The genome of a persistent giant algal virus encodes an unprecedented number of genes involved in energy metabolism

Romain Blanc-Mathieu[1,2], Håkon Dahle[3], Antje Hofgaard[4], David Brandt[5], Jörn Kalinowski[5], Hiroyuki Ogata[1] and Ruth-Anne Sandaa[6*]

1: Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011, Japan

2: Laboratoire de Physiologie Cellulaire & Végétale, CEA, Univ. Grenoble Alpes, CNRS, INRA, IRIG, Grenoble, France

3: Department of Biological Sciences and K.G. Jebsen Center for Deep Sea Research, University of Bergen, Bergen, Norway

4: Department of Biosciences, University of Oslo, Norway

5: Center for Biotechnology, Universität Bielefeld, Bielefeld, 33615, Germany

6: Department of Biological Sciences, University of Bergen, Bergen, Norway

*Corresponding author: Ruth-Anne Sandaa, +47 55584646, ruth.sandaa@uib.no

Abstract

Viruses have long been viewed as entities possessing extremely limited metabolic capacities. Over the last decade, however, this view has been challenged, as metabolic genes have been identified in viruses possessing large genomes and virions—the synthesis of which is energetically demanding. Here, we describe the 1.4-Mbp linear genome of *Prymnesium kappa* virus RF01 (PkV RF01), a giant virus of the *Mimiviridae* family. We found that this virus encodes an unprecedented number of proteins involved in energy metabolism, such as all four succinate dehydrogenase (SDH) subunits (A–D) as well as key enzymes in the β-oxidation pathway. The *SDHA* gene was transcribed upon infection, indicating that the viral SDH is actively used by the virus— potentially to modulate its host's energy metabolism. We detected orthologous *SDHA* and *SDHB* genes in numerous genome fragments from uncultivated marine *Mimiviridae* viruses, which suggests that the viral SDH is widespread in oceans. PkV RF01 was less virulent compared with other cultured prymnesioviruses, a phenomenon possibly linked to the metabolic capacity of this virus and suggestive of relatively long co-evolution with its hosts. Finally, we found that PkV RF01 is the only alga-infecting *Mimiviridae* virus encoding two aminoacyl-tRNA synthetases and enzymes corresponding to an entire base-excision repair pathway, as seen in heterotroph-infecting *Mimiviridae*. These *Mimiviridae* encoded-enzymes were found to be monophyletic and branching at the root of the eukaryotic tree of life. This placement suggests that the last common ancestor of *Mimiviridae* was endowed with a large, complex genome prior to the divergence of known extant eukaryotes.

Importance

Viruses on Earth are tremendously diverse in terms of morphology, functionality, and genomic composition. Over the last decade, the conceptual gap separating viruses and cellular life has tightened because of the detection of metabolic genes in viral genomes that express complex virus phenotypes upon infection. Here, we describe *Prymnesium kappa* virus RF01, a large alga-infecting virus with a unique morphology, an atypical infection profile, and an unprecedented number of genes involved in energy metabolism (such as the tricarboxylic acid cycle (TCA) cycle and the β-oxidation pathway). Moreover, we show that the gene corresponding to one of these enzymes (the succinate dehydrogenase subunit A) is transcribed during infection and is widespread among marine viruses. This discovery provides evidence that a virus has the potential to actively regulate energy metabolism with its own gene.

Key words: Algal viruses, *Mimiviridae*, persistent, co-evolution, metabolism, energy production, succinate dehydrogenase, β-oxidation and aminoacyl-tRNA synthetases

Introduction

In their essay "Varieties of Living Things: Life at the Intersection of Lineage and Metabolism," Dupré and O'Malley proposed to address Schrödinger's question "What is Life?" by "*describing a spectrum of biological entities that illustrates why no sharp dividing line between living and non-living things is likely to be useful*" (1). Microbiologists have contributed considerably to this descriptive effort, both by reporting the existence of viruses endowed with genes coding for functions once thought to be exclusive to cellular life and by concomitantly proposing that actively infecting viruses are a "living form" (2–4). Genes encoding elements for photosynthesis (5, 6), carbon metabolism (7), and nitrogen- (8) and sulfur-cycling (9) have been found in bacterial viruses, where they are used to maintain or augment cellular processes during infection and to redirect energy and resources towards viral production (8, 10, 11). Genes for protein synthesis, including translation initiation, elongation, and termination, and a range of aminoacyl-tRNA synthetases have been found in *Mimiviridae*, a group of giant viruses infecting single-celled eukaryotes (12–14). *Mimiviridae* and other large DNA viruses, including some bacterial viruses, also have tRNA genes (15, 16). Ribosomal proteins have recently been reported in viral genomes derived from metagenomes (17). Genes involved in other metabolic processes, such as fermentation (18), glycosylation (19), photosynthesis (20), and rhodopsin (21), are encoded in *Mimiviridae* and other related large eukaryotic DNA viruses. Metabolic genes are frequently observed within virus genomes (20, 22, 23); although they represent a tiny fraction of the viral gene pool, these genes have the potential to dramatically modify the phenotype of an actively infected cell and alter the ecological role of the host (7, 24, 25). The infected host in this state has been referred to as a virocell (2). One might expect that the interplay between viral genes

and host genes in virocells would become increasingly fine-tuned and complex during prolonged virus–host co-evolution, which also typically leads to lower virulence. Much of the complexity of virocells may still be undetected, as most currently isolated algal viruses are highly virulent, with several involved in rapid algal bloom termination events (26).

Viruses of the *Mimiviridae* family are known to infect heterotrophic and autotrophic microbial eukaryotes. This divide is also reflected in the phylogeny of these viruses, some of which are classified into two proposed sub-families: "Megavirinae" and "Mesomimivirinae" (27). The former contains viruses with genomes larger than 1 Mbp, all isolated from Amoebozoa, while the latter includes viruses with smaller genomes isolated from haptophyte algae of class Prymnesiophyceae. Several *Mimiviridae* members outside these two groups have been characterized to some extent as well, namely, viruses isolated from heterotrophs (*Cafeteria roenbergensis* virus, CroV; *Bodo saltans* virus, BsV; Choano virus), autotrophs (*Aureococcus anophagefferens* virus, AaV; Tetraselmis virus 1, TetV; *Pyramimonas orientalis* virus, PoV; *Prymnesium kappa* virus RF01, PkV RF01), a metazoan (Namao virus), and metagenomes (Klosneuviruses). The Mesomimivirinae sub-family includes viruses infecting bloom-forming hosts, such as *Phaeocystis pouchetii*, *Phaeocystis globosa*, and *Prymnesium parvum* (PpV, PgV Group I, and PpDVAV, respectively) (28–30); it also includes several viruses infecting *Haptolina ericina* and *Prymnesium kappa*, which normally do not form massive blooms but are present at low densities in seawater year round (31). In marine environments, viruses infecting low-density and non-bloom-forming algae may be the most common virus–host systems—that is, low-density hosts (non-blooming) and viruses that appear to have co-evolved in response to host growth strategy. Thus far, the only known

representatives of such viruses are *Prymnesium kappa* viruses RF01 (PkV RF01) and RF02 (PkV RF02), *Haptolina ericina* virus RF02 (HeV RF02), and *Chrysochromulina ericina* virus (CeV 01B, infecting *Haptolina ericina*) (32, 33). Together with PgV, all of these viruses, except for PkV RF01, belong to the sub-family Mesomimivirinae on the basis of their monophyletic relationship and, in the case of PgV and CeV, a shared genomic similarity (27). In contrast, phylogenetic analysis of two partially sequenced marker genes has placed PkV RF01 deep inside the *Mimiviridae* clade, and characterization of its life cycle has revealed an atypical infection profile (33). Here, we report new viral functions inferred from analysis of the genome sequence of PkV RF01. We found that this virus is less virulent than most other alga-infecting viruses and possesses an unprecedented number of energy-generating genes. We uncovered clues suggesting that members of *Mimiviridae* that potentially modulate the metabolism of their hosts are widespread in the ocean. Our findings of peculiar genomic features in a persistent virus provide new insights on virus–host coevolution and may stimulate further advances in modeling the history of their interaction.

Results and Discussion

PkV RF01 has an atypical morphology

The icosahedral PkV RF01 particle is approximately 400 nm in diameter (Fig. 1). Beneath the capsid, several convoluted inner membranes fill approximately 66% of the interior. Chloroform treatment of PkV RF01 drastically reduced the infectivity of the virus, suggesting that lipid-containing molecules are involved in the infection process (Appendix – Fig. S1). As no outer membrane was detected by cryo-electron tomography, the chloroform treatment may have affected lipid components in either

the capsid or the inner convoluted membranes. Internal lipid-containing membranes have been detected in several icosahedral-shaped double-stranded DNA viruses, including algal viruses belonging to families *Phycodnaviridae* and *Mimiviridae*, mimiviruses, and various bacteriophages (34–39). In all of these viruses, the inner membranes are suggested to play a role in the release of the viral nucleoprotein core or genome by fusing with the host plasma membrane (36, 38, 39). Inner membranes in currently described NCLDVs more or less adopt the icosahedral morphology defined by the outer layer of capsomers (40, 41). We detected several convoluted inner membranes in PkV RF01 that do not follow the structure of the capsid. To our knowledge, this structural inconsistency has not been previously detected in any double-stranded DNA viruses, which calls for further investigation to understand the assembly process of PkV RF01 and how it enters its host. Another striking feature of the PkV RF01 virion is an internal rod-shaped core (ca. 55 nm in diameter), which is filled with dense material and positioned in the center of the virus particle. Similar features have been observed in TEM images of large virus-like particles (VLPs) (300–700 nm) occurring in waste vacuoles of phaeodarian radiolarians collected from different oceans (42) and in zoospores of the green alga *Chlorococcus minutum* (43). To our knowledge, however, these features have not been described in isolated viruses thus far.

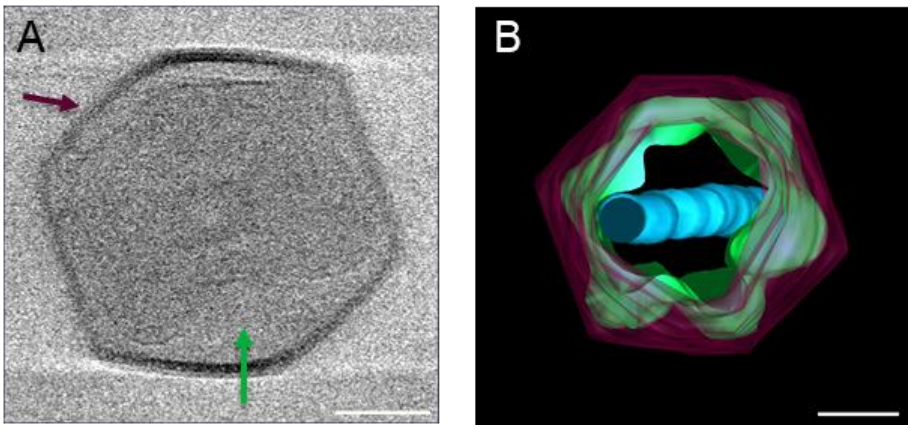

FIG 1 PkV RF01 morphology. (A) Screen shot of a cryo-electron tomogram of a PkV RF01 virion. (B) Composite image of 61 cryo-electron tomograms (–60 to 60°, imaged every 2°). Purple, capsid; green, inner membrane consisting of multiple irregular, convoluted membranes; blue, internal rod-shaped core filled with dense material. The full set of records is available on GitHub (see Data availability section). Scale bar, 100 nm.

PkV RF01 has an atypical infection strategy

Compared to other cultured algal viruses, PkV RF01 has an atypical infection strategy characterized by a more persistent relationship with its hosts. Only 2% of the total PkV RF01 viral particles produced during infection of *Haptolina ericina* UiO028 (He UiO028) are infectious (able to produce progeny) (Table 1). This infectivity is much lower than that of the other two prymnesioviruses, HeV RF02 and PkV RF02, which produced infectious progeny of 13% and 44%, respectively (Table 1). The portion of infectious particles of PkV RF01 is low also when compared to other algal viruses (44, 45). Further, the latent period of PkV RF01 is longer (ca. 24–32 h, (33)) in comparison with other prymnesioviruses (28, 29, 32, 33) and PkV RF01 is able to infect multi-species that is also an unusual trait among algal viruses (26).

TABLE 1 Infection parameters of *Prymnesium kappa* viruses RF01 and RF02 and *Haptolina ericina* virus RF02.

Viral species and hosts	Infectious progeny/mL (MPN)	Host cells/mL (FCM)[a]	Total VLP/mL (FCM)	Burst size (VLP)[b]	Infectivity (%)[c]	Infectious particles in a burst[d]
PkV RF01 (He UiO028)	2.9×10^6 (± 0.2)	4.9×10^5	1.8×10^8 (±0.9)	363	2	6
PkV RF02 (Pk RCC3423)	2.2×10^8 (± 0.2)	4.6×10^5	5.0×10^8 (±0.1)	1093	44	483
HeV RF02 (He UiO028)	5.8×10^7 (±0.2)	4.9×10^5	4.4×10^8 (±0.0)	907	13	119

VLP, virus-like particle; MPN, most probable number; FCM, flow cytometry.
[a]Measurement performed in duplicates
[b]The number of viral particles released from each host cell, estimated from the total number of host cells pre-infection and the total number of VLPs produced during the infection cycle.
[c]Estimated as the percentage of infectious progeny of all VLPs produced during the infection cycle.
[d]Number of infectious particles released per host cell.

Differences in viral infectious strategies results from the co-evolution between viruses and their hosts. The hosts of PkV RF01, PkV RF02, and HeV RF02 all belong to order Prymnesiales, whose members are normally present in low abundance but co-occur year round (*K*-strategists) (46). All these viruses have lower infectivity and longer latent periods compared with viruses infecting bloom-forming haptophytes (r-strategists). Long replication time and lower infectivity, as hosts becomes scarce, increases the chances of vertical transmission rather than horizontal transmission of the virus. Vertical parent-to-offspring transmission depends on host reproduction and should therefore evolve towards lower virulence to increase the virus fitness (47, 48). High virulence, on the other hand, may be supported by large, dense host populations because high host densities ensure successful horizontal transmission of viral progeny to new hosts (47, 49). The highly acute viruses infecting recurrent bloom-forming haptophytes, e.g. *Emiliania huxleyi* virus (EhV), *Phaeocystis pouchetii* virus (PpV), and *Phaeocystis globosa* virus (PgV), normally have short latent periods, which results in rapid lysis of the host (29, 30, 50). The infectivity of PgV and PpV is also high, between 60%–100% (44, 51).

Broad host range is another characteristic that might increase the chance of transmission in an environment with low host abundances. Generalist strategy is

associated with trade-offs in the form of "paying" a higher infectivity cost and decreasing the opportunity of transmission with longer replication times, higher decay rates and decreased infectivity (discussed in (52, 53)). This fits well with the two multi-species infecting haptophyte viruses, PkV RF01 and HeV RF02, that have reduced proportions of infectious particles and longer replication times (33), relative to other haptophyte viruses with restricted host ranges. In the ocean, persistent relationships—such as between PkV RF01 and its hosts—seem to be most common among viruses infecting unicellular algae; this has been demonstrated by several metabarcoding studies revealing the persistence of dominance of viral OTUs over several months (54, 55).

PkV RF01 has the largest genome among algal viruses

The genome of PkV RF01 was assembled as a linear DNA sequence of 1,421,182 bp. This size is more than twice that of the genome of TetV, which means that PkV RF01 has the largest reported genome of any virus infecting a photosynthetic organism (Fig. 2A). Evidence for the linear structure of this genome is the presence of ~5-kbp terminal inverted repeats. Despite being phylogenetically more closely related to alga-infecting *Mimiviridae*, the genome size of PkV RF01 is in the range of heterotroph-infecting *Mimiviridae*. The overall G+C content of PkV RF01 is 22.8%, which is low compared with other *Mimiviridae* (23%–41%). Similar to other *Mimiviridae*, the average G+C content of PkV RF01 in intergenic regions is relatively low, 17.8%. This lower G+C content may reflect an ongoing loss of G and C nucleotides, more prevalent in non-coding than coding regions because of weaker background selection in non-coding regions. The genome of PkV RF01 is predicted to contain 1,161 genes comprising 1,121 protein-coding DNA sequences (CDSs) and 40 tRNA genes corresponding to 13 amino acids (Appendix – Table S1). Most tRNA genes (30 out of

40) are clustered in three genomic regions lacking predicted CDSs, a feature also observed in other *Mimiviridae*. For example, all tRNAs of TetV (*n* = 10) and CroV (*n* = 22) are encoded consecutively on the same strand (18, 56). The average CDS length is 1,046 bp (minimum: 297; maximum: 1,493). Intergenic regions average 217 bp in length, with a cumulative sum of 244,005 bp, which corresponds to a gene density of 82.8%.

OrthoFinder identified 19 groups of paralogs, all containing two genes, except for one with six genes. This low number of identified paralogs indicates that gene duplication has not played an important role in the evolution of the genome size of this virus, at least not recently (rapid divergence may create distant paralogs that are difficult to identify). Amino acid sequences in the cluster of six paralogous genes identified by OrthoFinder could not be functionally annotated individually, but a HMM profile-to-profile comparison detected homology with the CTCF transcriptional regulator (Appendix – Fig. S2). Further comparison of this HMM profile against all *Mimiviridae* proteomes revealed that this protein is encoded by genes in the genomes of *Klosneuvirinae* (20, 12, and 9 gene copies each in Klosneuvirus, Indivirus, and Catovirus, respectively), with 10 copies found in each of the three PgVs, two copies each in CeV, OLPV2, and BsV, and one or two copies in some mimiviruses (no homologs found in *Phycodnaviridae*). CTCF is a DNA-binding protein highly conserved in bilaterian metazoans that functions in genomic organization and the control of gene expression (57). Viral-encoded CTCF has not been characterized, but studies have shown that herpesviruses recruit host CTCF proteins to modulate (enhance and repress) the transcription of their own genes (58, 59). Inspection of self-genome alignment dot plots revealed an unusual number of short repeats throughout the genome of PkV RF01 compared with other *Mimiviridae*

members (Appendix – Fig. S3). Of the 1,121 predicted CDSs, 567 (51%) exhibited sequence similarities (BLASTP *E*-value conservative cutoff of 1×10^{-5}) to protein sequences in the RefSeq database (Fig. 2A). Among them, 204 (36%) were most similar to eukaryotes, while 164 (29%) were closest to bacteria and 190 (34%) to NCLDVs. Among the CDSs most similar to NCLDVs, 138 (73%) corresponded to Mesomimivirinae and 38 (28%) to Megavirinae, with CeV and PgV being the most represented taxa (Fig. 2B). Among the closest homologs found in eukaryotic organisms, 24 were haptophytes (12 in *Emiliania huxleyi* and 12 in *Chrysochromulina* sp.). A sequence-based homology search of corrected nanopore reads and scaffolds composing the initial assembly against *Lavidaviridae* proteomes (BLASTX; matrix: BLOSUM45, *E*-value $< 1 \times 10^{-5}$) yielded no significant alignments against any major or minor *Lavidaviridae* capsid proteins, which suggests that virophages were absent from the sample used for sequencing.

