Genomic, proteomic, and phylogenetic analysis of spounaviruses indicates paraphyly of the order *Caudovirales*

Short title: Taxonomy of SPO1-like phages

Jakub Barylski[1], François Enault[2], Bas E. Dutilh[3,4], Margo B.P. Schuller[3], Robert A. Edwards[5], Annika Gillis[6], Jochen Klumpp[7], Petar Knezevic[8], Mart Krupovic[9], Jens H. Kuhn[10], Rob Lavigne[11], Hanna M. Oksanen[12], Matthew B. Sullivan[13], Johannes Wittmann[14], Igor Tolstoy[15], J. Rodney Brister[15], Andrew M. Kropinski[16], Evelien M. Adriaenssens[17*]

This paper is dedicated to Hans-Wolfgang Ackermann, a pioneer of prokaryotic virus electron microscopy and taxonomy, who died on February 12th, 2017, at the age of 80. He was involved in the early stages of this study, and his input is dearly missed.

[1]Adam Mickiewicz University, Institute of Experimental Biology, Department of Molecular Virology, Poznań, Poland, [2]Université Clermont Auvergne, CNRS, LMGE, F-63000, Clermont-Ferrand, France, [3]Theoretical Biology and Bioinformatics, Department of Biology, Utrecht University, Utrecht, The Netherlands, [4]Theoretical Biology and Bioinformatics, Department of Biology, Utrecht University, Utrecht, The Netherlands, [5]Departments of Biology and Computer Science, San Diego State University, San Diego, CA, USA, [6]Laboratory of Food and Environmental Microbiology, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, [7]Institute of Food, Nutrition and Health, ETH Zurich, Switzerland, [8]Department of Biology and Ecology, Faculty of Sciences, University of Novi Sad, Novi Sad, Serbia, [9]Unit of Molecular

Biology of the Gene in Extremophiles, Department of Microbiology, Institut Pasteur, Paris, France, [10]Integrated Research Facility at Fort Detrick, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Fort Detrick, Frederick, USA, [11]Laboratory of Gene Technology, KU Leuven, Belgium, [12]Department of Biosciences, University of Helsinki, Helsinki, Finland; and Institute of Biotechnology, University of Helsinki, Helsinki, Finland, [13]Departments of Microbiology and Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, USA, [14]Leibniz-Institut DSMZ—Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Braunschweig, Germany, [15]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA, [16]Departments of Food Science, Molecular and Cellular Biology; and Pathobiology, University of Guelph, Guelph, Ontario, Canada, [17]Institute of Integrative Biology, University of Liverpool, Biosciences Building, Crown Street, Liverpool L69 7ZB, United Kingdom; orcid.org/0000-0003-4826-5406

[*]evelien.adriaenssens@gmail.com (EMA)

Abstract

Since the mid-20th century, prokaryotic double-stranded DNA viruses producing tailed particles ("tailed phages") were grouped according to virion tail morphology. In the early 1980s, these viruses were classified into the families *Myoviridae*, *Siphoviridae*, and *Podoviridae*, later included in the order *Caudovirales*. However, recent massive sequencing of prokaryotic virus genomes revealed that caudovirads are extremely diverse. The official taxonomic framework does not adequately reflect caudovirad evolutionary relationships. Here, we reevaluate the classification of caudovirads using a particularly challenging group of viruses with large dsDNA genomes: SPO1-like viruses associated with the myovirid subfamily *Spounavirinae*. Our extensive genomic, proteomic, and phylogenetic analyses reveal that some of the currently established caudovirad taxa, especially at the family and subfamily rank, can no longer be supported. Spounavirins alone need to be elevated to family rank and divided into at least five major clades, a first step in an impending massive reorganization of caudovirad taxonomy.

Introduction

Prokaryotic virus taxonomy is the formal responsibility of the Bacterial and Archaeal Viruses Subcommittee of the International Committee on Taxonomy of Viruses (ICTV). In recent years, the Subcommittee has focused on classifying newly described double-stranded DNA viruses producing tailed particles ("tailed phages") into species and genera included in existing families in the order *Caudovirales* [1–5]. At the species rank, a similarity threshold of 95% nucleotide sequence identity is used for classification, i.e. viruses are classified into the same species if they shared ≥95% identity over the entire length of their genomes. Examination of wild populations of caudovirads infecting cyanobacteria demonstrated that such a 95% threshold robustly captures formally delineable species using population genetic metrics [6]. Extrapolation of such thresholds to viruses of the global surface oceans has resulted in population-scale ecological understanding of thousands of new virus "species" [7], which, however, do not have official status. At the genus rank, cohesive groups have been defined for viruses sharing a significant genome similarity(≥60% nucleotide identity), gene synteny, and a core gene set. This framework has helped to relatively rapidly establish official low-rank taxonomic positions for newly isolated and sequenced viruses [1,2,5,8,9].

The currently available ranks in virus taxonomy (in ascending order *species*, *genus*, *subfamily*, *family*, and *order*) limit the description of the full diversity of prokaryotic viruses. This limitation is particularly acute in the case of the order *Caudovirales*, which represents the most abundant and diverse group of viruses in the environment [10–12]. Indeed, the diversity of caudovirads greatly surpasses that of other bacterial, archaeal and eukaryotic double-stranded (ds)DNA viruses. A recent analysis of the dsDNA virosphere using a bipartite network approach, whereby viral genomes are connected via shared gene families, demonstrated that the global network of dsDNA viruses consists of at least 19 modules, 11

of which correspond to caudovirads [13]. The eight remaining modules encompass one or more families of eukaryotic or archaeal viruses. Consequently, each of the caudovirad modules could be considered to represent a separate family. Despite this remarkable sequence diversity, all caudovirads are currently classified into three families—*Myoviridae, Podoviridae*, and *Siphoviridae*. These families were historically established on the morphological, not genomic, features of their members, forming an artificial classification ceiling. In this study, the Subcommittee explored the diversity of the order *Caudovirales* based on the example of a large group of Bacillus phage SPO1-related viruses (*Myoviridae*: *Spounavirinae*), which forms a discrete caudovirad module [13,14].

The type virus of this group, Bacillus phage SPO1, was isolated in 1964 from soil in Osaka, Japan, using *Bacillus subtilis* as a host [15]. Bacillus phage SPO1 has been extensively studied ever since. Morphologically, the particle possesses an icosahedral head 87 nm in diameter and a 140-nm long (18.6-nm wide) contractile tail of "stacked-disk" appearance that is terminated by a complex baseplate structure [16,17]. The packaged genome is a 145.7-kb long, terminally redundant (13.1-kb) DNA molecule with thymine completely replaced by 5-hydroxymethyluracil (HMU) [18–20]. This genome encodes at least 204 proteins and five tRNAs [21].

In 1995, Bacillus phage SPO1 was assigned to the species *Bacillus phage SP8* (renamed *Bacillus phage SPO1* in 1996 and *Bacillus virus SPO1* in 2015), which in 1996 was included into a monospecific myovirid genus currently called *Spo1virus* [1,22,23]. The subfamily "*Spounavirinae*" was proposed in 2009 by Lavigne et al. to harbor Bacillus phage SPO1, Staphylococcus phage Twort, Staphylococcus phage K, Staphylococcus phage G1, Listeria phage P100, and Listeria phage A511 [4]. This subfamily became official in 2012 [24]. The unifying characteristics of members of this subfamily as described by Klumpp et al. are

that “(a) the[ir] host organisms are bacteria of the phylum Firmicutes; (b) [they] are strictly virulent myovirids; (c) all… feature common morphological properties; (d) [their genomes] consist of a terminally redundant, non-permuted dsDNA molecule of 127–157 kb in size; and (e) [they] share considerable amino acid homology” [20]. The inclusion requirement of a strictly lytic lifestyle became controversial when it was observed that a few related viruses (Bacillus phages Bcp1, Bp8p-T, and Bp8p-C) can persist in host cultures without causing the immediate lysis [25,26]. It remains unknown whether this persistence is due to virion entrapment inside bacillus spores or some other kind of semi-stable virus-host relationship.

By 2015, the subfamily *Spounavirinae* had been expanded to include five genera (*Kayvirus*, *P100virus*, *Silviavirus*, *Spo1virus*, and *Twortvirus*) and three unassigned or “floating” species (*Enterococcus virus phiEC24C*, *Lactobacillus virus Lb338-1*, and *Lactobacillus virus LP65*). It was already clear that *Bacillus virus SPO1* represented one of two major lineages within the subfamily. Consequently, spounavirins were divided in the Bacillus phage SPO1-like viruses, which have modified DNA (HMU) and which produce particles possessing generally shorter tails; and, the Staphylococcus phage Twort-like viruses, which feature non-modified DNA and produce particles with longer tails. Both Bacillus phage SPO1-like and Staphylococcus phage Twort-like virus particles feature double-ringed baseplates and visible capsomeres as a morphologic hallmark [20,27]. A third group, represented by Bacillus phages sharing limited similarity with Bacillus phage SPO1 but related to phage Bastille had already been proposed at that time but not yet officially recognized [28]. In 2016, additional genera of Bacillus phage SPO1-like viruses (henceforth “spouna-like viruses”) were established, but not all of these were immediately included in the subfamily *Spounavirinae* [1,2].

