Published online 26 October 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D491–D498***
doi: 10.1093/nar/gkw975

Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation

Ana Laura Grazziotin[1], Eugene V. Koonin[2] and David M. Kristensen[1,2,*]

[1]Department of Biomedical Engineering, College of Engineering, University of Iowa, Iowa City, IA 52242, USA and [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received August 30, 2016; Revised October 08, 2016; Editorial Decision October 10, 2016; Accepted October 21, 2016

ABSTRACT

Viruses are the most abundant and diverse biological entities on earth, and while most of this diversity remains completely unexplored, advances in genome sequencing have provided unprecedented glimpses into the virosphere. The Prokaryotic Virus Orthologous Groups (pVOGs, formerly called Phage Orthologous Groups, POGs) resource has aided in this task over the past decade by using automated methods to keep pace with the rapid increase in genomic data. The uses of pVOGs include functional annotation of viral proteins, identification of genes and viruses in uncharacterized DNA samples, phylogenetic analysis, large-scale comparative genomics projects, and more. The pVOGs database represents a comprehensive set of orthologous gene families shared across multiple complete genomes of viruses that infect bacterial or archaeal hosts (viruses of eukaryotes will be added at a future date). The pVOGs are constructed within the Clusters of Orthologous Groups (COGs) framework that is widely used for orthology identification in prokaryotes. Since the previous release of the POGs, the size has tripled to nearly 3000 genomes and 300 000 proteins, and the number of conserved orthologous groups doubled to 9518. User-friendly webpages are available, including multiple sequence alignments and HMM profiles for each VOG. These changes provide major improvements to the pVOGs database, at a time of rapid advances in virus genomics. The pVOGs database is hosted jointly at the University of Iowa at http://dmk-brain.ecn.uiowa.edu/pVOGs and the NCBI at ftp://ftp.ncbi.nlm.nih.gov/pub/kristensen/pVOGs/home.html.

INTRODUCTION

Although bacteria are the most plentiful and diverse cellular domain of life on earth, their abundance is vastly outweighed by the staggering number of viruses that infect them (1,2). Despite the presence of many thousands of completely sequenced virus genomes in NCBI sequence databases, the vast majority of viruses still remain to be characterized (3). Even so, the large number of sequences available creates considerable difficulties in navigating through all the data in search of biological meaning (4). The Prokaryotic Virus Orthologous Groups (pVOGs, formerly Phage Orthologous Groups, POGs (5–7)) resource aids researchers by providing clusters of orthologous genes in complete genomes of viruses that infect bacteria or archaea, using the microbial COG framework (8,9)—one of the oldest, most accurate, and most-often-used methods to computationally identify orthologs (10). Analogous to the COG database, each viral COG (VOG) represents all of the descendants of a single ancestral gene (orthologs) within the analyzed set of genomes. The pVOGs can be used for gene prediction and functional annotation in new virus genomes (11–13), metagenomic data (14–21) and host genomes containing proviruses (22,23). In addition to gene prediction and protein function, the pVOGs provide a third form of annotation, namely the distribution (presence or absence) of the constituent genes across viruses and cellular organisms, i.e. a 'phyletic pattern' or 'phylogenetic profile' (24–27). Beyond various forms of annotation, pVOGs have also been used to construct large-scale datasets for studies of viral phylogenetics (28), comparative genomics (29–33) and mathematical modelling of virus evolution (34–36).

The pVOGs provide evolutionary gene families from nearly 3000 complete genomes of viruses that infect bacteria or archaea as a pre-computed resource, thus providing immediate access to information that would otherwise incur a heavy computational cost. This is triple the size of the previous release (7), in both the number of genomes and genes. Taxonomic coverage of pVOGs database has

[*]To whom correspondence should be addressed. Tel: +1 319 335 5241; Email: david-kristensen@uiowa.edu

been broadened with the newly available genomes, including virus families not represented in the previous dataset, such as *Sphaerolipoviridae*, *Turriviridae* and *Guttaviridae* (Supplementary Figure S1). The number of pVOGs has also increased to more than double the previous number of 4542, now up to 9518 (Supplementary Figure S2). pVOGs are also now available as user-friendly webpages. Multiple routes of access are provided in the form of a list of all genomes in the database and a list of all pVOGs, from either of which pages showing a list of all pVOGs represented in individual genomes can be easily accessed, or the user can choose from among several forms of bulk downloads. For instance, multiple sequence alignments and HMM profiles for each pVOG are provided, with downloads available either for individual pVOGs, individual virus families, or the entire database in bulk.

CHANGES IN THE pVOGs DATABASE

Previous versions of the pVOGs resource have been described under the name POGs (Phage Orthologous Groups (5–7)). Later, after expanding beyond the original content of dsDNA tailed bacteriophages to also include archaeal viruses, and then even further to also include bacterial and archaeal viruses with ssDNA, dsRNA, or ssRNA genomes, this name was changed to Prokaryotic virus Orthologous Groups, but the acronym POGs was kept for historical continuity. However, that is no longer possible because another, unrelated database published in this journal very shortly after the initial POGs publication has already registered the POG name (37). Thus, we rename the individual POGs as VOGs (for viral COGs, similar to another database that existed previously at NCBI many years ago (38)), and the database as a whole as pVOGs (Prokaryotic Virus Orthologous Groups), to clearly indicate that eukaryotic viruses are not (yet) included.

Although the names of individual POGs were changed to VOGs, the numeric portion of the name was kept intact. For example, the old POG0001 is now re-named to VOG0001, and still represents the same V protein family present in the *Inoviridae* family of filamentous phages and a few other viruses (with appropriate updates—e.g. some genomes that were previously unclassified have since been assigned to the *T7likevirus* genus).

Since the previous release of this data set, the number of genomes and proteins has tripled, and the number of VOGs has doubled (Supplementary Figure S2). The coverage of virus types in the pVOGs database has also been both broadened and deepened with increased coverage of previously included viral groups (families, genera and species), as well as the addition of representatives from several new viral groups that were not present in the previous POGs dataset (Supplementary Table S1, Supplementary Table S2, and Supplementary Table S3).

