By the end of 2018, nearly 8000 complete tailed phage genomes were published online and a further 22,000 partial genomes were stored in databases gathered under the umbrella of the International Nucleotide Sequence Database Collaboration (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) have focused on classifying newly described phages into species and genera (Lavigne et al. 2008, 2009; Adriaenssens et al. 2013; Krupovic et al. 2016; Adriaenssens et al. 2017). However, once our attention shifted toward higher order relationships, we found that the ranks currently used in phage taxonomy (species, genus, subfamily, family, and order) are no longer sufficient for the description of phage
diversity. The limitation is particularly acute in the case of the order Caudovirales—arguably the most abundant and heterogeneous group of viruses (Paez-Espino et al. 2016; Roux et al. 2016; Nishimura et al. 2017a). Indeed, the diversity of caudoviruses surpasses that of any other virus taxon. A recent analysis of the gene content of the dsDNA virosphere demonstrated that the global network of dsDNA viruses consists of at least 19 modules, 11 of which correspond to caudoviruses (Iranzo et al. 2016). Each of the eight remaining modules encompasses one or more families of eukaryotic or archaeal viruses. Consequently, each of the 11 caudovirus modules could be considered a separate family. Despite this remarkable diversity, only one family, the Caudovirales, was formally recognized as related to it. When this subfamily was first described (Lavigne et al. 2009), the unifying characteristics of its members included: the hosts belong to the bacterial phylum Firmicutes; strictly virulent lifestyle; myovirion morphology (i.e., icosahedral capsid and long contractile tail); terminally redundant, nonpermuted dsDNA genome of 127–157 kbp in length; and “considerable amino acid similarity” (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 inclusion into the taxon. Since the initial description of the subfamily, the number of its members has grown significantly, and its taxonomic structure has been contested several times (Klumpp et al. 2010; Barylski et al. 2014; Iranzo et al. 2016; Krupovic et al. 2016; Bolduc et al. 2017a; Adriaenssens et al. 2017). Thus, we wanted not only to delineate a new family but also resolve its internal structure.

Unfortunately, there is no one-size-fits-all method for the classification of viruses at all taxonomic ranks. Virus taxonomy has always suffered from the lack of universal marker genes that could be used for phylogenetic reconstruction of the evolutionary relationships. Additionally, differing mutation rates between viral lineages, horizontal gene transfer, and genomic mosaicism limit usefulness of many of the available phylogenetic and phylogenomic methods that have become the gold standard in evolutionary biology (Davidson et al. 2015; Meier-Kolthoff and Göker 2017). Thus, our strategy for reclassification included a plethora of classification tools that employ very different approaches. Our analyses ranged from coarse-grained, high-throughput, holistic clustering methods where similarity is computed from comparison of all viral genes [vContact, GRAVITY (Bolund et al. 2017a; Aiewsakun et al. 2018; Aiewsakun and Simmonds 2018)] to detailed genome and proteome comparisons [Victor, Dice, GOAT and Phage Proteomic Tree (Rohwer and Edwards 2002; Mizuno et al. 2013; Meier-Kolthoff and Göker 2017)] and individual gene phylogenies [IQtree (Nguyen et al. 2015)]. This multifaceted approach allowed us to gradually descend from the definition of the new family to the study of its internal structure. Interestingly, despite the diversity of the applied methods their results turned out to be complementary and predominantly concordant. All methods painted a robust picture of the new family as a distinct and diverse taxon and supported the same general scheme for its structure (Table 1).

We emphasize that this reclassification is an essential step in the larger revision of the taxonomy of the order Caudovirales. The final goal of our group is a novel system that appropriately accommodates the genomic diversity of prokaryotic viruses and is consistent with taxonomy of eukaryotic viruses (Aiewsakun et al. 2018; Simmonds and Aiewsakun 2018).

**Materials and Methods**

For brevity and clarity’s sake, only the basic principles of previously published methods are summarized in the following section. A detailed description of each method used in this study can be found in Supplementary File 1 available on Dryad at http://dx.doi.org/10.5061/dryad.106q6g6.

### Creation of the “Herelleviridae” Data Set

Genome sequences of known spounaviruses were retrieved from the GenBank or (preferably) RefSeq databases based on literature data, and taxonomic classifications provided by the ICTV and the National Center for Biotechnology Information (NCBI). Records representing genomes of candidate spouna-related viruses were retrieved by searching the same databases with the tBLASTn algorithm (Altschul et al. 1990) using as queries terminase and major capsid proteins of type isolates of the original subfamily (Brister et al. 2015). After manual curation, the search yielded a set of 93 virus genomes (Supplementary Table S1.1 available on Dryad), which were reannotated using PROKKA (Seemann 2014) and used in the following analyses.
the reannotated genomes are available from Github.

capable of discerning relations even between divergent
phages, by using two cutting-edge virus clustering tools
is a clearly distinct group of viruses within the dsDNA

In Supplementary Table S1 available on Dryad) and
from NCBI (accession number information listed
on Dryad).

Metadata Resource (Supplementary Table S1.2 available
from the ICTV 2016 Master Species List 31V1.1 and Virus
additional genome set including well-described viruses

level are indicated between brackets.

Genera were renamed in 2018, taxonomy proposal 2018.007B.
The species listed here represent the 93 genome data set on which all analyses have been performed. Phage isolates at the subspecies or strain

To conduct interfamilial comparisons, we compiled an
additional genome set including well-described viruses
from the ICTV 2016 Master Species List 31V1.1 and Virus
Metadata Resource (Supplementary Table S1.2 available
on Dryad).

All original genome sequences are available from NCBI (accession number information listed in
Supplementary Table S1 available on Dryad) and the reannotated genomes are available from Github (github.com/evelenadr1/herelleviridae).

Definition of the New Herelleviridae Family Within the
dsDNA Virosphere

We examined whether or not the family, Herelleviridae,
is a clearly distinct group of viruses within the dsDNA
phages, by using two cutting-edge virus clustering tools
capable of discerning relations even between divergent taxa.

