.* (2013), doi:10.1093/nar/gkt157.

48. E. V. Koonin, T. G. Senkevich, V. V. Dolja, The ancient virus world and evolution of
cells. *Biol. Direct* (2006), doi:10.1186/1745-6150-1-29.

49. D. Prangishvili, P. Forterre, R. A. Garrett, Viruses of the Archaea: A unifying view.
Nat. Rev. Microbiol. (2006), , doi:10.1038/nrmicro1527.

50. P. Forterre, The origin of viruses and their possible roles in major evolutionary
transitions. *Virus Res.* (2006), , doi:10.1016/j.virusres.2006.01.010.

51. R. Catchpole, P. Forterre, Positively twisted: The complex evolutionary history of
Reverse Gyrase suggests a non-hyperthermophilic Last Universal Common Ancestor.
bioRxiv (2019), doi:10.1101/524215.

52. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J.
Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database
search programs. *Nucleic Acids Res.* (1997), , doi:10.1093/nar/25.17.3389.

53. J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, J. Walichiewicz, Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* (2005), doi:10.1159/000084979.
54. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* (2013), doi:10.1093/molbev/mst010.
55. L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg, The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* (2015), doi:10.1038/nprot.2015.053.
56. E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* (2004), doi:10.1002/jcc.20084.
57. V. Miele, S. Penel, L. Duret, Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*. **12** (2011), doi:10.1186/1471-2105-12-116.
58. L. T. Nguyen, H. A. Schmidt, A. Von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* (2015), doi:10.1093/molbev/msu300.
59. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. Von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* (2017), doi:10.1038/nmeth.4285.
60. H. C. Wang, B. Q. Minh, E. Susko, A. J. Roger, Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst. Biol.* (2018), doi:10.1093/sysbio/syx068.
61. S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* (2010), doi:10.1093/sysbio/syq010.
62. B. Q. Minh, M. A. T. Nguyen, A. Von Haeseler, Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* (2013), doi:10.1093/molbev/mst024.
63. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* (2007), doi:10.1093/bioinformatics/btl529.

Acknowledgments

General: This work used the computational and storage services (TARS cluster) provided by the IT department at Institut Pasteur, Paris.

Funding: This work was supported by the European Research Council (ERC) grant from the European Union's Seventh Framework Program (FP/2007-2013)/Project EVOMOBIL-ERC Grant Agreement no. 340440.

Author contributions: A.C.W. and P.F. designed the study. A.C.W., J.G., and V.D.C. performed bioinformatics experiments. A.C.W., J.G., M.G., V.D.C., and P.F. analyzed and interpreted the results. A.C.W., M.G., V.D.C. and P.F. wrote the manuscript.

Competing interests: The authors declare that they have no competing interests.

Data and materials availability: All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Figures and Tables

	MCP	pATPase	Domain
NCLDV	DJR	FtsK/HerA family	Eukarya
Polintonviruses	DJR	FtsK/HerA family	Eukarya
<i>Lavidaviridae</i>	DJR	FtsK/HerA family	Eukarya
<i>Adenoviridae</i>	DJR	ABC family	Eukarya
STIV	DJR	FtsK/HerA family	Bacteria/Archaea
Bam35	DJR	FtsK/HerA family	Bacteria
PRD1	DJR	FtsK/HerA family	Bacteria
Toil	DJR	FtsK/HerA family	Bacteria
Cellulomonas	DJR	FtsK/HerA family	Bacteria
PM2	DJR	FtsK/HerA family	Bacteria
FLiP	DJR	-	Bacteria
<i>Sphaerolipoviridae</i>	SJR	FtsK/HerA family	Bacteria/Archaea

Table 1. The PRD1-adenovirus lineage. This table summarizes the shared genes between different families of this lineage. Abbreviations: DJR, double jelly-roll; SJR: single jelly-roll; NCLDV, Nucleo-Cytoplasmic Large DNA Viruses; STIV, Sulfolobus turreted icosahedral virus; FLiP, Flavobacterium-infecting bacteriophage; MCP, major capsid protein; pATPase, packaging ATPase.

Fig. 1. Maximum likelihood (ML) single-protein trees of the two hallmark proteins of the viruses from the PRD1-adenovirus lineage. ML phylogenetic trees of (A) the major capsid protein (MCP) and (B) packaging ATPase (pATPase). The root of the ML phylogenetic tree was between the prokaryotic and eukaryotic members. The scale-bar indicates the average number of substitutions per site. Values on top and below branches represent support calculated by ultrafast bootstrap approximation (UFBoot; 1,000 replicates) and SH-like approximate likelihood ratio test (aLRT; 1,000 replicates), respectively. Only values superior to 70 are shown. The best-fit model for the MCP tree was LG + F + R4, which was chosen according to Bayesian Information Criterion (BIC) and the alignment has 103 sequences with 237 positions. The best-fit model for the pATPase tree was LG + R6, which was chosen according to Bayesian Information Criterion (BIC) and the alignment has 103 sequences with 171 positions. More detailed versions of the trees are shown in fig. S7 and S8.

Fig. 2. Maximum likelihood (ML) phylogenetic tree of the concatenated MCP and pATPase genes. The ML phylogenetic tree was rooted between the prokaryotic and eukaryotic members. The scale-bar indicates the average number of substitutions per site. Values on the branches represent transfer bootstrap expectation (TBE) support calculated by nonparametric bootstrap (1,000 replicates). The best-fit model was LG + F + R6, which was chosen according to Bayesian Information Criterion (BIC). The alignment has 107 sequences with 408 columns.

Fig. 3. Maximum likelihood (ML) phylogenetic tree of the concatenated MCP and pATPase genes of bacteriophages and archaeoviruses. The ML phylogenetic tree was rooted with the DJR bacterial cluster II. The scale-bar indicates the average number of substitutions per site. Values on the branches represent transfer bootstrap expectation (TBE) support calculated by nonparametric bootstrap (1,000 replicates). The best-fit model was LG + R4, which was chosen according to Bayesian Information Criterion (BIC). The alignment has 49 sequences with 468 columns.

Fig. 4. Hypothetical scenarios for the evolution of viruses of the PRD1-adenovirus lineage. (A) “Prokaryotic” ancestor hypothesis. Depending on the rooting, different possible topologies of the double jelly-roll viruses of this lineage are depicted (see fig. S10, S11, S12 and S13). In any case, double jelly-roll (DJR) eukaryoviruses originated from viruses, which were closely related to DJR bacterial cluster. (B) The enigma of the PRD1-adenovirus lineage. The presence of viruses from the PRD1-adenovirus lineage across the three domains of life suggests that these viruses could be very ancient. In this scenario, modern DJR archaeoviruses are more closely related to DJR bacteriophages. Ancient viruses which could infect archaea were lost after the divergence of Archaea and Eukarya. After that, modern DJR archaeoviruses were being transferred from DJR bacteriophages before the diversification of Archaea. Another possibility is some of the viruses which are closely related to DJR eukaryoviruses have yet to be found.