The basis of forming VOGs has changed from using domain-based orthology assignments to full-length proteins. All other COG-based resources currently use full-length proteins rather than domains (e.g. COGs (8,9), KOGs (39), arCOGs (40), NCVOGs (41), mimiCOGs (42), etc.), although those resources use manual curation to provide ‘domain-aware’ orthology predictions, avoiding grouping together of COGs due to domain recombination (10). This curation step was not implemented for pVOGs because multidomain proteins are rare in viruses, and the number of instances of domain recombination that would cause erroneous joining of multiple COGs is accordingly negligible (6). As an added bonus to consistency across COG resources, removal of the extremely computationally expensive domain splitting step, in conjunction with the automated procedure to map new viral COGs to existing groups, will allow for more frequent future updates of the pVOGs database.

STATISTICS OF THE pVOGs DATABASE

The pVOGs resource encompasses all publicly available complete genome assemblies of viruses that infect bacteria and archaea. As of May 2016, NCBI’s RefSeq database contained nearly 2000 complete genomes of prokaryotic viruses, which we then supplemented with a manually curated list of an additional 1000 complete genome entries from the Nucleotide (GenBank) database when a RefSeq entry was not available for a particular virus. Altogether then, the pVOGs database encompasses a comprehensive set of nearly 3000 complete genomes of viruses with prokaryotic hosts that are currently available. Most (>97%) represent phages that infect bacteria, although 77 complete genomes of archaeal viruses and four high-quality complete genome entries from uncultured viruses (e.g. crAssphage (43) and experimentally verified oceanic ssDNA phages (44)) are included as well. In total, these genomes contain nearly 300 000 protein-coding genes. To fill gaps left by automated annotation pipelines, we conservatively added >10 000 automated gene predictions (see below). These genes were then clustered into 9518 orthologous groups using the latest version of the COG algorithm (7,9,45,46). On average, about two-thirds of the proteins encoded in a virus genome are conserved in a VOG. The remaining one-third represents novel genes shared by fewer than three viruses, or cases of extreme evolutionary divergence where our automated detection procedure (see below) was unable to detect the low level of similarity. This coverage is not evenly distributed (Figure 1), but rather, the best-characterized viral families display a higher proportion of gene coverage than the poorly characterized families (especially archaeal viruses with few members). Some individual viruses even display coverage as high as 100%, e.g. *Mycobacterium phage* Ramsey for which every one of its genes is also found in at least two other distinct viruses. At the other end of the spectrum, some viruses do not have even a single gene that is shared with other viruses, such as the tiny *Leuconostoc phage* L5 with only five genes, none of which were annotated in GenBank.

Most of the individual VOGs in the pVOGs database are small, with 69% containing ≤10 members (Supplementary Figure S3). Furthermore, most VOGs are simple families of orthologs, with barely any paralogs: 95% of VOGs contain none, and <3% contain more than one. However, several notable exceptions exist for both features, such as several families that include typical mobile elements such as HNH endonucleases (47,48), and several very large families of VOGs shared across many viruses, such as the large

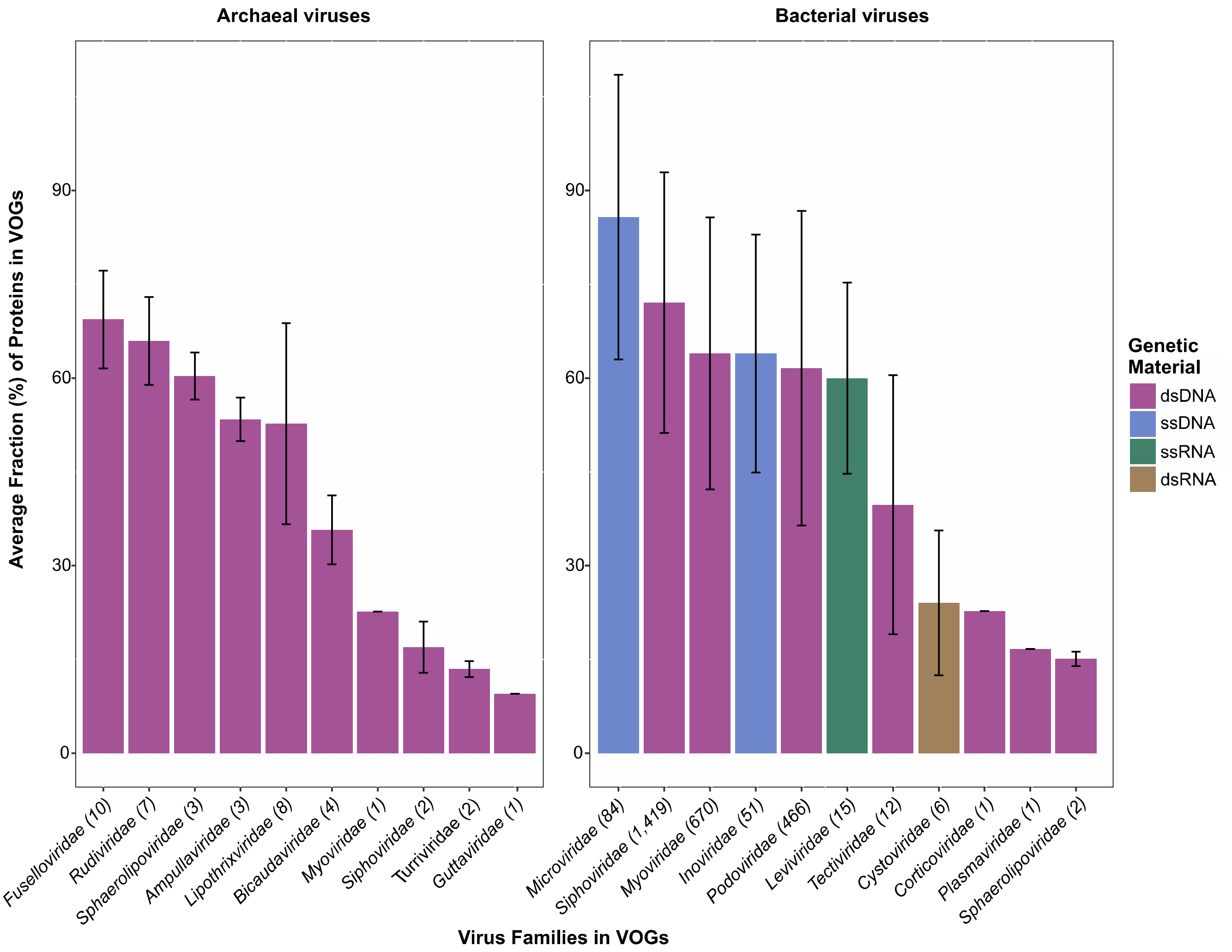


Figure 1. Coverage of protein sequences in VOGs for each viral family. The number of genomes in each family is shown in parentheses. Families with representatives covered by VOGs are shown. *Globuviridae* is not shown since its representative viruses had no proteins present in VOGs.