Using vConTACT v2.0, we constructed a monopartite
network of viral genomes by clustering gene families
based on BLAST hits between their protein products
as previously described (Bolduc et al. 2017a; Jang et al.
2019). In this framework, similarities between pairs of
genomes were calculated as a function of the shared
products and genome organization models calculated
profile hidden Markov models of the predicted gene

Using vConTACT v2.0, we constructed a monopartite
network of viral genomes by clustering gene families
based on BLAST hits between their protein products
as previously described (Bolduc et al. 2017a; Jang et al.
2019). In this framework, similarities between pairs of
genomes were calculated as a function of the shared
products and genome organization models calculated
profile hidden Markov models of the predicted gene

To conduct interfamilial comparisons, we compiled an
additional genome set including well-described viruses
from the ICTV 2016 Master Species List 31V1.1 and Virus
Metadata Resource (Supplementary Table S1.2 available
on Dryad).

All original genome sequences are available from NCBI (accession number information listed in
Supplementary Table S1 available on Dryad) and the reannotated genomes are available from Github (github.com/evelenadr1/herelleviridae).

Table 1. New classification of the 93 spounaviruses and spouna-like viruses in the new family Herelleviridae.

| Family        | Subfamily            | Genusa  | Speciesb |
|---------------|----------------------|---------|----------|
| Herelleviridae| Bastillevirinae      | Agateirus| Bacillus virus Agate, Bacillus virus BoBb, Bacillus virus Bp9pC (Bp9p-T) |
|               | Bequistrovirus (formerly B4evirus) |         | Bacillus virus ArosoBoore, Bacillus virus B4 (BOS), Bacillus virus Bigger bash, Bacillus virus Riley, Bacillus virus Speck, Bacillus virus Tull |
|               | Bastillevirinae      |         | Bacillus virus Bastille, Bacillus virus CAM003, Bacillus virus Eredi, Bacillus virus HaedAT |
|               | Carvellevirinae (formerly Be-431virus) |         | Bacillus virus Be431L, Bacillus virus Bep1, Bacillus virus BCPB2, Bacillus virus JBP96H |
|               | Nistauvirinae (formerly Nittievirus) |         | Bacillus virus Gras, Bacillus virus NIT1, Bacillus virus SPG24 |
|               | Turrombahvirinae     |         | Bacillus virus BCP78 (BCU4), Bacillus virus Turrombah |
|               | Wp8virinae           |         | Bacillus virus BPS13, Bacillus virus Hakama, Bacillus virus Mogetron (Eyak), Bacillus virus Wp8, Bacillus virus BPS18C |
|               | Unassigned           | Kochohovirus | Bacillus virus Mater, Bacillus virus Moonbeam, Bacillus virus SIOphle |
|               | Brockvirinae         |         | Enterococcus virus ECPA, Enterococcus virus EF24C (phaEFC24C-P2), Enterococcus virus EFLK1 |
|               | Jasinskavirinae      |         | Enteroxoccus virus EFDG1 |
|               | Unassigned           | Pecentomovirus (formerly P100virus) | Listeria virus AS1L, Listeria virus P100, Listeria virus List6, Listeria virus LMSF25 (LMTA-97, LMTA-94), Listeria virus LMTA44B, Listeria virus LMTA34, Listeria virus LPO65, Listeria virus LP06 (LP-125), Listeria virus LP08-2 (LP-124), Listeria virus AC20, Listeria virus WIL |
| Spounavirinae  | Simonovitchvirinae (formerly Cy52virus) |         | Bacillus virus CPS1, Bacillus virus IL, Bacillus virus Shanette |
|               | Okuhovirinae (formerly Sp1virus) |         | Bacillus virus Camphank, Bacillus virus SP101 |
| Twortvirinae   | Kageirus             |         | Staphylococcus virus G1, Staphylococcus virus G15, Staphylococcus virus J01, Staphylococcus virus K, Staphylococcus virus MCE2014, Staphylococcus virus P108, Staphylococcus virus Rodi, Staphylococcus virus S523, Staphylococcus virus S52-4, Staphylococcus virus SA2, Staphylococcus virus SM (692, A3R, A5W, F200W, IME-SA1, IME-SA18, IME-SA19, IME-SA2, ISP, MS6A, F4W, S4A, Staph.IN, Team1) |
|               | Sillianirus          |         | Staphylococcus virus Kenus (Romulus), Staphylococcus virus SA11 |
|               | Segunavirinae (formerly Sep1virus) |         | Staphylococcus virus IPLAC1C, Staphylococcus virus SEP1 |
|               | Tuortervirus         |         | Staphylococcus virus Teort |
|               | Unassigned           |         | Lactobacillus virus L6338 |
|               | Unassigned           |         | Lactobacillus virus LP65 |
|               | Unassigned           |         | Brochobrix virus A9 |

aGenera were renamed in 2018, taxonomy proposal 2018.007B.
bThe species listed here represent the 93 genome data set on which all analyses have been performed. Phage isolates at the subspecies or strain level are indicated between brackets.

Bacillus virus Bc431. Bacillus virus BCP82.
Bacillus virus JRP96H.
Bacillus virus Gras, Bacillus virus NIT1, Bacillus virus SPG24.
Bacillus virus BCP78 (BCU4), Bacillus virus Turrombah.
Bacillus virus BPS13, Bacillus virus Hakama, Bacillus virus Mogetron (Eyak), Bacillus virus Wp8, Bacillus virus BPS18C.
Bacillus virus Mater, Bacillus virus Moonbeam, Bacillus virus SIOphle.
Enterococcus virus ECPA, Enterococcus virus EF24C (phaEFC24C-P2), Enterococcus virus EFLK1.
Enteroxoccus virus EFDG1.
Listeria virus AS1L, Listeria virus P100, Listeria virus List6, Listeria virus LMSF25 (LMTA-97, LMTA-94), Listeria virus LMTA44B, Listeria virus LMTA34, Listeria virus LPO65, Listeria virus LP06 (LP-125), Listeria virus LP08-2 (LP-124), Listeria virus AC20, Listeria virus WIL.
Bacillus virus CPS1, Bacillus virus IL, Bacillus virus Shanette.
Bacillus virus Camphank, Bacillus virus SP101.
Staphylococcus virus G1, Staphylococcus virus G15, Staphylococcus virus J01, Staphylococcus virus K, Staphylococcus virus MCE2014, Staphylococcus virus P108, Staphylococcus virus Rodi, Staphylococcus virus S523, Staphylococcus virus S52-4, Staphylococcus virus SA2, Staphylococcus virus SM (692, A3R, A5W, F200W, IME-SA1, IME-SA18, IME-SA19, IME-SA2, ISP, MS6A, F4W, S4A, Staph.IN, Team1).
Staphylococcus virus Kenus (Romulus), Staphylococcus virus SA11.
Staphylococcus virus IPLAC1C, Staphylococcus virus SEP1.
Staphylococcus virus Teort.
Lactobacillus virus L6338.
Lactobacillus virus LP65.
Brochobrix virus A9.