In this study, we reevaluated the classification of spounavirins and spouna-like viruses. To this end, a well-defined set of 93 viruses was analyzed using complementary DNA and protein sequence analysis tools and phylogenetic methods. Our results indicate that the subfamily *Spounavirinae* fails to adequately reflect the diversity of its current members, and we therefore outline a better fitting classification scheme.

Results

General overview

To determine the phylogenetic relationship between 93 known and alleged spounavirins, we employed genomic, proteomic and marker gene-based comparative strategies. Regardless of the adopted phylogenetic approach applied, five separate, clear-cut clusters were identified. We believe that they clearly have a common origin and ought to come together under one caudoviral umbrella taxon. We propose to name this taxon "*Herelleviridae,*" in honor of the 100th anniversary of the discovery of prokaryotic viruses by Félix d'Hérelle (Table 1, Figs 1-3, S1 Table). The first cluster (here suggested to retain the name *Spounavirinae*) groups *Bacillus*-infecting viruses that are similar to Bacillus phage SPO1. The second cluster ("*Bastillevirinae,*" named after the type species *Bacillus virus Bastille* [28]) includes *Bacillus*-infecting viruses that have only limited similarity to Bacillus phage SPO1 and resemble Bacillus phage Bastille instead. The third cluster ("*Brockvirinae*," named in honor of Thomas D. Brock [1926–], an American microbiologist and educator known for his discovery of hyperthermophiles, who worked on Streptococcus phages early in his career) comprises currently unclassified viruses of enterococci that are similar to Enterococcus phage ϕEF24C. The fourth cluster ("*Twortvirinae*," named in honor of Frederick William Twort (1877–1950), the English bacteriologist who discovered prokaryotic

viruses in 1915) gathers staphylococci-infected viruses that are similar to Staphylococcus phage Twort, whereas the remaining cluster ("*Jasinskavirinae*," named in honor of Stanislawa Jasińska-Lewandowska (1921–1998), Polish scientist who was one of the first to study *Listeria* and their viruses) consists of viruses infecting *Listeria* that are similar to Listeria phage P100, the type isolate of the *P100virus* genus. The classification in five clusters left three viruses unassigned at this rank: Lactobacillus phage Lb338, Lactobacillus phage LP65, and Brochothrix phage A9.

These robust clusters can be further subdivided into smaller clades that correspond well with the currently accepted genera. The evidence supporting this suggested taxonomic re-classification is presented in the following sections.

158 **Table 1. Suggested new classification of the 93 spounavirins and spouna-like viruses in the new caudoviral family "*Herelleviridae*".**

Order	Family	Subfamily	Genus	Species[a]	Viruses
Caudovirales	*"Herelleviridae"*	**"***Bastillevirinae***"**	*Agatevirus*	*Bacillus virus Agate, Bacillus virus Bobb, Bacillus virus Bp8pC*	Bp8p-T
			B4virus	*Bacillus virus AvesoBmore, Bacillus virus B4, Bacillus virus Bigbertha, Bacillus virus Riley, Bacillus virus Spock, Bacillus virus Troll*	B5S
			Bastillevirus	*Bacillus virus Bastille, Bacillus virus CAM003, "Bacillus virus Evoli", "Bacillus virus HoodyT"*	
			Bc431virus	*Bacillus virus Bc431, Bacillus virus Bcp1, Bacillus virus BCP82, Bacillus virus JBP901*	
			Nit1virus	*Bacillus virus Grass, Bacillus virus NIT1, Bacillus virus SPG24*	
			Tsarbombavirus	*Bacillus virus BCP78, Bacillus virus*	BCU4

				TsarBomba	
			Wphvirus	*Bacillus virus BPS13*, *Bacillus virus Hakuna, Bacillus virus Megatron, Bacillus virus WPh, "Bacillus virus BPS10C*"	Eyuki
		"*Brockvirinae***"**	"*Kochikohdavirus*"	"*Enterococcus virus ECP3*", "*Enterococcus virus EF24C*", "*Enterococcus virus EFLK1*"	phiEFC24C-P2
			Unassigned	"*Enterococccus virus EFDG1*"	
		"*Jasinskavirinae***"**	*P100virus*	*Listeria virus A511*, *Listeria virus P100*	List-36, LMSP-25, AvB_LmoM_AG20, LP-125, LP-064, LP-083-2, LP-124, LP-125, LP-048, LMTA-34, LMTA-94, LMTA-148, LMTA-57, WIL-1
		Spounavirinae	*Cp51virus*	*Bacillus virus CP51, Bacillus virus JL, Bacillus virus Shanette*	
			Spo1virus	*Bacillus virus Camphawk*, *Bacillus*	

				virus SPO1	
			Unassigned	*"Bacillus virus Mater", "Bacillus virus Moonbeam", "Bacillus virus SIOphi"*	
		"Twortvirinae"	*Kayvirus*	*Staphylococcus virus G1, Staphylococcus virus G15, Staphylococcus virus JD7, Staphylococcus virus K, Staphylococcus virus MCE2014, Staphylococcus virus P108, Staphylococcus virus Rodi, Staphylococcus virus S253, Staphylococcus virus S25-4, Staphylococcus virus SA12, "Staphylococcus virus Sb1"*	676Z, A3R, A5W, Fi200W, IME-SA1, IME-SA118, IME-SA119, IME-SA2, ISP, MSA6, P4W, SA5, Staph1N, Team1
			Silviavirus	*Staphylococcus virus Remus, Staphylococcus virus SA11*	Romulus
			Sep1virus	*Staphylococcus virus IPLAC1C*,	

				Staphylococcus virus SEP1	
			Twortvirus	*Staphylococcus virus Twort*	
		Unassigned	Unassigned	*"Lactobacillus virus Lb338",* *"Lactobacillus virus LP65",* *"Brochothrix virus A9"*	

[a] The species listed here are representing the 93 genome dataset on which all analyses have been performed. Species ratified in 2017 and later have not been included here.

Genome-based analyses

BLASTn analysis revealed that the genomes of several viruses were similar enough to consider them strains of the same species (they shared >95% nucleotide identity, S1 Fig). The *Staphylococcus* viruses fell into four distinct, yet closely related groups corresponding to the established genera *Twortvirus, Sep1virus, Silviavirus*, and *Kayvirus* (S1 Fig). With the exception of Enterococcus phage EFDG1, all *Enterococcus* viruses clustered as a clade representing a new genus (here suggested to be named "*Kochikohdavirus*" after the place of origin of the type virus of the clade, Enterococcus phage ϕEF24C; [29,30]). The *Bacillus* viruses clustered into the established genera *Spo1virus*, *Cp51virus*, *Bastillevirus*, *Agatevirus*, *B4virus*, *Bc431virus*, *Nit1virus*, *Tsarbombavirus*, and *Wphvirus*, with three species remaining unassigned at the genus rank (Table 1). These results were also confirmed with VICTOR, a genome-BLAST distance phylogeny (GBDP) method (S2 Fig) and the Dice score (S3 Fig) [31], a tBLASTx-based measure that compares whole genome sequences at the amino acid level.