terminase subunit, a DNA packaging ATPase family found in about fourth fifths (83%) of the viruses in the order *Caudovirales* (tailed bacteriophages). VOG4543 uniquely represents an exception to both features at once, with 3.5-fold more proteins and a nearly 5-fold higher level of paralogy than any other VOG, in a group of helix-turn-helix DNA binding proteins found in an extremely diverse array of viruses (Figure 2). This VOG contains members of several protein families—including HNH endonucleases, transcriptional regulators such as cro/cI repressor, etc.—and altogether encompasses more than four fifths of the 3000 genomes, including both dsDNA and ssDNA, and with both bacterial and archaeal hosts. As such, from the standpoint of utility for the purposes of functional annotation, this VOG might better have been treated as a result of overlumping of too many distinctly related proteins. However, we refrained from manually splitting this group up, as was done for the microbial COG database, for two reasons: (i) consistency: although this group seems an obvious outlier, similar situations could lay hidden among the >9000 VOGs in the pVOGs database, especially among the many groups that, large or small, exclusively have annotations of 'hypothetical protein' and (ii) ease of future updates, especially because the vast majority of the genetic diversity present in the virosphere has yet to be explored, and therefore any procedure that includes manual annotation as a requirement for future progress will be difficult to keep up to date.

pVOGs DATABASE ACCESS AND WEB INTERFACE

The pVOGs database is primarily accessed through the homepage at the University of Iowa at http://dmk-brain.ecn.uiowa.edu/pVOGs or the NCBI at ftp://ftp.ncbi.nlm.nih.gov/pub/kristensen/pVOGs/home.html. Access to the pVOGs database is provided via two main interactive avenues: a list of all genomes (Figure 3A) and all VOGs (Figure 3B) in the pVOGs database. In the former, a user can choose a particular virus genome to investigate (Figure 4A), examine the conserved gene families it shares with other viruses, and choose an individual VOG (Figure 4B) to find

