Analysis of Spounaviruses as a Case Study for the Overdue Reclassification of Tailed Bacteriophages

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; evelien.adriaenssens@gmail.com, evelien.adriaenssens@liv.ac.uk; corresponding author*

ABSTRACT

It is almost a cliché that tailed bacteriophages of the order *Caudovirales* are the most abundant and diverse viruses in the world. Yet, their taxonomy still consists of a single order with just three families: *Myoviridae*, *Siphoviridae*, and *Podoviridae*. Thousands of newly discovered phage genomes have recently challenged this morphology-based classification, revealing that tailed bacteriophages are genomically even more diverse than once thought. Here, we evaluate a range of methods for bacteriophage taxonomy by using a particularly challenging group as an example, the Bacillus phage SPO1-related viruses of the myovirid subfamily *Spounavirinae*. Exhaustive phylogenetic and phylogenomic analyses indicate that the spounavirins are consistent with the taxonomic rank of family and should be divided into at least five subfamilies. This work is a case study for virus genomic taxonomy and the first step in an impending massive reorganization of the tailed bacteriophage taxonomy.

KEYWORDS

Virus taxonomy; bacteriophage taxonomy; *Caudovirales*; spounaviruses; phylogenetics; phylogenomics

By the end of 2017, 3,033 complete genomes of tailed phages were available in the National Center for Biotechnology Information (NCBI) RefSeq database and a further 18,753 partial genomes were found in International Nucleotide Sequence Database Collaboration databases (Karsch-Mizrachi et al. 2012; O'Leary et al. 2016). The classification of this massive group is the formal responsibility of the Bacterial and Archaeal Viruses Subcommittee of the International Committee on the Taxonomy of Viruses (ICTV). In recent years, we (the Subcommittee) focused on classifying newly described phages into species and genera, within established viral families (Lavigne et al. 2008, 2009, Adriaenssens et al. 2015, 2017; Krupovic et al. 2016). However, once our attention shifted towards higher order relationships, we found that the ranks available in virus taxonomy (*species*, *genus*, *subfamily*, *family*, and *order*) were no longer sufficient for the description of phage diversity. The limitation is particularly acute in the case of the order *Caudovirales*—the most abundant and diverse group of viruses (Paez-Espino et al. 2016; Roux et al. 2016; Nishimura et al. 2017). Indeed, the diversity of caudovirads surpasses that of any other virus taxon. 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 (Iranzo et al. 2016). Each of eight remaining modules encompasses one or more families of eukaryotic or archaeal viruses. Consequently, each of the 11 caudovirad modules could be considered a separate family. Despite this remarkable diversity, all caudovirads are currently classified into three families - *Myoviridae, Podoviridae*, and *Siphoviridae*. These families were historically established on morphological features alone, forming an artificial classification ceiling.

In this study, the Subcommittee explored the diversity of the order *Caudovirales* on the example of the *Spounavirinae* subfamily, a large group of myoviruses that forms one of the above-mentioned caudovirad modules (Iranzo et al. 2016; Bolduc et al. 2017). The subfamily 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 (Lavigne et al. 2009). The unifying characteristics of members of this subfamily are: the host belongs to the bacterial phylum *Firmicutes*; strictly virulent lifestyle; myovirion morphology; terminally redundant, non-permuted dsDNA genome 127–157 kb in length; and "considerable amino acid homology" (Klumpp et al. 2010). The strictly virulent lifestyle of these viruses has been somewhat disputed (Schuch and Fischetti 2009; Yuan et al. 2015) but still remains a rule of thumb for the taxon inclusion. Since the inception of the subfamily, the number of its members has grown significantly, and its taxonomic structure was contested several times (Klumpp et al. 2010; Barylski et al. 2014; Iranzo et al. 2016; Krupovic et al. 2016; Adriaenssens et al. 2017; Bolduc et al. 2017). At present, the *Spounavirinae* subfamily includes six genera (*Kayvirus, P100virus, Silviavirus, Spo1virus, Tsarbombavirus and Twortvirus*) and three unassigned species (*Enterococcus virus phiEC24C, Lactobacillus virus Lb338-1 and Lactobacillus virus LP65*).

Here, we reevaluated the current classification of spounavirins and related viruses and outlined a better fitting scheme, in the process also reaffirming the need for major changes in phage taxonomy that will better accommodate the observed genomic diversity.

Materials & Methods

Creation of the Dataset

Genome sequences of known spounavirins and spouna-like viruses were retrieved from GenBank or (preferably) RefSeq databases based on literature data, ICTV and taxonomic classifications provided by the NCBI. Records representing genomes of candidate spounavirins were retrieved by searching the same databases with the tBLASTx algorithm using as a queries terminase and major capsid proteins of Bacillus phage SPO1, Staphylococcus phage Twort, Bacillus phage Bastille, Listeria phage A511, Enterococcus phage φEF24C, and Lactobacillus phage LP65 [type isolates of the original subfamily, (Altschul et al. 1990; Brister et al. 2015)]. Sequences were manually curated and pre-clustered using Cluster Analysis of Sequences (CLANS; E-value cut-off 1e-10) to confirm their spounaviral affiliation (Frickey and Lupas 2004). This search yielded a set of 93 complete virus genomes, which were used in the following analyses (Supplementary Table 1).

The genomes were re-annotated using PROKKA with the settings --kingdom Viruses, --E-value 1e-6 (Seemann 2014). All original genome sequences are available from NCBI (accession number information listed in Supplementary Table 1) and the reannotated genomes from Github (github.com/evelienadri/herelleviridae).

Genome-based Analyses

Gegenees (Ågren et al. 2012) was used to analyze genome similarities (fragment length 200 bp; step length 100 bp). Pairwise identities between all genomes under study were determined using BLASTn and tBLASTx algorithms with default parameters (Camacho et al. 2009). Symmetrical identity scores (% SI) were calculated for each pairwise comparison using the formula:

$$\% \text{ SI} = 2.0 \text{ x } \frac{HL\ x\ HI}{QL + SL}$$

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:

$$\text{Distance} = \sqrt[2]{1.0\ -\ \%SI \div 100}$$

The resulting distance matrix was hierarchically clustered (complete linkage) using the hclust function of R (Development Core Team 2008). Trees were visualized using Itol (Letunic and Bork 2007).

Additionally, pairwise comparisons of the nucleotide sequences using VICTOR, a Genome-BLAST Distance Phylogeny (GBDP) method, were conducted under settings recommended for prokaryotic viruses (Meier-Kolthoff et al. 2014; Meier-Kolthoff and Göker

2017). 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 post-processing (Lefort et al. 2015) for each of the formulas D0, D4, and D6, respectively. Trees were visualized with FigTree (Rambaud 2007). Taxon demarcations at the species, genus and family rank were estimated with the OPTSIL program (Göker et al. 2009), the recommended clustering thresholds (Meier-Kolthoff and Göker 2017), and an F value (fraction of links required for cluster fusion) of 0.5 (Meier-Kolthoff et al. 2014).