To conduct interfamilial comparisons, we compiled an
additional genome set including well-described viruses
from the ICTV 2016 Master Species List 31V1.1 and Virus
Metadata Resource (Supplementary Table S1.2 available
on Dryad).

All original genome sequences are available from NCBI (accession number information listed in
Supplementary Table S1 available on Dryad) and the reannotated genomes are available from Github (github.com/evelenadr1/herelleviridae).

Definition of the New Herelleviridae Family Within the
dsDNA Virosphere

We examined whether or not the family, Herelleviridae,
is a clearly distinct group of viruses within the dsDNA
phages, by using two cutting-edge virus clustering tools
capable of discerning relations even between divergent taxa.

Using vConTACT v2.0, we constructed a monopartite
network of viral genomes by clustering gene families
based on BLAST hits between their protein products
as previously described (Bolduc et al. 2017a; Jang et al.
2019). In this framework, similarities between pairs of
genomes were calculated as a function of the shared
protein families. The network was visualized with
Cytoscape (version 3.5.1; http://cytoscape.org/) with
networks sharing more proteins clustered more closely
together (detailed information in Supplementary File 1
available on Dryad).

The second method used is 'Genome Relationships
Applied to Virus Taxonomy' or GRAVITy (GitHub:
Paewsakun/GRAVITy [Aiewsakun et al. 2018; Aiewsakun and Simmonds 2018]). This framework
created a dendrogram of viruses, based on protein
profile hidden Markov models of the predicted gene
products and genome organization models calculated
into a composite generalized Jaccard (CGJ) score
representing the difference between two viruses on a
scale from 1 to 0 (detailed information in Supplementary File 1 available on Dryad).

We also investigated the clustering of the family within the Caudovirales order on the VIPtree server (Nishimura et al. 2017b), which uses the Phage Proteomic Tree approach described below and detailed in in Supplementary File 1 available on Dryad.

**Exploration of the Intrafamilial Relationship**

After demarcation of the family, we proceeded with analysis of its internal structure, using the defined set of 93 genomes described above. In the process, we compared a collection of the classification tools, gathering the phylogenetic signal from the different types of data (whole genome sequences, complete proteomes, marker genes, and gene order).

**Genome-Based Analyses**

Nucleotide sequence-based grouping of phages was conducted using VICTOR (Virus Classification and Tree Building Online Resource), a Genome-BLAST Distance Phylogeny (GBDP) method (Meier-Kolthoff et al. 2014; Meier-Kolthoff and Göker 2017). The program calculates intergenomic distances from BLAST+ hits using GBDP (including 100 pseudobootstrap replicates) and used them to infer a balanced minimum evolution tree with branch support via FASTME including subtree pruning and regrafting postprocessing (for details of the algorithm design, see Meier-Kolthoff et al. 2014; Meier-Kolthoff and Göker 2017). The analysis was conducted under settings recommended for prokaryotic viruses.

To reevaluate and interpret results of the VICTOR clustering, we compared the genome sequences using the Gegenees tool with default parameters (Camacho et al. 2009; Ågren et al. 2012). The program calculated symmetrical identity (SI) scores for each pairwise comparison based on BLASTn hits and a genome length.

To check if the translated local alignment of the whole genomes will be more sensitive to a phylogenetic signal at higher taxonomic ranks, we followed the Dice methodology proposed previously (Mizuno et al. 2013). The Dice score was calculated based on all reciprocal tBLASTn hits between pairs of genomes with ≥30% identity, alignment length ≥30 amino acids, and E-value ≤0.01. Pairs of scores were used to construct a distance matrix, which in turn was converted to the final tree using the BioNJ algorithm (Gascuel 1997).

Again, to evaluate and interpret this result, we calculated SI scores between all translated genome sequences using Gegenees. This time, we applied tBLASTx as the alignment algorithm with the other settings left on default values.

**Proteome-Based Analyses**

The Phage Proteomic Tree was constructed as described previously (Rohwer and Edwards 2002). In brief, 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 converted to distances as described in Rohwer and Edwards (2002), and the distances thus obtained were used to generate the final tree using the neighbor joining algorithm.

**Identification of Protein Clusters**

In order to comprehensively define the gene content in *herellitevirus* genomes, we applied two independent, yet complementary methods of identifying orthologous clusters.

An initial set of orthologous protein clusters (OPCs) was constructed using the GET_HOMOLOGUES software suite, which utilizes several independent clustering methods (Contreras-Moreira and Vinuesa 2013). To capture as many evolutionary relationships as possible, a greedy COGtriangles algorithm (Kristensen et al. 2010) was applied with a 50% sequence identity threshold, 50% coverage threshold, and an E-value cutoff equal to 1e-10. The results were converted into an orthologue matrix with the “compare_clusters” script (part of the GET_HOMOLOGUES suite) (Felsenstein 1989).

A second method was based on assignment of the genes to a predefined pVOG (prokaryotic Virus Orthologous Group) set described previously (Graziotin et al. 2017) and available at http://dmk-brain.ecn.uiowa.edu/pVOGs/. In brief, protein-coding genes in the 93 analyzed genomes were identified using Prodigal V2.6.3 in anonymous mode (Hyatt et al. 2010). Then, the gene products were assigned to the respective orthologue group by HMMsearch (E-value <10^-2) against the database of Hidden Markov Models (HMMs) created for every of 9518 pVOG alignments using HMMbuild of HMMer v3.1b2 (Finn et al. 2011).

**Analysis of Gene Synteny**

To investigate a genomic synteny-based classification signal, we implemented a method developed at the University of Utrecht, 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). The tool used the pVOG assignments described above to generate a synteny profile of every genome (in fact, this pVOGs methodology is integral part of the GOAT pipeline).