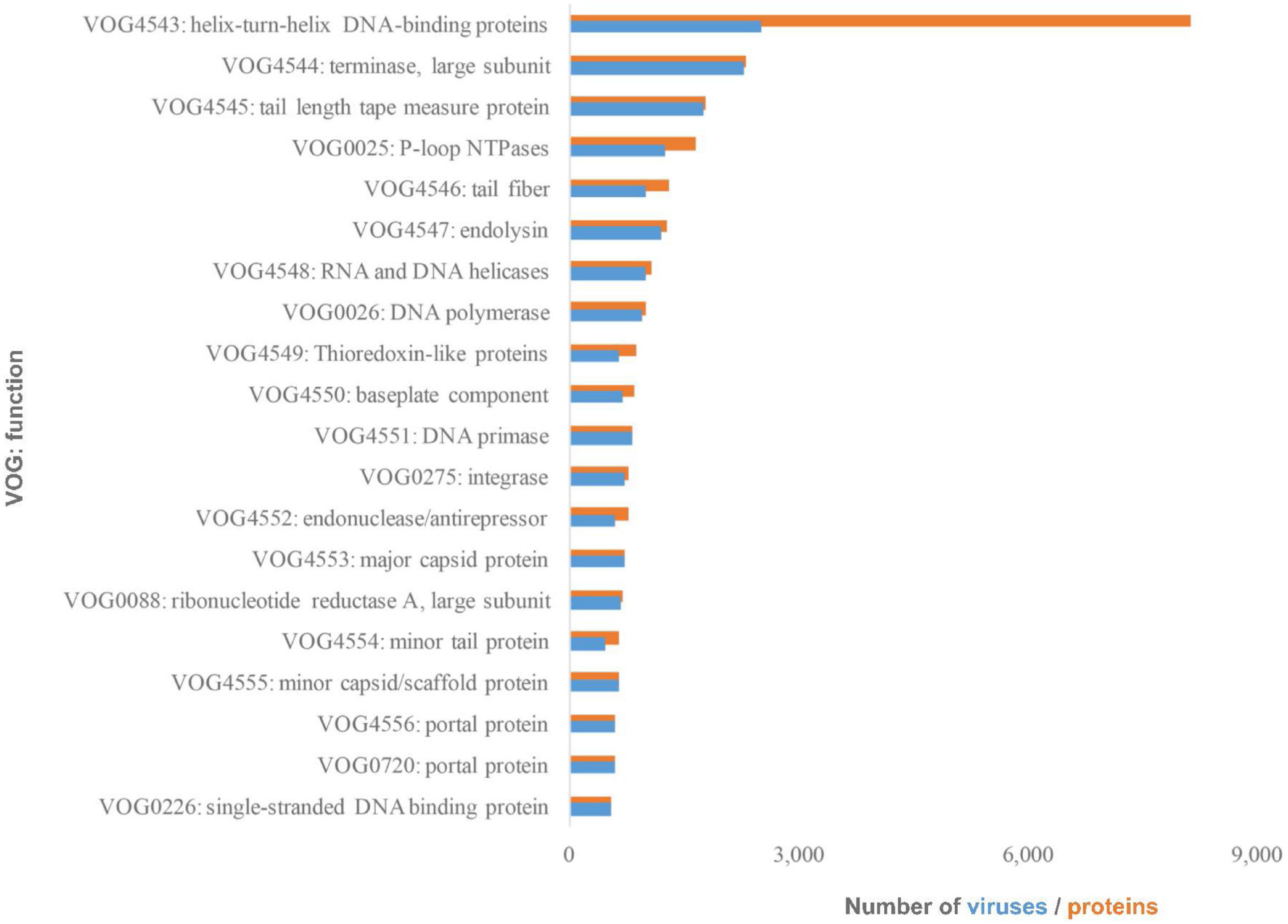


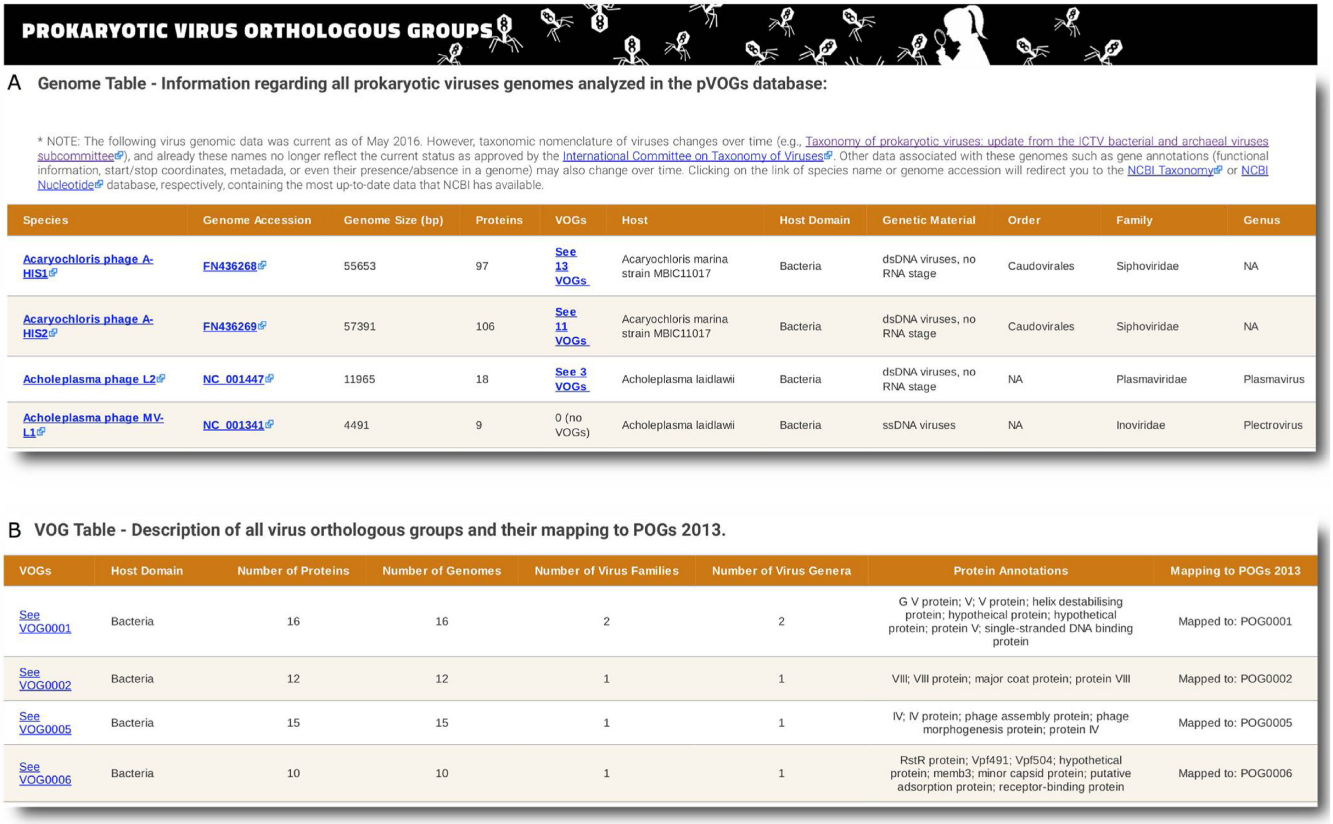
Figure 2. Sizes and functions of the largest VOGs. A single function for these was manually defined and reported based on the consistency of individual protein annotations within each VOG.

more information; via the latter avenue, the user can choose an individual VOG directly by name or protein annotation. From an individual VOG page (Figure 4B), links to the other VOGs conserved in each of its' member genomes are provided. A third, non-interactive avenue of access is provided in the form of bulk data files accessible from the downloads page, available for the pVOGs database in its entirety as well as for each viral family. All data available on the webpages are also provided as text files for programmatic access (for large-scale studies), and can be accessed either from individual VOG pages or the downloads page. In addition, multiple sequence alignments and HMM profiles for each VOG are provided, and are again accessible either from individual VOG pages or the downloads page. A tutorial page is also available with screenshots that further illustrate each type of webpage and helps users to navigate the pVOGs database.

pVOGs DATABASE CONSTRUCTION

Genomes

All complete genome assemblies of viruses that infect bacteria were downloaded as of May 2016 from NCBI's Nucleotide database (GenBank), as described previously (7). For example, for phages, the following Entrez query was used to obtain RefSeq entries: Viruses[Organism] NOT cellular organisms[ORGN] AND srcdb_refseq[PROP] AND vhost bacteria[filter] AND 'complete genome' [All Fields]; a similar query was used for archaeal viruses; for non-RefSeq entries, the query was modified to include NOT srcdb_refseq[PROP] NOT mRNA[filter]. The resulting set was manually curated to remove records described as clones, mutants, partial sequences, plasmids, CDS, or unverified, as well as prophage or proviral. After further examination, exceptions were made for four uncultured virus genomes obtained from metagenomic studies because these viruses have been validated experimentally (43,44,49). For duplicate virus genomes with identical taxonomy IDs shared between RefSeq and non-RefSeq entries, the former was kept and all of the latter were discarded; when there was no RefSeq record, all of the non-RefSeq entries were retained (no automated way exists of choosing the single 'best' representative). This curation process yielded 2912 bacterial viruses, 77 archaeal viruses and the four experimentally-verified uncultured viruses. Of these 2993 entries, 1986 were from RefSeq, and the rest from GenBank entries present in the nucleotide database.