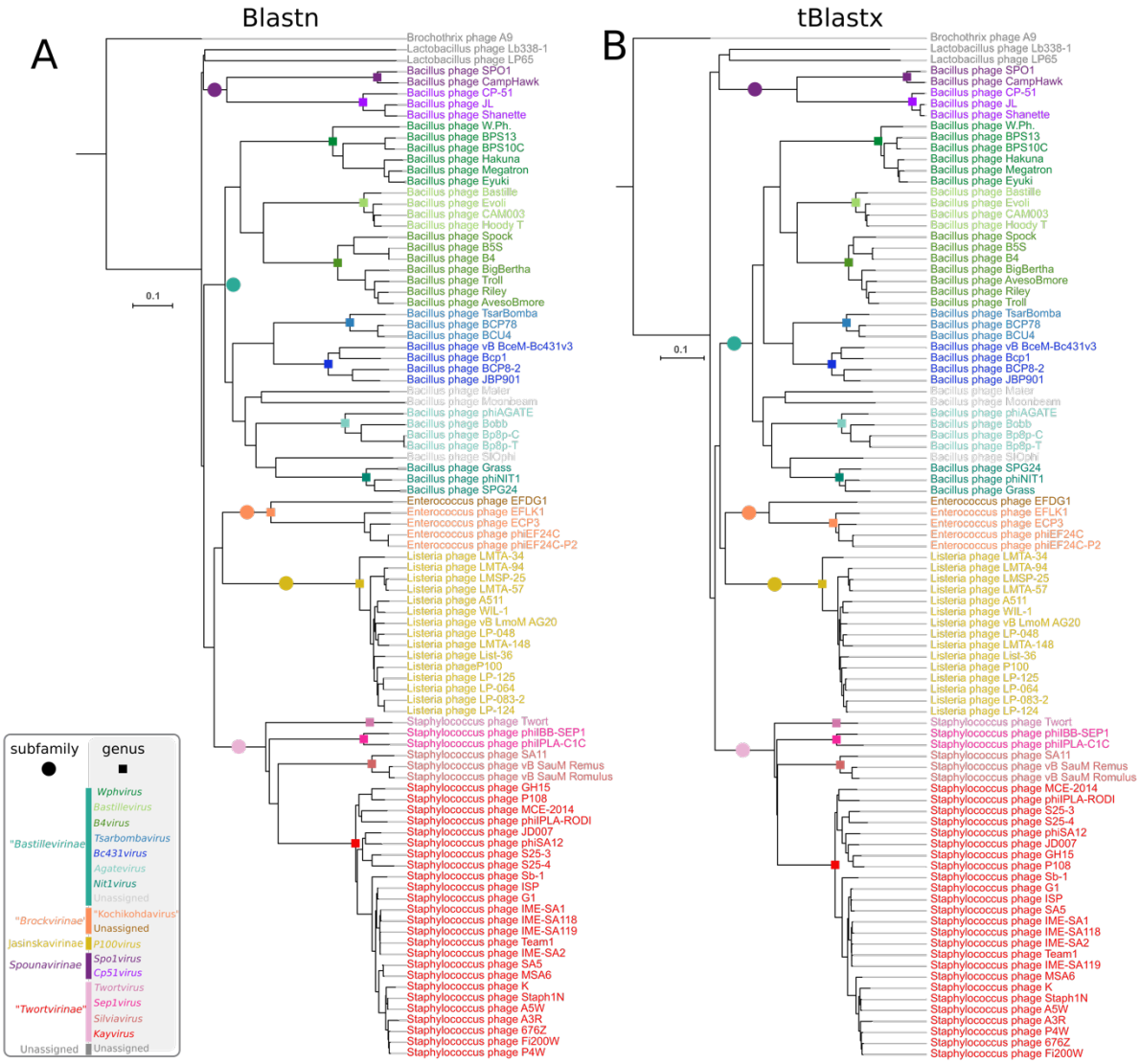

Fig 1: Genome-based clustering trees of 93 spounavirin and spouna-like viruses. Clustering was performed using nucleotide similarities (BLASTn, A) or translated nucleotide similarities (tBLASTx, B). Genomes were compared in a pairwise fashion using Gegenees, transformed into a distance matrix, clustered using R and visualized as trees using Itol. The trees were rooted at Brochotrix phage A9. Genera are delineated with colored squares and suggested subfamilies with colored circles.

The patterns coalesced at a higher taxonomic level when the genomes were analyzed using tBLASTx (S4 Fig). The *Enterococcus* viruses clustered into a single group sharing 41% genome identity, whereas the *Bacillus* viruses fell into two major groups, a group combining the genera *Spo1virus* and *Cp51virus*, and the remainder. All *Staphylococcus*

viruses clustered above ≈36% genome identity, whereas *Listeria* viruses grouped with more than 79% genome identity. Overall, all these genomes were related at the level of at least 15% genome identity. *Lactobacillus* and *Brochothrix* viruses remained genomic orphans, peripherally related to the remainder of the viruses in this assemblage.

Predicted proteome-based analyses

The virus proteomic tree shows four robust groupings mainly determined by the hosts that the viruses infect, corresponding largely with the suggested subfamilies (Fig 2). Viruses that infect *Bacillus* fell into two groups as described before, represented by the revised *Spounavirinae* subfamily and the suggested new subfamily "*Bastillevirinae*." Similarly, the *Listeria* and *Staphylococcus* viruses formed their own clusters, "*Jasinskavirinae*" and "*Twortvirinae*", respectively. This clustering suggests that the major *Bacillus, Listeria,* and *Staphylococcus* virus groups are represented, but that further representatives are required from the under-sampled groups. The suggested "*Brockvirinae*" subfamily is under-sampled, and the grouping observed in the tree is not as well-supported as the other clusters.

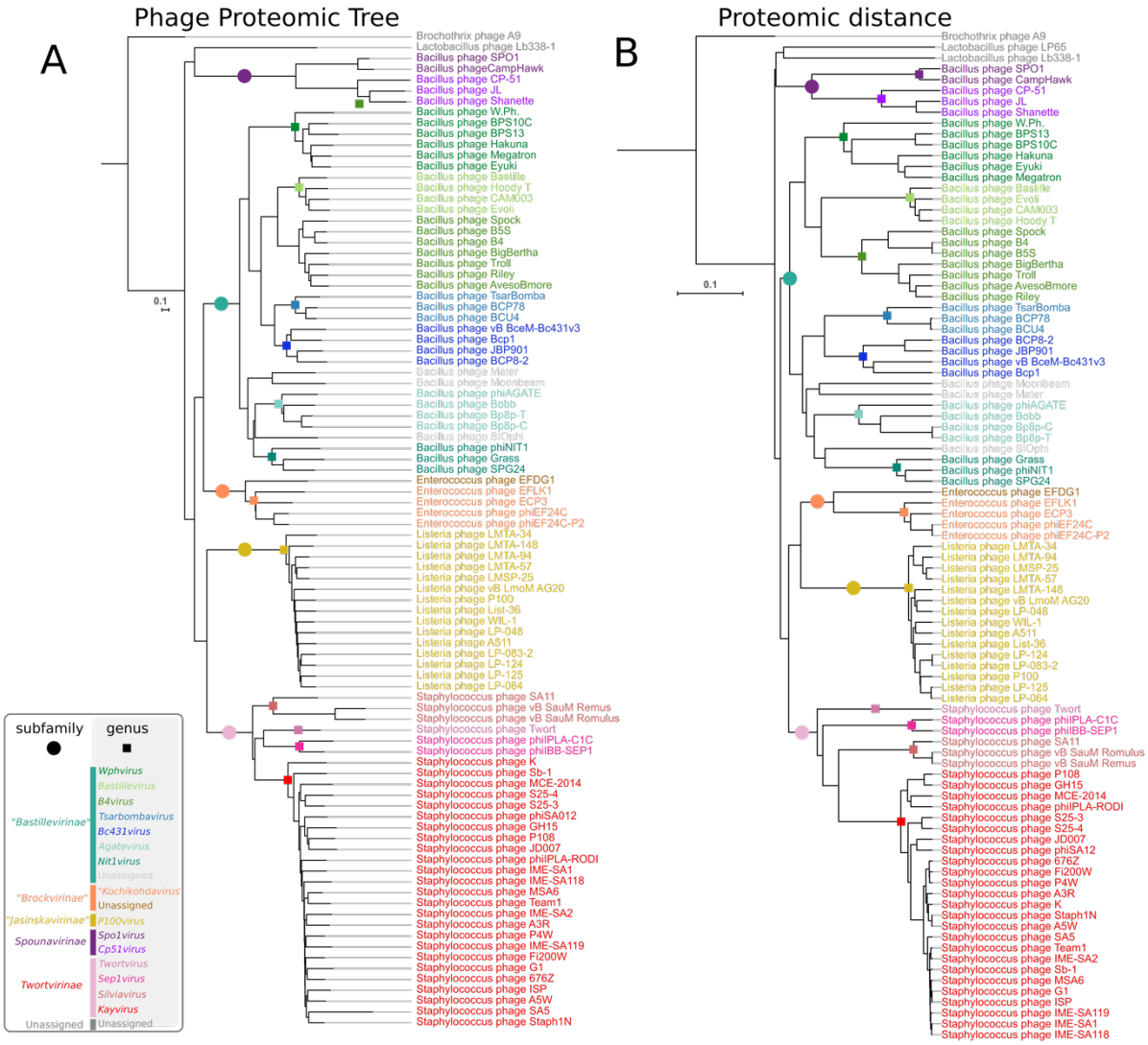

Fig 2: Predicted proteome-based clustering trees of 93 spounavirin and spouna-like viruses. Clustering was performed using the Phage Proteomics Tree approach (A) and proteomic distance (B). Distances were calculated pairwise between all sets of predicted proteomes, clustered with R and visualized using Itol. The trees were rooted at Brochotrix phage A9. Genera are delineated with colored squares and suggested subfamilies with colored circles.