The algorithm accounted for gene replacements and low similarity between genes by using an all-vs-all similarity matrix between pVOG pairs based on HMM–HMM similarity (HH-suite 2.0.16) (Soding 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 semiglobal 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, the gene order was transcribed 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:

\[ DAB = 1 - \frac{GAB + GBA}{GAA + GBB} \]

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 converted to a tree with BioNJ (Gascuel 1997).

**Marker Protein Phylogenies**

Based on the OPC and pVOG clusters defined above, which respectively identified 14 and 38 core protein clusters (Supplementary Table S2 available on Dryad), we chose 10 consistently-predicted protein groups (encoded by genes with well-defined boundaries and without introns) for inclusion as phylogenetic marker. The selected clusters included: DNA helicase cluster, tail sheath protein, two different groups of virion proteins (including the major capsid protein cluster), and six clusters with no known function. The members of these clusters were aligned using Clustal Omega with default parameters (Sievers et al. 2011). The resulting alignments were analyzed with the IQ-TREE pipeline, which includes the ModelFinder tool that determines the most suitable model of sequence evolution for the alignment, the main algorithm that constructs a maximum-likelihood tree and ultrafast bootstrap (UFBOOT)—an UFBOOT subroutine that calculates the support of the branches (Nguyen et al. 2015; Chernomor et al. 2016; Kalyaanamoorthy et al. 2017; Hoang et al. 2018). The same program was used to generate the approximation of the “species tree” based on the concatenated alignments of all markers. In this case, the partitioned model of the alignment was also calculated using the ModelFinder module of IQ-TREE and the analysis was run in 100 replicates to select the final tree with best log-likelihood score.

**Visualization and Comparison of the Results**

All trees were rooted at Brochothrix phage A9—a phage that consistently appeared as a distant outlier in all obtained topologies (to facilitate comparisons) and visualized using Geneious tree viewer. The taxon coloring and the legend was added using Inkscape.

Topological distances between different trees were calculated as Robinson–Foulds metrics (Robinson and Foulds 1981) with IQ-TREE and detected differences were visualized as tanglegrams generated using Neighbor Net-based heuristics in Dendroscope 3.5.9 (Huson and Scornavacca 2012).

**RESULTS**

**Definition of the Candidate “Herelleviridae” Family**

Recently, several studies have shown the paraphyly of the families constituting the order Caudovirales (Ianzo et al. 2016; Bolduc et al. 2017a; Aiewsakun et al. 2018). We created a monopartite network of all dsDNA viruses in the NCBI RefSeq using vContACT v2.0 (Bolduc et al. 2017a, Bolduc et al. under revision) showing the phages related to SPO1 as a clearly defined, interrelated cluster (Fig. 1a). The distinctness of the cluster was confirmed with the GRAVity pipeline (Fig. 1b), which showed that subfamily classifications in the order Caudovirales are clustered at the same distance as the new tailed phage family Ackermanniaviridae and as eukaryotic virus families (Aiewsakun et al. 2018; Aiewsakun and Simmonds 2018). A further comparison of all dsDNA viruses using the Phage Proteomic Tree method on the VIPTree server showed that myoviruses, siphoviruses, and podoviruses were interspersed with each other, but SPO1-related phages formed a distinct and coherent clade (Supplementary Fig. S1 available on Dryad). These results clearly indicate that the SPO1-related viruses are distinct and form a cohesive group. Based on this evidence, we propose that this group of viruses represents a new family, and we suggest the name Herelleviridae, in honor of the 100th anniversary of the discovery of prokaryotic viruses by Félix d’Hérelle.

**Exploration of the Intrafamilial Relationship**

After delineating the family, we proceeded with the investigation of the relationships between its members. Regardless of the approach applied, we found five clearly-separated clusters interpreted by us as potential subfamilies (Figs. 1b, 2, and 3, Supplementary Figs. S2–S4 available on Dryad, Table 1, Supplementary Table S1 available on Dryad). 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) includes Bacillus-infecting viruses that most closely resemble phage Bastille. The third cluster (Brockvirinae) comprises viruses of enterococci that are similar to Enterococcus phage ϕEF24C. The fourth cluster (Twortvirinae) gathers staphylococci-infected viruses that are similar to Staphylococcus phage Twort. The remaining cluster (Jasinskavirinae) consists of viruses infecting Listeria that are similar to Listeria phage P100. The classification left
a) Network representation of predicted protein content similarity of dsDNA viruses generated with vConTACT v2.0. Viruses are represented as circles (nodes) connected with each other (edges) based on a significant number of shared protein clusters, with more similar genomes displayed closer together on the network. The genomes belonging to the new family *Herelleviridae* are indicated with a circle. Genomes previously assigned to the subfamily *Spounavirinae* are indicated in pink.

b) Clustering of dsDNA bacteriophages that possess subfamily assignments in the order *Caudovirales* generated with GRAVITY, darker colors in the heatmap represent higher degrees of similarity between genomes. The phages are clustered using UPGMA into a dendrogram, showing bootstrap values (100 pseudoreplicates) on each branch.
three viruses with no genus and subfamily assignment: Lactobacillus phage Lb338, Lactobacillus phage LP65, and Brochothrix phage A9.

Five subfamily-rank clusters can be further subdivided into smaller clades that correspond well with the currently accepted genera (Table 1). The evidence supporting this suggested taxonomic reclassification is presented in the following sections.

**Genome-Based Analyses**

The genome-based analyses used to identify close relationships between phage genomes provide powerful information for species and genus demarcation. We performed an all-against-all BLASTn analysis with Gegenees (Ågren et al. 2012), revealing that the genomes of several viruses were similar enough to consider them strains of the same species (they shared >95% nucleotide identity, Table 1, Supplementary Table S1, Fig. S2 available on Dryad). We could delineate clear groups with significant nucleotide similarity, proposed as genus-rank taxa, at similarities greater than 50%. Using the BLAST-based phylogenetics framework VICTOR (Meier-Kolthoff and Göker 2017), we were able to confirm that the existing genera form well-supported clades (Fig. 2a).

Similar patterns emerged at the translated nucleotide level when the genomes were analyzed using the tBLASTx-based Dice method (Fig. 2b) (Mizuno et al. 2013). An all-against-all comparison at the translated nucleotide level (tBLASTx) with Gegenees showed an overall low level of similarity (15%) within the newly proposed family and allowed us to start delineating the subfamily level at approximately 25% translated genome similarity (Supplementary Fig. S2 available on Dryad). However, the subfamily boundaries were not always clear using these methods. For example, the members...
of the Brockvirinae subfamily shared 20–25% similarity at the translated nucleotide level with the twortviruses and Jasinkaviruses.