Proteome-based Analyses

The Phage Proteomic Tree was constructed as described previously (Rohwer and Edwards 2002) 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 (Larkin et al. 2007). Alignments were scored using open-source PROTDIST from the phylogeny inference package (PHYLIP) (Felsenstein 1989). Alignment scores were averaged and weighted as described previously (Rohwer and Edwards 2002) resulting in the final tree.

Orthologous protein clusters (OPCs) were constructed using GET_HOMOLOGUES software, which utilizes several independent clustering methods (Contreras-Moreira and Vinuesa 2013). 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 (Kristensen et al. 2010). The results were converted into an orthologue matrix with the “compare_clusters” script (part of the GET_HOMOLOGUES suite) (Felsenstein 1989).

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

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

with Ui being the number of genes of i not found in j and Mi being the number of genes in i (Kislyuk et al. 2011). The resulting distance matrix was hierarchically clustered (complete linkage) using the hclust function of R (Development Core Team 2008). Trees were visualized using Itol (Letunic and Bork 2007).

Multiple alignments were generated for each OPC using Clustal Omega (Sievers et al. 2011). For each cluster, the amino acid identity between all protein pairs inside a cluster was determined using multiple alignment. For all genome pairs, the AAI (Konstantinidis and Tiedje 2005) was then computed and transformed into distance using the formula:

$$\text{Distance} = \frac{100-AAI}{100}$$

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 distance (dist) function of the seqinR package (Charif et al. 2005)was preferred to PROTDIST of the PHYLIP package (Felsenstein 1989) 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 (Mizuno et al. 2013). The total summed bit-scores of all tBLASTx hits with ≥30% identity, alignment length ≥30 amino acids, and E-value ≤0.01 was converted to a distance DAB as follows:

$$DAB = 1 - \frac{SAB + SBA}{SAA + SBB}$$

in which SAB and SBA represent the summed bit-scores between tBLASTx searches of A versus B, and B versusA, respectively, while SAA and SBB represent the summed tBLASTx bit-scores of the self-queries of A and B, respectively. The resulting distance matrix was clustered with BionJ (Gascuel 1997).

To investigate a genomic synteny-based classification signal, we developed a gene order-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 (Hyatt et al. 2010), and assigned them to the latest pVOGs (Grazziotin et al. 2017)). 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, (Finn et al. 2011)). 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) (Söding et al. 2005)). 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:

$$\text{DAB} = 1 - \frac{GAB + GBA}{GAA + GBB}$$

in which GAB and GBA represent the optimal GOAT score between A and B, and B and A, respectively, while GAA and GBB represent the GOAT scores of the self-alignments of A and B, respectively. This pairwise distance matrix was clustered with BionJ (Gascuel 1997).

Prokka re-annotated genomes were used to create pan-, core-, and accessory genomes of all selected spounavirins and spouna-like viruses (Seemann 2014). The annotations were analyzed using Roary (Page et al. 2015) with a 50% length BLASTp identity threshold for homologous genes. Roary functions as follows: CD-HIT (Fu et al. 2012) was used to pre-cluster protein sequences and perform an all-vs-all comparison of protein sequences with BLASTp to identify orthologs and paralogs within the genomes. Markov cluster algorithm (MCL) (Enright et al. 2002) was then used to cluster the genomes based on the presence and absence of the accessory genes. The gene presence-absence output table from Roary was then imported into R and pairwise shared gene contents were calculated for each combination of genomes using a custom R-script (available from github.com/evelienadri/herelleviridae/tree/master). The resulting tree file was visualized using FigTree v1.4.3 (Rambaud 2007).

Single Protein 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 (Sievers et al. 2011). Resulting alignments were analyzed with ProtTest 3.4 (Darriba et al. 2011) to determine a suitable protein evolution model (only variations of models compatible with downstream software like JTT (Jones et al. 1992) and WAG (Whelan and Goldman 2001) were considered). Estimated models were used to generate phylograms with FastTree 2.1.7 (Price et al. 2010). 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 (Price et al. 2010; Darriba et al. 2011; Liu et al. 2011).

RESULTS

General Overview

To determine the phylogenetic relationship between 93 known and alleged spounavirins, we used 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 these clusters have a common origin and ought to come together under one umbrella taxon. We suggest 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 and Supplementary Table 1). The first cluster (here suggested to retain the name *Spounavirinae*) groups *Bacillus*-infecting viruses that are similar to Bacillus phage SPO1. The second cluster includes *Bacillus*-infecting viruses that resemble phage Bastille instead (named "*Bastillevirinae*" after the type species (Barylski et al. 2014)). The third cluster ("*Brockvirinae*," named in honor of Thomas D. Brock, a microbiologist known for 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, the bacteriologist who discovered prokaryotic viruses in 1915) gathers staphylococci-infected viruses that are similar to Staphylococcus phage Twort. The remaining cluster ("*Jasinskavirinae*," named in honor of Stanisława Jasińska-Lewandowska who was one of the first to study *Listeria* and its viruses) consists of viruses infecting *Listeria* that are similar to Listeria phage P100. 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.

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, Supplementary Fig. 1). The Staphylococcus viruses fell into four distinct, yet closely related groups corresponding to the established genera Twortvirus, Sep1virus, Silviavirus, and Kayvirus (Supplementary Fig. 1). 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 (Uchiyama et al. 2008a, 2008b)). 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 the Virus Classification and Tree Building Online Resource (VICTOR), a genome-BLAST distance phylogeny (GBDP) method (Supplementary Fig. 2) (Meier-Kolthoff and Göker 2017) and the Dice score (Supplementary Fig. 3), a tBLASTx-based measure that compares whole genome sequences at the amino acid level (Mizuno et al. 2013).

The patterns coalesced at a higher taxonomic level when the genomes were analyzed using tBLASTx (Supplementary Fig. 4). 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.

Table 1. Suggested new classification of the 93 spounavirins and spouna-like viruses in the new family “Herelleviridae.”[a]

Order	Family	Subfamily	Genus	Species
Caudovirales	*“Herelleviridae”*	“*Bastillevirinae*”	*Agatevirus*	*Bacillus virus Agate, Bacillus virus Bobb, Bacillus virus Bp8pC* (Bp8p-T)
			B4virus	*Bacillus virus AvesoBmore, Bacillus virus B4* (B5S), *Bacillus virus Bigbertha, Bacillus virus Riley, Bacillus virus Spock, Bacillus virus Troll*
			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* (BCU4), *Bacillus virus TsarBomba*
			Wphvirus	*Bacillus virus BPS13, Bacillus virus Hakun*), *Bacillus virus Megatron* (Eyuki), *Bacillus virus WPh, "Bacillus virus BPS10C"*
		“*Brockvirinae*”	“*Kochikohdavirus*”	“*Enterococcus virus ECP3*”, “*Enterococcus virus EF24C*” (phiEFC24C-P2), “*Enterococcus virus EFLK1*”
			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* (Romulus), *Staphylococcus virus SA11*
			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 names ratified in 2017 and later are not included. Phage isolates at the subspecies or strain level are indicated between brackets. Non-ICTV-ratified taxa are indicated between quotation marks.