Among 1,296 singleton proteins and 2,070 protein clusters defined using the orthologous protein clusters (OPC) approach, we identified 12 clusters common for all viruses (Table 2, S2 Table). Classification of the viral proteins using prokaryotic virus orthologous groups (pVOGs) showed that 38 pVOGs were shared between all 93 virus genomes (Table 2, S3 Table). This finding was in stark contrast with the results from core

genome analysis using Roary, which revealed only one core gene (the tail tube protein gene). Upon closer inspection of the gene annotations, we found that these analyses might have been confounded by the presence of introns and inteins in many of the core genes (S5–S6 Figs). Indeed, many genes of spounavirins and related viruses are invaded by mobile introns or inteins [32,33]. These gaps in coding sequences challenge gene prediction tools and introduce additional bias in similarity-based cluster algorithms.

Table 2: Core genes with putative annotated functions identified in all 93 spounavirin and spouna-like virus genomes.

Putative function of the core gene identified[a]	**pVOG**[b] **/ OPC**[b] **ID**	**Identification method**
DnaB-like helicase	VOG0025, OPC6121	OPC, pVOG
Baseplate J-like protein	VOG4691, VOG4644, OPC6132	OPC, pVOG
Tail sheath protein	VOG0067, OPC6142	OPC, pVOG
Terminase large subunit	VOG0051, OPC6160	pVOG
Major capsid protein	VOG0061, OPC6148	OPC, pVOG
Prohead protease	VOG4568, OPC6150	pVOG
Portal protein	VOG4556, OPC6151	OPC, pVOG
DNA primase	VOG4551	pVOG
DNA polymerase I	VOG0668, OPC6097	OPC, pVOG
RNA polymerase	VOG0118	pVOG
Recombination exonuclease	VOG4575	pVOG

Recombination endonuclease	VOG0083	pVOG
Tail tape measure protein	VOG0069	pVOG
Tail tube protein	VOG0068, OPC6141	OPC, pVOG, Roary

[a] The full lists of orthologous proteins and pVOGs are available in S2 Table and S3 Table, respectively.

[b] pVOG, prokaryotic virus orthologous group; OPC, orthologous protein clusters.

The pairwise comparison of the predicted proteome content of the viruses revealed a very low overall relatedness at the protein level (S7 Fig S7). The majority of viruses shared less than 10% of their proteins. However, at the suggested new subfamily rank, we observed obvious virus groups sharing their proteomes. The *Enterococcus* viruses ("*Brockvirinae*") shared over 35% of their protein content. The members of the *Bacillus* virus genera *Spo1virus* and *Cp51virus* of the subfamily *Spounavirinae* (sensu stricto) had approximately 20% of their proteins in common, whereas the *Bacillus* virus genera *Bastillevirus*, *B4virus*, *Bc431virus*, *Agatevirus*, *Nit1virus*, *Tsarbombavirus*, and *Wphvirus* ("*Bastillevirinae*") and the *Staphylococcus* virus genera *Kayvirus*, *Silviavirus*, and *Twortvirus* ("*Twortvirinae*") shared over 25% and over 30% of their predicted proteomes, respectively.

Genomic fluidity is a measure of the dissimilarity of genomes evaluated at the gene level [34]. Accordingly, the genomic fluidity results followed those obtained using proteome content analysis (S8 Fig). Despite a high genomic fluidity for most of these viruses, the newly suggested subfamilies and genera were all supported.

The topology of the dendrogram obtained using the average amino acid identity (AAI) approach also supported the suggested new taxonomic scheme (S8 Fig). The AAI was greater than 35% within each subfamily and greater than 67% within each genus. The AAI of

all viruses analyzed in this study was not lower than 22%. The members of the genus *Wphvirus* had the lowest AAI (76%) and the lowest AAI for a pair of proteomes (67% between Bacillus phage W.Ph. and Bacillus phage Eyuki) but surprisingly they had a mid-range genomic fluidity (0.15), suggesting that the protein sequences of wphviruses might have evolved rapidly.

The pangenome of the spounavirins and spouna-like viruses as calculated using Roary [35], consisting of 4,182 genes, was further analyzed by clustering the genomes based on the presence or absence of the accessory genes (S9 Fig). The obtained tree supported the current division of the viruses into approved genera and the suggested new subfamilies.

Many virus genomes are thought to be highly modular, with recombination and horizontal gene transfer potentially resulting in "mosaic genomes" [36,37]. By clustering the spounavirin and spouna-like virus genomes based solely on the gene order of their genomes, we investigated whether the gene synteny was preserved (S10 Fig). The results revealed that genomic rearrangements leave a measurable evolutionary signal in all lineages, since the genomic architecture analysis clustered all viruses according the suggested taxa with the potential exception of Bacillus phage Moonbeam [38]. However, we did not observe the high modularity that may be expected with rampant mosaicism. The lack of rampant mosaicism supports the recent findings by Bolduc et al. that at most about 10% of reference virus genomes have a high degree of mosaicism [14]. Thus, while the gene order in viruses belonging to the newly suggested family "*Herelleviridae*" is not necessarily strictly conserved, we observed a clear evolutionary pattern that is consistent with the sequence-based approaches tested in this study.

Single protein phylogenies

The phylogenetic trees based on comparisons of the major capsid, tail sheath, and DnaB-like helicase proteins are presented in Fig 3. All nine phylograms based on conserved proteins (OPC) are depicted in the trees in S11 Fig. For nearly all single marker trees, the topologies supported the suggested taxonomic scheme. Generally, each taxon is represented as a separate branch on the dendrogram. Notable exceptions could be found in two trees based on hypothetical proteins (OPCs 10357 and 10386). The first (10357) places revised the subfamily *Spounavirinae* as a subclade of "*Bastillevirinae*" and the second (10386) shuffled silviaviruses and kayviruses. This result may indicate that some degree of horizontal gene transfer occurs between groups, which share common hosts.

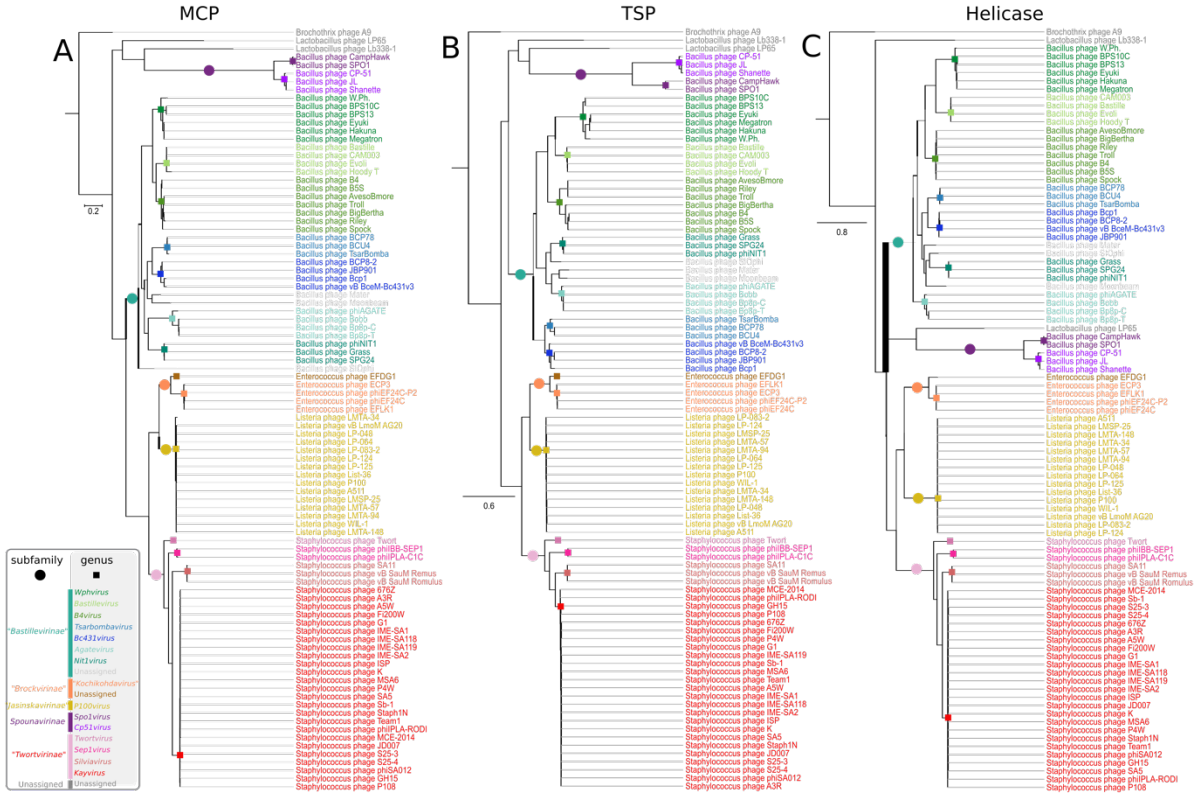

Fig 3: Single gene phylogenies of the major capsid protein (MCP, A), tail sheath protein (TSP, B) and helicase (C) amino acid sequences of 93 spounavirin and spouna-like viruses. Amino acid sequences were aligned with Clustal Omega and trees were generated using

FastTree maximum likelihood with Shimidaira-Hasegawa tests. The scale bar represents the number of substitutions per site. The trees were rooted at Brochotrix phage A9. Genera are delineated with colored squares and suggested subfamilies with colored circles.