**Proteome-Based Analyses**

As proteome-based analyses rely on genome annotation, they are sensitive to bias introduced by different annotation methods, and the results of such analyses should, therefore, be interpreted with caution. To mitigate this, we reannotated all genomes with the same automated pipeline as described above (M&M, Supplementary File 1 available on Dryad).

We inferred a Virus (Phage) Proteomic Tree using only the members of the new family to assess its internal structure (Fig. 3a). This showed clearly-defined clusters at the subfamily and genus rank, but revealed longer than expected branch lengths for phages that had very similar genomes, implying that this method should not be used for fine-grained taxonomic classification.

Among 1296 singleton proteins (proteins without recognizable homologues in the analyzed genomes) and 2070 protein clusters defined using the OPC approach, we identified 14 clusters common for all viruses belonging to the new family “Herelleviridae” (Table 2, Supplementary Table S2 available on Dryad). Classification of the viral proteins using pVOGs showed that 38 pVOGs were shared between all 93 virus genomes, with 14 pVOGs functionally annotated (Table 2, Supplementary Table S2 available on Dryad). 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. Indeed, many genes of spounaviruses and related viruses are invaded by mobile introns or inteins (Goodrich-Blair et al. 1990; Lavigne and Vanderstegen 2013). These gaps in coding sequences challenge standard gene prediction tools and introduce additional bias in similarity-based cluster algorithms. Because of these insertions as confounding factors, we used a subset of 10 core genes for further phylogenetic analysis.
resulting in “mosaicism” (Juhala et al. 2000; Krupovic recombination and horizontal gene transfer potentially marker proteins identified from the OPC analysis high degree of mosaicism (Bolduc et al. 2017a). at most, about 10% of reference virus genomes have a be expected with rampant mosaicism. The lack of we did not observe the high modularity that might genomic rearrangements leave a measurable of the central part of its genome compared with the other herelleviruses. From this overall picture, we can (Cadungog et al. 2015), which showed an inversion almost all viruses clustered according the proposed taxa. results obtained using sequence-based methods, with (Fig. 3b). The clustering results proved comparable with genome structure and potential effects of recombination (e.g., whole genome phylogenies) may be necessary to marker loci, additional sources of phylogenetic signal (e.g., whole genome phylogenies) may be necessary to properly interpret any result.  

**Analysis of Gene Synteny**

Viral genomes are thought to be highly modular, with recombination and horizontal gene transfer potentially resulting in “mosaicism” (Juhala et al. 2000; Krupovic et al. 2011). By clustering the herelleviruses based solely on the gene order, we investigated plasticity of their genome structure and potential effects of recombination (Fig. 3b). The clustering results proved comparable with results obtained using sequence-based methods, with almost all viruses clustered according the proposed taxa. The potential exception was Bacillus phage Moonbeam (Cadungog et al. 2015), which showed an inversion of the central part of its genome compared with the other herelleviruses. From this overall picture, we can infer that genomic rearrangements leave a measurable evolutionary signal in all lineages, but do not shuffle genomes of related viruses beyond recognition. Thus, we did not observe the high modularity that might be expected with rampant mosaicism. The lack of considerable mosaicism supports recent findings that, at most, about 10% of reference virus genomes have a high degree of mosaicism (Bolduc et al. 2017a).  

**Marker Protein Phylogenies**

We used the amino acid sequences of concatenated marker proteins identified from the OPC analysis (Table 2) to generate a phylogenetic tree that is able to identify the evolutionary relationships at the genus and subfamily rank within the new family Herelleviridae (Fig. 4). This tree supported all proposed new taxa but was unable to differentiate between the different species. Branches representing subfamilies and genera were particularly well-supported (UFBOOT support above 99%). Additionally, nearly all topologies of single marker trees (Supplementary Fig. S3 available on Dryad) fitted well in the suggested taxonomic structure. The only notable deviation from the proposed classification scheme could be found in the Tail tube protein tree (VOG0068–OPC6141, Supplementary Fig. S3 available on Dryad). It shuffled members of the genus Silviavirus into the Kayivirus clade and also mixed the genera Nitunavirus and Agatevirus with unclassified phages. This may indicate that the evolutionary signal contained in this marker is insufficient to resolve related genera. Alternatively, the inconsistencies may be explained by the effect of horizontal gene transfer or convergent evolution introducing additional noise in our data. Regardless of the true reason of this inconsistency, it should be stressed that with a small number of available marker loci, additional sources of phylogenetic signal (e.g., whole genome phylogenies) may be necessary to properly interpret any result.  

**Comparison of the Results Obtained Using Different Methods**

Virus classification methods in general suffer from a low signal-to-noise ratio. This “noise” may be introduced in the data by horizontal gene transfer and differences in mutation rates in different viral lineages. To get a measure of the discrepancies between the methods used above, we calculated the normalized Robinson–Foulds distances (representing the fraction of data partitions that are present only in one of the analyzed trees, Supplementary Table S3 available on Dryad) and created tanglegrams for the visual comparison of topologies (Supplementary Fig. S4 available on Dryad). Trees obtained using different methods differed considerably (normalized Robinson–Foulds metric in range 0.16–0.58) but topological distances between them were comparable to distances between single marker trees (and in most cases smaller, see Supplementary Table S3 available on Dryad). Interestingly, for the herelleviruses, most of the noise becomes averaged at the genus rank, meaning that the grouping at this rank and above remains almost the same regardless of the classification method employed. The only significant inconsistencies compared with the proposed taxonomic classification were observed in the GOAT analysis and one single-marker tree (i.e., tail tube protein tree, VOG0068–OPC6141). Both of these deviations concerned a single genus or even unclassified species and they did not follow any commonpattern.
FIGURE 4. Maximum-likelihood tree based on concatenated alignment of 10 marker proteins generated using IQ-tree. The scale bar represents the number of substitutions per site, branch support values were calculated from 1000 ultrafast bootstrap (UFBOOT) replicates. The trees were rooted at Brochothrix phage A9 to facilitate comparison. Branches corresponding to genera and subfamilies are delineated with colored squares and circles, respectively.
DISCUSSION

The rapid expansion of phage genomics and metagenomics has left taxonomy behind. There are more than 8000 publicly-available caudovirus genomes, but only 873 have been officially classified by the ICTV (Davison 2017). The remaining genomes are provisionally stashed in the NCBI database with "unclassified" bins attached to the order Caudovirales or its associated families (Brister et al. 2015; Adams et al. 2017; Simmonds et al. 2017). One of the main problems is that the level of sequence divergence is so high that it often leaves no detectable sequence similarity between disparate members of the same order. Thus, not a single reliable phage-specific or even Caudovirales-specific marker gene could be defined. In addition, a classification system based on a single marker would be highly prone to instances of horizontal gene transfer. Indeed, there is no commonly recognized general phage classification tool and all of the currently used phylogenetic approaches have their critical limitations as described in this study.