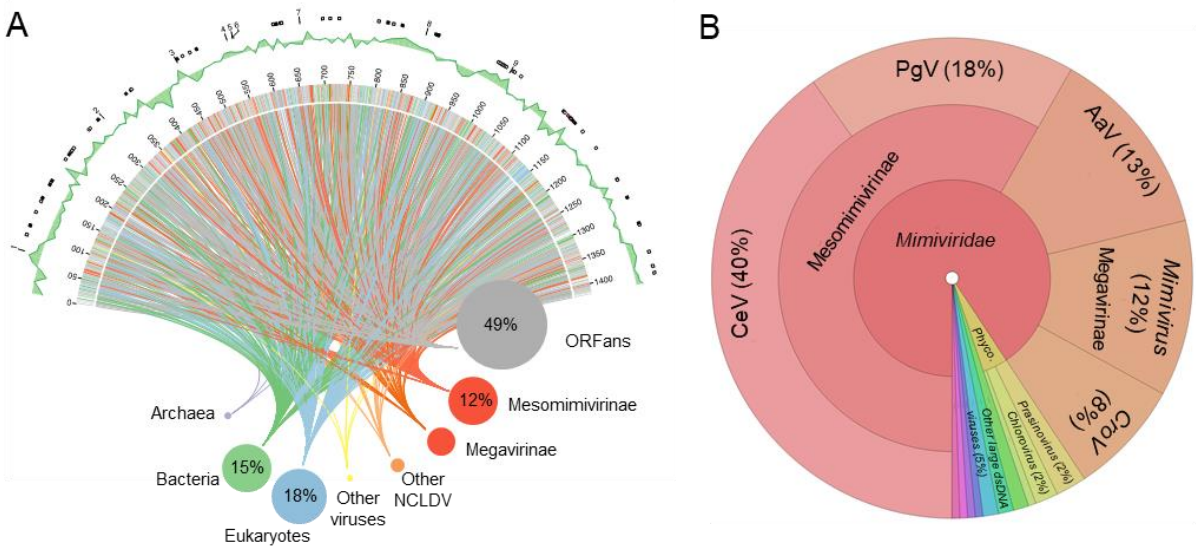

FIG 2 Structure and gene taxonomic composition of the PkV RF01 genome sequence. (A) Rhizome and genomic features of the PkV RF01 genome. As illustrated by the rhizome (inner part of the figure), ORFans comprise the largest set of PkV RF01 genes, while a substantial portion (12%) have their best BLAST hits (RefSeq) against "Mesomimivirinae." Colors indicate taxonomic origin. Intergenic regions are white. Percentage hits per taxonomic group higher than 5% of total genes are indicated. In the outermost ring, rectangles indicate the positions of glycosyltransferases (white), lipid-related enzymes (black), and succinate dehydrogenase genes (red), and the numbers correspond to *Mimiviridae* key enzymes (1 and 3: DNA-directed RNA polymerase II subunits 1 and 2, respectively; 2: DNA mismatch repair protein MutS7; 4: Packaging ATPase; 5: VLTF3, 6: Major capsid protein; 7: Eukaryotic translation initiation factor 4E; 8: Asparagine synthase; 9: DNA polymerase family B). The ring adjacent to the outermost ring shows GC skew over a 10-KB window. (B) Taxonomic breakdown of 194 genes with best hits to virus genes. Taxa represented by fewer than 2% of the genes are Pacmanvirus, *Caudovirales*, *Poxviridae*, *Baculoviridae*, *Marseilleviridae*, Pithovirus, and Kaumoebavirus.

A previous analysis of PkV RF01 family-B DNA polymerase (PolB) and the major capsid protein (MCP) placed this virus in family *Mimiviridae* (33). We also recently reported that the PkV RF01 genome has additional NCLDV core genes, such as A32-like virion packing ATPase (NCVOG0249) and RNApol (RNA pol subunit I [NCVOG0274] and subunit II [NCVOG0271]), and orthologous genes that are specific to *Mimiviridae*, namely, MutS7 (NCVOG2626) and asparagine synthase

(AsnS, NCVOG0061) (60). Phylogenetic reconstruction using five NCLDV core genes confirmed the deep branching of PkV RF01 within the *Mimiviridae* family and suggested that PkV RF01, along with TetV and AaV, is more closely related to Mesomimivirinae than to Megavirinae (Fig. 3A). In support of this evolutionary relationship, PkV RF01 has an additional copy of the second largest RNA polymerase subunit gene (*rpb2*). This *rpb2* duplication is shared with all other *Mimiviridae* that infect algae, including Mesomimivirinae members, AaV (whose second copy is very short), and TetV. This additional *rpb2* copy is not found in other *Mimiviridae.* Phylogenetic analysis indicates that these two *rbp2* copies were present in the ancestor of alga-infecting *Mimiviridae* (Fig. 3B). The *rpb2* duplication is a useful feature to discriminate between the two main clades (autotroph versus heterotroph-infecting viruses) within the *Mimiviridae* family.

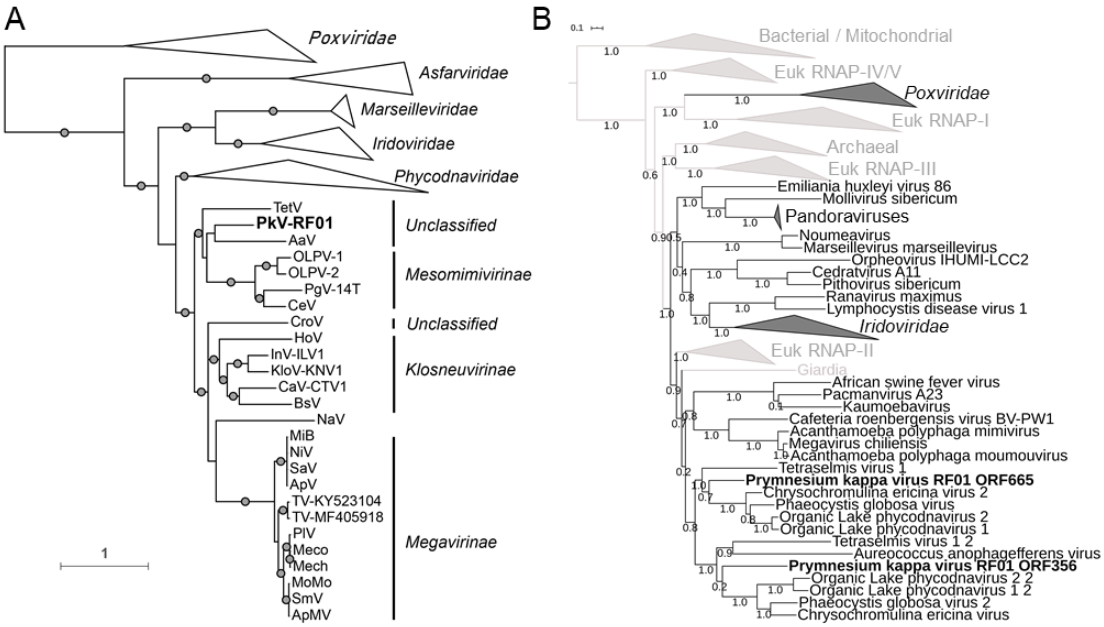

FIG 3 Phylogenetic evidence for PkV RF01 as a distant relative of "Mesomimivirinae." (A) Bayesian phylogenetic tree of NCLDVs reconstructed from a concatenated alignment of five core nucleocytoplasmic virus orthologous genes. Gray circles indicate branches statistically supported by posterior probabilities above 0.95. The tree was rooted using *Poxviridae* as outgroup. The scale bar indicates substitutions per site. TetV, Tetraselmis Virus 1; AaV, *Aureococcus anophagefferens* virus; OLPV-1, Organic Lake phycodnavirus 1; OLPV-2, Organic Lake phycodnavirus 2; PgV-14T, *Phaeocystis globosa* virus; CeV, *Chrysochromulina ericina* virus; CroV, *Cafeteria roenbergensis* virus BV-PW1; HoV, Hokovirus HKV1; InV-ILV1, Indivirus ILV1; KloV-KNV1, Klosneuvirus KNV1; CaV-CTV1, Catovirus CTV1; BsV, *Bodo saltans* virus; NaV, Namao Virus; MiB, Mimivirus_Bombay; NiV, Niemeyer virus; SaV, Samba virus; ApV, *Acanthamoeba polyphaga* virus; TV-KY523104, Tupanvirus KY523104; TV-MF405918, Tupanvirus MF405918; PlV, Powai Lake megavirus; Meco, Megavirus courdo 7; Mech, Megavirus chilensis; MoMo, Moumouvirus Monve; SmV, Saudi moumouvirus; ApMV, *Acanthamoeba polyphaga* moumouvirus. (B) Maximum likelihood phylogenetic tree of cellular and NCLDV DNA-directed RNA polymerase subunit beta (RPB2). Tools, evolutionary models, and parameters used for tree reconstruction are reported in Appendix 1 – Table S3.

Out of 1,121 predicted protein-coding genes in the genome of PkV RF01, only about a third could be annotated with some functional description based on their sequence homology with characterized proteins. Such a small percentage is typical of eukaryotic viruses with large genomes. A total of 339 proteins (30%) showed

significant sequence similarity with proteins in the Cluster of Orthologous Gene (COG) database (61) (Appendix – Fig. S4). The distribution of COG functions associated with these hits was dominated by "Posttranslational modification, protein turnover, chaperones" (43 proteins) and "Cell wall/membrane/envelope biogenesis" (42 proteins), which is approximately two times more proteins than in other *Mimiviridae* members except for Tupanvirus (Appendix – Fig. S5). Among other well-represented categories, numbers of proteins in "Replication, recombination and repair" (36 proteins) and "Transcription" (23 proteins) were similar to those of other *Mimiviridae*, while the categories of "Translation, ribosomal structure and biogenesis" (25 proteins) and "Amino acid transport and metabolism" (20 proteins) were respectively in the same range or higher than those of heterotroph-infecting *Mimiviridae* (mimiviruses, BsV, and CroV). Interestingly, 24, 17, and 9 PkV RF01 proteins were respectively assigned to the categories of "Lipid transport and metabolism", "Carbohydrates transport and metabolism," and "Energy production and conservation," all much higher compared with other *Mimiviridae* viruses.

Similar to other *Mimiviridae*, PkV RF01 encodes several genes involved in DNA repair, transcription, and translation (Appendix – Results). Notably, this virus has the full set of enzymes required for the base excision repair (BER) pathway, which is also the case for all *Mimiviridae* members except for those with smaller genomes (PgV, CeV, and AaV). PkV RF01 BER enzymes are closer (i.e., have a greater alignment score) to heterotrophic *Mimiviridae* than to cellular homologs, thus suggesting that this pathway was present in the last common ancestor of *Mimiviridae*. According to a previous phylogenetic analysis, *Mimiviridae* BER enzymes are monophyletic with regard to *Mimiviridae* and have not recently been acquired from eukaryotes (62).

Unlike alga-infecting *Mimiviridae*, PkV RF01 encodes two amino-acyl tRNA synthetases (aaRS): an isoleucyl-tRNA synthetase (IleRS; ORF 480) and an asparaginyl-tRNA synthetase (AsnRS; ORF 764). Both of these synthetases are found in most lineages of heterotroph-infecting *Mimiviridae* (AsnRS is missing from CroV and BsV, and IleRS is missing from *Mimivirus* lineage A). Phylogenetic analyses of these two proteins revealed a deep branching of viral homologs, which formed a monophyletic clade well separated from cellular homologs (Fig. 4). Although the phylogenetic relationship among viruses on these trees does not exactly reflect their classification (e.g., Orpheovirus within Megavirinae), the recovered topologies suggest an ancient history for these genes rather than a recent acquisition from cellular organisms. The IleRS tree (Fig. 4A) is in agreement with previous phylogenetic analyses of cellular organisms with the known split of bacterial lineages (63, 64). In our tree, *Mimiviridae* and Orpheovirus form a monophyletic group branching at the root of eukaryotic organisms, which suggests that the IleRS gene was acquired before the divergence of extant eukaryotic phyla or from a now-extinct or as yet unknown eukaryotic phylum at the root of the eukaryotic tree. Similarly, the AsnRS tree supports the presence of this enzyme in the ancestor of *Mimiviridae* (Fig. 4B). Alternatively, rapid evolution of more recently acquired genes could generate these observed topologies in a phenomenon known as the long-branch attraction bias (65). Nonetheless the branches for viruses were not particularly long compared to the branches for eukaryotes and the evolutionary model used in our phylogenetic analyses is known to minimize this bias (compared to more traditional models) (66, 67). Therefore, it seems more likely that viral genes were recruited from proto-eukaryotes before the diversification of modern eukaryotes.

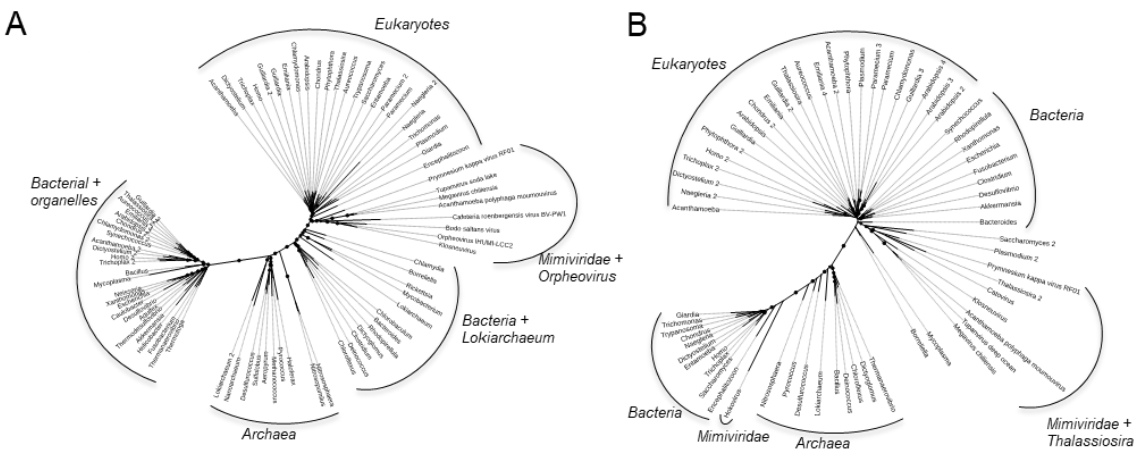

FIG 4 Bayesian phylogenetic trees of two viral amino-acyl tRNA synthetases and their cellular homologs. (A) Isoleucine tRNA synthases. (B) Aspartyl tRNA synthetases. Branches supported by posterior probability (PP) values >70% are indicated by circles whose diameters are proportional to the PP value.

A viral-encoded succinate dehydrogenase and energy production genes

We found six predicted protein-coding genes (ORFs 893 to 900) related to energy production in an 8,026-bp region (Fig. 5A). Four ORFs (ORFs 893 and 898–900) were predicted to code for all four subunits (SDHA, D, C, and B) of a functional succinate dehydrogenase (SDH, or Electron Transport Chain Complex II) of the oxidative phosphorylation pathway (Fig. 5B). In eukaryotes, all four subunits of this enzyme are encoded in the nuclear genome. This enzyme acts in the mitochondrial respiratory chain and participates in both the TCA cycle and the respiratory electron transfer chain. In the TCA cycle, this succinate dehydrogenase oxidizes succinate to fumarate, while its activity in the inner mitochondrial membrane involves the reduction of a FAD cofactor followed by electron transfer through three Fe–S centers to ubiquinone (Fig. 5C).

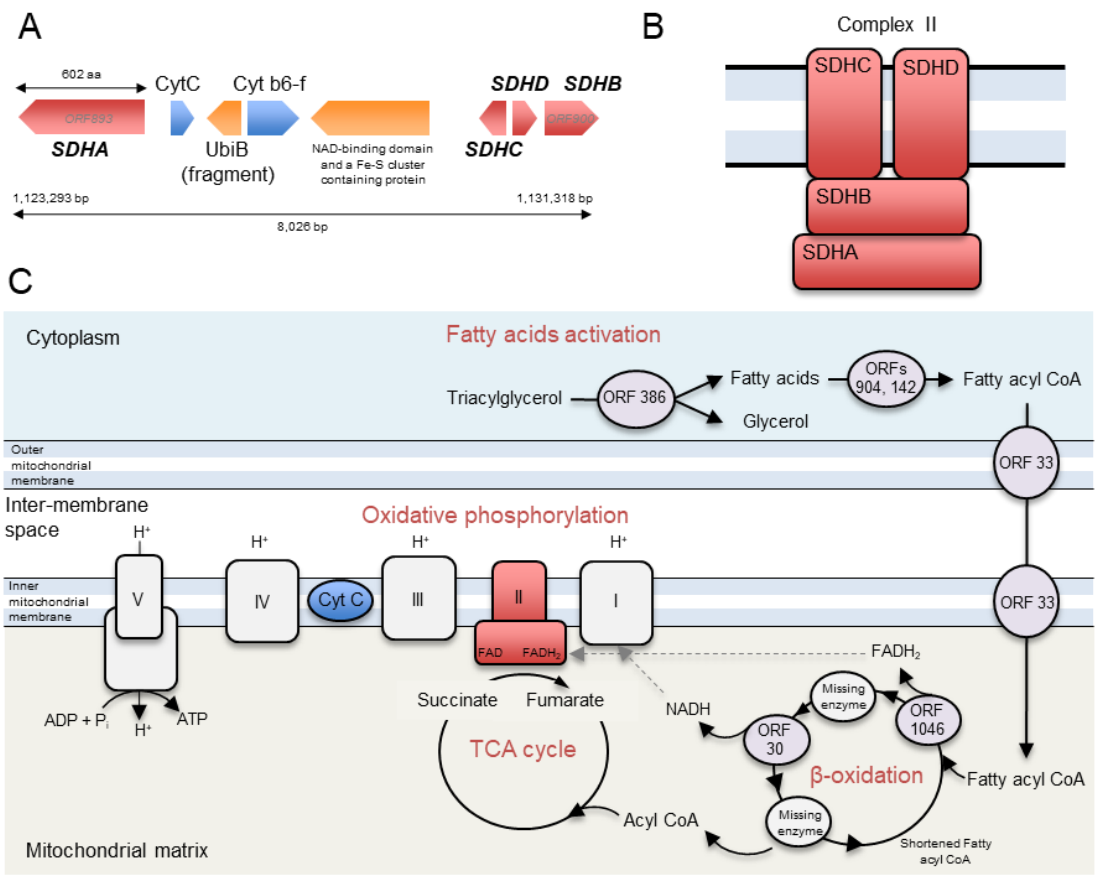

FIG 5 Genes in PkV RF01 predicted to encode enzymes of oxidative phosphorylation and β-oxidation pathways. (A) Gene organization in the succinate dehydrogenase-containing region. (B) Schematic representation of the canonical enzymatic complex II in the mitochondrial membrane. (C) Location of succinate dehydrogenase in the TCA cycle and electron transport chain as known in plants and a schematic reconstruction of the PkV RF01-encoded β-oxidation metabolic pathway.