PROKARYOTIC VIRUS ORTHOLOGOUS GROUPS

A Genome Table - Information regarding all prokaryotic viruses genomes analyzed in the pVOGs database:

* NOTE: The following virus genomic data was current as of May 2016. However, taxonomic nomenclature of viruses changes over time (e.g., Taxonomy of prokaryotic viruses: update from the ICTV bacterial and archaeal viruses subcommittee), and already these names no longer reflect the current status as approved by the International Committee on Taxonomy of Viruses. Other data associated with these genomes such as gene annotations (functional information, start/stop coordinates, metadada, or even their presence/absence in a genome) may also change over time. Clicking on the link of species name or genome accession will redirect you to the NCBI Taxonomy or NCBI Nucleotide database, respectively, containing the most up-to-date data that NCBI has available.

Species	Genome Accession	Genome Size (bp)	Proteins	VOGs	Host	Host Domain	Genetic Material	Order	Family	Genus
Acaryochloris phage A-HIS1	FN436268	55653	97	See 13 VOGs	Acaryochloris marina strain MBIC11017	Bacteria	dsDNA viruses, no RNA stage	Caudovirales	Siphoviridae	NA
Acaryochloris phage A-HIS2	FN436269	57391	106	See 11 VOGs	Acaryochloris marina strain MBIC11017	Bacteria	dsDNA viruses, no RNA stage	Caudovirales	Siphoviridae	NA
Acholeplasma phage L2	NC_001447	11965	18	See 3 VOGs	Acholeplasma laidlawii	Bacteria	dsDNA viruses, no RNA stage	NA	Plasmaviridae	Plasmavirus
Acholeplasma phage MV-L1	NC_001341	4491	9	0 (no VOGs)	Acholeplasma laidlawii	Bacteria	ssDNA viruses	NA	Inoviridae	Plectrovirus

B VOG Table - Description of all virus orthologous groups and their mapping to POGs 2013.

VOGs	Host Domain	Number of Proteins	Number of Genomes	Number of Virus Families	Number of Virus Genera	Protein Annotations	Mapping to POGs 2013
See VOG0001	Bacteria	16	16	2	2	G V protein; V; V protein; helix destabilising protein; hypotheical protein; hypothetical protein; protein V; single-stranded DNA binding protein	Mapped to: POG0001
See VOG0002	Bacteria	12	12	1	1	VIII; VIII protein; major coat protein; protein VIII	Mapped to: POG0002
See VOG0005	Bacteria	15	15	1	1	IV; IV protein; phage assembly protein; phage morphogenesis protein; protein IV	Mapped to: POG0005
See VOG0006	Bacteria	10	10	1	1	RstR protein; Vpf491; Vpf504; hypothetical protein; memb3; minor capsid protein; putative adsorption protein; receptor-binding protein	Mapped to: POG0006

Figure 3. Screenshot of pVOGs webpages showing the main access to database content: (**A**) genome table with information about all virus genomes in the database; (**B**) VOG table describing a list of all VOGs present in the pVOGs database, protein annotations of sequences present in each VOG and their mapping to the previous POGs.

Genes

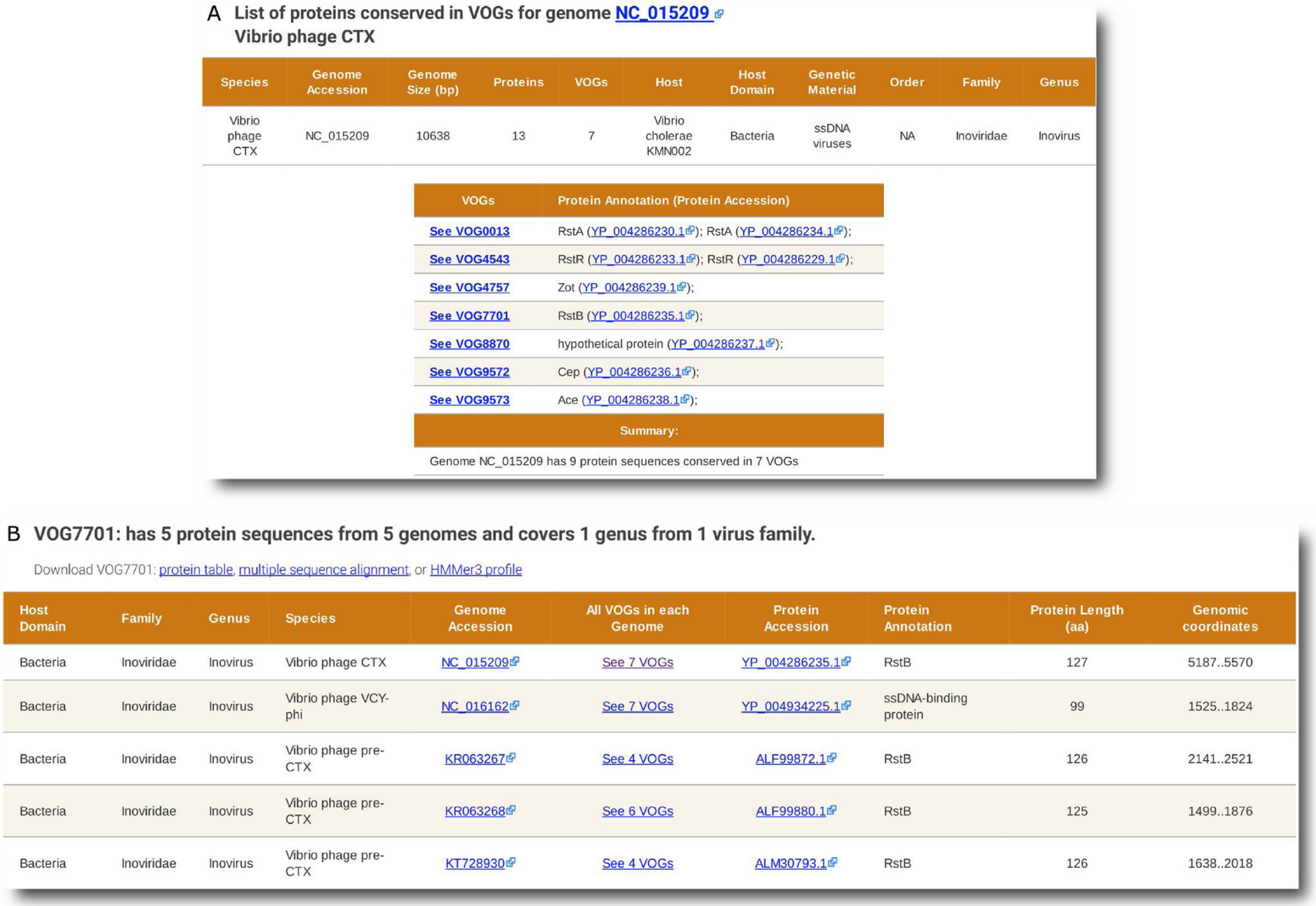
In addition to the 286 242 protein-coding genes that were already present in the genome records, we conservatively added 10 353 *de novo* gene predictions made using the non-supervised heuristic GeneMarkS program (50). Genes were only added if they met the following criteria: (i) not overlapping with an existing gene by more than 50 bp and (ii) not oriented in the opposite direction of the three upstream and three downstream gene neighbors. This procedure was manually verified on several tens of well-annotated genomes from curated databases (such as T4, T7 and lambda), as described previously (6,7).