For that reason, above the family rank we had to rely on high-throughput network and clustering analyses (vConTACT, GRAVity, and VipTree) that are capable of discerning the groups of taxa that are comparable, even if phylogenetic signal is sparse. These methods can analyze significant subsets of the viral genomic space in a reasonable time, outcompeting traditional phylogenetic approaches in terms of speed. They are, however, still expensive computationally and need to be recalculated when new data become available (Bolduc et al. 2017b). Moreover, these high-throughput methods do not attempt to model the process that gave rise to the observed data, but rather calculate arbitrary distance matrices from local similarities and use them to define groupings. Thus, the relation between the calculated distance and the divergence time remains unclear and the results of these methods should be taken with a grain of salt, especially in less divergent taxa or at the lower taxonomic ranks.

After defining the new family Herelleviridae, we applied a combination of genome and proteome analyses, gene synteny assessments, and multimarker gene phylogenies to establish its internal taxonomic structure. It has to be stressed that the results of most of these methods should be treated as approximations of phylogenetic reconstruction. Many of them suffer from similar methodological drawbacks as the abovementioned high-throughput clustering techniques, lacking proper theoretical support of their algorithms. Only the maximum-likelihood analysis of (a) marker sequence(s) allows for rigorous, statistically sound phylogenetic inference under a well-defined model of sequence evolution. Unfortunately, if the number of available marker loci is small, this method becomes vulnerable to the noise introduced by horizontal gene transfer (Davidson et al. 2015). More importantly, this approach is heavily influenced by the gene annotation. This may be a crucial disadvantage as the quality of database records is often debatable and computational reannotation of analyzed genomes does not always yield valid, comparable results.

On the other hand, these drawbacks can be easily circumvented by methods analyzing whole genome sequences (DICE, VICTOR, BLAST). Obviously, they are annotation-independent and mitigate the effects of horizontal gene transfer by averaging the signal across the total genome length. Unfortunately, if the untranslated nucleotide sequence of the virus is used, rapid decay of the similarity should be expected above the genus rank (e.g., Supplementary Fig. S2 available on Dryad). Above that rank, nucleotide sequence similarities were virtually undetectable, but sequence translations (DICE coefficient) or protein sequences (Phage Proteomic Tree) were still considerably similar. Thus, nucleotide sequence-based approaches capture small differences (e.g., silent mutations) between closely related genomes and may be well suited for species and strain demarcation but gradually lose sensitivity with each consecutive taxonomic rank.

To the best of our knowledge, the GOAT algorithm is the only method explicitly aimed at capturing the signal associated with genomic rearrangements in fluid genomes of viruses. Unfortunately, the evolutionary process that is responsible for the observed variations is even less studied than whole genome similarity metrics and we cannot rule out that this algorithm may be disproportionately susceptible to the effects of genome rearrangement events. However, it is ideally suited to pinpoint just those kinds of genomic rearrangements and mutations that are missed by other methods. Thus, it can provide unique data on structural dynamics of the studied genomes but in its present form should not be treated as the primary classification tool.

Bearing in mind all the advantages and limitations of the classification tools utilized here, and the convergence of their results for the analyzed taxa, we recommend an "ensemble of methods" approach similar to the one we used as a method of choice for the phage taxonomy. We suggest that future classification efforts should implement at least one well established phylogenetic method (e.g., maximum-likelihood analysis of concatenated marker genes/proteins) and at least one whole genome-based annotation-independent method to account for annotation inconsistencies, rearrangements and mosaicism. Additional approaches may be used, especially if methods of choice produce inconclusive or discordant results but should always be used with regard to their limitations.

All evidence considered, we suggest that the SPO1-related phages should be removed from the family Mygorviridae and given a family rank. Hence, we proposed establishing a new family Herelleviridae, containing five subfamilies: Spounavirinae (sensu stricto), Bastillevirinae, Twortvirinae, Janikavirinae, and Brockeirinae, each comprising the genera listed in Table 1. The suggested classification corresponds well with host taxonomy and leaves only 3% of viruses within the new family
unassigned. These unassigned viruses may represent clades at the rank of genus or even subfamily that are still undersampled.

Removing spouanviruses from the family Myoviridae to form the new Herelleviridae family is a major change in phage taxonomy. We envisage this detachment from their original taxon will be followed by abolishment of the Podoviridae, Myoviridae, and Siphoviridae and creation of new “phylogenomic” families, based on current subfamily-rank clades, which will faithfully reflect the genetic relationships between bacterial viruses. In our opinion, these changes are necessary to accommodate the observed diversity of tailed phages. It is worth stressing that this change does not remove the historically established caudovirus morphotypes: myovirids forming virions with contractile tails, siphovirids with long noncontractile tails, and podovirids with short noncontractile tails. Nevertheless, by disconnecting morphotype and taxonomy, related clades can be grouped across different morphotypes. Such an approach would solve the problems of the mvuviruses 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). Finally, abolishing the current morphology-based classification of tailed phages will remove the major barrier in classifying phages from metagenomic sequence data.

SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.j06q96.