SDH genes have recently been reported in viral genomes assembled from environmental samples for which functional experiments cannot be done (68). In a RT-PCR experiment using primers specific for the PkV RF01 gene for SDHA (hereafter, *vSDHA*), we detected transcripts of this gene in samples collected 24, 72, and 96 h post infection (Fig. 6). The *vSDHA* primers were tested on an uninfected culture to ensure that only the viral version of the *SDH*A gene was amplified (Appendix – Fig. S6). The MCP gene of PkV RF01 was used both for protocol optimization and later as an internal positive control (Appendix – Materials and Methods; Fig. S7). Although the transcription of the viral *SDHA* suggests that the

viral SDH is functional, we can only speculate on the possible role of this enzyme during infection. One possibility is that the viral SDH sustains the carbohydrate metabolism of infected cells (i.e., virocells) to supply building blocks of viral particles such as amino acids and to support proper replication of this large virus. Another possibility is that PkV RF01 uses its SDH as a part of an arms race with its host to turn on the TCA cycle after the host had turned it off to counter viral replication, or more simply to boost the energy metabolism of the virocells to augment the fitness of the host and/or to maximize virus production efficiency.

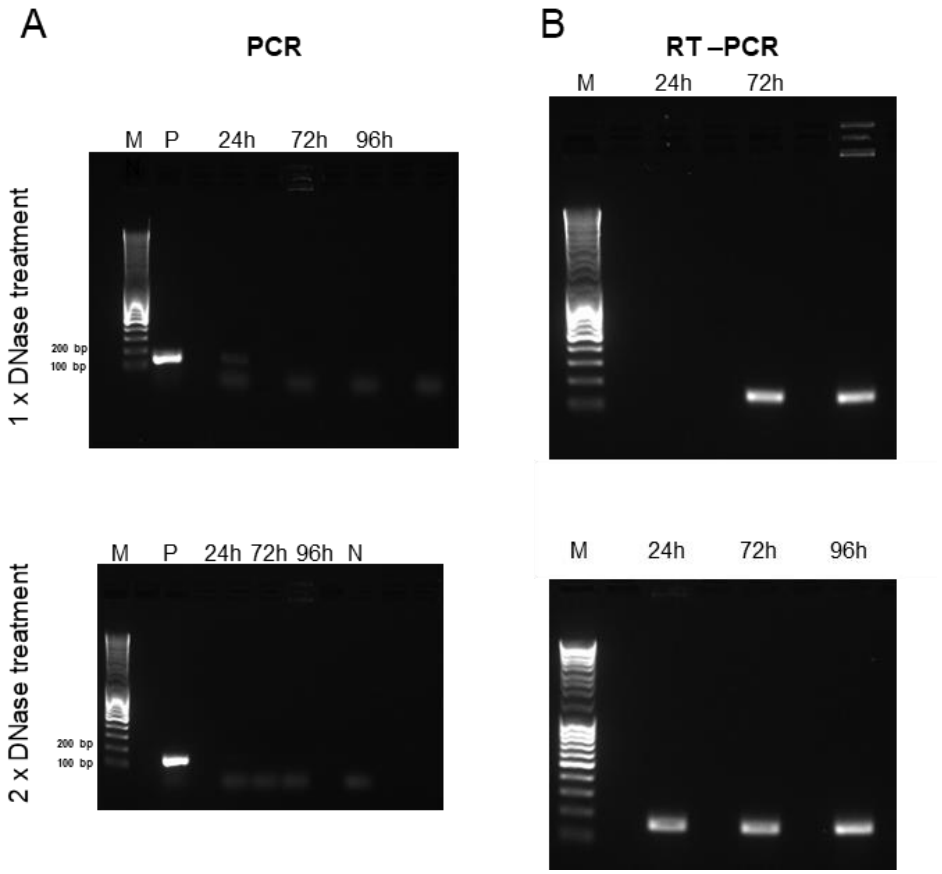

FIG 6 The viral SDHA gene is transcribed during infection. Gels of PCR and RT-PCR in combination with a TURBO DNA-free kit. Samples were taken 24, 72, and 96 h after infection. (A) PCR with *vSDHA*-specific primers was used to check for the presence of genomic DNA after RNA isolation and 1× and 2× DNase treatments. P, positive control (PKV RF01 genomic DNA); N, negative control (sd$H_2$0). (B) RT-PCR of RNA samples using *vSDHA*-specific primers. M, DNA marker (MassRuler DNA Ladder Mix, Thermo Fisher, 80 to 10,000 bp).

The discovery of the viral SDH prompted us to search for other potential viral-encoded SDHA and SDHB homologs in marine metagenomes. These two subunits (SDHA and SDHB) form the catalytic core containing the redox cofactors that participate in electron transfer to ubiquinone; they are thus more conserved than SDHC and SDHD subunits. To test for the presence of this viral SDH in other viruses, we searched for *vSDHA* and *B* in marine metagenomes of the *Tara* Oceans expedition. The 50 most-similar and non-redundant SDHA and B sequences predicted from 101 *Tara* Oceans genome fragments were most likely derived from *Mimiviridae* viruses (Fig. 7). Indeed, out of 1,113 genes predicted from these 101 genome fragments, 681 were annotated at some taxonomic level, of which 449 were predicted to be cellular and 157 viral. Of the 157 viral genes, 146 and 130 had their last common ancestor in *Mimiviridae* and Mesomimivirinae, respectively. A total of 32 of the 101-genome fragments contained at least one gene predicted to be of *Mimiviridae* origin, and the larger the genome fragment, the more *Mimiviridae* genes it was found to encode (Fig. 7A). Functional analysis indicated that 12 of the 1,113 predicted genes were NCLDV hallmark genes (encoding five VLTF3s, two capsid proteins, two PCNAs, two helicases, and one PolB). The high proportion of unknown genes and genes annotated as *Mimiviridae* in the 101 *Tara* Oceans genome fragments encoding SDHA or SDHB strongly suggests that these fragments belong to *Mimiviridae* viruses. This finding demonstrates that the presence of SDH is not restricted to PkV RF01 and is arguably widespread among marine *Mimiviridae*. According to phylogenetic analyses of cellular and viral SDHA and SDHB, the viral homologs form a monophyletic group that branches deeply within eukaryotic lineages (Fig. 7B-C). Long-branch attraction bias could generate such topologies but, as explained above for the IleRS and AsnRS, it is more likely that the viral SDHA and SDHB were

acquired at an early stage in the radiation of eukaryotic lineages. The transcription of *vSDHA* and its occurrence in marine environments calls for further investigation to understand the biological role and co-evolutionary significance of this viral SDH.

Other genes related to energy production were detected in the 8,026 bp-long region. ORF 894 and ORF 896, respectively corresponding to cytochrome *c* (CytC) and cytochrome b6-f complex iron-sulfur (Cyt b6-f) subunits, showed high sequence conservation with *Chrysochromulina* sp. CCMP291 proteins (78% and 59% amino acid [aa] identities, respectively). CytC is a short protein (~100 aa) involved in the oxidative phosphorylation pathway, where it accommodates the transfer of electrons between the coenzymes Q-cytochrome *c* reductase (complex III) and cytochrome *c* oxidase (complex IV). The presence of Cyt b6-f between oxidative phosphorylation genes is puzzling because the cytochrome b6-f complex is involved in photosynthesis. The core of the chloroplast b6f complex, however, is similar to the analogous respiratory cytochrome bc(1) complex. The other two predicted ORFs in this region are similar to ubiquinone biosynthesis protein UbiB (ORF 895) or contain a NAD-binding domain and a Fe-S cluster (ORF 897) and may thus be associated with electron transport as well. ORF 897 has two distant (25%–31% aa identity) homologs in the PkV RF01 genome (ORF 456 and ORF 625).

Some other genes were predicted to encode enzymes involved in pyruvate metabolism. ORF 79 has sequence homology with L-lactate dehydrogenases; it might thus catalyze the conversion of lactate to pyruvate, an intermediary compound serving as a starting point for several major metabolic pathways, such as glycolysis, gluconeogenesis, and the TCA cycle. ORF 727 was predicted to code for an isochorismate hydrolase that also produces pyruvate from isochorismate. ORF 24 and ORF 726 share sequence homology with phosphoenolpyruvate synthase and a partial

pyruvate kinase, respectively. The former catalyzes the conversion of pyruvate to phosphoenolpyruvate (PEP), while the latter catalyzes the reverse reaction. Formation of PEP is an initial step in gluconeogenesis.

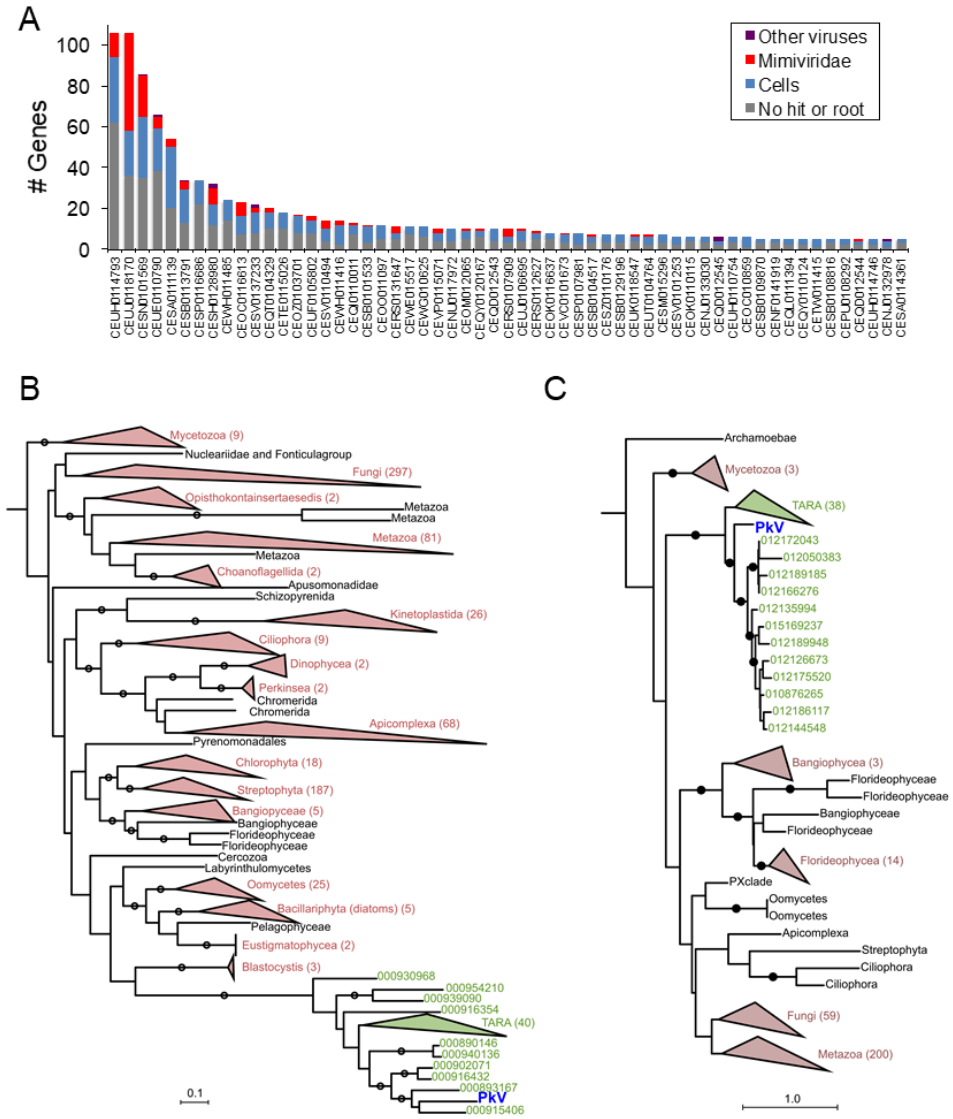

FIG 7 Origin of PkV RF01 SDHA and SDHB and their most similar homologs in *Tara* Oceans metagenomes. (A) Taxonomy of genes predicted in *Tara* Oceans metagenome assembled-genome fragments encoding the 50 SDHAs and SDHBs most similar to PkV RF01 genes (for genome fragments having at least five predicted genes). (B and C) Phylogenetic trees of viral and cellular SDHAs (B) and SDHBs (C). Clades in green contain PkV RF01 SDHA or SDHB and their 50 most similar hits identified in *Tara* Oceans metagenomes (predicted to be *Mimiviridae* homologs from A). Red, eukaryotic phyla; black, unclassified eukaryotes. Trees are rooted with Proteobacteria and Firmicutes homologs (not shown). Circles indicate branches with posterior probability support ≥ 50%.

A nearly complete viral-encoded β-oxidation pathway

In this study, 22 predicted genes were inferred to code for proteins involved in lipid synthesis or degradation, including key enzymes of the *β*-oxidation pathway (Appendix – Table S2). Several genes were predicted to code for lipase-like proteins (ORFs 386, 481, 635, 653, and 690), including a triacylglycerol lipase (ORF 386) that can break down triacylglycerol into glycerol and fatty acids. Glycerol and fatty acids can be used as a starting point for ATP production—by glycolysis and *β*-oxidation, respectively. In the *β*-oxidation pathway, fatty acids are fully oxidized to produce acetyl-CoA, which can then enter the TCA cycle to yield NADH and FADH2; these latter two products can funnel through to the electron transport chain to produce ATP (Fig. 5C). Each *β*-oxidation cycle itself also produces NADH and FADH2 cofactors. We found that PkV RF01 encodes key *β*-oxidation enzymes. First, two distantly related ORFs (ORF 142 and ORF 904 sharing 22% aa identity) have sequence homology with a long-chain fatty acyl-CoA synthetase. This enzyme catalyzes the formation of fatty acyl-CoA in the cytosol. Fatty acyl-CoA can be imported to mitochondria using a (carnitine) CoA-transferase also encoded in PkV RF01 (ORF 33). Once in the mitochondrial matrix, fatty acyl-CoA serves as a substrate on which an acyl-CoA dehydrogenase (ORF 1046) oxidizes the fatty acyl-CoA and reduces a FAD cofactor to produce a FADH2 cofactor. We identified a 2,4-dienoyl-CoA reductase (ORF 30) that may facilitate the next oxidation step to produce a NADH cofactor. FADH2 and NADH molecules produced by a *β*-oxidation cycle can both be oxidized in the electron transport chain to generate ATP. The enzymes involved in the two intermediate steps following each oxidation, either an enoyl-CoA hydratase or a β-ketothiolase, were not detected in our analysis.

Most of these genes have no homologs in reference viral genomes, and, to our knowledge, this is the first report of a virus possessing proteins directly involved in lipid-based energy production. By diverting host lipid machinery, interactions of viruses with lipids or lipid based-structures have long been known to have structural or signaling roles at different stages of the virus life cycle, such as entry, genome replication, morphogenesis, and exit (69–71). More recently, several studies on human viruses (two herpesviruses and one RNA virus) have shown that the metabolic state of an infected cell can be shifted toward energy generation to support viral replication (70). These studies have highlighted the increasing abundance—up to 48 h after HCV infection—of enzymes involved in β-oxidation, amino acid catabolism, and the TCA cycle (72) and an increase in cellular β-oxidation following the release of free fatty acids caused by Dengue virus-induced autophagy (73). Among algal viruses, EhV remodels the transcription of host lipid genes for fatty acid synthesis to support viral assembly (74) and also to generate triacylglycerols stored in the virion and available as an energy pool in later infection phases (75). Besides diverting the host metabolism, EhV encodes seven proteins involved in the sphingolipid biosynthesis pathway (76). This pathway produces a viral sphingolipid that is a central component of EhV lipid membranes and that can also act as a signaling lipid and induce programed cell death during the lytic infection phase (77). EhV also encodes a triglyceride lipase (with detectable homology to predicted PkV RF01 lipases ORF 635 and ORF653) that is highly expressed during late infection concomitantly with significant up-regulation of host β-oxidation genes (74). These examples and our observations of several genes involved in β-oxidation clearly show that viruses can introduce new metabolism-related genes, sometimes representing

entire pathways, into the host, most likely to satisfy the high metabolic requirement of these giant viruses.