Discussion

Using the conventional definition of "tailed phage" families, *Myoviridae*, *Podoviridae*, and *Siphoviridae* (order *Caudovirales*), researchers effectively classified caudovirads for decades [39,40]. However, the classification of these viruses, defined by a traditional morphology-based approach, has been contested with the advent of high-throughput sequencing. The steadily increasing number of available genomes and debates on the impact of horizontal gene transfer, which marked the late 1990s and early 2000s, resulted in a decade-long moratorium on the introduction of any new taxonomy for prokaryotic viruses [41]. The increasing discrepancy between the official taxonomic framework and the emerging high-volume genomics and metagenomics research left ≈90% of prokaryotic viruses known from genome sequences unclassified beyond the family rank, i.e. they were classified as orphan species in a family. Consequently, prokaryotic virus diversity was vastly underappreciated, and virus genome curation remained in disarray. Recently, rather than implementing a 'repeal and replace' strategy for prokaryotic virus taxonomy, the Committee has introduced a holistic system, involving virus particle morphology, overall DNA and protein sequence identities, and phylogeny—an approach used for classification of all other viruses [1,2].

Successful modern taxonomic approaches must scale up to accommodate the increasing pace of prokaryotic virus discovery and be effective across the 4,470 putative complete prokaryotic virus sequences currently deposited in GenBank and other databases [42]. These scaling requirements have remained problematic, and although there are more

306 than 3,400 publicly available caudovirad genomes, only 873 have been officially classified by
307 the ICTV by now [43]. The remaining genomes are provisionally stashed within
308 “unclassified” bins attached to the order *Caudovirales* or its member families, either
309 because they still need to be classified in a species/genus/subfamily, or because they may
310 be unidentified isolates of already classified viruses.

311 The growth in the number of prokaryotic virus genome sequences now supports the
312 application of a range of genomic analyses for robust taxonomic classification [44,45].
313 Meanwhile, phylogenomic approaches are yielding to network-based approaches to better
314 reflect the evolution of viral genomes [9,14]. These network-based approaches help to
315 organize the viral sequence space into statistically-resolvable “viral clusters.” These clusters
316 are approximately equivalent to ICTV-recognized genera and provide a taxonomic
317 description that better reflects evolutionary relationships. Given the high computational
318 demand of network-based approaches, however, the development of centralized resources
319 and authority-monitored cyberinfrastructure, such as the iVirus platform on Cyverse [46] or
320 the Joint Genome Institute Integrated Microbial Genomes Virus resource IMG/VR [47], will
321 have to assist the prokaryotic virus community with large dataset computation and
322 classification.

323 Based on the results of this study, we suggest that the group of spounavirin and
324 spouna-like viruses be removed from the family *Myoviridae* and be given a family rank.
325 Hence, we propose establishing the suggested family “*Herelleviridae*”, in the order
326 *Caudovirales* next to a smaller *Myoviridae* and the established *Podoviridae* and *Siphoviridae*
327 families. The new family would contain five subfamilies: *Spounavirinae* (sensu stricto),
328 “*Bastillevirinae*”, “*Twortvirinae*”, “*Jasinkavirinae*”, and “*Brockvirinae*”, each comprising the
329 ICTV-established genera listed in Table 1 (with additional information in S1 Table).

This study represents the first example of a true taxonomic assessment from an 'ensemble of methods'. The *de facto* taxon splitting suggested here results from the observed diversity of prokaryotic viruses. We are encouraged that the combination of genome BLAST analyses, virus proteomic trees, core protein clusters, genomic synteny (GOAT), and single gene phylogenies yields consistent and complementary results, showing the robustness of the suggested taxa. In addition, the suggested genera correspond well with the taxonomy of the hosts (*Bacillus, Listeria, Staphylococcus*, and *Enterococcus*) indicating broader microbiological consistency. Moreover, only approximately 3% of viruses are left as unassigned at the genus and subfamily rank at this time within this group. These unassigned viruses may represent clades at the genus and subfamily rank that are still under-sampled.

This work demonstrates the usefulness of genome-based classification at a higher taxonomic rank and its ability to accommodate the complex viral diversity. Substitution of the families *Podoviridae*, *Myoviridae*, and *Siphoviridae* with a set of new families which more faithfully reflect the true genetic relationships of the viruses would clarify the taxonomic situation. However, this change does not remove the historically established virus morphotypes observed in the nature among caudovirads: myovirids forming particles with contractile tails, siphovirids forming particles with long non-contractile tails, and podovirids forming particles with short non-contractile ones. By disconnecting morphotype and family classification of caudovirads, taxonomically related clades can be grouped across morphotypes. This grouping includes the muviruses suggested to be classified in the family "*Saltoviridae*" [48] and potentially the broad set of Escherichia phage lambda-related viruses that are currently distributed among the families *Siphoviridae* and *Podoviridae* [49].

353 We believe that abolishment of the *Podoviridae*, *Myoviridae* and *Siphoviridae*
354 families will soon be followed by the “upgrade” of existing viral taxonomy with additional
355 taxon ranks required to accommodate the observed diversity in an orderly manner.
356

Materials and methods

Creation of the dataset

Genome sequences of known spounavirins and spouna-like viruses were retrieved from GenBank or (preferably) RefSeq databases in based on literature data, ICTV and taxonomic classifications provided by the National Center for Biotechnology Information (NCBI). Records representing genomes of candidate spounavirins and spouna-like viruses were retrieved by searching the same databases with the tBLASTx algorithm using terminase and major capsid proteins of several type virus isolates as a query (i.e., Bacillus phage SPO1, Staphylococcus phage Twort, Bacillus phage Bastille, Listeria phage A511, Enterococcus phage ϕEF24C, and Lactobacillus phage LP65) [50,51]. Sequences were manually curated and pre-clustered using CLANS (E-value cut-off 1e-10) to confirm their spounaviral affiliation [52]. This search yielded a set of 93 complete virus genomes, which were used in the following analyses (S1 Table).

The coding sequences in the genomes were re-annotated using PROKKA with the settings --kingdom Viruses, --E-value 1e-6 [53]. All genome sequences are available from NCBI (accession number information listed in S1 Table) or from Github (github.com/evelienadri/herelleviridae).

Genome-based analyses

Gegenees [54] was used in BLASTn and tBLASTx modes (fragment length 200 bp; step length 100 bp) to analyze virus genome nucleotide similarities. Pairwise identities between all genomes under study were determined using BLASTn and tBLASTx algorithms with default parameters [55]. Symmetrical identity scores (% SI) were calculated for each pairwise comparison using the formula

$$\% SI = 2.0 \times \frac{HL \; x \; HI}{QL + SL} \quad (1)$$

in which the HL is defined as the hit length of the BLAST hit, HI is defined as the percentage hit identity, QL is defined as the query length, and SL is defined as the subject length. Symmetrical identity scores were converted into distances using the formula

$$Distance = \sqrt[2]{1.0 \; - \; \%SI \div 100} \quad (2)$$

The resulting distance matrix was hierarchically clustered (complete linkage) using the hclust function of R [56]. Trees were visualized using Itol [57].

All pairwise comparisons of the nucleotide sequences using VICTOR, a Genome-BLAST Distance Phylogeny (GBDP) method, were conducted under settings recommended for prokaryotic viruses [58,59]. The resulting intergenomic distances (including 100 replicates each) were used to infer a balanced minimum evolution tree with branch support via FASTME including subtree pruning and regrafting (SPR) post-processing [60] for each of the formulas D0, D4, and D6, respectively. Trees were visualized with FigTree [61]. Taxon demarcations at the species, genus and family rank were estimated with the OPTSIL program [62], the recommended clustering thresholds [59], and an F value (fraction of links required for cluster fusion) of 0.5 [58].