Orthologous gene clusters

Prior to forming orthologous gene clusters (those shared between at least three 'distinct' viruses), closely related genomes that effectively represent multiple isolates of the same virus were merged. This was done to reduce the bias in the dataset caused by the non-uniform representation of different viruses, by preventing the formation of a VOG from only trivially related isolates of the same virus (6,7). In the current dataset, this procedure clustered the 2993 genomes into 1795 lineages. Next, an all-against-all BLAST (51) search was performed for all proteins in all genomes, and the standard COGtriangles algorithm was used to construct Clusters of Orthologous Groups (COGs) of the virus proteins (9,45). Briefly, this procedure begins by identifying symmetric best matches shared between three genomes, and merges such triangles into larger groups whenever they share a common 'side' (a two-way symmetric best-match). This approach groups together even fast-evolving orthologs as long as sufficient sequence similarity exists to detect the signal, with matches having *e*-value >10 or covering <50% of the query or target length discarded to ensure high quality. In 1.5% of the proteins, this approach failed to distinguish between membership in multiple COGs, which is an indication of unresolved paralogy or domain recombination.

Mapping

To ensure continuity of the updated pVOGs database with the previous dataset of virus orthologs (7) and measure the differences resulting from splitting proteins into domains vs. using full-length proteins, the clustering of viral proteins into orthologous groups in the 'pVOGs 2016' dataset was compared to that in the older 'POGs 2013' dataset (Table 1) for a reduced dataset shared between the two versions. Given that several virus genome entries that were present in GenBank in the older POGs dataset were now replaced with entries from RefSeq, and during this change several protein annotations were also changed, genomic coordinates shifted, and gene start and stop locations altered, a recip-



A List of proteins conserved in VOGs for genome NC_015209
Vibrio phage CTX

Species	Genome Accession	Genome Size (bp)	Proteins	VOGs	Host	Host Domain	Genetic Material	Order	Family	Genus
Vibrio phage CTX	NC_015209	10638	13	7	Vibrio cholerae KMN002	Bacteria	ssDNA viruses	NA	Inoviridae	Inovirus

VOGs	Protein Annotation (Protein Accession)
See VOG0013	RstA (YP_004286230.1); RstA (YP_004286234.1);
See VOG4543	RstR (YP_004286233.1); RstR (YP_004286229.1);
See VOG4757	Zot (YP_004286239.1);
See VOG7701	RstB (YP_004286235.1);
See VOG8870	hypothetical protein (YP_004286237.1);
See VOG9572	Cep (YP_004286236.1);
See VOG9573	Ace (YP_004286238.1);
Summary:	
Genome NC_015209 has 9 protein sequences conserved in 7 VOGs	

B VOG7701: has 5 protein sequences from 5 genomes and covers 1 genus from 1 virus family.

Download VOG7701: protein table, multiple sequence alignment, or HMMer3 profile

Host Domain	Family	Genus	Species	Genome Accession	All VOGs in each Genome	Protein Accession	Protein Annotation	Protein Length (aa)	Genomic coordinates
Bacteria	Inoviridae	Inovirus	Vibrio phage CTX	NC_015209	See 7 VOGs	YP_004286235.1	RstB	127	5187..5570
Bacteria	Inoviridae	Inovirus	Vibrio phage VCY-phi	NC_016162	See 7 VOGs	YP_004934225.1	ssDNA-binding protein	99	1525..1824
Bacteria	Inoviridae	Inovirus	Vibrio phage pre-CTX	KR063267	See 4 VOGs	ALF99872.1	RstB	126	2141..2521
Bacteria	Inoviridae	Inovirus	Vibrio phage pre-CTX	KR063268	See 6 VOGs	ALF99880.1	RstB	125	1499..1876
Bacteria	Inoviridae	Inovirus	Vibrio phage pre-CTX	KT728930	See 4 VOGs	ALM30793.1	RstB	126	1638..2018

Figure 4. Screenshot of pVOGs webpages showing additional types of information available: (**A**) individual genome table showing a list of VOGs present in a particular genome and respective protein annotations; (**B**) individual VOG table, describing detailed information about each VOG and protein content. This page also provides tabular files, multiple sequence alignments and HMM profiles for downloading.

rocal best match procedure was used to recover protein annotations for those that were changed. In this enhanced intersection of genomes and proteins shared between the two versions, the orthology assignment by the COG procedure was measured and counted as either: (i) an identical match; (ii) partial overlap (e.g. a fusion of multiple old POGs into a single VOG) or (iii) completely disjoint (e.g. multiple VOGs representing only subsets of multiple old POGs, due to disjoint domain content).

A high degree of agreement was found between the datasets (Table 1), with >92% of VOGs belonging to the first two match categories. For these, the numeric portion of the names of the original POGs were assigned to the current VOGs, to ensure consistency between versions. In cases where a current VOG matched multiple older POGs, and did so with 100% identity to the union of those clusters, the name of the largest original cluster was inherited, and a note added describing all of the names of the original clusters. New VOGs not observed in the previous dataset, as well as those with only partial overlap with one or more previous POGs, were sequentially assigned new names, starting at the point where the older POGs left off (i.e. VOG4543 and above).