High representation of glycosyltransferases

Compared with other viruses, PkV RF01 was found to encode an unusually high number of glycosyltransferases (GTs) as well as other carbohydrate-active enzymes (Appendix – Results). Automated annotation of GTs (and other carbohydrate-active enzymes) in reference viral proteomes using dbCAN2 (78) revealed that the largest number of GT domains was encoded by PkV RF01 (n = 48), followed by CeV (n = 13), *Mimivirus* members, and CroV and AaV (n = 8–10) (Appendix – Fig. S8). We uncovered 48 GT domains encoded in 40 ORFs, 8 of which were predicted to encode more than one GT domain. These domains correspond to 16 different GT families. Most domains were inferred to be functional, as 31 out of 48 covered at least 70% of the dbCAN2 reference domain, with coverage ranging from 44% to 99%. GTs were found scattered across the genome of PkV RF01 but with some local clustering (Fig. 2A), the latter indicating possible involvement in the same pathway. GT32 was the most represented domain, with 11 proteins (as annotated by dbCAN2) and potentially three additional proteins (ORFs 40, 84, and 861). Eight proteins possessed a GT25 domain that can catalyze the transfer of various sugars onto a growing lipopolysaccharide chain during its biosynthesis. Among these eight predicted ORFs, four contained an additional non-overlapping GT domain (two GT2s, one GT6, and one GT60). Functional analyses of GTs in mimiviruses (or in related *Paramecium bursaria* Chlorella viruses) have demonstrated that some of these enzymes are functional, being able to modify viral collagen-like proteins (79) and polymerize sugars (80). Conservation between PkV RF01 GTs and functionally characterized GTs in viruses and cells is absent or extremely low, which precludes any predictions

as to the specific roles of these enzymes in the PkV RF01 life cycle. Nevertheless, this putative glycosylation-conducive autonomy possibly allows the virus to infect a variety of hosts, as the virus can modify its own glycans, which are used for host recognition, independently of the host system (81). In alpha-, flavi-, and herpes-viruses, fusion is mediated by viral glycoproteins (36).

Conclusions

The haptophyte virus PkV RF01 has been previously shown to have a longer replication cycle and a broader host range compared with other prymnesioviruses and most other algal viruses. Here, we revealed that PkV RF01 has atypical virion morphology and that infections yield several orders of magnitude fewer infectious particles than other tested prymnesioviruses. In-depth phylogenetic analysis using genes conserved in NCLDVs confirmed that PkV RF01 belongs to *Mimiviridae* but is deeply diverged from existing members, although closer to alga-infecting *Mimiviridae* than heterotroph-infecting ones. Unlike other alga-infecting *Mimiviridae*, however, PkV RF01 has a large genome (1.4 Mb) and contains genes coding for two aminoacyl-tRNA synthetases and the complete BER pathway. All these features are conserved in most heterotrophic *Mimiviridae* and therefore must have been lost in other alga-infecting *Mimiviridae*. This outlier virus features an unprecedentedly high number of genes involved in energy metabolism and glycosylation machinery that may enable its longer replication cycle and broader host range compared with other algal viruses. These genomic and phenotypic features are suggestive of a persistent infection behavior that probably evolved in response to the host growth strategy. Because of nutrient limitations, these persistent systems of slow-growing but

ubiquitous hosts with less virulent viruses may represent the most common type of virocells in oceans.

Materials and Methods

PkV RF01 culturing and characterization

Detailed materials and methods are provided in Appendix – Materials and Methods. The details include information on culturing and infection, virus infectivity, sensitivity to chloroform, cryo-electron tomography, purification of viral particles and DNA isolation, and amplification and RT-PCR of *vSDHA*.

In brief, exponentially growing cultures of He UiO028 were infected with PkV RF01 at a virus-to-alga ratio of 10 to facilitate a one-step virus growth cycle (33). Infection was followed by flow cytometry counting (82, 83) to measure total VLPs and host cells. Furthermore, virus infectivity was determined by MPN (84) using He UiO028 as host and compared with that of two other prymnesioviruses, HeV RF02 and PkV RF02 (33), propagated on He UiO028 and *Prymnesium kappa* RCC3423, respectively. Sensitivity to chloroform, which suggests the presence of a lipid membrane or lipid molecules in the capsid, was tested in triplicate samples of exponentially growing HeUiO028 cells infected with a 1:10 volume of chloroform-treated PkV RF01virus.

The structure of PkV RF01 was determined by cryo-electron tomography using a 200-kV transmission electron microscope (Talos F200C, Thermo Scientific) equipped with a Ceta 16M camera. Tilt series were recorded at 45,000× magnification and –7-µm defocus between –60° to 60° by 2° increments. Finally, reconstruction, segmentation, and visualization of tomograms was performed with the software package IMOD v4.9 (85).

596 Transcription of *vSDHA* was determined in an infected He UiO028 and PkV
597 RF01 culture using an uninfected He UiO028 culture as a control. Samples were
598 collected at 24, 72, and 96 h post infection from both cultures. RNA was extracted
599 using the RNeasy Plus Universal Mini kit (Qiagen), with gDNA removed in an extra
600 step using a TURBO DNA-free kit (Ambion). RT-PCRs were performed on the
601 isolated mRNA using a SuperScript III One-Step RT-PCR with Platinum *Taq* DNA
602 Polymerase system (Thermo Fisher).
603 DNA for PkV RF01 genome sequencing was isolated from 2 L of lysed He
604 UiO028 culture. Algal debris and bacteria were removed by low-speed centrifugation.
605 Viral particles were concentrated by ultracentrifugation at 25,000 rpm in a Beckman
606 Coulter Optima L90K ultracentrifuge for 2 h and further purified by Optiprep gradient
607 centrifugation (86). Isolation of high-quality DNA for sequencing was performed
608 using the protocol described in (87) with some modifications. Total DNA was
609 purified using a Zymo Genomic DNA Clean & Concentrator Kit-10 (Zymo Research,
610 Irvine, CA, USA).

611 **Genome sequencing, assembly, and annotation**

612 DNA isolated from PkV RF01 was subjected to Illumina TruSeq PCR-free library
613 preparation (insert size: 350 bp) and sequenced on an Illumina MiSeq instrument in
614 paired-end mode (2 × 300 bp), thereby yielding approximately 1.9 million reads. In
615 addition, a ligation-based $1D^2$ nanopore library (LSK-308) was constructed and
616 sequenced using an Oxford Nanopore MinION Mk1b device and a FLO-MIN107 flow
617 cell, which resulted in 825 long reads with an N50 of 13.6 kbp comprising 9.89 Mbp in
618 total. These data were assembled in a two-step process using Newbler (88) and Consed
619 (89), and the consensus sequence was polished using Nanopolish (90) and Pilon (91).
620 Genes were predicted using GeneMarkS (92) and tRNAscan-SE (93). Predicted protein

sequences were searched against various public databases using BLASTP, and significant alignments served as the basis for manual investigation on the GenomeNet (https://www.genome.jp) server to assign putative functions. Appendix – Materials and Methods contains further details on the genome assembly and annotation as well as specific methods used for reconstruction of each phylogenetic tree.

Data availability

Raw sequence reads and PkV RF01 genome sequence were deposited at the European Bioinformatics Institute (EMBL-EBI) (https://www.ebi.ac.uk) under project name PRJEB37450. The complete video records of a cryo-electron tomogram of a PkV RF01 virion and sequence data as well as curated gene annotation table as reported in this study are available at https://github.com/RomainBlancMathieu/PkV-RF01.

Acknowledgements

The recording of tilt series was performed with the help of Sebastian Schultz at the Unit of Cellular Electron Microscopy, the Norwegian Radium Hospital. Initial sequencing (MiSeq and Pacbio) of PkV RF01 total DNA was performed at the Norwegian Sequencing Center (https://www.sequencing.uio.no/). We thank Hilde M. K. Stabell and Solveig Siqveland, Department of Biological Sciences, University of Bergen, Norway, for technical assistance with molecular biology experiments as well as Christian Rückert, Bielefeld University, for support in manual finishing of genome assembly. This work was supported by the Research Council of Norway project entitled "Uncovering the key players for regulation of phytoplankton function and structure: lesson to be learned from algal virus-haptophyte coexistence" (VirVar, project number 294364 to RAS). Additional funding was provided by the European Union Horizons 2020 research and innovation program, grant agreement no. 685778 ("Virus-X") to RAS and DB. This work was also supported by the Future Development Funding Program of the Kyoto University Research Coordination Alliance. HO was supported by JSPS/KAKENHI (No. 18H02279), and Scientific Research on Innovative Areas from the Ministry of Education, Culture, Science, Sports and Technology (MEXT) of Japan (Nos. 16H06429, 16K21723, 16H06437). The Super Computer System, Institute for Chemical Research, Kyoto University, provided computational time. We thank Barbara Goodson, from Edanz Group (www.edanzediting.com/ac), for editing the English text of a draft of this manuscript.

Competing interests

Authors declare having no competing interests.

References

1. Dupré JO. 2009. Varieties of Living Things: Life at the Intersection of Lineage and Metabolism. Philosophy & Theory in Biology 1.

2. Forterre P. 2012. Virocell Concept, The, p. a0023264. *In* John Wiley & Sons, Ltd (ed.), eLS. John Wiley & Sons, Ltd, Chichester, UK.

3. Raoult D, Forterre P. 2008. Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol 6:315–319.

4. Claverie J-M. 2006. Viruses take center stage in cellular evolution. Genome Biology 7:110.

5. Mann NH, Cook A, Millard A, Bailey S, Clokie M. 2003. Bacterial photosynthesis genes in a virus. Nature 424:741–741.

6. Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. 2004. Transfer of photosynthesis genes to and from Prochlorococcus viruses. PNAS 101:11013–11018.

7. Hurwitz BL, Hallam SJ, Sullivan MB. 2013. Metabolic reprogramming by viruses in the sunlit and dark ocean. Genome Biol 14:R123.

8. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, Chisholm SW. 2011. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci USA 108:E757-764.

9. Anantharaman K, Duhaime MB, Breier JA, Wendt KA, Toner BM, Dick GJ. 2014. Sulfur Oxidation Genes in Diverse Deep-Sea Viruses. Science 344:757–760.

10. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. 2005. Photosynthesis genes in marine viruses yield proteins during host infection. Nature 438:86–89.

11. Fridman S, Flores-Uribe J, Larom S, Alalouf O, Liran O, Yacoby I, Salama F, Bailleul B, Rappaport F, Ziv T, Sharon I, Cornejo-Castillo FM, Philosof A, Dupont CL, Sánchez P, Acinas SG, Rohwer FL, Lindell D, Béjà O. 2017. A myovirus encoding both photosystem I and II proteins enhances cyclic electron flow in infected Prochlorococcus cells. Nat Microbiol 2:1350–1357.

12. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie J-M. 2004. The 1.2-megabase genome sequence of Mimivirus. Science 306:1344–1350.

13. Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, Daims H, Horn M, Wagner M, Jensen GJ, Kyrpides NC, Koonin EV, Woyke T. 2017. Giant viruses with an expanded complement of translation system components. Science 356:82–85.

14. Abrahão J, Silva L, Silva LS, Khalil JYB, Rodrigues R, Arantes T, Assis F, Boratto P, Andrade M, Kroon EG, Ribeiro B, Bergier I, Seligmann H, Ghigo E, Colson P, Levasseur A, Kroemer G, Raoult D, Scola BL. 2018. Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. Nat Commun 9:1–12.

15. Miller ES, Heidelberg JF, Eisen JA, Nelson WC, Durkin AS, Ciecko A, Feldblyum TV, White O, Paulsen IT, Nierman WC, Lee J, Szczypinski B, Fraser CM. 2003. Complete Genome Sequence of the Broad-Host-Range Vibriophage KVP40: Comparative Genomics of a T4-Related Bacteriophage. J Bacteriol 185:5220–5233.

16. Yoshikawa G, Askora A, Blanc-Mathieu R, Kawasaki T, Li Y, Nakano M, Ogata H, Yamada T. 2018. Xanthomonas citri jumbo phage XacN1 exhibits a wide host range and high complement of tRNA genes. Sci Rep 8:4486.

17. Mizuno CM, Guyomar C, Roux S, Lavigne R, Rodriguez-Valera F, Sullivan MB, Gillet R, Forterre P, Krupovic M. 2019. Numerous cultivated and uncultivated viruses encode ribosomal proteins. Nature Communications 10:752.

18. Schvarcz CR, Steward GF. 2018. A giant virus infecting green algae encodes key fermentation genes. Virology 518:423–433.

19. Piacente F, Gaglianone M, Laugieri ME, Tonetti MG. 2015. The Autonomous Glycosylation of Large DNA Viruses. International Journal of Molecular Sciences 16:29315–29328.

20. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh D, Denef VJ, McMahon KD, Konstantinidis KT, Eloe-Fadrosh EA, Kyrpides N, Woyke T. 2020. Giant virus diversity and host interactions through global metagenomics. Nature 1–7.

720 21. Needham DM, Yoshizawa S, Hosaka T, Poirier C, Choi CJ, Hehenberger E,
721 Irwin NAT, Wilken S, Yung C-M, Bachy C, Kurihara R, Nakajima Y, Kojima K,
722 Kimura-Someya T, Leonard G, Malmstrom RR, Mende DR, Olson DK, Sudo Y,
723 Sudek S, Richards TA, DeLong EF, Keeling PJ, Santoro AE, Shirouzu M,
724 Iwasaki W, Worden AZ. 2019. A distinct lineage of giant viruses brings a
725 rhodopsin photosystem to unicellular marine predators. PNAS 116:20574–
726 20583.

727 22. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT,
728 Solonenko N, Lara E, Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral
729 M, Searson S, Cruaud C, Alberti A, Duarte CM, Gasol JM, Vaqué D, Bork P,
730 Acinas SG, Wincker P, Sullivan MB. 2016. Ecogenomics and potential
731 biogeochemical impacts of globally abundant ocean viruses. Nature
732 537:689–693.

733 23. Nishimura Y, Watai H, Honda T, Mihara T, Omae K, Roux S, Blanc-Mathieu R,
734 Yamamoto K, Hingamp P, Sako Y, Sullivan MB, Goto S, Ogata H, Yoshida T.
735 2017. Environmental Viral Genomes Shed New Light on Virus-Host
736 Interactions in the Ocean. mSphere 2.

737 24. Forterre P. 2013. The virocell concept and environmental microbiology.
738 ISME J 7:233–236.

739 25. Rosenwasser S, Ziv C, Creveld SG van, Vardi A. 2016. Virocell Metabolism:
740 Metabolic Innovations During Host–Virus Interactions in the Ocean. Trends
741 in Microbiology 24:821–832.

742 26. Coy SR, Gann ER, Pound HL, Short SM, Wilhelm SW. 2018. Viruses of
743 Eukaryotic Algae: Diversity, Methods for Detection, and Future Directions.
744 Viruses 10:487.

745 27. Gallot-Lavallée L, Blanc G, Claverie J-M. 2017. Comparative Genomics of
746 Chrysochromulina Ericina Virus and Other Microalga-Infecting Large DNA
747 Viruses Highlights Their Intricate Evolutionary Relationship with the
748 Established Mimiviridae Family. Journal of Virology 91:e00230-17.

749 28. Jacobsen A, Bratbak G, Heldal M. 1996. Isolation and characterization of a
750 virus infecting phaeocystis pouchetii (prymnesiophyceae)1. Journal of
751 Phycology 32:923–927.

752 29. Baudoux A-C, Brussaard CPD. 2005. Characterization of different viruses
753 infecting the marine harmful algal bloom species Phaeocystis globosa.
754 Virology 341:80–90.

755 30. Wagstaff BA, Vladu IC, Barclay JE, Schroeder DC, Malin G, Field RA. 2017.
756 Isolation and Characterization of a Double Stranded DNA Megavirus
757 Infecting the Toxin-Producing Haptophyte Prymnesium parvum. Viruses 9.

758 31. Egge ES, Johannessen TV, Andersen T, Eikrem W, Bittner L, Larsen A,
759 Sandaa R-A, Edvardsen B. 2015. Seasonal diversity and dynamics of
760 haptophytes in the Skagerrak, Norway, explored by high-throughput
761 sequencing. Mol Ecol 24:3026–3042.

762 32. Sandaa RA, Heldal M, Castberg T, Thyrhaug R, Bratbak G. 2001. Isolation and
763 characterization of two viruses with large genome size infecting

764 Chrysochromulina ericina (Prymnesiophyceae) and Pyramimonas orientalis
765 (Prasinophyceae). Virology 290:272–280.

766 33. Johannessen TV, Bratbak G, Larsen A, Ogata H, Egge ES, Edvardsen B,
767 Eikrem W, Sandaa R-A. 2015. Characterisation of three novel giant viruses
768 reveals huge diversity among viruses infecting Prymnesiales (Haptophyta).
769 Virology 476:180–188.

770 34. Huiskonen JT, Kivelä HM, Bamford DH, Butcher SJ. 2004. The PM2 virion has
771 a novel organization with an internal membrane and pentameric receptor
772 binding spikes. Nat Struct Mol Biol 11:850–856.

773 35. Yan X, Chipman PR, Castberg T, Bratbak G, Baker TS. 2005. The Marine Algal
774 Virus PpV01 Has an Icosahedral Capsid with T=219 Quasisymmetry. J Virol
775 79:9236–9243.

776 36. Huiskonen JT, Butcher SJ. 2007. Membrane-containing viruses with
777 icosahedrally symmetric capsids. Curr Opin Struct Biol 17:229–236.

778 37. King AM, Lefkowitz E, Adams MJ, Carstens EB. 2011. Virus Taxonomy: Ninth
779 Report of the International Committee on Taxonomy of Viruses. Elsevier.

780 38. Peralta B, Gil-Carton D, Castaño-Díez D, Bertin A, Boulogne C, Oksanen HM,
781 Bamford DH, Abrescia NGA. 2013. Mechanism of Membranous Tunnelling
782 Nanotube Formation in Viral Genome Delivery. PLOS Biology 11:e1001667.

783 39. Philippe C, Krupovic M, Jaomanjaka F, Claisse O, Petrel M, le Marrec C. 2018.
784 Bacteriophage GC1, a Novel Tectivirus Infecting Gluconobacter Cerinus, an
785 Acetic Acid Bacterium Associated with Wine-Making. Viruses 10.

786 40. Yan X, Yu Z, Zhang P, Battisti AJ, Chipman PR, Bajaj C, Bergoin M, Rossmann
787 MG, Baker TS. 2009. The Capsid Proteins of a Large, Icosahedral dsDNA
788 Virus. J Mol Biol 385:1287–1299.

789 41. Klose T, Rossmann MG. 2014. Structure of large dsDNA viruses. Biol Chem
790 395:711–719.

791 42. Gowing MM. 1993. Large virus-like particles from vacuoles of phaeodarian
792 radiolarians and from other marine samples. Marine Ecology Progress
793 Series 101:33–43.

794 43. Gromov BV, Mamkaeva KA. 1981. A virus infection in the synchronized
795 population of the *Chlorococcum minutum* zoospores. Algological
796 Studies/Archiv für Hydrobiologie, Supplement Volumes 252–259.