FUNDING

This work was supported by the National Science Centre, Poland (2016/23/D/NZ2/001435) to J.B.; the Netherlands Organization for Scientific Research (NWO) (Vidi 864.14.004) to B.E.D. (MBFS and BED); the US National Science Foundation (DUE-132809 and MCB-133080) to R.A.E.; the University of Helsinki and Academy of Finland funding for Instruct-IT to H.M.O.; the Chargé de Recherches fellowship from the National Fund for Scientific Research, FNRS, Belgium to A.G.; the EUed Horizon 2020 Framework Programme for Research and Innovation, “Virus-X” (685778) to E.E., the Gordon and Betty Moore Foundation Investigator Award (GBMF#5790) to M.B.S.; the Battelle Memorial Institute’s prime contract with the US National Institute of Allergy and Infectious Diseases (NIAID) (Contract number HHSN27220070016d) to J.H.K.; the Intramural Research Program of the NIH, GOA grant ‘Phage Biosystems’ of the KULeuven to J.R.B. and I.T.; and the Biotechnology and Biological Sciences Research Council Institute Strategic Programme in Gut Microbes and Health BB/R012490/1 and its constituent project BBS/E/F/000PR10353 to E.M.A. 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.

ACKNOWLEDGMENTS

We thank Laura Bollinger, Integrated Research Facility at Fort Detrick, for technical writing services.

REFERENCES

Adams M.J., Leftowitz E.J., King A.M., Harrach B., Harrison R.L., Knowles N.J., Kropinski A.M., Krupovic M., Kuhn J.H., Mushegain A.R., Nibert M.L., Sabanadzovic S., Sanfaçon H., Siddell S.G., Simmonds P., Varsani A., Zehnbauer 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. doi: 10.1007/s00705-016-3215-y.

Adriaenssens E.M., Ackermann H.-W., Anary H., Blasdel B., Connerton I.F., Coulondre D., Griffiths M.W., Hooten S.P., Kutter E.M., Kropinski A.M., Lee J.-H., Maes M., Pickard D., Ryu S., Sephehrizadeh Z., Shahrbabak S.S., Toribio A.L., Lavigne R. 2012. A suggested new bacteriophage genus: “Viunalikevirus”. Arch. Virol. 157:2035–2046. doi: 10.1007/s00705-012-1360-5.

Adriaenssens E.M., Edwards R., Nash J.H.E., Mahadevan F., 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. doi: 10.1016/j.virol.2014.10.016.

Adriaenssens E.M., Krupovic M., Knezevic P., Ackermann H.W., Barylski J., Brister J.R., Clokie M.R., Duffy S., Duttilh B.E., Edwards R.A., Eustait F., Jang H.B., Klumpp J., Kropinski A.M., Lavigne R., Poranen M.M., Prangshivli D., Ramniak 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. doi: 10.1007/s00705-016-3173-4.

Adriaenssens E.M., Wittmann J., Kuhn J.H., Turner D., Sullivan M.B., Duttilh B.E., Jang H.B., van Zyl J.L., Klumpp J., Lobocka M., Moreno Switt A.J., Rumnieks J., Edwards R.A., Uchihama J., Allen-Azarabi A., Zehnbauer F.M., Petty N.K., Kropinski A.M., Barylski J., Gillis A., Clokie M.R.J., Prangshivli D., Lavigne R., Aziz R.K., Duffy S., Krupovic M., Poranen M.M., Kaczevic P., Enault F., Tong Y., Oksanen H.M., Brister J.R. 2018. Taxonomy of prokaryotic viruses: 2017 update from the ICTV bacterial and archaeal viruses subcommittee. Arch. Virol. 163:1125–1129. doi: 10.1007/s00705-018-3727-z.

Ägren J., Sundström A., Hästström T., Segerman B. 2012. Gegenes: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. PLoS One. 7:e39107. doi: 10.1371/journal.pone.0039107.

Aiewsakun P., Adriaenssens E.M., Lavigne R., Kropinski A.M., Simmonds P. 2018. Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy. J. Gen. Virol. 99:1331–1343. doi: 10.1099/jgv.0.001110.

Aiewsakun P., Simmonds P. 2018. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. Microbiome. 6:38. doi: 10.1186/s40168-018-0422-7.

Aartschot S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410. doi: 10.1016/S0022-2836(05)80354-2.
insights into the phylogeny of the subfamily Spumavirinae. PLoS One. 9:e86632. doi: 10.1371/journal.pone.0086632.

Bolduc B, Jang H.B, Droulier G, You Q.Z, Roux S, Sullivan M.B. 2017. vCONTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. PeerJ. 5:e5243. doi: 10.7717/peerj.5242.

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

Brister J.R, Ako-adijie D, Rao Y, Binkova O. 2015. NCBI Viral Genome Resource. Nucleic Acids Res. 43:D571-D577. doi: 10.1093/nar/gku1207.

Cadungog J.N., Khatemi B.E., Hernandez A.C., Kuty Everett B. 2015. Genus “Saltoviridae” within the family “Caudovirales” Caudovirales: Virology. 468–470:421–443. doi:10.1016/j.virology.2015.07.016.

Cameron C, Coulouris G, Avagyan V, Shanoski J, Couroucé R, McWilliam H.D., Lopez I. 2019. Morpheus: a fast and accurate tree search method. Bioinformatics. 35:i296-i304. doi:10.1093/bioinformatics/btz598.

Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a resource for comparative genomics of orthologous proteins. Bioinformatics. 29:1676-1683. doi:10.1093/bioinformatics/btt207.

Darriba D, Taboada L.G., Doallo R, Posada D. 2019. TreeMap: a fast tool for drawing phylogenetic trees from alignments. Bioinformatics. 35:288-290. doi:10.1093/bioinformatics/btz316.

Duffy S., Clokie M.R., Barylski J., Ackermann H.W., Kuhn J.H. 2016. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. ISME J. 10:786-797. doi: 10.1038/ismej.2015.128.

Eichler J., Eichler E. 2014. Genomic sequences of bacteriophages HK97 and HK022: a perspective on future sequencing. Virology. 451:200-214. doi:10.1016/j.virol.2013.12.022.

Eichler J., Valentine J., Sanger F., Akhtar A., Klonowski P. 2014. Prokaryotic genome annotation: the iVirus project. Expert Rev. Mol. Diagn. 14:271-283. doi:10.1586/14737159.2014.915777.

Eickbush T.H., Fiers W. 1990. Self-splicing group I intron in the DNA polymerase gene of Bacillus subtilis bacteriophage SPO1. Cell. 63:417–424. doi: 10.1016/0092-8674(90)90174-D.