Availability

The pVOGs database is freely and publicly available at the University of Iowa at http://dmk-brain.ecn.uiowa.edu/pVOGs and the NCBI at ftp://ftp.ncbi.nlm.nih.gov/pub/kristensen/pVOGs/home.html. Both locations will be updated as changes are made in the future, although the latter FTP site will host only simple webpages devoid of active content such as JavaScript and/or CGI whereas the former HTTP server will have active content added as it is developed. The previous version (still called POGs) will remain available at its original location on the NCBI FTP site at ftp://ftp.ncbi.nlm.nih.gov/pub/kristensen/thousandgenomespogs/. All queries and comments regarding the pVOGs database should be directed to DMK.

FUTURE DEVELOPMENTS

In accordance with the name ‘Prokaryotic Virus Orthologous Groups’, only the subset of viruses that infect bacterial and archaeal hosts are currently available in the pVOGs database, although in the future we anticipate the addition of eukaryotic virus families. Algorithmic improvements to orthology identification are also planned to span the some-

Table 1. Mapping of the current pVOGs to the older POGs

	POG 2013	pVOGs 2016
Original datasets		
Number of clusters	4542	9518
Number of genomes in clusters	1018	2976
Number of conserved proteins in clusters	58 276	195 002
Reduced intersection dataset		
Number of clusters	4201	3549
Number of genomes	982	982
Number of conserved proteins	48 246	48 246
Number of exactly identical clusters	2773	2773
Number of overlapping clusters	1153	509
Number of disjoint clusters	275	267
Agreement between datasets (%)[a]	94%	92%

[a] Percentage of cluster agreement between reduced intersection datasets of POG 2013 and pVOGs 2016.

times wide distances between VOGs that are homologous but are not detected by simple BLAST searches that lack the increased sensitivity of profile HMM approaches. In the more immediate future, the webpages at the Iowa HTTP webserver will be enhanced to allow for easier interactivity with the large-scale data by using JavaScript functionality (such as sorting, filtering and per-column searching of the large tables). Additional ease-of-use improvements that are currently under development will also continue to be released when they become ready, including additional statistics for each VOG, such as the Viral Quotient as a helpful measure of the 'virusness' of a particular gene family (propensity to be found in virus genomes versus non-prophage regions of cellular ones).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

D.M.K. and A.L.G. are supported by the Department of Biomedical Engineering at the University of Iowa (Iowa City, USA). E.V.K. is supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health (NIH) (Bethesda, USA). Funding for open access charge: NIH.
Conflict of interest statement. None declared.

REFERENCES

1. Breitbart,M. and Rohwer,F. (2005) Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.*, **13**, 278–284.
2. Wommack,K.E. and Colwell,R.R. (2000) Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.*, **64**, 69–114.
3. Brister,J.R., Ako-Adjei,D., Bao,Y. and Blinkova,O. (2015) NCBI viral genomes resource. *Nucleic Acids Res.*, **43**, D571–D577.
4. Brister,J.R., Le Mercier,P. and Hu,J.C. (2012) Microbial virus genome annotation-mustering the troops to fight the sequence onslaught. *Virology*, **434**, 175–180.
5. Liu,J., Glazko,G. and Mushegian,A. (2006) Protein repertoire of double-stranded DNA bacteriophages. *Virus Res.*, **117**, 68–80.
6. Kristensen,D.M., Cai,X. and Mushegian,A. (2011) Evolutionarily conserved orthologous families in phages are relatively rare in their prokaryotic hosts. *J. Bacteriol.*, **193**, 1806–1814.
7. Kristensen,D.M., Waller,A.S., Yamada,T., Bork,P., Mushegian,A.R. and Koonin,E.V. (2013) Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *J. Bacteriol.*, **195**, 941–950.
8. Galperin,M.Y., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.
9. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
10. Kristensen,D.M., Wolf,Y.I., Mushegian,A.R. and Koonin,E.V. (2011) Computational methods for gene orthology inference. *Brief Bioinform*, **12**, 379–391.
11. Santamaria,R.I., Bustos,P., Sepulveda-Robles,O., Lozano,L., Rodriguez,C., Fernandez,J.L., Juarez,S., Kameyama,L., Guarneros,G., Davila,G. *et al.* (2014) Narrow-host-range bacteriophages that infect Rhizobium etli associate with distinct genomic types. *Appl. Environ. Microbiol.*, **80**, 446–454.
12. Hochstein,R.A., Amenabar,M.J., Munson-McGee,J.H., Boyd,E.S. and Young,M.J. (2016) Acidianus tailed spindle virus: a new archaeal large tailed spindle virus discovered by culture-independent methods. *J. Virol.*, **90**, 3458–3468.
13. Dang,V.T. and Sullivan,M.B. (2014) Emerging methods to study bacteriophage infection at the single-cell level. *Front. Microbiol.*, **5**, 724.
14. Kato,H., Mori,H., Maruyama,F., Toyoda,A., Oshima,K., Endo,R., Fuchu,G., Miyakoshi,M., Dozono,A., Ohtsubo,Y. *et al.* (2015) Time-series metagenomic analysis reveals robustness of soil microbiome against chemical disturbance. *DNA Res.*, **22**, 413–424.
15. Waller,A.S., Yamada,T., Kristensen,D.M., Kultima,J.R., Sunagawa,S., Koonin,E.V. and Bork,P. (2014) Classification and quantification of bacteriophage taxa in human gut metagenomes. *ISME J.*, **8**, 1391–1402.
16. Vazquez-Castellanos,J.F., Garcia-Lopez,R., Perez-Brocal,V., Pignatelli,M. and Moya,A. (2014) Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics*, **15**, 37.
17. Jeffries,T.C., Ostrowski,M., Williams,R.B., Xie,C., Jensen,R.M., Grzymski,J.J., Senstius,S.J., Givskov,M., Hoeke,R., Philip,G.K. *et al.* (2015) Spatially extensive microbial biogeography of the indian ocean provides insights into the unique community structure of a pristine coral atoll. *Sci. Rep.*, **5**, 15383.
18. Bellas,C.M., Anesio,A.M. and Barker,G. (2015) Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. *Front. Microbiol.*, **6**, 656.
19. Mizuno,C.M., Rodriguez-Valera,F., Kimes,N.E. and Ghai,R. (2013) Expanding the marine virosphere using metagenomics. *PLoS Genet.*, **9**, e1003987.
20. Laffy,P.W., Wood-Charlson,E.M., Turaev,D., Weynberg,K.D., Botte,E.S., van Oppen,M.J., Webster,N.S. and Rattei,T. (2016) HoloVir: a workflow for investigating the diversity and function of viruses in invertebrate holobionts. *Front. Microbiol.*, **7**, 822.
21. Kristensen,D.M., Mushegian,A.R., Dolja,V.V. and Koonin,E.V. (2010) New dimensions of the virus world discovered through metagenomics. *Trends Microbiol.*, **18**, 11–19.
22. Rosenwald,A.G., Murray,B., Toth,T., Madupu,R., Kyrillos,A. and Arora,G. (2014) Evidence for horizontal gene transfer between Chlamydophila pneumoniae and Chlamydia phage. *Bacteriophage*, **4**, e965076.