797 44. Bratbak G, Jacobsen A, Heldal M, Nagasaki K, Thingstad F. 1998. Virus
798 production in Phaeocystis pouchetii and its relation to host cell growth and
799 nutrition. Aquatic Microbial Ecology 16:1–9.

800 45. Zimmerman AE, Bachy C, Ma X, Roux S, Jang HB, Sullivan MB, Waldbauer JR,
801 Worden AZ. 2019. Closely related viruses of the marine picoeukaryotic alga
802 Ostreococcus lucimarinus exhibit different ecological strategies. Environ
803 Microbiol 21:2148–2170.

804 46. Thomsen HA, Buck KR, Chavez FP. 1994. Haptophytes as components of
805 marine phytoplankton. DTU Research Database 187–208.

806 47. Berngruber TW, Froissart R, Choisy M, Gandon S. 2013. Evolution of
807 Virulence in Emerging Epidemics. PLOS Pathogens 9:e1003209.

808 48. Day T, Proulx SR. 2004. A General Theory for the Evolutionary Dynamics of
809 Virulence. The American Naturalist 163:E40–E63.

810 49. King AA, Shrestha S, Harvill ET, Bjørnstad ON. 2009. Evolution of Acute
811 Infections and the Invasion-Persistence Trade-Off. The American Naturalist
812 173:446–455.

813 50. Castberg T, Thyrhaug R, Larsen A, Sandaa R-A, Heldal M, Etten JLV, Bratbak
814 G. 2002. Isolation and Characterization of a Virus That Infects Emiliania
815 Huxleyi (haptophyta)1. Journal of Phycology 38:767–774.

816 51. Brussaard CPD, Bratbak G, Baudoux A-C, Ruardij P. 2007. Phaeocystis and
817 its interaction with viruses. Biogeochemistry 83:201–215.

818 52. Leggett HC, Buckling A, Long GH, Boots M. 2013. Generalism and the
819 evolution of parasite virulence. Trends Ecol Evol (Amst) 28:592–596.

820 53. Woolhouse MEJ, Taylor LH, Haydon DT. 2001. Population Biology of
821 Multihost Pathogens. Science 292:1109–1112.

822 54. Johannessen TV, Larsen A, Bratbak G, Pagarete A, Edvardsen B, Egge ED,
823 Sandaa R-A. 2017. Seasonal Dynamics of Haptophytes and dsDNA Algal
824 Viruses Suggest Complex Virus-Host Relationship. Viruses 9.

825 55. Gran-Stadniczeñko S, Krabberød AK, Sandaa R-A, Yau S, Egge E, Edvardsen
826 B. 2019. Seasonal Dynamics of Algae-Infecting Viruses and Their Inferred
827 Interactions with Protists. Viruses 11:1043.

56. Fischer MG, Allen MJ, Wilson WH, Suttle CA. 2010. Giant virus with a remarkable complement of genes infects marine zooplankton. Proc Natl Acad Sci USA 107:19508–19513.

57. Heger P, Marin B, Bartkuhn M, Schierenberg E, Wiehe T. 2012. The chromatin insulator CTCF and the emergence of metazoan diversity. PNAS 109:17507–17512.

58. Pentland I, Parish JL. 2015. Targeting CTCF to Control Virus Gene Expression: A Common Theme amongst Diverse DNA Viruses. Viruses 7:3574–3585.

59. Washington SD, Musarrat F, Ertel MK, Backes GL, Neumann DM. 2018. CTCF Binding Sites in the Herpes Simplex Virus 1 Genome Display Site-Specific CTCF Occupation, Protein Recruitment, and Insulator Function. J Virol 92.

60. Sandaa R-A, Dahle H, Brussaard CPD, Ogata H, Blanc-Mathieu R. Algal viruses belonging to a subgroup within the Mimiviridae familyEncyclopedia of Virology, 4th ed. Bamford, D., M. Zuckerman, Elsevier, Academic.

61. Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36.

62. Blanc-Mathieu R, Ogata H. 2016. DNA repair genes in the Megavirales pangenome. Current Opinion in Microbiology 31:94–100.

63. Wolf YI, Aravind L, Grishin NV, Koonin EV. 1999. Evolution of Aminoacyl-tRNA Synthetases—Analysis of Unique Domain Architectures and

850 Phylogenetic Trees Reveals a Complex History of Horizontal Gene Transfer
851 Events. Genome Res 9:689–710.

852 64. Cvetesic N, Dulic M, Bilus M, Sostaric N, Lenhard B, Gruic-Sovulj I. 2016.
853 Naturally Occurring Isoleucyl-tRNA Synthetase without tRNA-dependent
854 Pre-transfer Editing. J Biol Chem 291:8618–8631.

855 65. Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods
856 Will be Positively Misleading. Systematic Zoology 27:401–410.

857 66. Moreira D, López-García P. 2015. Evolution of viruses and cells: do we need
858 a fourth domain of life to explain the origin of eukaryotes? Philos Trans R
859 Soc Lond B Biol Sci 370.

860 67. Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch
861 attraction artefacts in the animal phylogeny using a site-heterogeneous
862 model. BMC Evolutionary Biology 7:S4.

863 68. Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO.
864 2020. Dynamic genome evolution and complex virocell metabolism of
865 globally-distributed giant viruses. 1. Nature Communications 11:1710.

866 69. Ono A. 2010. Viruses and Lipids. Viruses 2:1236–1238.

867 70. Heaton NS, Randall G. 2011. Multifaceted roles for lipids in viral infection.
868 Trends Microbiol 19:368–375.

869 71. Lange PT, Lagunoff M, Tarakanova VL. 2019. Chewing the Fat: The
870 Conserved Ability of DNA Viruses to Hijack Cellular Lipid Metabolism.
871 Viruses 11:119.

872 72. Diamond DL, Syder AJ, Jacobs JM, Sorensen CM, Walters K-A, Proll SC,
873 McDermott JE, Gritsenko MA, Zhang Q, Zhao R, Metz TO, Camp DG, Waters
874 KM, Smith RD, Rice CM, Katze MG. 2010. Temporal Proteome and Lipidome
875 Profiles Reveal Hepatitis C Virus-Associated Reprogramming of
876 Hepatocellular Metabolism and Bioenergetics. PLoS Pathog 6.

877 73. Heaton NS, Randall G. 2010. Dengue virus induced autophagy regulates lipid
878 metabolism. Cell Host Microbe 8:422–432.

879 74. Rosenwasser S, Mausz MA, Schatz D, Sheyn U, Malitsky S, Aharoni A,
880 Weinstock E, Tzfadia O, Ben-Dor S, Feldmesser E, Pohnert G, Vardi A. 2014.
881 Rewiring Host Lipid Metabolism by Large Viruses Determines the Fate of
882 Emiliania huxleyi, a Bloom-Forming Alga in the Ocean[C][W][OPEN]. Plant
883 Cell 26:2689–2707.

884 75. Malitsky S, Ziv C, Rosenwasser S, Zheng S, Schatz D, Porat Z, Ben-Dor S,
885 Aharoni A, Vardi A. 2016. Viral infection of the marine alga Emiliania
886 huxleyi triggers lipidome remodeling and induces the production of highly
887 saturated triacylglycerol. New Phytologist 210:88–96.

888 76. Wilson WH, Schroeder DC, Allen MJ, Holden MTG, Parkhill J, Barrell BG,
889 Churcher C, Hamlin N, Mungall K, Norbertczak H, Quail MA, Price C,
890 Rabbinowitsch E, Walker D, Craigon M, Roy D, Ghazal P. 2005. Complete

891 Genome Sequence and Lytic Phase Transcription Profile of a
892 Coccolithovirus. Science 309:1090–1092.

893 77. Vardi A, Van Mooy BAS, Fredricks HF, Popendorf KJ, Ossolinski JE, Haramaty
894 L, Bidle KD. 2009. Viral glycosphingolipids induce lytic infection and cell
895 death in marine phytoplankton. Science 326:861–865.

896 78. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y.
897 2018. dbCAN2: a meta server for automated carbohydrate-active enzyme
898 annotation. Nucleic Acids Res 46:W95–W101.

899 79. Luther KB, Hülsmeier AJ, Schegg B, Deuber SA, Raoult D, Hennet T. 2011.
900 Mimivirus collagen is modified by bifunctional lysyl hydroxylase and
901 glycosyltransferase enzyme. J Biol Chem 286:43701–43709.

902 80. Rommel AJ, Hülsmeier AJ, Jurt S, Hennet T. 2016. Giant mimivirus R707
903 encodes a glycogenin paralogue polymerizing glucose through α- and β-
904 glycosidic linkages. Biochem J 473:3451–3462.

905 81. Parakkottil Chothi M, Duncan GA, Armirotti A, Abergel C, Gurnon JR, Van
906 Etten JL, Bernardi C, Damonte G, Tonetti M. 2010. Identification of an l-
907 Rhamnose Synthetic Pathway in Two Nucleocytoplasmic Large DNA
908 Viruses. J Virol 84:8829–8838.

909 82. Marie D, Brussaard CPD, Thyrhaug R, Bratbak G, Vaulot D. 1999.
910 Enumeration of Marine Viruses in Culture and Natural Samples by Flow
911 Cytometry. Appl Environ Microbiol 65:45–52.

912 83. Brussaard CPD. 2004. Optimization of procedures for counting viruses by
913 flow cytometry. Appl Environ Microbiol 70:1506–1513.

914 84. Suttle CA, Chan AM. 1993. Marine cyanophages infecting oceanic and coastal
915 strains of Synechococcus: abundance, morphology, cross-infectivity and
916 growth characteristics. Marine ecology progress series
917 https://doi.org/10.3354/meps092099.

918 85. Kremer JR, Mastronarde DN, McIntosh JR. 1996. Computer visualization of
919 three-dimensional image data using IMOD. J Struct Biol 116:71–76.

920 86. Lawrence JE, Steward GF. 2010. Purification of viruses by centrifugation, p.
921 166–181. *In* Wilhelm, S, Weinbauer, M, Suttle, C (eds.), Manual of Aquatic
922 Viral Ecology. American Society of Limnology and Oceanography.

923 87. Sandaa R-A, E. Storesund J, Olesin E, Lund Paulsen M, Larsen A, Bratbak G,
924 Ray JL. 2018. Seasonality Drives Microbial Community Structure, Shaping
925 both Eukaryotic and Prokaryotic Host–Viral Relationships in an Arctic
926 Marine Ecosystem. Viruses 10.

927 88. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J,
928 Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV,
929 Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI,
930 Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM,
931 Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers
932 EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ,
933 Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA,
934 Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM.

935 2005. Genome sequencing in microfabricated high-density picolitre
936 reactors. Nature 437:376–380.

937 89. Gordon D, Green P. 2013. Consed: a graphical editor for next-generation
938 sequencing. Bioinformatics 29:2936–2937.

939 90. Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome
940 assembled de novo using only nanopore sequencing data. Nat Methods
941 12:733–735.

942 91. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA,
943 Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: An Integrated Tool for
944 Comprehensive Microbial Variant Detection and Genome Assembly
945 Improvement. PLOS ONE 9:e112963.

946 92. Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training
947 method for prediction of gene starts in microbial genomes. Implications for
948 finding sequence motifs in regulatory regions. Nucleic Acids Res 29:2607–
949 2618.

950 93. Lowe TM, Chan PP. 2016. tRNAscan-SE On-line: integrating search and
951 context for analysis of transfer RNA genes. Nucleic Acids Res 44:W54-57.

952

Appendix

Materials and Methods 2
Culturing and infection 2
Infectious progeny 2
Sensitivity to chloroform 2
Cryo-electron tomography 2
Purification of viral particles and DNA isolation 3
Genome assembly 3
Phylogenetic analyses 3
Five core genes, SDHA, and SDHB 3
Rpb2, IleRS, and AsnRS 4
Gene prediction and functional and taxonomic annotation 4
Taxonomic and functional analysis of vSDHA homologs in OM-RGCv1 5
PCR and RT-PCR optimization 5
PCR amplification and RT-PCR analysis of *vSDHA* 5
Results 6
DNA repair enzymes 6
Transcription 6
Translation 6
Other carbohydrate-active enzymes 7
Figures 8
Fig. S1. Reduction of PkV RF01 infectivity with chloroform. 8
Fig. S2. A putative CTCF DNA-binding protein in PkV RF01. 9
Fig. S3. High density of exact small repeats in the PkV RF01 genome. 10
Fig. S4. COG functional distribution of 339 proteins encoded by PkV RF01. 11
Fig. S5. Comparative COG functional distribution among *Mimiviridae* members. 12
Fig. S6. PCR optimization and confirmation of the *SDHA* gene in the PkV RF01 genome. 13
Fig. S7. PCR and RT-PCR optimization using an internal control gene (*mcp*). 14
Fig. S8. Comparative distribution of glycosyltransferase domains among viruses. 15
Tables 16
Table S1. Type and position of 40 predicted tRNA genes in the PkV RF01 genome 16
Table S2. Genes related to lipid metabolism 17
Table S3. Details of tree reconstruction using PhyloBayes 17
Table S4. Forward and reverse PCR primers for amplification of *vSDHA* and *MCP* 17
References 18

Materials and Methods

Culturing and infection

All algal host cultures were grown in liquid IMR/2 medium consisting of 70% aged seawater, 30% distilled water (25 PSU), and additional selenite (10 nM final concentration). The cultures were kept at 14°C and partially synchronized using a 14:10 h light: dark cycle with irradiance of 100 µmol photons m^{-2} s^{-2} supplied by white fluorescent tubes. Viruses were produced by adding freshly produced viral lysate (ca. 2 x 10^8 VLP/mL), propagated three time on the host before added to exponentially growing host cultures (ca. 5x10^5 cells/mL) in a ratio of 1:10 volume. Infection was followed by flow cytometry (FCM) (1, 2) for 72 h by counting viral particles and host cells, as described in (3). Burst size was calculated as the number of viral particles released from each host cell, estimated from the total number of host cells pre-infection and the total number of VLPs produced during the infection cycle (3).

Infectious progeny

The percentage of viral infectious progeny was determined by comparing the most probable number (MPN; endpoint dilution (2)) and flow cytometric total counts of viral particles produced during infection. The number of infectious particles released in a burst was determined based on the percentage of viral infectivity produced during the infection cycle and the burst size. Infectivity was tested using *Haptolina ericin*a UiO028 as a host, and also compared with two other prymnesioviruses, HeV RF02 and PkV RF02 (3), propagated on He UiO028 and *Prymnesium kappa* RCC3423, respectively.

Briefly, 10× dilution were prepared from fresh viral lysate and added to exponentially growing host cells in 96-well microtiter plates (eight replicates for each dilution). The plates were incubated for 7 days under normal incubation conditions. Cell lysis was measured by monitoring *in situ* fluorescence on a plate reader (PerkinElmer EnSpire™ 2300 Multilabel Reader) at 460/680 nm. Numbers of infectious particles were estimated from the proportion of lysed wells using the MPN_ver4.xls excel spreadsheet from (4).

Sensitivity to chloroform

The effect of chloroform on infectivity, used to infer the presence of a lipid membrane or lipid molecules in the capsid, was tested by adding 50% (v/v) chloroform to PkV RF01 lysate. After mixing, the chloroform phase was separated from the solution by centrifugation at 4,000 *g* for 5 min. The tubes were incubated at 37°C for 2 h with the lids open to allow evaporation of any remaining chloroform.

Triplicates of exponentially growing He UiO028 cells (1,6 x 10^5 cells /mL) were incubated with 1:10 volumes of chloroform-treated viruses (ca. 2 x 10^8 VLP/mL). The incubation was followed for 7 days by counting host cells by FCM (2). Host cells in chloroform-treated or untreated medium at the same ratio used with the viral lysate were used as controls. Virus propagation was confirmed in lysed cultures by FCM.

Cryo-electron tomography

A small drop of concentrated PkV RF01 (8x109) was deposited on a glow-discharged, 200-mesh copper grid with holey carbon film (R2/1 Cu 200, Quantifoil Micro Tools GmbH, Germany). The sample was blotted with filter paper and immediately plunge frozen in liquid ethane. Grids were transferred under liquid nitrogen to a cryo-transfer tomography holder (Fishione Instruments, USA) and inserted in a 200-kV transmission electron microscope (Thermo Scientific Talos F200C) equipped with a Ceta 16M camera. Tilt series were recorded

at 45,000× magnification and –7 µm defocus between –60° to 60° in 2° increments. Finally, reconstruction, segmentation, and visualization of the tomograms was performed with IMOD v4.9 software (5).

Purification of viral particles and DNA isolation

Exponentially growing He UiO028 cultures (2 L) were infected with 20 mL of PkV RF01 and inspected visually for lysis. An uninfected culture (100 mL) was used as a control. Lysed algal cultures were checked for viruses by FCM counting. Lysed cultures were first centrifuged to remove algal debris and some bacteria (5,500 rpm for 15 min). Viruses were then pelleted by ultracentrifugation at 25,000 rpm in a Beckman Coulter Optima L90K ultracentrifuge for 2 h. The pellets were resuspended in SM buffer (0.1 M NaCl, 8 mM $MgSO_4 \cdot 7H_20$, 50 mM Tris-HCl, and 0.005% glycerin). Viral particles were further purified by Optiprep gradient centrifugation (6). Fractions were checked for viruses by FCM and for infectivity by infection of He UiO028.

Isolation of high-quality DNA for sequencing was done by following the protocol of (7) with some modifications. Viral particles were disrupted by one round of heating to 90°C for 2 min and then chilling on ice for 2 min. Disodium ethylenediaminetetraacetic acid and proteinase K at a final concentration of 20 mM and 100 µg mL^{-1}, respectively, were then added before incubation of the samples for 10 min at 55°C. Sodium dodecyl sulfate at a final concentration of 0.5% (w/v) was subsequently added, and samples were incubated for an additional 1 h at 55°C. Double-stranded DNA was then purified from the lysates using a Zymo Genomic DNA Clean & Concentrator Kit-10 (Zymo Research, Irvine, CA, USA) according to the manufacturer's protocols. To avoid shearing DNA, gentle pipetting and mixing (accomplished by turning the tubes instead of vortexing) were performed in all steps.