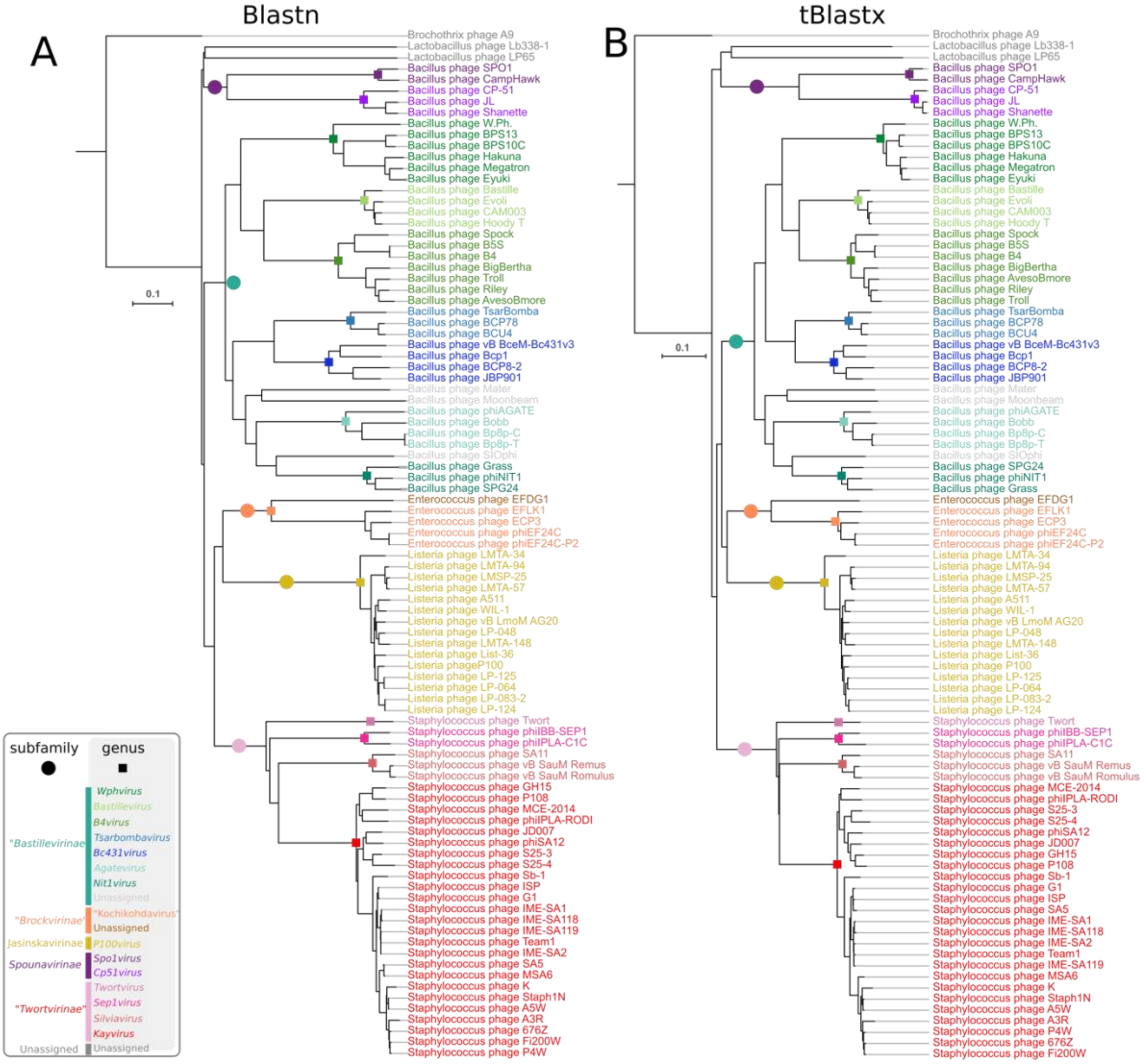

Figure 1. Genome-based clustering trees of 93 spounavirin and spouna-like viruses. A) Clustering was performed using nucleotide similarities (BLASTn) or B) translated nucleotide similarities (tBLASTx,). Genomes were compared in a pairwise fashion using Gegenees, transformed into a distance matrix, clustered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

Proteome-based Analyses

The virus proteomic tree showed five robust groupings corresponding 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 was not as well-supported as the other clusters.

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, Supplementary Table 2). Classification of the viral proteins using prokaryotic virus orthologous groups (pVOGs) showed that 38 pVOGs were shared between all 93 virus genomes (Table 2, Supplementary Table 3). 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 (Supplementary Figs. 5–6). Indeed, many genes of spounavirins and related viruses are invaded by mobile introns or inteins (Goodrich-Blair et al. 1990; Lavigne and Vandersteegen 2013). These gaps in coding sequences challenge gene prediction tools and introduce additional bias in similarity-based cluster algorithms.