23. Busby,B., Kristensen,D.M. and Koonin,E.V. (2013) Contribution of phage-derived genomic islands to the virulence of facultative bacterial pathogens. *Environ. Microbiol.*, **15**, 307–312.
24. Kazlauskas,D., Krupovic,M. and Venclovas,C. (2016) The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. *Nucleic Acids Res.*, **44**, 4551–4564.
25. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
26. Glazko,G.V. and Mushegian,A.R. (2004) Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biol.*, **5**, R32.
27. Huerta-Cepas,J., Szklarczyk,D., Forslund,K., Cook,H., Heller,D., Walter,M.C., Rattei,T., Mende,D.R., Sunagawa,S., Kuhn,M. *et al.* (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.*, **44**, D286–D293.
28. Glazko,G., Makarenkov,V., Liu,J. and Mushegian,A. (2007) Evolutionary history of bacteriophages with double-stranded DNA genomes. *Biol. Direct*, **2**, 36.
29. Glazko,G. and Mushegian,A. (2010) Measuring gene expression divergence: the distance to keep. *Biol. Direct*, **5**, 51.
30. Moyer,E., Hagenauer,M., Lesko,M., Francis,F., Rodriguez,O., Nagarajan,V., Huser,V. and Busby,B. (2016) MetaNetVar: Pipeline for applying network analysis tools for genomic variants analysis. *F1000Res*, **5**, 674.
31. Koonin,E.V., Krupovic,M. and Yutin,N. (2015) Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses. *Ann. N. Y. Acad. Sci.*, **1341**, 10–24.
32. Koonin,E.V. and Dolja,V.V. (2014) Virus world as an evolutionary network of viruses and capsidless selfish elements. *Microbiol. Mol. Biol. Rev.*, **78**, 278–303.
33. Kristensen,D.M., Saeed,U., Frishman,D. and Koonin,E.V. (2015) A census of alpha-helical membrane proteins in double-stranded DNA viruses infecting bacteria and archaea. *BMC Bioinformatics*, **16**, 380.
34. Iranzo,J., Krupovic,M. and Koonin,E.V. (2016) The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. *MBio*, **7**, e00978-16.
35. Iranzo,J., Puigbo,P., Lobkovsky,A.E., Wolf,Y.I. and Koonin,E.V. (2016) Inevitability of genetic parasites. *Genome Biol. Evol.* ,**8** ,2856-2869.
36. Iranzo,J., Lobkovsky,A.E., Wolf,Y.I. and Koonin,E.V. (2015) Immunity, suicide or both? Ecological determinants for the combined evolution of anti-pathogen defense systems. *BMC Evol. Biol.*, **15**, 43.
37. Walker,N.S., Stiffler,N. and Barkan,A. (2007) POGs/PlantRBP: a resource for comparative genomics in plants. *Nucleic Acids Res.*, **35**, D852–D856.
38. Bao,Y., Federhen,S., Leipe,D., Pham,V., Resenchuk,S., Rozanov,M., Tatusov,R. and Tatusova,T. (2004) National center for biotechnology information viral genomes project. *J. Virol.*, **78**, 7291–7298.
39. Koonin,E.V., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Krylov,D.M., Makarova,K.S., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N., Rao,B.S. *et al.* (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.*, **5**, R7.
40. Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2015) Archaeal clusters of orthologous genes (arCOGs): an update and application for analysis of shared features between thermococcales, methanococcales, and methanobacteriales. *Life (Basel)*, **5**, 818–840.
41. Yutin,N., Wolf,Y.I., Raoult,D. and Koonin,E.V. (2009) Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virol. J.*, **6**, 223.
42. Yutin,N., Colson,P., Raoult,D. and Koonin,E.V. (2013) Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. *Virol. J.*, **10**, 106.
43. Dutilh,B.E., Cassman,N., McNair,K., Sanchez,S.E., Silva,G.G., Boling,L., Barr,J.J., Speth,D.R., Seguritan,V., Aziz,R.K. *et al.* (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.*, **5**, 4498.
44. Tucker,K.P., Parsons,R., Symonds,E.M. and Breitbart,M. (2011) Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *ISME J.*, **5**, 822–830.
45. Kristensen,D.M., Kannan,L., Coleman,M.K., Wolf,Y.I., Sorokin,A., Koonin,E.V. and Mushegian,A. (2010) A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*, **26**, 1481–1487.
46. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
47. Kala,S., Cumby,N., Sadowski,P.D., Hyder,B.Z., Kanelis,V., Davidson,A.R. and Maxwell,K.L. (2014) HNH proteins are a widespread component of phage DNA packaging machines. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 6022–6027.
48. Chevalier,B.S. and Stoddard,B.L. (2001) Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic Acids Res.*, **29**, 3757–3774.
49. Cantalupo,P.G., Calgua,B., Zhao,G., Hundesa,A., Wier,A.D., Katz,J.P., Grabe,M., Hendrix,R.W., Girones,R., Wang,D. *et al.* (2011) Raw sewage harbors diverse viral populations. *MBio*, **2**, e00180-11.
50. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.
51. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.