Genome assembly

Isolated DNA from PkV RF01 was subjected to Illumina TruSeq PCR-free library preparation (insert size 350 bp). The generated library was sequenced on an Illumina MiSeq instrument in paired-end mode (2 × 300 bp) to yield approximately 1.9 million reads, which corresponds to about 400× coverage. Reads were assembled into 2,498 contigs of 500 bp or more with a total assembly size of 4.75 Mb using Newbler (8). In addition, a ligation-based $1D^2$ nanopore library (LSK-308) was constructed and sequenced using an Oxford Nanopore MinION Mk1b device and a FLO-MIN107 flow cell, which resulted in 825 long reads with an N50 of 13.6 kb and a total of 9.89 Mb. To improve the assembly, short-read contigs were manually bridged with the long reads. Manual assembly using Consed (9) yielded a linear genome sequence of 1.4 Mb with inverted terminal repeats. After assembly, the consensus was polished using Nanopolish (10) and Pilon (11).

Phylogenetic analyses

Five core genes, SDHA, and SDHB

The phylogenetic position of PkV RF01 was inferred from concatenated protein alignments of five core nucleocytoplasmic virus orthologous genes (NCVOGs) (12): D5-like helicase-primase (NCVOG0023), DNA polymerase elongation subunit family B (NCVOG0038), DNA or RNA helicases of superfamily II (NCVOG0076), packaging ATPase (NCVOG0249), and Poxvirus Late Transcription Factor VLTF3-like (NCVOG0262). Sequences were obtained from the NCVOG database (ftp.ncbi.nlm.nih.gov/pub/wolf/COGs/NCVOG/) (13). Additional sequences were obtained from genomes retrieved from GenBank and annotated with HMMER v3.12b using the hmmsearch (14) command with hidden Markov models available in Schults et al. (2017) (15). Sequences from each NCVOG were aligned independently using

MAFFT L-INS-i (16). The alignments were trimmed with trimAl v1.2 in *gapyout* mode (17) prior to concatenation using a custom Python script. Bayesian phylogenetic trees were inferred with PhyloBayes 1.7 (18) using the CAT model and a GTR substitution matrix. Four chains were run for 34,500–35,500 generations. The *bpcomp* command was used to check for convergence. One chain was discarded, and a consensus tree was constructed using the remaining three chains.

For phylogenetic analyses of succinate dehydrogenase subunits, top hits of PkV RF01 SDHA and SDHB were retrieved from UniProt (https://www.uniprot.org/) using online PHMMR searches (https://www.ebi.ac.uk/Tools/hmmer/search/phmmer) and also from the *Tara* Oceans project using online BLASTP searches (http://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/) (Villar et al., 2018). Alignments generated with MAFFT L-INS-i were filtered with trimAl in gapyout mode. Maximum-likelihood phylogenies were inferred with RAxML 8.2.9 (19) using the PROTCATALG model and automatic bootstrapping with the following options: '-N autoMRE -f a -n autoresult'. Phylogenetic trees of PkV RF01, SDHA, and SDHB were visualized using iTOL (20).

Rpb2, IleRS, and AsnRS

To reconstruct a phylogenetic tree based on the second largest RNA polymerase subunit, homologs were recruited by comparing Mimivirus Rpb2 against all proteins of viruses and selected organisms in the KEGG database using the GenomeNet BLASTP tool (https://www.genome.jp/). Organisms were manually selected from the KEGG list to ensure broad taxonomic coverage of the tree of life. The retrieved amino acid sequences were aligned using MAFFT-LINSI (16) and then trimmed using trimAl (17) with the following parameters: '-resoverlap 0.5 -seqoverlap 70 -gt 0.8 -st 0.001 -cons 50'. The tree was reconstructed using FastTree (21) as implemented in the GenomeNet TREE tool (https://www.genome.jp/tools-bin/ete). Isoleucine tRNA synthase and aspartyl tRNA synthetase viral and cellular homologs were retrieved and aligned in the same way. Trees were searched using PhyloBayes MPI (22) with the non-homogeneous CAT+GTR model (23). Details of the PhyloBayes runs for each tree are provided in Table S3.

Gene prediction and functional and taxonomic annotation

GeneMarkS with the option 'virus' (24) predicted 1,121 open reading frames (ORFs) in the fully assembled genome sequence of PkV RF01, while tRNAscan-SE (25) predicted 41 tRNAs. PkV RF01 CDS amino acid sequences were searched against Virus-Host DB (26), RefSeq (27), UniRef90 (28), and COG (29) databases using BLASTP with an *E*-value of 1×10^{-5} as the significant similarity threshold and against the Conserved Domain Database (30) using RPS-BLAST with an *E*-value threshold of 1×10^{-2}. The 10 best hits for each database were compiled in a single file and manually inspected to transfer annotations of subject sequences to our query. In ambiguous cases, such as distant homologs (often seen in viral genomes) or unclear or contradictory annotations of subject sequences, the query was searched against KEGG genes (31) to allow extensive manual checking using GenomeNet tools (https://www.genome.jp/; alignment quality, length comparison to canonical genes, and links with KEGG orthology). We automatically annotated glycosyltransferases (GTs) and other carbohydrate-active enzymes (glycoside hydrolases, GHs; polysaccharide lyases, PLs; carbohydrate esterases, CEs; and auxiliary activities, AAs) in PkV RF01 and all viral genomes in Virus-Host DB (as of June 2018) using the *hmm* option of the dbCAN2 pipeline and its profile database (32). We retained hits with *E*-values $< 1 \times 10^{-5}$ and domain coverage > 35%, which corresponded to default settings. Sequence-based identification of paralogous genes in the PkV RF01 proteome was conducted using OrthoFinder (33) with default

parameters. Transfer of functional information onto the largest group of paralogs, which contained six proteins (with all other groups containing only two), was conducted as follows: sequences were aligned using MAFFT-LINSI and searched against the "uniclust30" database using the HHPred toolkit (34) at https://toolkit.tuebingen.mpg.de. We considered hits with probabilities > 99%.

Taxonomic and functional analysis of vSDHA homologs in OM-RGCv1

We searched PkV RF01 SDHA and SDHB against OM-RGCv1 (35) using the Ocean Gene Atlas (36) BLAST-based tool and kept the top 50 hits with significant *E*-values for further analysis. We then collected genome fragments (contigs) encoding these 50 SDHAs and 50 SDHBs by searching via BLASTN for identical hits over full *SDHA* or *SDHB* lengths against *Tara* ocean assemblies (downloaded from EBI) used to construct OM-RGCv1 (35). We predicted ORFs in these genome fragments using GeneMarkS. The resulting 1,113 amino acid sequences were functionally annotated by searching against Pfam protein families (37) using profile HMM scan (38) and also taxonomically using a last common ancestor strategy as in (39); in brief, protein sequences were searched against a database composed of UniRef cells, MMETSP (40) and Virus-Host DB (26) data using DIAMOND (41). Selected hits were then used to derive the last common ancestor of the query using a NCBI taxonomic tree re-wired to reflect the taxonomy of NCLDVs.

PCR and RT-PCR optimization

We designed specific primers (Table S4) targeting a 256-bp region of the *mcp* gene to use both as an internal control in the RT-PCR and to confirm that our protocols were optimized. For each PCR, a negative control (sterile distilled H_2O) was included. PCR amplifications were carried out in 50-µL total volumes containing 1 µL of template using a DNA HotStarTaq Master Mix kit (Qiagen). The cycling protocol was as follows: 15 min at 95°C, followed by 35 cycles of 30 s at 94°C, 30 s at 59°C, and 30 s at 72°C, with a final extension of 12 min at 72°C.

RT-PCRs were performed using the SuperScript III One-Step RT-PCR with Platinum *Taq* DNA Polymerase system (Thermo Fisher). Cycling conditions were as follows: 16 min at 55°C and 2 min at 94°C, followed by 40 cycles of 15 s at 94°C, 30 s at 49°C, and 30 s at 68°C, and a final extension of 5 min at 68°C.

All PCR products were checked for the correct size on a 1.5% agarose gel stained with GelRed (Biotium). PCR products were further checked by sequencing using BigDye v3.1 (Thermo Fisher) for cycle sequencing (Sekvenseringslaboratoriet, UiB, Norway).

PCR amplification and RT-PCR analysis of *vSDHA*

To investigate whether the *vSDHA* gene is transcribed during infection, an infected culture of He_UiO028 plus PkV RF01 as well as an uninfected He UiO028 culture (control) were set up as described above. Samples were collected at 24, 72, and 96 h post infection from both cultures. RNA was extracted using an RNeasy Plus Universal Mini kit (Qiagen), with gDNA removed in an extra step using a TURBO DNA-free kit (Ambion).

Specific primers were designed to target a 150-bp region of the *vSDHA* gene (Table S4). For each PCR, two negative controls (sterile distilled H_2O and extracted DNA from He028) were included. As positive controls for the transcription, we used primers targeting the *mcp* gene (see above). As a positive PCR control, we used genomic PkV RF01 DNA. PCR amplifications were conducted in 50-µL total volumes containing 1 µL of template DNA

using an ExTaq kit (Takara). The cycling protocol was as follows: 5 min at 94°C, followed by 35 cycles of 30 s at 94°C, 30 s at 59°C, and 30 s at 72°C, with a final extension of 12 min extension at 72°C.

RT-PCRs were performed using a SuperScript III One-Step RT-PCR with Platinum Taq DNA Polymerase system (Thermo Fisher). Cycling conditions were as follows: 16 min at 55°C and 2 min at 94°C, followed by 40 cycles of 15 s at 94°C, 30 s at 49°C, and 30 s at 68°C, with a final extension of 5 min at 68°C. PCR products were checked as described above.

Results

DNA repair enzymes

NCLDV members are known to encode several genes corresponding to major DNA repair pathways, with some members encompassing full, or nearly full, pathway representation, such as the base excision repair (BER) pathway of Mimivirus (42). While some of these genes were acquired relatively recently from cellular organisms, others are connected with the early evolutionary history of these viruses (43). PkV RF01 also encodes a set of enzymes needed to facilitate the BER pathway, namely, two DNA glycosylases (ORF 196 and ORF 871), an apurinic-apyrimidinic (AP) endonuclease (ORF 935), a family-X DNA polymerase (ORF 630), and a NAD-dependent DNA ligase (ORF 741). These enzymes are conserved in all *Mimiviridae* members except for those with relatively small genomes (PgV, CeV, and AaV). In PgV and CeV, but not AaV and PkV RF01, the family-X DNA polymerase and DNA ligase are fused. The presence of these BER enzymes in PkV RF01 suggests that this viral BER pathway was already present in the last common ancestor of *Mimiviridae*. As in all other *Mimiviridae*, PkV RF01 encodes MutS7, a key enzyme of the mismatch repair pathway, and the MutS8 homolog—thus far only observed in PgV and CeV. Enzymes involved in other DNA repair pathways, such as XPG/Rad2 endonuclease (present in *Mimivirus* and CroV) and the fused Mre11-Rad50 DNA break repair protein (present in Mimivirus members only), were not found in PkV RF01.

Transcription

Similar to other *Mimiviridae*, PkV RF01 encodes several transcription-related genes, including those coding for several subunits of eukaryotic DNA-dependent RNA polymerase type II (RPB1, RPB2 [×2], RPB3, RPB5, RPB6, RPB7, RPB9, and RPB10, the latter not present in other *Mimiviridae* besides BsV and HKV1); transcription initiation (ORF306), elongation (ORF856), late (ORF436), and termination (ORF275) factors; a TATA-box-like binding protein (ORF709); and a mRNA capping enzyme (ORF42)—but not the polyA polymerase seen in other *Mimiviridae*. PkV RF01 also encodes a cold-shock protein (ORF 1082) shared with CeV and PgV that may prevent the formation of secondary structures in mRNA at low temperature and thus facilitate the initiation of translation (44).

Translation

A striking feature of *Mimiviridae* is the high prevalence of genes coding for translation-associated proteins. In particular, amino-acyl tRNA synthetases (aaRSs) are found in all heterotroph-infecting *Mimiviridae*—ranging from 1 in CroV to 20 in Tupanvirus—but are missing from alga-infecting *Mimiviridae*, although CeV encodes a catalytic domain for Asn/Asp-RNAt synthetase. In addition to encoding the two aaRSs reported in the main text, we found that PkV RF01 encodes a catalytic domain for Asn/Asp-RNAt (ORF762) also found in Prochlorococcus phages P-SSM2, 5, and 7. PkV RF01 encodes other translation-related proteins also present in other *Mimiviridae*, namely, the translation initiation factor 4F cap

binding subunit (eIF-4E; ORF 512), the helicase subunit (eIF-4A; ORF 229), and the peptide chain release factor 1 (eRF1; ORF 651). The latter protein is not orthologous to chain release factors in other *Mimiviridae*, as it exhibits much higher sequence conservation with cellular or even *Marseilleviridae* homologs. PkV RF01 shares with PgV the translation elongation factor eEF3 (ORF 489), whose sequence in both viral species is highly similar to haptophyte homologs. In addition, PkV RF01 encodes the translation elongation factor EF-1alpha (ORF946) that is absent in other *Mimiviridae* members but present in AaV. EF-1alpha of PkV RF01 was most likely acquired from haptophytes, as it has a high amino acid identity (57%) to that of *E. huxlei*, whereas the AaV homolog is most closely related (62% aa identity) to that of diatoms. PkV RF01 also encodes the translation initiation factor 1A (EIF1A) (ORF 62) with no detectable homolog in viruses (*E*-value = 1×10^{-5} in BLASTP searches against the Virus-Host DB database). While translation-associated genes may predate the divergence of known extant eukaryotic lineages, the above examples demonstrate that some other genes were acquired much more recently, consistent with the "accordion model" of episodic gain and loss of genes in the evolution of these viruses (15, 45).

Other carbohydrate-active enzymes

Other carbohydrate-active enzymes in the PkV RF01 genome include seven glycoside hydrolases (GHs), four carbohydrate esterases (CEs), one polysaccharide lyase (PL), one carbohydrate-binding module (CBM), and a putative sugar fermentation stimulation protein A (ORF 1003) possibly involved in maltose metabolism. These numbers are not excessively high compared with other viruses. Other detected ORFs were homologous to enzymes involved in carbohydrate transport and metabolism, notably a transketolase (ORF 528) involved in the pentose phosphate pathway in all organisms and in the Calvin cycle of photosynthetic organisms. Finally, we detected a 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 2 (ORF 539) and a mannose-1-phosphate guanylyltransferase/mannose-6-phosphate isomerase (ORF 836) respectively involved in fructose and mannose metabolism.

Figures

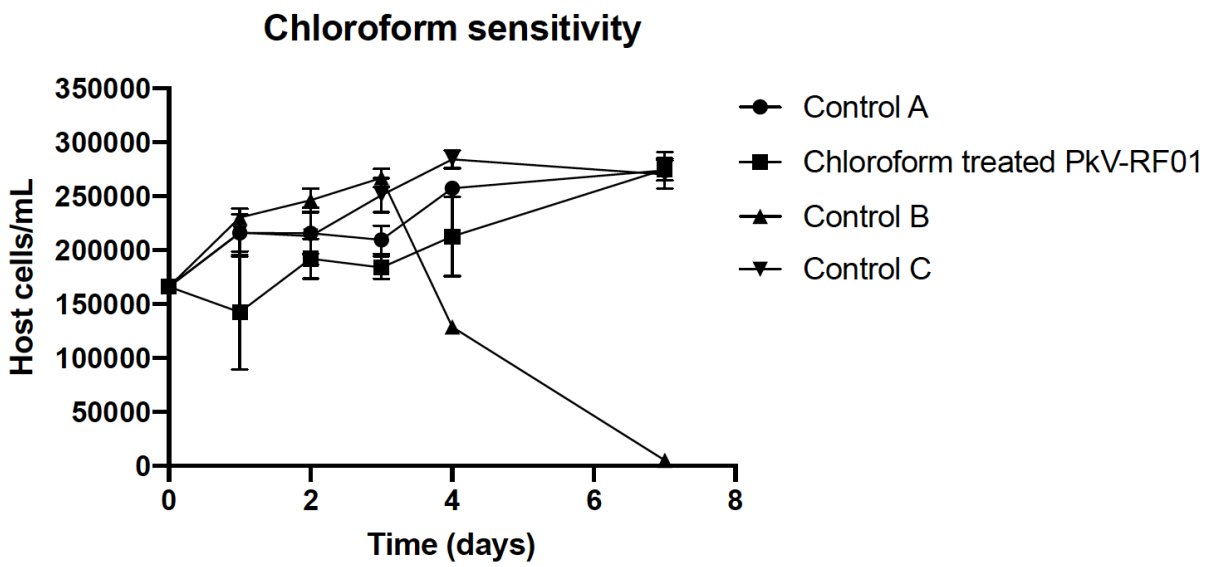

Fig. S1. Reduction of PkV RF01 infectivity with chloroform.
Experiments were set up in triplicate, and host cells were counted by flow cytometry. Chloroform-treated PkV RF01 was added to exponentially growing He UiO028 cells in a 1:10 volume ratio. Controls were He UiO028 cells incubated with chloroform-treated medium (Control A), untreated PkV RF01 (Control B), and untreated medium (Control C). SDs are indicated with error bars.