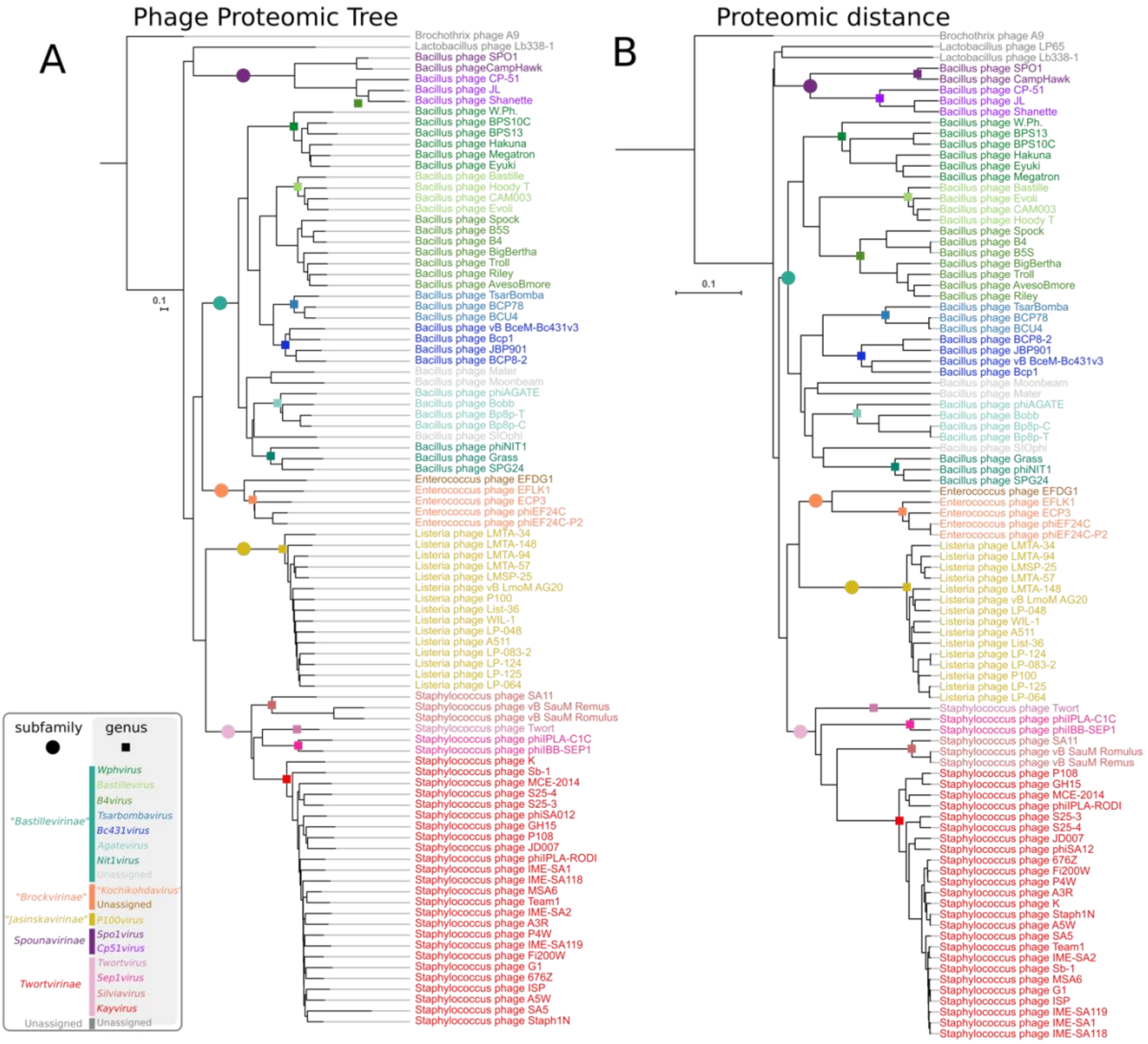

Figure 2. Predicted proteome-based clustering trees of 93 spounavirin and spouna-like viruses. A) Clustering was performed using the Phage Proteomics Tree approach and B) proteomic distance. Distances were calculated pairwise between all sets of predicted proteomes, clustered with R, and visualized using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

The pairwise comparison of the predicted proteome content of the viruses revealed a very low overall similarity at the protein level (Supplementary Fig. 7). Most 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 (Kislyuk et al. 2011). Accordingly, the genomic fluidity results followed those obtained using proteome content analysis (Supplementary Fig. 8). 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 (Supplementary Fig. 8). 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.

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 Supplementary Tables 2 and 3, respectively.

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

The pangenome of the spounavirins and spouna-like viruses (4,182 genes) as calculated using Roary (Page et al. 2015) was further analyzed by clustering the genomes based on the presence or absence of the accessory genes (Supplementary Fig. 9). 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" (Juhala et al. 2000; Krupovic et al. 2011). By clustering the spounavirin and spouna-like virus genomes based solely on the gene order of their genomes, we investigated whether gene synteny was preserved (Supplementary Fig. 10). 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. The potential exception was Bacillus phage Moonbeam (Cadungog et al. 2015). However, we did not observe the high modularity that may be expected with rampant mosaicism. The lack of considerable mosaicism supports the recent findings by Bolduc et al. that, at most, about 10% of reference virus genomes have a high degree of mosaicism (Bolduc et al. 2017). 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 Figure 3. All nine phylograms based on OPC are included in the trees in Supplementary Figure 11. 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 protein places the revised subfamily *Spounavirinae* as a subclade of "*Bastillevirinae*" and the second protein shuffled viruses from the genera *Silviavirus* and *Kayvirus*. This result may indicate that some degree of horizontal gene transfer occurs between groups, which share common hosts.

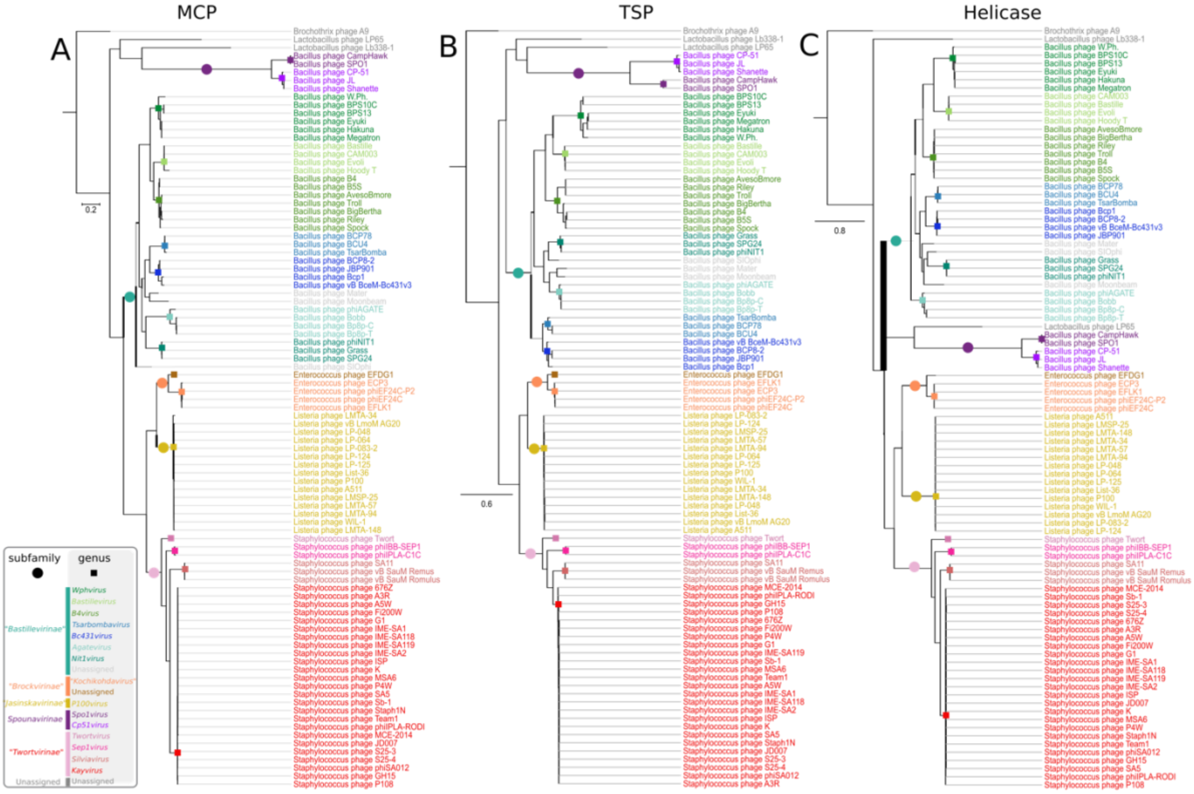

Figure 3. Phylogenetic trees based on comparisons of major protein clusters of amino acid sequences of spounavirin and spouna-like viruses. Amino acid sequences from A) the major capsid protein, B) tail sheath protein, and C) helicase were aligned with Clustal Omega, and trees were generated using FastTree 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 and suggested subfamilies are delineated with colored squares and colored circles, respectively.