Felsenstein J. 1998. PHYLIP—Phylogeny inference package—v3.2. Cladistics. 14:63-69. doi:10.1111/j.1096-0031.1999.tb00562.x.

Ford M.E., Youlton A., Hatfull G.F., Hendrix R.W. 2000. Taxonomic assignment of bacteriophages HK97 and HK022: a perspective on future sequencing. Virology. 280:105-119. doi:10.1006/viro.2000.0416.

Fujita Y., Kajitani K., Nakamura Y., Yamanaka Y., Ikenaga K. 2015. Tree-based gene prediction and translation initiation site identification. BMC Bioinformatics. 11:119. doi: 10.1186/1471-2105-16-119.

Fusco P.A., McWilliam H., Valentin F., Wallace I.M., Aasen S., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947-2948. doi:10.1093/bioinformatics/btm404.

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. doi:10.1093/oxfordjournals.molbev.a020141.

Gibson T.J., Jeffay L., Mahadevan P. 2017. Self-splicing group I introns in the DNA polymerase gene of Bacillus subtilis bacteriophage SPO1. J. Virol. 91:1-3. doi:10.1128/JVI.00040-16.

Gouy M., Guigo R.,3.8 A., Mangeat P.A., McWilliam H., Valentin F., Wallace I.M., Aasen S., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947-2948. doi:10.1093/bioinformatics/btm404.

Gouy M., Guigo R.,3.8 A., Mangeat P.A., McWilliam H., Valentin F., Wallace I.M., Aasen S., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947-2948. doi:10.1093/bioinformatics/btm404.

Gouy M., Guigo R.,3.8 A., Mangeat P.A., McWilliam H., Valentin F., Wallace I.M., Aasen S., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947-2948. doi:10.1093/bioinformatics/btm404.

Gouy M., Guigo R.,3.8 A., Mangeat P.A., McWilliam H., Valentin F., Wallace I.M., Aasen S., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947-2948. doi:10.1093/bioinformatics/btm404.

Gouy M., Guigo R.,3.8 A., Mangeat P.A., McWilliam H., Valentin F., Wallace I.M., Aasen S., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947-2948. doi:10.1093/bioinformatics/btm404.

Gouy M., Guigo R.,3.8 A., Mangeat P.A., McWilliam H., Valentin F., Wallace I.M., Aasen S., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947-2948. doi:10.1093/bioinformatics/btm404.

Gouy M., Guigo R.,3.8 A., Mangeat P.A., McWilliam H., Valentin F., Wallace I.M., Aasen S., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947-2948. doi:10.1093/bioinformatics/btm404.

Gouy M., Guigo R.,3.8 A., Mangeat P.A., McWilliam H., Valentin F., Wallace I.M., Aasen S., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947-2948. doi:10.1093/bioinformatics/btm404.

Gouy M., Guigo R.,3.8 A., Mangeat P.A., McWilliam H., Valentin F., Wallace I.M., Aasen S., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947-2948. doi:10.1093/bioinformatics/btm404.

Gouy M., Guigo R.,3.8 A., Mangeat P.A., McWilliam H., Valentin F., Wallace I.M., Aasen S., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947-2948. doi:10.1093/bioinformatics/btm404.

Gouy M., Guigo R.,3.8 A., Mangeat P.A., McWilliam H., Valentin F., Wallace I.M., Aasen S., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947-2948. doi:10.1093/bioinformatics/btm404.

Gouy M., Guigo R.,3.8 A., Mangeat P.A., McWilliam H., Valentin F., Wallace I.M., Aasen S., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947-2948. doi:10.1093/bioinformatics/btm404.

Gouy M., Guigo R.,3.8 A., Mangeat P.A., McWilliam H., Valentin F., Wallace I.M., Aasen S., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23:2947-2948. doi:10.1093/bioinformatics/btm404.
Nishimura Y., Yoshida T., Kuronishi M., Uehara H., Ogata H., Goto S. 2017b. ViPTree: the viral proteomic tree server. Bioinformatics. 33:2379–2380. doi: 10.1093/bioinformatics/btx157.

O'Leary N.A., Wright M.W., Brister J.R., Clufo S., Haddad D., McVeigh R., Rajput B., Robbertse B., Smith-White B., Ako-Adjei D., Astashyn A., Badretdin A., Bao Y., Blinokova O., Brover V., Chetverian V., Choi J., Cox E., Ermoslavea 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., Varsam 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. doi: 10.1093/nar/gkv1189.

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. doi: 10.1038/nature19094.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147. doi: 10.1016/0025-5564(81)90043-2.

Rohwer F., Edwards R. 2002. The Phage Proteomic Tree: a genome-based taxonomy for phage. J. Bacteriol. 184:4529–4535. doi: 10.1128/JB.184.16.4529.

Roux S., Brum J.R., Dutillie B.E., Sunagawa S., Duhaime M.B., Loy A., Poulos B.T., Solonenko N., Lara E., Poujain J., Pesant S., Kandels-Lewis S., Dimier C., Picheral M., Searson S., Cruaud C., Alberti A., Duarte C.M., Gasol J.M., Vaqué D., Taro Oceans Coordinators, Berth P., Acinas S.G., Wincker P., Sullivan M.B. 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature. 537:689–693. doi: 10.1038/nature19366.

Schuch R., Fischetti V.A. 2009. The secret life of the anthrax agent Bacillus anthracis: bacteriophage-mediated ecological adaptations. PLoS One. 4:e6532. doi: 10.1371/journal.pone.0006532.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 30:2668–2669. doi: 10.1093/bioinformatics/btu153.

Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Renner 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. doi: 10.1038/msb.2011.75.

Simmonds P., Adams M.J., Benko M., Breitbart M., Brister J.R., Carstens E.B., Davison A.J., Delwart E., Corbalenya 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., Rossinck M.J., Sabanadzovic S., Sullivan M.B., Suttle C.A., 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. doi: 10.1038/nrmicro.2016.177.

Simmonds P., Aiewsakun P. 2018. Virus classification—where do you draw the line? Arch. Virol. 163:2037–2046. doi: 10.1007/s00705-018-3938-z.

Söding J., Biegert A., Lupas A.N. 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33:W244–W248. doi: 10.1093/nar/gki408.

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. doi: 10.1128/AEM.02889-14.