5YEL_A Transcriptional repressor CTCF/DNA Complex; zinc fingers, insulators, enhancers, promoters; HET: ZN; 2.96A {Homo sapiens}; Related PDB entries: 5YEL_B

Probability: 99.76 E-value: 2.4E-21 Score: 171.17 Aligned Cols: 166 Identities: 15% Similarity: 0.282

```
Q ss_pred              CceeCCCCCCccCCHHHHHHHHHHccccC--CCCC--CCCccCChHHHHHHHHHHHHHhhhcCCccCC--
Q gene_22        1     MVFECLICDYKTEIKFNYQKHLKTNKHLL--NFKK--CGKKYENVEKSINDVEKSINLFENNTNKISN--  64 (425)
Q Consensus      1     kpy~C~~C~k~F~s~~~L~~H~~~hh~~~~~~~~~~C~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~--  64 (425)
                       |||.|.+|++.|.+...|..|++.+++..  ...+  |...+.......           .+...+..
T Consensus      1     k~y~C~~C~~~f~~~~~l~~H~~~~~~~~~~~~C~~C~~~f~~~~~l~~~~~~~~~~~~H~~~~~~~~  59 (176)
T 5YEL_A         1     KPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHCDTVIARKSDLG-----------VHLRKQHSYI  59 (176)
T ss_dssp              CCCBCSSSCCBCSSHHHHHHHHHHC----CCCCBCSSSSCBCSSHHHHH-----------HHHHHSSSCC
T ss_pred              CCccCcCcCCcccchHHHHHHHHHHHhhccCcCeeCCCCchhhcChHHHH-----------HHHHHhcCcc

Q ss_pred              CCCEeCCCCCCcCCCHHHHHHHHHHcCCCCcccccHHHHhhccCCCCCcceeCCCCCCccCCHHHHHHHH
Q gene_22        65    THLFSCQWCKRNFGRNDSLKRHLRSCDKYNSQKFSEIITNYQNNSTNDISHVCSICFKQCSNKYNLKRHY  134 (425)
Q Consensus      65    ~~~f~C~~C~k~F~s~~~L~~H~r~hh~~~~~~~~~~~~~~~~~~~~~~~~C~~C~k~F~s~~~L~~H~  134 (425)
                       ..+|.|+.|++.|.....|..|+..++....                  +. |+.|+..|.....|..|+
T Consensus      60    ~~~~~C~~C~~~f~~~~~l~~H~~~~~~~~~~-------------------~~-C~~C~~~~~~~~~l~~H~  110 (176)
T 5YEL_A         60    EQGKKCRYCDAVFHERYALIQHQKSHKNEKR-----------------FK-CDQCDYASRQERHMIMHK  110 (176)
T ss_dssp              SSCEECSSSSCEESSHHHHHHHHGGGTTSCC-----------------EE-CSSSSCEESSHHHHHHHH
T ss_pred              ccCCCCCcccchhhcCHHHHHHHHHhccCCCC-----------------ee-CCCCCCeecCHHHHHHHH

Q ss_pred              HHHcccccccccccccCCchhHHHhhhhccccccCccccCCCC----ceeCCCCCCccCCHHHHHHHHHHc
Q gene_22        135   NRCQEINEEKNVEKSMIKPFLKATIVSTVKHNNYTTEFKDH----EFTCKYCNNKYSHQRSLSRHYNSCA  200 (425)
Q Consensus      135   ~~~h~~~~~~~~~~~~C~~C~~~~~~~~~l~~H~~~h~~~k-----f~C~~Cgk~F~~~s~L~rH~r~~H  200 (425)
                       ..+++...      +.|+.|+..|.....|..|...+.+..    +|.|+.|++.|...++|.+|++.||
T Consensus      111   ~~~~~~~~~~~~~~~~C~~C~~~f~~~~~l~~H~~~~~~~~~~~~~~~C~~C~~~f~~~~~l~~H~~~h~  174 (176)
T 5YEL_A         111   RTHTGEKP------YACSHCDKTFRQKQLLDMHFKRYHDPNFVPAAFVCSKCGKTFTRRNTMARHADNCA  174 (176)
T ss_dssp              HHHCC-CC------CBCSSSSCBCSSSGGGGHHHHHHSCC--CCCCCCCSSSCCCCSSHHHHHHHHTCC
T ss_pred              HHhcCCCC------ccCcccchhhCCHHHHHHHHHHhCCCCCCCccccCCCcCCEeCCHHHHHHHHHcC

Q ss_pred              CC
Q gene_22        201   ER  202 (425)
Q Consensus      201   ~~  202 (425)
                       +.
T Consensus      175   ~~  176 (176)
T 5YEL_A         175   GP  176 (176)
T ss_dssp              --
T ss_pred              CC
```

Fig. S2. A putative CTCF DNA-binding protein in PkV RF01.

A HHpred alignment between a profile resulting from the alignment of six PkV RF01 paralogous sequences detected by OrthoFinder and a database profile for the transcriptional repressor CTFC/DNA complex are shown. Gene_22 (for PkV RF01) and 5YEL_A are representative sequences for these two profiles.

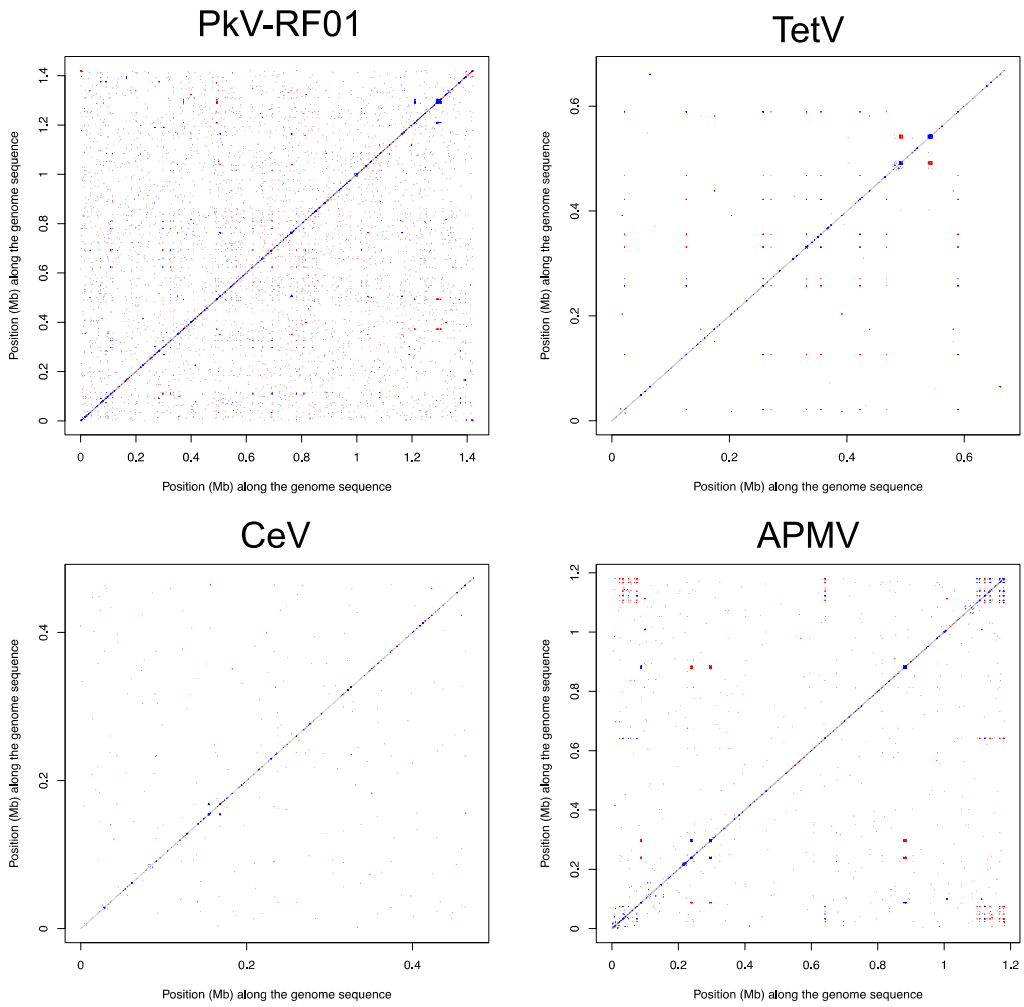

Fig. S3. High density of exact small repeats in the PkV RF01 genome.
Dot plots show exact direct (blue) or inverted (red) genomic repeats of at least 20 nucleotides in PkV RF01, another unclassified *Mimiviridae* (TetV), a member of "Mesomimivirinae" (CeV), and a "Megavirinae" member (APMV). Self-alignments were removed.

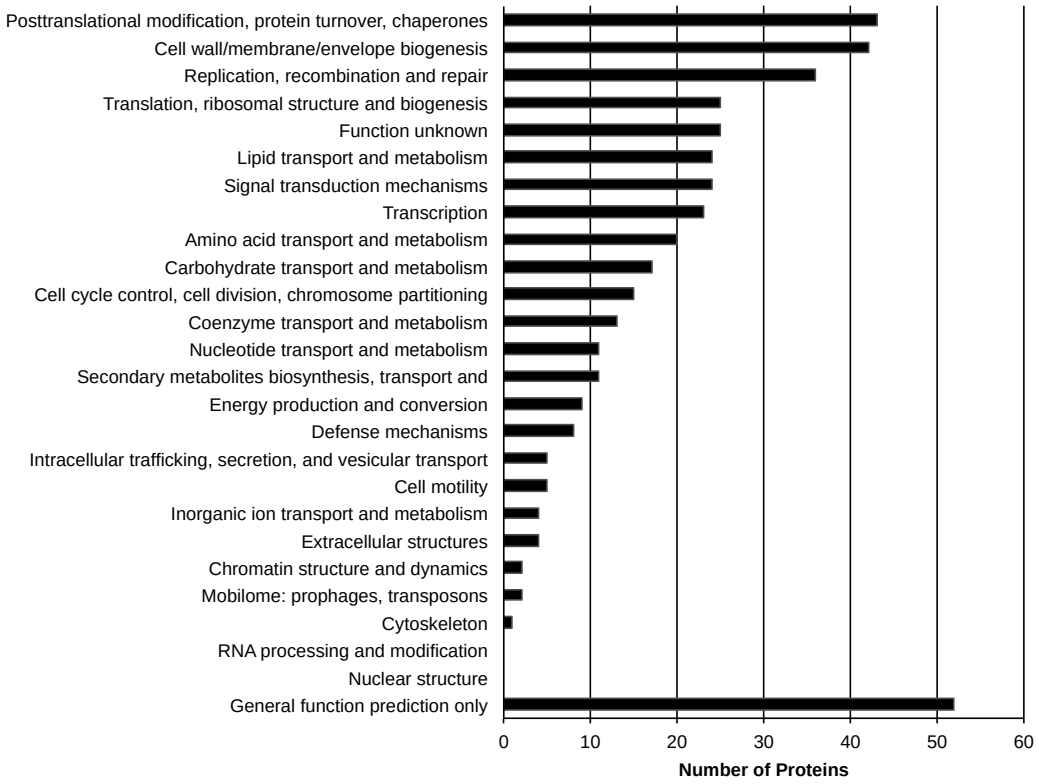

Fig. S4. COG functional distribution of 339 proteins encoded by PkV RF01.

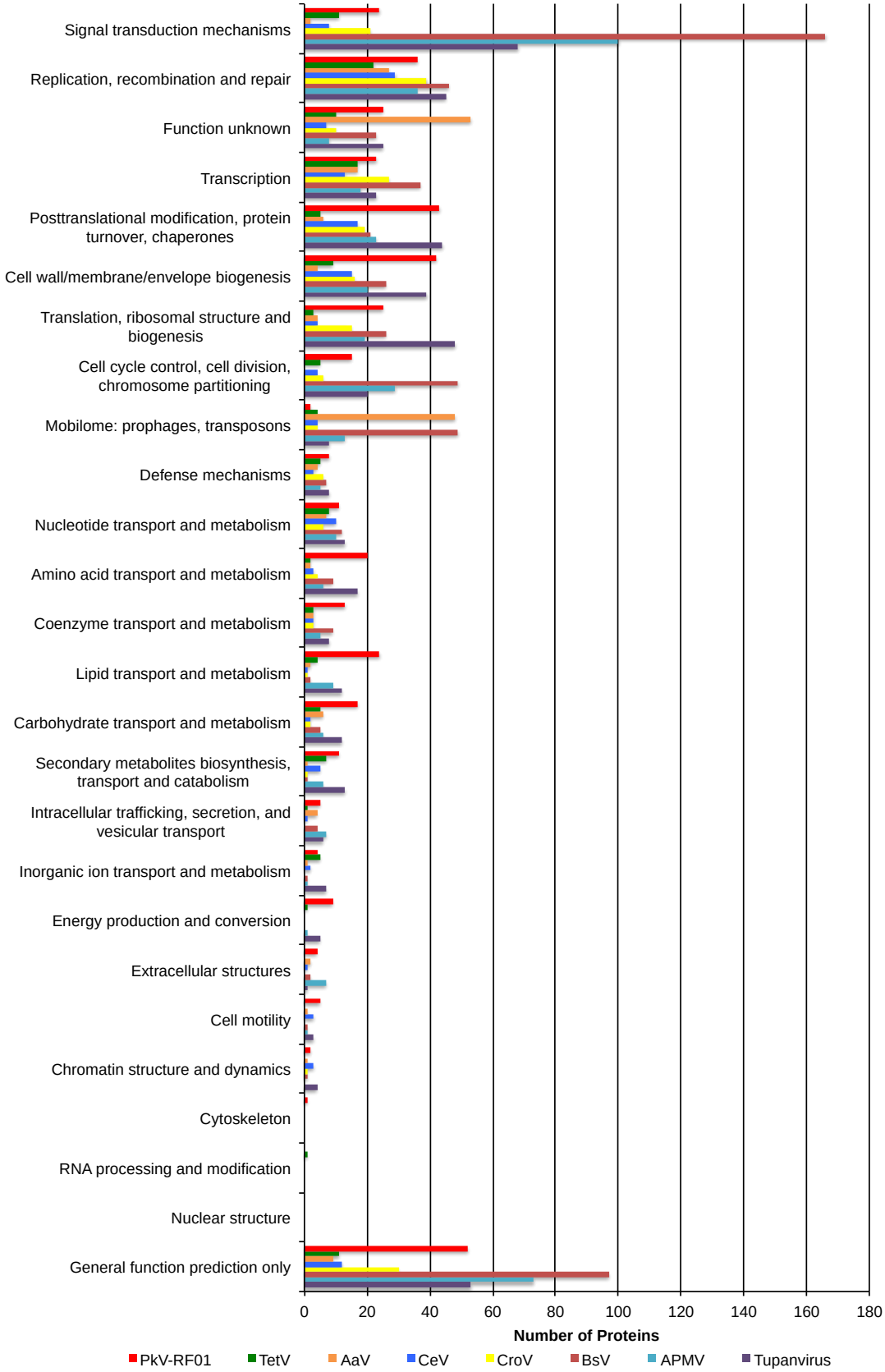

Fig. S5. Comparative COG functional distribution among *Mimiviridae* members. COG sequences were automatically searched against the proteomes of each virus using BLASTP with an *E*-value of 1×10^{-5} as the significant similarity threshold.

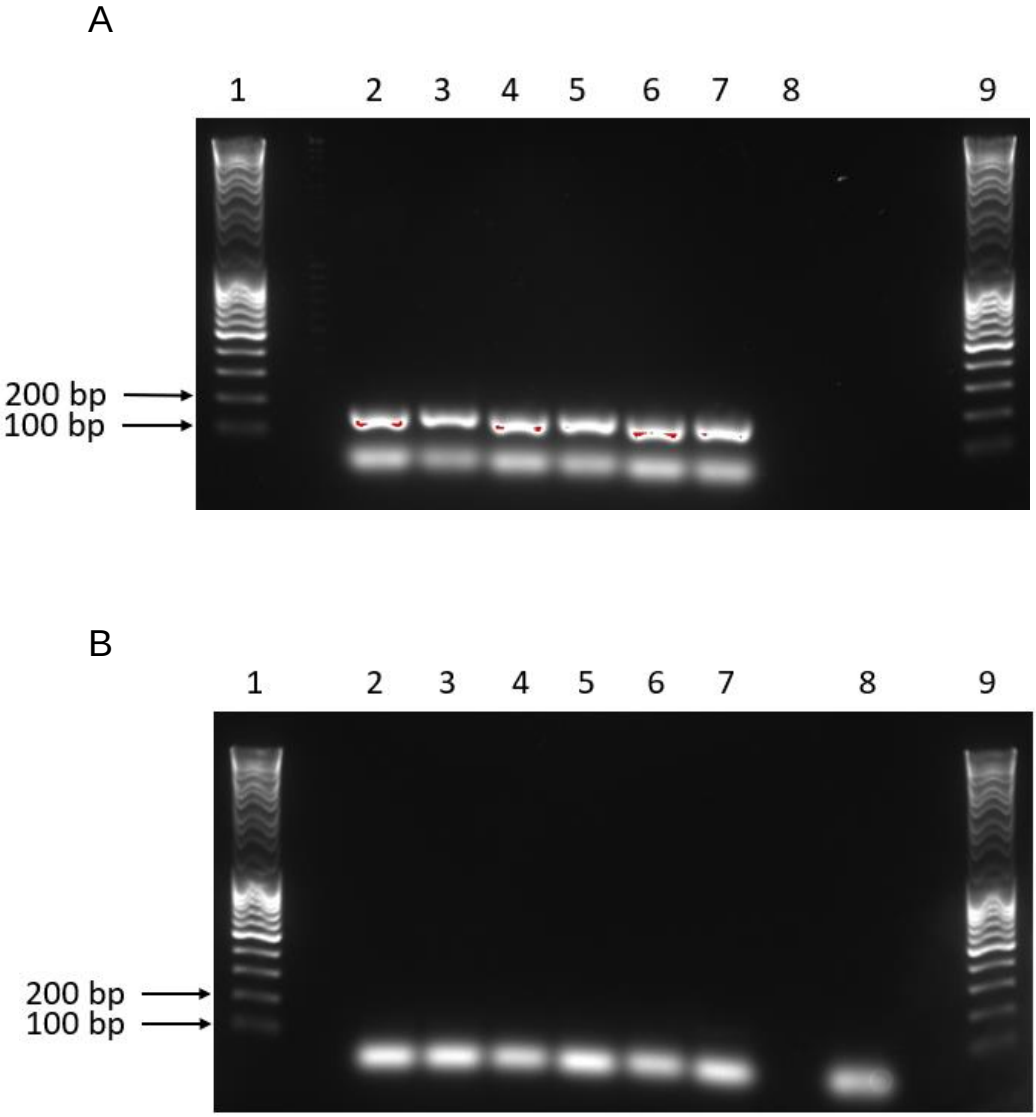

Fig. S6. PCR optimization and confirmation of the *SDHA* gene in the PkV RF01 genome.
(A–B) Results of PCR with *SDHA* primers using genomic PkV RF01 DNA (A) and genomic He UiO028 DNA (B) as templates. Lanes 1 and 9, DNA ladder; 2–7, optimization of the PCR annealing temperature from 55°C (2) to 60°C (7); 8, negative control (sdH_2O).