DISCUSSION

Taxonomic methods must constantly develop to keep up with the ever-increasing pace of virus discovery. In the rapidly expanding field of phage studies, this requirement proved to be problematic, and although there are more than 3000 publicly available caudovirad genomes, only 873 have been officially classified by the ICTV (Davison 2017). The remaining genomes are provisionally stashed within "unclassified" bins attached to the order or associated families (Adams et al. 2017; Simmonds et al. 2017). We believe that this work is an important step toward solving the problem of these "phage orphans". This study represents the first example of a true taxonomic assessment from an "ensemble of methods". We are encouraged that the combination of genome sequence analyses, virus proteomic trees, core protein clusters, gene order genomic synteny (GOAT), and single gene phylogenies yields consistent and complementary results. Convergence of the results reasserts the usefulness of genome-based classification at a higher taxonomic rank and the ability of these methods to accommodate viral diversity.

All evidence considered, we suggest that the spounavirins should be removed from the family *Myoviridae* and given a family rank. Hence, we propose establishing a new family "*Herelleviridae*", in the order *Caudovirales* next to a smaller *Myoviridae* family. The new family would contain five subfamilies: *Spounavirinae* (sensu stricto), "*Bastillevirinae*", "*Twortvirinae*", "*Jasinkavirinae*", and "*Brockvirinae*", each comprising the genera listed in Table 1 (with additional information in S1 Table). The suggested classification corresponds well with the taxonomy of the hosts and leaves only 3% of viruses within the new family unassigned at the genus and subfamily rank. These unassigned viruses may represent clades at the genus and subfamily rank that are still under-sampled.

We believe that detachment of spounavirins from their original taxon will soon be followed by abolishment of the *Podoviridae*, *Myoviridae* and *Siphoviridae* families, in combination with the addition of new taxon ranks (e.g., class) required to accommodate the observed diversity of tailed phages. Substitution of the current families with a set of new "phylogenomic" ones will more faithfully reflect the genetic relationships of these viruses. This change does not remove the historically established virus morphotypes observed 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, taxonomically related clades can be grouped across different morphotypes. Such an approach would solve the problems of the muviruses that are suggested to be classified in the family "*Saltoviridae*" (Hulo et al. 2015) and potentially the broad set of Escherichia phage lambda-related viruses that are currently distributed among the families *Siphoviridae* and *Podoviridae* (Grose and Casjens 2014).

References

Adams M.J., Lefkowitz E.J., King A.M.Q., Harrach B., Harrison R.L., Knowles N.J., Kropinski A.M., Krupovic M., Kuhn J.H., Mushegian A.R., Nibert M.L., Sabanadzovic S., Sanfaçon H., Siddell S.G., Simmonds P., Varsani A., Zerbini F.M., Orton R.J., Smith D.B., Gorbalenya A.E., Davison A.J. 2017. 50 years of the International Committee on Taxonomy of Viruses: progress and prospects. Arch. Virol. 162:1441–1446.

Adriaenssens E.M., Edwards R., Nash J.H.E., Mahadevan P., Seto D., Ackermann H.-W., Lavigne R., Kropinski A.M. 2015. Integration of genomic and proteomic analyses in the classification of the *Siphoviridae* family. Virology. 477:144–154.

Adriaenssens E.M., Krupovic M., Knezevic P., Ackermann H.-W., Barylski J., Brister J.R., Clokie M.R.C., Duffy S., Dutilh B.E., Edwards R.A., Enault F., Jang H. Bin, Klumpp J., Kropinski A.M., Lavigne R., Poranen M.M., Prangishvili D., Rumnieks J., Sullivan M.B., Wittmann J., Oksanen H.M., Gillis A., Kuhn J.H. 2017. Taxonomy of prokaryotic viruses: 2016 update from the ICTV bacterial and archaeal viruses subcommittee. Arch. Virol. 162:1153–1157.

Ågren J., Sundström A., Håfström T., Segerman B. 2012. Gegenees: Fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. PLoS One. 7:e39107.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Barylski J., Nowicki G., Goździcka-Józefiak A. 2014. The discovery of phiAGATE, a novel phage infecting *Bacillus pumilus,* leads to new insights into the phylogeny of the subfamily *Spounavirinae*. PLoS One. 9:e86632.

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

Brister J.R., Ako-adjei D., Bao Y., Blinkova O. 2015. NCBI Viral Genomes Resource. Nucleic Acids Res. 43:D571–D577.

Cadungog J.N., Khatemi B.E., Hernandez A.C., Kuty Everett G.F. 2015. Complete genome sequence of *Bacillus megaterium* myophage Moonbeam. Genome Announc. 3:e01428-14.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421.

Charif D., Thioulouse J., Lobry J.R., Perrière G. 2005. Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinformatics. 21:545–547.

Contreras-Moreira B., Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl. Environ. Microbiol. 79:7696–7701.

Darriba D., Taboada G.L., Doallo R., Posada D. 2011. ProtTest 3: Fast selection of best-fit models of protein evolution. Bioinformatics. 27:1164–1165.

Davison A.J. 2017. Journal of General Virology – Introduction to "ICTV Virus Taxonomy Profiles." J. Gen. Virol. 98:1–1.

Development Core Team R. 2008. R: A language and environment for statistical computing.

Available from http://www.r-project.org.

Enright A.J., Van Dongen S., Ouzounis C.A. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575–1584.

Felsenstein J. 1989. PHYLIP - Phylogeny inference package - v3.2. Cladistics. 5:164–166.

Finn R.D., Clements J., Eddy S.R. 2011. HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res. 39:29–37.

Frickey T., Lupas A. 2004. CLANS: A Java application for visualizing protein families based on pairwise similarity. Bioinformatics. 20:3702–3704.

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

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14:685–695.

Göker M., García-Blázquez G., Voglmayr H., Tellería M.T., Martín M.P. 2009. Molecular taxonomy of phytopathogenic fungi: A case study in Peronospora. PLoS One. 4:8–10.

Goodrich-Blair H., Scarlato V., Gott J.M., Xu M.-Q., Shub D.A. 1990. A self-splicing group I intron in the DNA polymerase gene of *Bacillus subtilis* bacteriophage SPO1. Cell. 63:417–424.

Grazziotin A.L., Koonin E. V, Kristensen D.M. 2017. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. 45:D491–D498.