Proteomic tree

The Phage Proteomic Tree was constructed as described previously [63] and detailed at https://github.com/linsalrob/PhageProteomicTree/tree/master/spounavirus. Briefly, the protein sequences were extracted and clustered using BLASTp. These clusters were refined by Smith-Waterman alignment using CLUSTALW version 2 [64]. Alignments were scored using PROTDIST from the PHYLIP package [65]. Alignment scores were averaged and weighted as described previously [63] resulting in the final tree.

Core protein clusters

Orthologous proteins were clustered using GET_HOMOLOGUES software, which utilizes several independent clustering methods [66]. To capture as many evolutionary relationships as possible, a greedy COGtriangles algorithm was applied with a 50% sequence identity threshold, 50% coverage threshold, and an E-value cut-off equal to 1e-10 [67]. The results were converted into an orthologue matrix with the "compare_clusters" script (part of the GET_HOMOLOGUES suite) [65].

The orthologous protein clusters (OPCs) defined above were used to compute the genomic fluidity for each pair of genomes. For two genomes i and j:

$$\text{Fluidity}(i,j) = \frac{Ui+Uj}{Mi+Mj} \quad (3)$$

with Ui being the number of genes of i not found in j and Mi being the number of genes in i [34]. The resulting distance matrix was hierarchically clustered (complete linkage) using the hclust function of R [56]. Trees were visualized using Itol [57].

Multiple alignments were generated for each OPC using Clustal Omega [68]. For each cluster, the amino acid identity between all protein pairs inside a cluster were determined using multiple alignment. For all genome pairs, the AAI [69] was then computed and transformed into distance using the formula:

$$\text{Distance} = \frac{100 - AAI}{100} \quad (4)$$

The resulting distance matrix was clustered and visualized as described above.

OPCs and multiple alignments for each cluster were used to determine a distance similar to the distance used to generate the Phage Proteomic Tree. To estimate protein distances, in this case, the dist function of the seqinR package [70]was preferred to PROTDIST of the PHYLIP package [65] as the resulting distances are between 0 and 1. Proteomic distances were then computed using the same formula as for the Phage Proteomic Tree. The results were clustered and visualized as described above.

The Dice score is based on reciprocal BLAST searches between all pairs of genomes A and B [31]. The total summed bitscoresof all tBLASTx hits with ≥30% identity, alignment length ≥30 amino acids, and E-value ≤0.01 was converted to a distance DAB as follows:

$$\text{DAB} = 1 - \frac{SAB + SBA}{SAA + SBB} \quad (5)$$

In which SAB SAB and SBA represent the summed bitscores between tBLASTx searches of A versus B, and B versusu A, respectively, while SAA and SBB represent the summed tBLASTx bitscores of the self-queries of A and B, respectively. The resulting distance matrix was clustered with BionJ [71].

To investigate a genomic synteny-based classification signal, we developed a geneorder-based metric built on dynamic programming, the Gene Order Alignment Tool (GOAT, Schuller et al.: Python scripts are available on request, manuscript in preparation). GOAT first identified protein-coding genes in the 93 spounavirin and spouna-like virus genomes using Prodigal V2.6.3 in anonymous mode [72], and assigned them to the latest pVOGs [73]). pVOG alignments (9,518) were downloaded (http://dmk-brain.ecn.uiowa.edu/pVOGs/) and converted to profiles of hidden Markov models (HMM) using HMMbuild (HMMer 3.1b2, [74]). Proteins were assigned to pVOGs using HMMsearch (E-value <10-2) and used to generate a synteny profile of every genome. GOAT accounted for gene replacements and distant homology by using an all-vs-all similarity matrix between pVOG pairs based on HMM-HMM similarity (HH-suite 2.0.16) [75]). Distant HHsearch similarity scores between protein families were calculated as the average of reciprocal hits and used as substitution scores in the gene order alignment. The GOAT algorithm identified the optimal gene order alignment score between two virus genomes by implementing semi-global dynamic programming alignment based only on the order of pVOGs identified on every virus genome. To account for virus genomes being cut at arbitrary positions during sequence assembly, GOAT transmutes the gene order at all possible positions and in both sense and antisense directions in search of the optimal alignment score. The optimal GOAT alignment score GAB between every pair of virus genomes A and B, was converted to a distance DAB as follows:

$$DAB = 1 - \frac{GAB + GBA}{GAA + GBB} \quad (6)$$

463 in which GAB and GBA represent the optimal GOAT score between A and B, and B and A,
464 respectively, while GAA and GBB represent the GOAT scores of the self-alignments of A and
465 B, respectively. This pairwise distance matrix was clustered with BionJ [71].
466 Prokka re-annotated genomes were used to create pan-, core-, and accessory
467 genomes of all selected spounavirins and spouna-like viruses [53]. The annotations were
468 analyzed using Roary [35] with a 50% length BLASTp identity threshold for homologous
469 genes. Roary functions as follows: CD-HIT [76] was used to pre-cluster protein sequences
470 and perform an all-vs-all comparison of protein sequences with BLASTp to identify orthologs
471 and paralogs within the genomes. MCL [77] was then used to cluster the genomes based on
472 the presence and absence of the accessory genes. The resulting tree file was visualized using
473 FigTree v1.4.3 [61]. The tree was rooted in Brochothrix phage A9. The gene presence-
474 absence output table from Roary was then imported into R and using a custom R-script
475 (available from github.com/evelienadri/herelleviridae/tree/master). Pairwise shared gene
476 contents were calculated for each combination of genomes.
477

Single gene phylogenies

Based on the OPC and pVOG analyses, we chose nine well-annotated protein clusters present in all 93 spounavirins and spouna-like viruses. Selected clusters included: DNA helicases, major capsid proteins, tail sheath proteins, two different groups of baseplate proteins, and four clusters with no known function. The members of these clusters were aligned using Clustal Omega with default parameters [47]. Resulting alignments were analyzed with ProtTest 3.4 [59] to determine a suitable protein evolution model (only variations of models compatible with downstream software like JTT and WAG were considered). Estimated models were used to generate phylograms with FastTree 2.1.7 [60]. The program implements the approximately maximum-likelihood method with Shimodaira-Hasegawa tests to generate the tree and calculate support of the splits. This approach is much faster than “traditional” maximum-likelihood methods with negligible accuracy loss [59–61].

Figure legends

Fig 1: Genome-based clustering trees of 93 spounavirin and spouna-like viruses. Clustering was performed using nucleotide similarities (BLASTn, A) or translated nucleotide similarities (tBLASTx, B). Genomes were compared in a pairwise fashion using Gegenees, transformed into a distance matrix, clustered using R and visualized as trees using Itol. The trees were rooted at Brochotrix phage A9. Genera are delineated with colored squares and suggested subfamilies with colored circles.

Fig 2: Predicted proteome-based clustering trees of 93 spounavirin and spouna-like viruses. Clustering was performed using the Phage Proteomics Tree approach (A) and proteomic distance (B). Distances were calculated pairwise between all sets of predicted proteomes, clustered with R and visualized using Itol. The trees were rooted at Brochotrix phage A9. Genera are delineated with colored squares and suggested subfamilies with colored circles.

Fig 3: Single gene phylogenies of the major capsid protein (MCP, A), tail sheath protein (TSP, B) and helicase (C) amino acid sequences of 93 spounavirin and spouna-like viruses. Amino acid sequences were aligned with Clustal Omega and trees were generated using FastTree maximum likelihood with Shimidaira-Hasegawa tests. The scale bar represents the number of substitutions per site. The trees were rooted at Brochotrix phage A9. Genera are delineated with colored squares and suggested subfamilies with colored circles.

Supporting information

S1 Fig. Heatmap of the blastn-based nucleotide similarities between pairs of genomes as calculated with Gegenees at default parameters.

S2 Fig. Genome-blast Distance Phylogeny as calculated using VICTOR.

S3 Fig. Heatmap of the DICE coefficient calculated between each pair of genomes.

S4 Fig. Heatmap of the tblastx-based nucleotide similarities between pairs of genomes as calculated with Gegenees at default parameters.

S5 Fig. Heatmap of the pairwise comparison of all genomes visualized as percentage of shared orthologous proteins (OPCs) as calculated on original GenBank files.

S6 Fig. Heatmap of the pairwise comparison of all genomes visualized as percentage of shared orthologous proteins (OPCs) as calculated on reannotated genomes.

S7 Fig. Heatmap of the pairwise comparison of all genomes visualized as percentage of shared proteins as calculated with Roary on reannotated genome files.

S8 Fig. Clustering trees of genomic fluidity and amino acid identity calculated pairwise between all genomes using orthologous protein clusters.

S9 Fig. Accessory genome clustering tree, calculated based on the presence and absence of accessory genes in each genome.