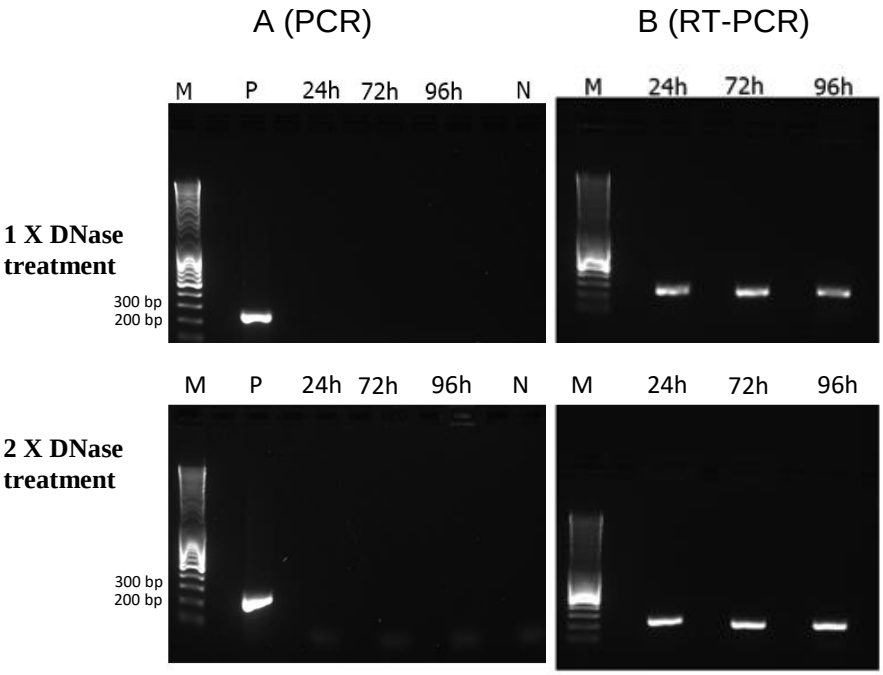

Fig. S7. PCR and RT-PCR optimization using an internal control gene (*mcp*). PCR and RT-PCR were carried out after removal of genomic DNA using a TURBO DNA-free kit. Samples were taken 24, 72, and 96 h after infection. Two different protocols, both provided in the TURBO DNA-free kit manual, were used to optimize the reactions. (A) PCR check for the presence of genomic DNA after RNA isolation and DNase treatments P, positive control (PkV RF01 genomic DNA); N, negative control (sd$H_2$0). (B) Result of RT-PCR of samples harvested 24, 72 and 96 h post infection. M, DNA marker (MassRuler DNA Ladder Mix, Thermo Fisher, 80 to 10,000 bp).

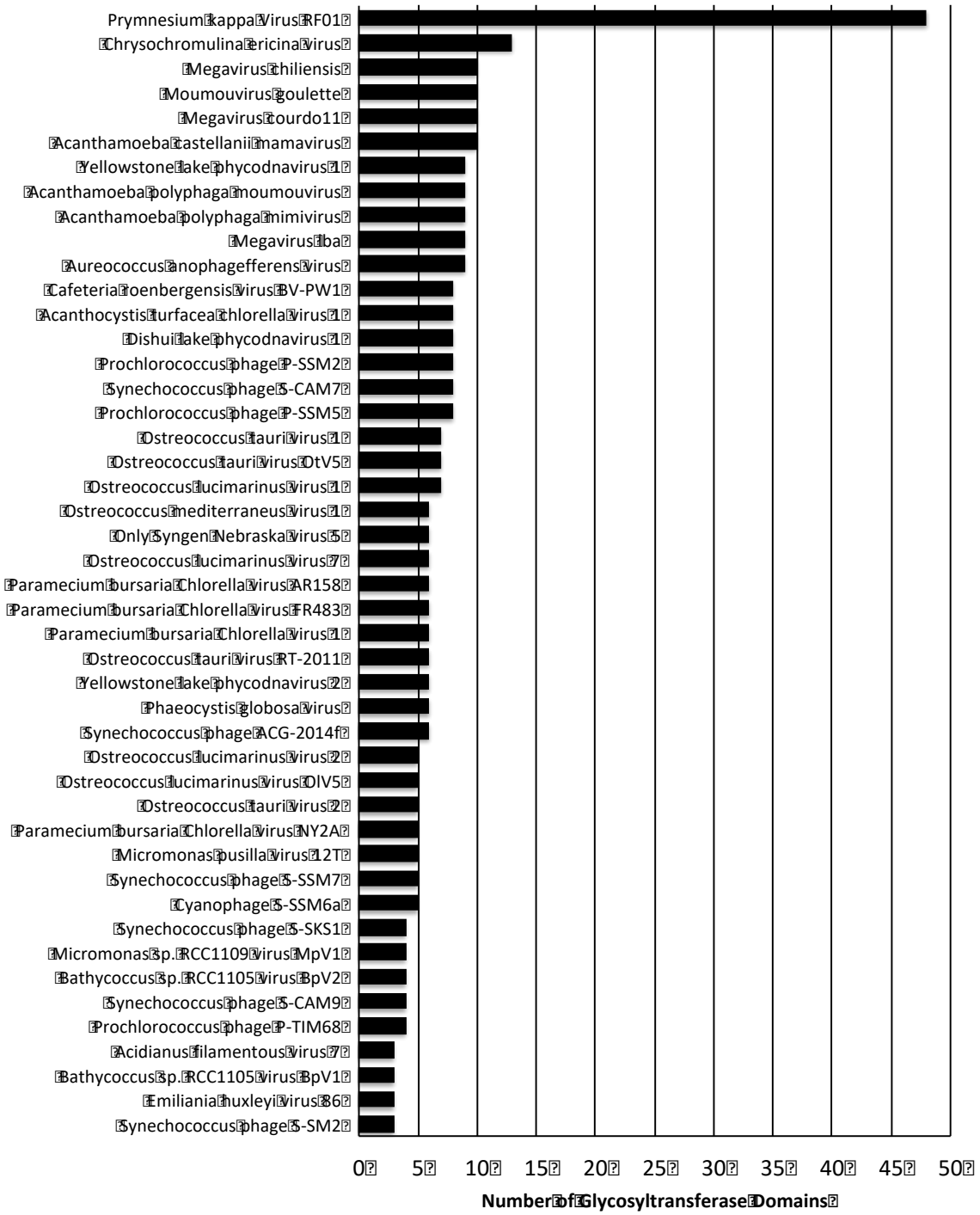

Fig. S8. Comparative distribution of glycosyltransferase domains among viruses.

Tables

Table S1. Type and position of 40 predicted tRNA genes in the PkV RF01 genome

Anti codon	Begin	End	Strand
CAT	3556	3483	-
NNN	90515	90393	-
AGT	165083	165153	+
TAT	165371	165444	+
TAA	165508	165590	+
TTT	165592	165666	+
TTT	165672	165744	+
CAT	165763	165835	+
GTT	165838	165911	+
TTG	166072	166144	+
AAT	166165	166238	+
TGT	686381	686452	+
GCT	787944	787863	-
TGA	930076	930157	+
TGC	941356	941284	-
TGG	956727	956656	-
TCC	956801	956731	-
AAT	1342428	1342564	+
CAA	1388099	1388180	+
CAT	1388185	1388257	+
GTT	1388263	1388336	+
TTG	1388494	1388565	+
CTT	1388664	1388736	+
CTT	1388946	1389018	+
TCT	1389231	1389303	+
TCT	1389516	1389588	+
TCT	1389801	1389873	+
CTT	1390089	1390163	+
TCT	1393421	1393493	+
TTC	1393523	1393594	+
CAA	1393697	1393778	+
TAT	1393782	1393855	+
TAA	1393963	1394044	+
TAT	1394048	1394121	+
TAA	1394229	1394310	+
TAT	1394314	1394387	+
TAA	1394495	1394576	+
TAT	1394580	1394653	+
TAA	1394761	1394841	+
CAT	1417619	1417692	+

The distribution of tRNA types is as follows: $tRNA_{Ile}$ and $tRNA_{Leu}$, seven each; $tRNA_{Lys}$, five; $tRNA_{Arg}$ and $tRNA_{Met}$, four each; $tRNA_{Asn}$, $tRNA_{Gln}$, $tRNA_{Ser}$, and $tRNA_{Thr}$, two each; and $tRNA_{Ala}$, $tRNA_{Glu}$, $tRNA_{Gly}$, $tRNA_{Pro}$, and undetermined, one each.

Table S2. Genes related to lipid metabolism

ORF	Length (aa)	Annotation	KEGG orthology	Pathway
ORF30	276	2,4-dienoyl-CoA reductase, mitochondrial [EC:1.3.1.34]	K13236	Beta oxidation
ORF33	888	Putative CoA-transferase	no significant hit	Beta oxidation
ORF121	225	glycerophosphoryl diester phosphodiesterase	K01126	Glycerophospholipids metabolisms
ORF138	2116	Fatty acid synthase (FASN)	K00665	Fatty acid biosynthesis
ORF142	523	Long-chain-fatty-acid--CoA ligase ACSBG [EC:6.2.1.3]	K15013	Fatty acid degradation /biosynthesis / Beta Oxidation
ORF175	1303	Acetyl-CoA carboxylase / biotin carboxylase 1 [EC:6.4.1.2 6.3.4.14 2.1.3.15]	K11262	Fatty acid biosynthesis
ORF236	410	Glutaryl-CoA dehydrogenase [EC:1.3.8.6]	K00252	Fatty acid degradation
ORF293	330	Lysophospholipase like	no significant hit	not assigned
ORF357	313	Lysophospholipase like	no significant hit	not assigned
ORF386	293	Triacylglycerol lipase [EC:3.1.1.3]	K01046	Glycerolipid metabolism
ORF481	503	Lipase like	no significant hit	not assigned
ORF635	282	Lipase-like	no significant hit	not assigned
ORF653	292	Lipase-like	no significant hit	not assigned
ORF690	260	Lipase-like	no significant hit	not assigned
ORF774	327	Lysophospholipid Acyltransferases [EC:2.3.1.22]	K14457	Glycerolipid metabolism
ORF694	336	Lipase esterase (Carbohydrate esterase CE10)	no significant hit	not assigned
ORF695	335	Lipase esterase (Carbohydrate esterase CE10)	no significant hit	not assigned
ORF886	516	Stearoyl-CoA desaturase (Delta-9 desaturase) [EC:1.14.19.1]	K00507	Biosynthesis of unsaturated fatty acids
ORF902	2083	Fatty acid synthase (FASN)	K00665	Fatty acid biosynthesis
ORF904	678	Long-chain-fatty-acid--CoA ligase ACSBG [EC:6.2.1.3]	K15013	Fatty acid degradation /biosynthesis / Beta Oxidation
ORF1016	422	Cyclopropane-fatty-acyl-phospholipid synthase [EC:2.1.1.79]	k00574	not assigned
ORF1046	652	Acyl-CoA dehydrogenase	K06445	Fatty acid degradation / Beta oxidation

*ORF138 and ORF902 share 32% aa identity
**ORF142 and ORF904 share 22% aa identity
***ORF694 and ORF695 share 92% aa identity

Table S3. Details of tree reconstruction using PhyloBayes

Protein	Burn-in	Sub-sampling	Chain	#Trees	Maxdiff	Meandiff
Five NCLDV-core proteins*	2000	2	1 3 4 2	16,680 16,725 16,483 16,430	0.32 Discarded	0.006 Discarded
Rpb2	Reconstructed using FastTree v2.1 default settings					
SDHA	Reconstructed using RaxML					
SDHB	Reconstructed using RaxML					
AsnRS	10,000	3	1	15,485	0.27	0.02
			2	15,066		
			3	15,627		
IleRS	10,000	3	1	10,757	0.16	0.009
			2	10,819		
			3	10,856	Discarded	Discarded

* D5-like helicase-primase (NCVOG0023), DNA polymerase elongation subunit family B (NCVOG0038), DNA or RNA helicases of superfamily II (NCVOG0076), packaging ATPase (NCVOG0249), and Poxvirus Late Transcription Factor VLTF3 like (NCVOG0262).

Table S4. Forward and reverse PCR primers for amplification of *vSDHA* and *MCP*

Primer name	Sequence (5´- 3´)	PCR product (bp)
vSDHA-F1	ATGTGCCGAGAAGCTCCTAA	154
vSDHA-R1	CTGCACAGGCTGTTCGATAA	
PkV-RF01-MCP-F	GATGAACCTTGCCCACAACT	256
PkV-RF01-MCP-F	GTGCATGGTACGTTTTCGTG	

References

1. Marie D, Brussaard CPD, Thyrhaug R, Bratbak G, Vaulot D. 1999. Enumeration of Marine Viruses in Culture and Natural Samples by Flow Cytometry. Appl Environ Microbiol 65:45–52.

2. Brussaard CPD. 2004. Optimization of procedures for counting viruses by flow cytometry. Appl Environ Microbiol 70:1506–1513.

3. Johannessen TV, Bratbak G, Larsen A, Ogata H, Egge ES, Edvardsen B, Eikrem W, Sandaa R-A. 2015. Characterisation of three novel giant viruses reveals huge diversity among viruses infecting Prymnesiales (Haptophyta). Virology 476:180–188.

4. Jarvis B, Wilrich C, Wilrich P-T. 2010. Reconsideration of the derivation of Most Probable Numbers, their standard deviations, confidence bounds and rarity values. J Appl Microbiol 109:1660–1667.

5. Kremer JR, Mastronarde DN, McIntosh JR. 1996. Computer visualization of three-dimensional image data using IMOD. J Struct Biol 116:71–76.

6. Lawrence JE, Steward GF. 2010. Purification of viruses by centrifugation, p. 166–181. *In* Wilhelm, S, Weinbauer, M, Suttle, C (eds.), Manual of Aquatic Viral Ecology. American Society of Limnology and Oceanography.

7. Sandaa R-A, E. Storesund J, Olesin E, Lund Paulsen M, Larsen A, Bratbak G, Ray JL. 2018. Seasonality Drives Microbial Community Structure, Shaping both Eukaryotic and Prokaryotic Host–Viral Relationships in an Arctic Marine Ecosystem. Viruses 10.

8. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin

BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.

9. Gordon D, Green P. 2013. Consed: a graphical editor for next-generation sequencing. Bioinformatics 29:2936–2937.

10. Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 12:733–735.

11. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE 9:e112963.

12. Yutin N, Wolf YI, Raoult D, Koonin EV. 2009. Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. Virol J 6:223.

13. Yutin N, Wolf YI, Koonin EV. 2014. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. Virology 466–467:38–52.

14. Eddy SR. 2011. Accelerated Profile HMM Searches. PLOS Comput Biol 7:e1002195.

15. Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, Daims H, Horn M, Wagner M, Jensen GJ, Kyrpides NC, Koonin EV, Woyke T. 2017. Giant viruses with an expanded complement of translation system components. Science 356:82–85.

16. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780.

17. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinforma Oxf Engl 25:1972–1973.

18. Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

19. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinforma Oxf Engl 30:1312–1313.

20. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W245.

21. Price MN, Dehal PS, Arkin AP. 2009. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Mol Biol Evol 26:1641–1650.

22. Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol 62:611–615.

23. Lartillot N, Philippe H. 2004. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Mol Biol Evol 21:1095–1109.

24. Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29:2607–2618.

25. Lowe TM, Chan PP. 2016. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res 44:W54-57.

26. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, Hingamp P, Goto S, Ogata H. 2016. Linking Virus Genomes with Host Taxonomy. Viruses 8:66.

27. Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:D61–D65.

28. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31:926–932.

29. Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36.

30. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222–D226.

31. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457–D462.

32. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. 2018. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 46:W95–W101.

33. Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157.

34. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol 430:2237–2243.

35. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Tara Oceans coordinators, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P. 2015. Ocean plankton. Structure and function of the global ocean microbiome. Science 348:1261359.

36. Villar E, Vannier T, Vernette C, Lescot M, Cuenca M, Alexandre A, Bachelerie P, Rosnet T, Pelletier E, Sunagawa S, Hingamp P. 2018. The Ocean Gene Atlas: exploring the biogeography of plankton genes online. Nucleic Acids Res 46:W289–W295.

37. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan

D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432.

38. Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755–763.

39. Carradec Q, Pelletier E, Silva CD, Alberti A, Seeleuthner Y, Blanc-Mathieu R, Lima-Mendez G, Rocha F, Tirichine L, Labadie K, Kirilovsky A, Bertrand A, Engelen S, Madoui M-A, Méheust R, Poulain J, Romac S, Richter DJ, Yoshikawa G, Dimier C, Kandels-Lewis S, Picheral M, Searson S, Jaillon O, Aury J-M, Karsenti E, Sullivan MB, Sunagawa S, Bork P, Not F, Hingamp P, Raes J, Guidi L, Ogata H, Vargas C de, Iudicone D, Bowler C, Wincker P. 2018. A global ocean atlas of eukaryotic genes. Nat Commun 9:373.

40. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, Beszteri B, Bidle KD, Cameron CT, Campbell L, Caron DA, Cattolico RA, Collier JL, Coyne K, Davy SK, Deschamps P, Dyhrman ST, Edvardsen B, Gates RD, Gobler CJ, Greenwood SJ, Guida SM, Jacobi JL, Jakobsen KS, James ER, Jenkins B, John U, Johnson MD, Juhl AR, Kamp A, Katz LA, Kiene R, Kudryavtsev A, Leander BS, Lin S, Lovejoy C, Lynn D, Marchetti A, McManus G, Nedelcu AM, Menden-Deuer S, Miceli C, Mock T, Montresor M, Moran MA, Murray S, Nadathur G, Nagai S, Ngam PB, Palenik B, Pawlowski J, Petroni G, Piganeau G, Posewitz MC, Rengefors K, Romano G, Rumpho ME, Rynearson T, Schilling KB, Schroeder DC, Simpson AGB, Slamovits CH, Smith DR, Smith GJ, Smith SR, Sosik HM, Stief P, Theriot E, Twary SN, Umale PE, Vaulot D, Wawrik B, Wheeler GL, Wilson WH, Xu Y, Zingone A, Worden AZ. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of

Eukaryotic Life in the Oceans through Transcriptome Sequencing. PLOS Biol 12:e1001889.

41. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60.

42. Redrejo-Rodríguez M, Salas ML. 2014. Repair of base damage and genome maintenance in the nucleo-cytoplasmic large DNA viruses. Virus Res 179:12–25.

43. Blanc-Mathieu R, Ogata H. 2016. DNA repair genes in the Megavirales pangenome. Curr Opin Microbiol 31:94–100.

44. Keto-Timonen R, Hietala N, Palonen E, Hakakorpi A, Lindström M, Korkeala H. 2016. Cold Shock Proteins: A Minireview with Special Emphasis on Csp-family of Enteropathogenic Yersinia. Front Microbiol 7.

45. Filée J. 2015. Genomic comparison of closely related Giant Viruses supports an accordion-like model of evolution. Front Microbiol 6.