Grose J.H., Casjens S.R. 2014. Understanding the enormous diversity of bacteriophages: The tailed phages that infect the bacterial family *Enterobacteriaceae*. Virology. 468–470:421–443.

Hulo C., Masson P., Le Mercier P., Toussaint A. 2015. A structured annotation frame for the transposable phages: A new proposed family “Saltoviridae” within the *Caudovirales*.

Virology. 477:155–163.

Hyatt D., Chen G.L., LoCascio P.F., Land M.L., Larimer F.W., Hauser L.J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 11.

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

Jones D.T., Taylor W.R., Thornton J.M. 1992. The rapid generation of mutation data matrices from protein sequences. Bioinformatics. 8:275–282.

Juhala R.J., Ford M.E., Duda R.L., Youlton A., Hatfull G.F., Hendrix R.W. 2000. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J. Mol. Biol. 299:27–51.

Karsch-Mizrachi I., Nakamura Y., Cochrane G. 2012. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 40:D33–D37.

Kislyuk A.O., Haegeman B., Bergman N.H., Weitz J.S. 2011. Genomic fluidity: An integrative view of gene diversity within microbial populations. BMC Genomics. 12:32.

Klumpp J., Lavigne R., Loessner M.J., Ackermann H.-W. 2010. The SPO1-related bacteriophages. Arch. Virol. 155:1547–61.

Konstantinidis K.T., Tiedje J.M. 2005. Towards a genome-based taxonomy for prokaryotes. J. Bacteriol. 187:6258–6264.

Kristensen D.M., Kannan L., Coleman M.K., Wolf Y.I., Sorokin A., Koonin E. V, Mushegian A. 2010. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. Bioinformatics. 26:1481–7.

Krupovic M., Dutilh B.E., Adriaenssens E.M., Wittmann J., Vogensen F.K., Sullivan M.B., Rumnieks J., Prangishvili D., Lavigne R., Kropinski A.M., Klumpp J., Gillis A., Enault F., Edwards R.A., Duffy S., Clokie M.R.C., Barylski J., Ackermann H.-W., Kuhn J.H.

2016. Taxonomy of prokaryotic viruses: update from the ICTV bacterial and archaeal viruses subcommittee. Arch. Virol. 161:1095–1099.

Krupovic M., Prangishvili D., Hendrix R.W., Bamford D.H. 2011. Genomics of bacterial and archaeal viruses: Dynamics within the prokaryotic virosphere. Microbiol. Mol. Biol. Rev. 75:610–635.

Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947–2948.

Lavigne R., Darius P., Summer E.J., Seto D., Mahadevan P., Nilsson A.S., Ackermann H.W., Kropinski A.M. 2009. Classification of *Myoviridae* bacteriophages using protein sequence similarity. BMC Microbiol. 9:224.

Lavigne R., Seto D., Mahadevan P., Ackermann H.-W., Kropinski A.M. 2008. Unifying classical and molecular taxonomic classification: analysis of the *Podoviridae* using BLASTP-based tools. Res. Microbiol. 159:406–414.

Lavigne R., Vandersteegen K. 2013. Group I introns in *Staphylococcus* bacteriophages. Future Virol. 8:997–1005.

Lefort V., Desper R., Gascuel O. 2015. FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. Mol. Biol. Evol. 32:2798–2800.

Letunic I., Bork P. 2007. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. Bioinformatics. 23:127–128.

Liu K., Linder C.R., Warnow T. 2011. RAxML and FastTree: Comparing two methods for large-scale maximum likelihood phylogeny estimation. PLoS One. 6:e27731.

Meier-Kolthoff J.J.P., Hahnke R.L., Petersen J., Scheuner C., Michael V., Fiebig A., Rohde C., Rohde M., Fartmann B., Goodwin L.A., Chertkov O., Reddy T.B.K., Pati A., Ivanova N.N., Markowitz V.V., Kyrpides N.C.N., Woyke T., Göker M., Klenk H.-P.H.-