S10 Fig. Heatmap and clustering tree calculated by the Gene Order Alignment Tool and visualized as a distance matrix between all genome pairs.

S11 Fig. Maximum Likelihood trees of single gene phylogenies using protein clusters present in all 93 genomes.

S1 Table. Overview of the 93 phage genomes used in this study.

S2 Table. Complete list of all orthologous proteins identified in the set of 93 spounavirin and spouna-like virus genomes.

S3 Table. Complete list of prokaryotic virus orthologous groups identified in the set of 93 spounavirin and spouna-like virus genomes.

Acknowledgments

We thank Laura Bollinger, Integrated Research Facility at Fort Detrick, for technical writing services. EMA would like to thank PH Nel for assistance with R scripting. JB was supported by the National Science Centre (Poland SONATA 12 grant number 2016/23/D/NZ2/00435). MBPS and BED were supported by the Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004. RAE was supported by grants DUE-132809 and MCB-1330800 from the US National Science Foundation. HMO was supported by University of Helsinki funding for Instruct research infrastructure (Virus and Macromolecular Complex Production, ICVIR). AG was supported by a *Chargé de Recherches* fellowship from the National Fund for Scientific Research (FNRS, Belgium). FE was supported by the EUed Horizon 2020 Framework Programme for Research and Innovation ('Virus-X', project no. 685778). MBS would like to acknowledge the Gordon and Betty Moore Foundation Investigator Award (GBMF#3790) for funding.

Competing interests

All authors, except for MBPS, are members of the Bacterial and Archaeal Viruses Subcommittee of the International Committee on the Taxonomy of Viruses (ICTV). The authors declare no conflicts of interest.

References

1. Krupovic M, Dutilh BE, Adriaenssens EM, Wittmann J, Vogensen FK, Sullivan MB, et al. Taxonomy of prokaryotic viruses: update from the ICTV bacterial and archaeal viruses subcommittee. Arch Virol. Springer Vienna; 2016;161: 1095–1099. doi:10.1007/s00705-015-2728-0

2. Adriaenssens EM, Krupovic M, Knezevic P, Ackermann H-W, Barylski J, Brister JR, et al. Taxonomy of prokaryotic viruses: 2016 update from the ICTV bacterial and archaeal viruses subcommittee. Arch Virol. Springer Vienna; 2017;162: 1153–1157. doi:10.1007/s00705-016-3173-4

3. Lavigne R, Seto D, Mahadevan P, Ackermann H-W, Kropinski AM. Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools. Res Microbiol. 2008;159: 406–414. doi:10.1016/j.resmic.2008.03.005

4. Lavigne R, Darius P, Summer EJ, Seto D, Mahadevan P, Nilsson AS, et al. Classification of Myoviridae bacteriophages using protein sequence similarity. BMC Microbiol. 2009;9: 224. doi:10.1186/1471-2180-9-224

5. Adriaenssens EM, Edwards R, Nash JHE, Mahadevan P, Seto D, Ackermann H-W, et al. Integration of genomic and proteomic analyses in the classification of the Siphoviridae family. Virology. 2015;477: 144–154. doi:10.1016/j.virol.2014.10.016

6. Gregory AC, Solonenko SA, Ignacio-Espinoza JC, LaButti K, Copeland A, Sudek S, et al. Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. BMC Genomics. BMC Genomics; 2016;17: 930. doi:10.1186/s12864-016-3286-x

7. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, et al. Patterns and ecological drivers of ocean viral communities. Science (80-). 2015;348: 1261498–

1261498. doi:10.1126/science.1261498

8. Holmfeldt K, Solonenko N, Shah M, Corrier K, Riemann L, Verberkmoes NC, et al. Twelve previously unknown phage genera are ubiquitous in global oceans. Proc Natl Acad Sci U S A. 2013;110: 12798–12803. doi:10.1073/pnas.1305956110

9. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Reticulate representation of evolutionary and functional relationships between phage genomes. Mol Biol Evol. 2008;25: 762–777. doi:10.1093/molbev/msn023

10. Nishimura Y, Watai H, Honda T, Mihara T, Omae K, Roux S, et al. Environmental viral genomes shed new light on virus-host interactions in the ocean. mSphere. 2017;2: e00359-16. doi:10.1128/mSphere.00359-16

11. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. Nature. 2016;536: 425–430. doi:10.1038/nature19094

12. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature. 2016;537: 689–693. doi:10.1038/nature19366

13. Iranzo J, Krupovic M, Koonin E V. The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. MBio. 2016;7: e00978-16. doi:10.1128/mBio.00978-16

14. Bolduc B, Jang H Bin, Doulcier G, You Z-Q, Roux S, Sullivan MB. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. PeerJ. 2017;5: e3243. doi:10.7717/peerj.3243

15. Okubo S, Strauss B, Stodolsky M. The possible role of recombination in the infection of competent Bacillus subtilis by bacteriophage deoxyribonucleic acid. Virology.

1964;24: 552–562. doi:10.1016/0042-6822(64)90207-7

16. Parker ML, Eiserling FA. Bacteriophage SPO1 structure and morphogenesis I. Tail structure and length regulation. J Virol. 1983;46: 239–249.

17. Duda RL, Hendrix RW, Huang WM, Conway JF. Shared architecture of bacteriophage SPO1 and herpesvirus capsids. Curr Biol. 2006;16: R11–R13. doi:10.1016/j.cub.2006.02.014

18. Hemphill HE, Whiteley HR. Bacteriophages of Bacillus subtilis. Bacteriol Rev. 1975;39: 257–315.

19. Okubo S, Yanagida T, Fujita DJ, Olsson-Wilhelm BM. The genetics of bacteriophage SPO1. Biken J. 1972;15: 81–97.

20. Klumpp J, Lavigne R, Loessner MJ, Ackermann H-W. The SPO1-related bacteriophages. Arch Virol. 2010;155: 1547–61. doi:10.1007/s00705-010-0783-0

21. Stewart CR, Casjens SR, Cresawn SG, Houtz JM, Smith AL, Ford ME, et al. The genome of Bacillus subtilis bacteriophage SPO1. J Mol Biol. 2009;388: 48–70. doi:10.1016/j.jmb.2009.03.009

22. Murphy FA, Fauquet CM, Bishop DHL, Ghabrial SA, Jarvis AW, Martelli GP, et al. Sixth report of the International Committee on Taxonomy of Viruses. Arch Virol Suppl. 1995;10: 590.

23. Ackermann H-W, Elzanowski A, Fobo G, Stewart G. Relationships of tailed phages: a survey of protein sequence identity. Arch Virol. 1995;140: 1871–1884. doi:10.1007/BF01384350

24. Adams MJ, Carstens EB. Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2012). Arch Virol. 2012;157: 1411–1422. doi:10.1007/s00705-012-1299-6

635 25. Schuch R, Fischetti VA. The secret life of the anthrax agent Bacillus anthracis:
636 Bacteriophage-mediated ecological adaptations. PLoS One. 2009;4: e6532.
637 doi:10.1371/journal.pone.0006532
638 26. Yuan Y, Peng Q, Wu D, Kou Z, Wu Y, Liu P, et al. Effects of actin-like proteins encoded
639 by two Bacillus pumilus phages on unstable lysogeny, revealed by genomic analysis.
640 Appl Environ Microbiol. 2015;81: 339–350. doi:10.1128/AEM.02889-14
641 27. Klumpp J, Dorscht J, Lurz R, Bielmann R, Wieland M, Zimmer M, et al. The terminally
642 redundant, nonpermuted genome of Listeria bacteriophage A511: a model for the
643 SPO1-like myoviruses of gram-positive bacteria. J Bacteriol. 2008;190: 5753–65.
644 doi:10.1128/JB.00461-08
645 28. Barylski J, Nowicki G, Goździcka-Józefiak A. The discovery of phiAGATE, a novel phage
646 infecting Bacillus pumilus, leads to new insights into the phylogeny of the subfamily
647 Spounavirinae. Dąbrowska K, editor. PLoS One. 2014;9: e86632.
648 doi:10.1371/journal.pone.0086632
649 29. Uchiyama J, Rashel M, Maeda Y, Takemura I, Sugihara S, Akechi K, et al. Isolation and
650 characterization of a novel Enterococcus faecalis bacteriophage phiEF24C as a
651 therapeutic candidate. FEMS Microbiol Lett. 2008;278: 200–206. doi:10.1111/j.1574-
652 6968.2007.00996.x
653 30. Uchiyama J, Rashel M, Takemura I, Wakiguchi H, Matsuzaki S. In silico and in vivo
654 evaluation of bacteriophage phiEF24C, a candidate for treatment of Enterococcus
655 faecalis infections. Appl Environ Microbiol. 2008;74: 4149–4163.
656 doi:10.1128/AEM.02371-07
657 31. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. Expanding the marine virosphere
658 using metagenomics. Rocha EPC, editor. PLoS Genet. 2013;9: e1003987.