P.H., Pagani I., Liolios K., Jansson J., Chen I., Smirnova T., Nosrat B., Markowitz V.V., Kyrpides N.C.N., Lapage S., Sneath P., Lessel E., Skerman V., Seeliger H., Clark W., Blattner F., Plunkett G., Bloch C., Perna N., Burland V., Riley M., Vides J.C., Glasner J., Rode C., Mayhew G., Gregor J., Davis N., Kirkpatrick H., Goeden M., Rose D., Mau B., Shao Y., Wu D., Hugenholtz P., Mavromatis K., Pukall R., Dalin E., Ivanova N.N., Kunin V., Goodwin L.A., Wu M., Tindall B., Hooper S., Pati A., Lykidis A., Spring S., Anderson I., D P., Escherich T., Skerman V., McGowan V., Sneath P., Kauffmann F., Ã˜rskov F., Ã˜rskov I., Filannino P., Azzi L., Cavoski I., Vincentini O., Rizzello C., Gobbetti M., Cagno R. Di, Schumann P., Pukall R., Farnleitner A., Kreuzinger N., Kavka G., Grillenberger S., Rath J., Mach R., Tee T., Chowdhury A., Maranas C., Shanks J., Wen M., Bond-Watts B., Chang M., Rosano G., Ceccarelli E., Donovan C., Bramkamp M., Kuzminov A., Kang Z., Zhang C., Zhang J., Jin P., Zhang J., Du G., Chen J., Whitfield C., Roberts I., Cooper K., Mandrell R., Louie J., Korlach J., Clark T., Parker C., Huynh S., Chain P., Ahmed S., Carter M., Allocati N., Masulli M., Alexeyev M., Ilio C. Di, Kaper J., Nataro J., Mobley H., Auch A., Jan M. Von, Klenk H.-P.H.-P.H., Göker M., Meier-Kolthoff J.J.P., Auch A., Klenk H.-P.H.-P.H., Göker M., Meier-Kolthoff J.J.P., Klenk H.-P.H.-P.H., Göker M., Göker M., Cleland D., Saunders E., Lapidus A., Nolan M., Lucas S., Hammon N., Deshpande S., Cheng J.-F., Tapia R., Han C., Goodwin L.A., Pitluck S., Liolios K., Pagani I., Ivanova N.N., Mavromatis K., Pati A., Chen A., Palaniappan K., Land M., Hauser L., Chang Y.Y.-J., Jeffries C., Detter J., Beck B., Woyke T., Bristow J., Eisen J., Markowitz V.V., Welch R., Scheutz F., Strockbine N., Koser S., Topley W., Wilson G., Huys G., Cnockaert M., Janda J., Swings J., Field D., Garrity G., Gray T., Morrison N., Selengut J., Sterk P., Tatusova T., Thomson N., Allen M., Angiuoli S., Ashburner M., Axelrod N., Baldauf S., Ballard S., Boore J., Cochrane G., Cole J., Dawyndt P., Vos P. De, Pamphilis C. de, Edwards R., Faruque N., Feldman R., Gilbert J., Gilna P., Glockner F., Goldstein P., Guralnick R., Haft D., Hancock D., Field D., Amaral-Zettler L., Cochrane G., Cole J., Dawyndt P., Garrity G., Gilbert J., Glöckner F., Hirschman L., Karsch-Mizrachi I., Klenk H.-P.H.-P.H., Knight R., Kottmann R., Kyrpides N.C.N., Meyer F., Gil I.S., Sansone S.-A., Schriml L., Sterk P., Tatusova T., Ussery D., White O., Wooley J., Woese C., Kandler O., Weelis M., Garrity G., Bell J., Lilburn T., Garrity G., Bell J., Lilburn T., Williams K., Kelly D., Brenner D., Castellani A., Chalmers A., Ashburner M., Ball C., Blake J., Botstein D., Butler H., Cherry J., Davis A., Dolinski K., Dwight S., Eppig J., Harris M., Hill D., Issel-Tarver L., Kasarskis A., Lewis S., Matese J., Richardson J., Ringwald M., Rubin G., Sherlock G., Consortium G., Vaas L., Sikorski J., Michael V., Göker M., Klenk H.-P.H.-P.H., Vaas L., Sikorski J., Hofner B., Fiebig A., Buddruhs N., Klenk H.-P.H.-P.H., Göker M., Chang Y.Y.-J., Feingold D., Boer H., Maaheimo H., Koivula A., Penttila M., Richard P., Xiao Z., Xu P., Göker M., Klenk H.-P.H.-P.H., Mavromatis K., Land M., Brettin T., Quest D., Copeland A., Clum A., Goodwin L.A., Woyke T., Lapidus A., Klenk H.-P.H.-P.H., Cottingham R., Kyrpides N.C.N., Markowitz V.V., I-M A.C., Palaniappan K., Chu K., Szeto E., Grechkin Y., Ratner A., Jacob B., Huang J., Williams P., Huntemann M., Anderson I., Mavromatis K., Ivanova N.N., Kyrpides N.C.N., Gemeinholzer B., Dröge G., Zetzsche H., Haszprunar G., Klenk H.-P.H.-P.H., Güntsch A., Berendsohn W., Wägele J., Zerbino D., Birney E., Gordon D., Abajian C., Green P., Hyatt D., Chen G., LoCascio P., Land M., Larimer F., Hauser L., Mavromatis K., Ivanova N.N., Chen I., Szeto E., Markowitz V.V., Kyrpides N.C.N., Finn D., Clements J., Eddy S., Lowe T., Eddy S., Nawrocki E., Kolbe D., Eddy S., Markowitz V.V., Ivanova N.N., Chen I., Chu K., Kyrpides N.C.N., Bland C., Ramsey T., Sabree F., Lowe M., Brown K., Kyrpides N.C.N., Hugenholtz P., Edgar R., Wayne L., Brenner D., Colwell R., Grimont P., Kandler O., Krichevsky M., Moore L., Moore W., Murray R., Stackebrandt E., Starr M., Truper H., Tindall B., Rosselló-Móra R., Busse H., Ludwig W., Kämpfer P., Kaas R., Friis C., Ussery D., Aarestrup F., Clermont O., Bonacorsi S., Bingen E., Clermont O., Gordon D., Brisse S., Walk S., Denamur E., Clermont O., Christenson J., Denamur E., Gordon D., Sahl J., Morris C., Rasko D., Patil K., McHardy A., Thorne J., Kishino H., Meier-Kolthoff J.J.P., Auch A., Klenk H.-P.H.-P.H., Göker M., Letunic I., Bork P., Desper R., Gascuel O., Lukjancenko O., Wassenaar T., Ussery D., Touchon M., Hoede C., Tenaillon O., Barbe V., Baeriswyl S., Bidet P., Bingen E., Bonacorsi S., Bouchier C., Bouvet O., Calteau A., Chiapello H., Clermont O., Cruveiller S., Danchin A., Diard M., Dossat C., Karoui M. El, Frapy E., Garry L., Ghigo J., Gilles A., Johnson J., Bouguenec C. Le, Lescat M., Mangenot S., Martinez-Jehanne V., Matic I., Nassif X., Oztas S., Zuo G., Xu Z., Hao B., Abt B., Han C., Scheuner C., Lu M., Lapidus A., Nolan M., Lucas S., Hammon N., Deshpande S., Cheng J.-F., Tapia R., Goodwin L.A., Pitluck S., Mavromatis K., Mikhailova N., Huntemann M., Pati A., Chen A., Palaniappan K., Land M., Hauser L., Brambilla E.-M., Rohde M., Spring S., Gronow S., Göker M., Woyke T., Bristow J., Eisen J., Markowitz V.V., Abt B., Göker M., Scheuner C., Han C., Lu M., Misra M., Lapidus A., Nolan M., Lucas S., Hammon N., Deshpande S., Chang J.-F., Tapia R., Goodwin L.A., Pitluck S., Liolios K., Pagani I., Ivanova N.N., Mavromatis K., Mikhailova N., Huntemann M., Pati A., Chen A., Palaniappan K., Land M., Hauser L., Brambilla E.-M., Rohde M., Spring S., Gronow S., Anderson I., Scheuner C., Göker M., Mavromatis K., Hooper S., Porat I., Klenk H.-P.H.-P.H., Ivanova N.N., Kyrpides N.C.N., Frank O., Pradella S., Rohde M., Scheuner C., Klenk H.-P.H.-P.H., Göker M., Petersen J., Göker M., Scheuner C., Klenk H.-P.H.-P.H., Stielow J., Menzel W., Spring S., Scheuner C., Lapidus A., Lucas S., Rio T. Del, Tice H., Copeland A., Cheng J.-F., Chen F., Nolan M., Saunders E., Pitluck S., Liolios K., Ivanova N.N., Mavromatis K., Lykidis A., Pati A., Chen A., Palaniappan K., Land

M., Hauser L., Chang Y.Y.-J., Jeffries C., Goodwin L.A., Detter J., Brettin T., Rohde M., Göker M., Woyke T., Bristow J., Stackebrandt E., Scheuner C., Göker M., Schumann P., Verbarg S., Göker M., Scheuner S., Schumann P., Stackebrandt E., Altschul S., Madden T., Schaffer A., Zhang J., Zhang Z., Miller W., Lipman D., Li L., Stoeckert C., Roos D., Edgar R., Thompson J., Thierry J.-C., Poch O., Castresana J., Meusemann K., Reumont B. von, Simon S., Roeding F., Strauss S., Kuck P., Ebersberger I., Walzl M., Pass G., Breuers S., Achter V., Haeseler A. von, Burmester T., Hadrys H., Wagele J., Misof B., Felsenstein J., Fitch W., Goloboff P., Stamatakis A., Pattengale N., Alipour M., Bininda-Emonds O., Moret B., Stamatakis A., Swofford D., Klenk H.-P.H.-P.H., Göker M., Enright A., Dongen S. van, Ouzounis C., Albuquerque L., Rainey F., Nobre M.F., Costa M. da, Fricke W., McDermott P., Mammel M., Zhao S., Johnson T., Rasko D., Fedorka-Cray P., Pedroso A., Whichard J., Leclerc J., White D., Cebula T., Ravel J., Brinkkötter A., Klöss H., Alpert C., Lengeler J., Göker M., Garcáa-BlÃ¡zquez G., Voglmayr H., Telleráa M., Martán M., Staley J., Krieg N., Tindall B., Kampfer P., Euzeby J., Oren A., Lan R., Reeves P. 2014. Complete genome sequence of DSM 30083T, the type strain (U5/41T) of *Escherichia coli*, and a proposal for delineating subspecies in microbial taxonomy. Stand. Genomic Sci. 9:2.

Meier-Kolthoff J.P., Göker M. 2017. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. Bioinformatics. 33:3396–3404.

Mizuno C.M., Rodriguez-Valera F., Kimes N.E., Ghai R. 2013. Expanding the marine virosphere using metagenomics. PLoS Genet. 9:e1003987.

Nishimura Y., Watai H., Honda T., Mihara T., Omae K., Roux S., Blanc-Mathieu R., Yamamoto K., Hingamp P., Sako Y., Sullivan M.B., Goto S., Ogata H., Yoshida T. 2017. Environmental viral genomes shed new light on virus-host interactions in the ocean. mSphere. 2:e00359-16.

O'Leary N.A., Wright M.W., Brister J.R., Ciufo S., Haddad D., McVeigh R., Rajput B., Robbertse B., Smith-White B., Ako-Adjei D., Astashyn A., Badretdin A., Bao Y., Blinkova O., Brover V., Chetvernin V., Choi J., Cox E., Ermolaeva O., Farrell C.M., Goldfarb T., Gupta T., Haft D., Hatcher E., Hlavina W., Joardar V.S., Kodali V.K., Li W., Maglott D., Masterson P., McGarvey K.M., Murphy M.R., O'Neill K., Pujar S., Rangwala S.H., Rausch D., Riddick L.D., Schoch C., Shkeda A., Storz S.S., Sun H., Thibaud-Nissen F., Tolstoy I., Tully R.E., Vatsan A.R., Wallin C., Webb D., Wu W., Landrum M.J., Kimchi A., Tatusova T., DiCuccio M., Kitts P., Murphy T.D., Pruitt K.D. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44:D733–D745.

Paez-Espino D., Eloe-Fadrosh E.A., Pavlopoulos G.A., Thomas A.D., Huntemann M., Mikhailova N., Rubin E., Ivanova N.N., Kyrpides N.C. 2016. Uncovering Earth's virome. Nature. 536:425–430.

Page A.J., Cummins C.A., Hunt M., Wong V.K., Reuter S., Holden M.T.G., Fookes M., Falush D., Keane J.A., Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 31:3691–3693.

Price M.N., Dehal P.S., Arkin A.P. 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One. 5:e9490.

Rambaud. 2007. FigTree. Available from http://tree.bio.ed.ac.uk/software/figtree/.

Rohwer F., Edwards R. 2002. The Phage Proteomic Tree: a genome-based taxonomy for phage. J. Bacteriol. 184:4529–4535.

Roux S., Brum J.R., Dutilh B.E., Sunagawa S., Duhaime M.B., Loy A., Poulos B.T., Solonenko N., Lara E., Poulain J., Pesant S., Kandels-Lewis S., Dimier C., Picheral M., Searson S., Cruaud C., Alberti A., Duarte C.M., Gasol J.M., Vaqué D., Bork P., Acinas S.G., Wincker P., Sullivan M.B. 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature. 537:689–693.

Schuch R., Fischetti V.A. 2009. The secret life of the anthrax agent Bacillus anthracis: Bacteriophage-mediated ecological adaptations. PLoS One. 4:e6532.

Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 30:2068–2069.

Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson J.D., Higgins D.G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7:539.

Simmonds P., Adams M.J., Benkő M., Breitbart M., Brister J.R., Carstens E.B., Davison A.J., Delwart E., Gorbalenya A.E., Harrach B., Hull R., King A.M., Koonin E. V, Krupovic M., Kuhn J.H., Lefkowitz E.J., Nibert M.L., Orton R., Roossinck M.J., Sabanadzovic S., Sullivan M.B., Suttle C.A., Tesh R.B., van der Vlugt R.A., Varsani A., Zerbini F.M. 2017. Consensus statement: Virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. 15:161–168.

Söding J., Biegert A., Lupas A.N. 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33:W244-2488.

Uchiyama J., Rashel M., Maeda Y., Takemura I., Sugihara S., Akechi K., Muraoka A., Wakiguchi H., Matsuzaki S. 2008a. Isolation and characterization of a novel *Enterococcus faecalis* bacteriophage phiEF24C as a therapeutic candidate. FEMS Microbiol. Lett. 278:200–206.

Uchiyama J., Rashel M., Takemura I., Wakiguchi H., Matsuzaki S. 2008b. In silico and in vivo evaluation of bacteriophage phiEF24C, a candidate for treatment of *Enterococcus faecalis* infections. Appl. Environ. Microbiol. 74:4149–4163.

Whelan S., Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–699.

Yuan Y., Peng Q., Wu D., Kou Z., Wu Y., Liu P., Gao M. 2015. Effects of actin-like proteins encoded by two *Bacillus pumilus* phages on unstable lysogeny, revealed by genomic analysis. Appl. Environ. Microbiol. 81:339–350.

FUNDING

This work was supported by the National Science Centre, Poland (grant number 2016/23/D/NZ2/00435) to JB; the Netherlands Organization for Scientific Research (NWO) (Vidi grant number 864.14.004) to MBPS and BED; the US National Science Foundation (grant numbers DUE-132809 and MCB-1330800) to RAE; the University of Helsinki funding for Instruct research infrastructure (Virus and Macromolecular Complex Production, ICVIR) to HMO; the Chargé de Recherches fellowship from the National Fund for Scientific Research, FNRS, Belgium to AG; the EUed Horizon 2020 Framework Programme for Research and Innovation, 'Virus-X' (grant number 685778) to FE; the Gordon and Betty Moore Foundation Investigator Award (grant number GBMF#3790) to MBS; the Battelle Memorial Institute's prime contract with the US National Institute of Allergy and Infectious Diseases (NIAID) (Contract number HHSN272200700016I) to JHK. The content of this publication does not necessarily reflect the views or policies of the US Department of Health and Human Services or of the institutions and companies affiliated with the authors.

ACKNOWLDEGMENTS

We thank Laura Bollinger, Integrated Research Facility at Fort Detrick, for technical writing services. E.M.A. would like to thank P.H. Nel for assistance with R scripting.