659 doi:10.1371/journal.pgen.1003987
660 32. Goodrich-Blair H, Scarlato V, Gott JM, Xu M-Q, Shub DA. A self-splicing group I intron
661 in the DNA polymerase gene of bacillus subtilis bacteriophage SPO1. Cell. 1990;63:
662 417–424. doi:10.1016/0092-8674(90)90174-D
663 33. Lavigne R, Vandersteegen K. Group I introns in Staphylococcus bacteriophages.
664 Future Virol. 2013;8: 997–1005. doi:10.2217/fvl.13.84
665 34. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: An integrative view
666 of gene diversity within microbial populations. BMC Genomics. 2011;12: 32.
667 doi:10.1186/1471-2164-12-32
668 35. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid
669 large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31: 3691–3693.
670 doi:10.1093/bioinformatics/btv421
671 36. Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, Hendrix RW. Genomic sequences
672 of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid
673 bacteriophages. J Mol Biol. 2000;299: 27–51. doi:10.1006/jmbi.2000.3729
674 37. Krupovic M, Prangishvili D, Hendrix RW, Bamford DH. Genomics of bacterial and
675 archaeal viruses: Dynamics within the prokaryotic virosphere. Microbiol Mol Biol Rev.
676 2011;75: 610–635. doi:10.1128/MMBR.00011-11
677 38. Cadungog JN, Khatemi BE, Hernandez AC, Kuty Everett GF. Complete genome
678 sequence of Bacillus megaterium myophage Moonbeam. Genome Announc. 2015;3:
679 e01428-14. doi:10.1128/genomeA.01428-14
680 39. Ackermann H-W, DuBow MS. Viruses of prokaryotes. Boca Raton, FL, USA: CRC Press;
681 1987.
682 40. Ackermann H-W. Classification of bacteriophages. In: Calendar R, editor. The

Bacteriophages. 2nd ed. New York, NY, USA: Oxford University Press; 2006. pp. 8–17.

41. Tolstoy I, Kropinski AM, Brister JR. Bacteriophage taxonomy: an evolving discipline. In: Azeredo J, Sillankorva S, editors. Methods in Molecular Biology: Bacteriophage Therapy. Springer Nature; doi:Forthcoming

42. Karsch-Mizrachi I, Nakamura Y, Cochrane G. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 2012;40: D33–D37. doi:10.1093/nar/gkr1006

43. Davison AJ. Journal of General Virology – Introduction to "ICTV Virus Taxonomy Profiles." J Gen Virol. 2017;98: 1–1. doi:10.1099/jgv.0.000686

44. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, et al. Consensus statement: Virus taxonomy in the age of metagenomics. Nat Rev Microbiol. 2017;15: 161–168. doi:10.1038/nrmicro.2016.177

45. Adams MJ, Lefkowitz EJ, King AMQ, Harrach B, Harrison RL, Knowles NJ, et al. 50 years of the International Committee on Taxonomy of Viruses: progress and prospects. Arch Virol. 2017;162: 1441–1446. doi:10.1007/s00705-016-3215-y

46. Bolduc B, Youens-Clark K, Roux S, Hurwitz BL, Sullivan MB. iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. ISME J. 2017;11: 7–14. doi:10.1038/ismej.2016.89

47. Paez-Espino D, Chen IMA, Palaniappan K, Ratner A, Chu K, Szeto E, et al. IMG/VR: A database of cultured and uncultured DNA viruses and retroviruses. Nucleic Acids Res. 2017;45: D457–D465. doi:10.1093/nar/gkw1030

48. Hulo C, Masson P, Le Mercier P, Toussaint A. A structured annotation frame for the transposable phages: A new proposed family "Saltoviridae" within the Caudovirales. Virology. 2015;477: 155–163. doi:10.1016/j.virol.2014.10.009

707 49. Grose JH, Casjens SR. Understanding the enormous diversity of bacteriophages: The
708 tailed phages that infect the bacterial family Enterobacteriaceae. Virology. 2014;468–
709 470: 421–443. doi:10.1016/j.virol.2014.08.024
710 50. Brister JR, Ako-adjei D, Bao Y, Blinkova O. NCBI Viral Genomes Resource. Nucleic Acids
711 Res. 2015;43: D571–D577. doi:10.1093/nar/gku1207
712 51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool.
713 J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2
714 52. Frickey T, Lupas A. CLANS: A Java application for visualizing protein families based on
715 pairwise similarity. Bioinformatics. 2004;20: 3702–3704.
716 doi:10.1093/bioinformatics/bth444
717 53. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30:
718 2068–2069. doi:10.1093/bioinformatics/btu153
719 54. Ågren J, Sundström A, Håfström T, Segerman B. Gegenees: Fragmented alignment of
720 multiple genomes for determining phylogenomic distances and genetic signatures
721 unique for specified target groups. PLoS One. 2012;7: e39107.
722 doi:10.1371/journal.pone.0039107
723 55. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:
724 architecture and applications. BMC Bioinformatics. 2009;10: 421. doi:10.1186/1471-
725 2105-10-421
726 56. Development Core Team R. R: A language and environment for statistical computing
727 [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2008. Available:
728 http://www.r-project.org
729 57. Letunic I, Bork P. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree
730 display and annotation. Bioinformatics. 2007;23: 127–128.

731 doi:10.1093/bioinformatics/btl529
732 58. Meier-Kolthoff JP, Hahnke RL, Petersen J, Scheuner C, Michael V, Fiebig A, et al.
733 Complete genome sequence of DSM 30083T, the type strain (U5/41T) of Escherichia
734 coli, and a proposal for delineating subspecies in microbial taxonomy. Stand Genomic
735 Sci. 2014;9: 2. doi:10.1186/1944-3277-9-2
736 59. Meier-Kolthoff JP, Göker M. VICTOR: genome-based phylogeny and classification of
737 prokaryotic viruses. Bioinformatics. 2017;33: 3396–3404.
738 doi:10.1093/bioinformatics/btx440
739 60. Lefort V, Desper R, Gascuel O. FastME 2.0: A comprehensive, accurate, and fast
740 distance-based phylogeny inference program. Mol Biol Evol. 2015;32: 2798–2800.
741 doi:10.1093/molbev/msv150
742 61. Rambaud. FigTree [Internet]. 2007 [cited 1 Aug 2015]. Available:
743 http://tree.bio.ed.ac.uk/software/figtree/
744 62. Göker M, García-Blázquez G, Voglmayr H, Tellería MT, Martín MP. Molecular
745 taxonomy of phytopathogenic fungi: A case study in Peronospora. PLoS One. 2009;4:
746 8–10. doi:10.1371/journal.pone.0006319
747 63. Rohwer F, Edwards R. The Phage Proteomic Tree: a genome-based taxonomy for
748 phage. J Bacteriol. Am Soc Microbiol; 2002;184: 4529–4535.
749 doi:10.1128/JB.184.16.4529
750 64. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al.
751 Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23: 2947–2948.
752 doi:10.1093/bioinformatics/btm404
753 65. Felsenstein J. PHYLIP - Phylogeny inference package - v3.2. Cladistics. 1989;5: 164–
754 166. doi:10.1111/j.1096-0031.1989.tb00562.x

66. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. 2013;79: 7696–7701. doi:10.1128/AEM.02411-13

67. Kristensen DM, Kannan L, Coleman MK, Wolf YI, Sorokin A, Koonin E V, et al. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. Bioinformatics. 2010;26: 1481–7. doi:10.1093/bioinformatics/btq229

68. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7: 539. doi:10.1038/msb.2011.75

69. Konstantinidis KT, Tiedje JM. Towards a genome-based taxonomy for prokaryotes. J Bacteriol. 2005;187: 6258–6264. doi:10.1128/JB.187.18.6258-6264.2005

70. Charif D, Thioulouse J, Lobry JR, Perrière G. Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinformatics. 2005;21: 545–547. doi:10.1093/bioinformatics/bti037

71. Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 1997;14: 685–695. doi:10.1017/CBO9781107415324.004

72. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11. doi:10.1186/1471-2105-11-119

73. Grazziotin AL, Koonin E V, Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. 2017;45: D491–D498. doi:10.1093/nar/gkw975

74. Finn RD, Clements J, Eddy SR. HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res. 2011;39: 29–37. doi:10.1093/nar/gkr367

75. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005;33: W244-2488. doi:10.1093/nar/gki408

76. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–3152. doi:10.1093/bioinformatics/bts565

77. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30: 1575–1584. doi:doi: 10.1093/nar/30.7.1575