vConTACT: an iVirus tool to classify viruses that infect *Archaea* and *Bacteria*

Benjamin Bolduc [1] , **Ho Bin Jang** [1] , **Guilhem Doulcier** [2] , **Zhi-Qiang You** [3] , **Simon Roux** [1] , **Matthew B Sullivan** Corresp. [1, 4]

[1] Department of Microbiology, Ohio State University, Columbus, Ohio, United States
[2] Department of Biology École Normale Supérieure, PSL Research University, Paris, France
[3] Department of Chemistry and Biochemistry, Ohio State University, Columbus, Ohio, United States
[4] Department of Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, Ohio, United States

Corresponding Author: Matthew B Sullivan
Email address: mbsulli@gmail.com

Taxonomic classification of archaeal and bacterial viruses is challenging, yet also fundamental for developing a predictive understanding of microbial ecosystems. Recent identification of hundreds of thousands of new viral genomes and genome fragments, whose hosts remain unknown, require a paradigm shift away from traditional classification approaches and towards the use of genomes for taxonomy. Here we revisited the use of genomes and their protein content as a means for developing a viral taxonomy for bacterial and archaeal viruses. A network-based analytic was optimized and benchmarked against authority-accepted taxonomic assignments and found to be largely concordant. Exceptions were manually examined and found to represent areas of viral genome ‘sequence space’ that are under-sampled or prone to excessive gene flow. While both cases are poorly resolved by genome-based taxonomic approaches, the former will improve as viral sequence space is better sampled and the latter are uncommon. Finally, given the largely robust taxonomic capabilities of this approach, we sought to enable researchers to easily and systematically classify new viruses. Thus, we established a tool, vConTACT, as an app at iVirus, where it operates as a fast, highly scalable, user-friendly app within the free and powerful CyVerse cyberinfrastructure.

vConTACT: an iVirus tool to classify viruses that infect *Archaea* and *Bacteria*

Benjamin Bolduc[1&], Ho Bin Jang[1&], Guilhem Doulcier[3,4], Zhi-Qiang You[5], Simon Roux[1] & Matthew B. Sullivan*[1,2]

[1]Department of Microbiology, The Ohio State University, Columbus, OH 43210

[2]Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH 43210

[3]École normale supérieure, PSL Research University, IBENS, F-75005, Paris, France.

[4]ESPCI, PSL Research University, CBI, F-75005, Paris, France.

[5]Department of Chemistry and Biochemistry, The Ohio State University, Columbus, OH 43210

[&] These authors contributed equally to this work.

Corresponding Author:

Matthew B. Sullivan[1,2]

Riffe Building Rm 914, 496 W 12th Ave Columbus, OH 43210, USA

Email address: mbsulli@gmail.com

Abstract.

Taxonomic classification of archaeal and bacterial viruses is challenging, yet also fundamental for developing a predictive understanding of microbial ecosystems. Recent identification of hundreds of thousands of new viral genomes and genome fragments, whose hosts remain unknown, require a paradigm shift away from traditional classification approaches and towards the use of genomes for taxonomy. Here we revisited the use of genomes and their protein content as a means for developing a viral taxonomy for bacterial and archaeal viruses. A network-based analytic was optimized and benchmarked against authority-accepted taxonomic assignments and found to be largely concordant. Exceptions were manually examined and found to represent areas of viral genome 'sequence space' that are under-sampled or prone to excessive gene flow. While both cases are poorly resolved by genome-based taxonomic approaches, the former will improve as viral sequence space is better sampled and the latter are uncommon. Finally, given the largely robust taxonomic capabilities of this approach, we sought to enable researchers to easily and systematically classify new viruses. Thus, we established a tool, vConTACT, as an app at iVirus, where it operates as a fast, highly scalable, user-friendly app within the free and powerful CyVerse cyberinfrastructure.

Introduction.

Classification of viruses that infect Archaea and Bacteria remains challenging in virology. Official viral taxonomy is handled by the International Committee for the Taxonomy of Viruses (ICTV) and organizes viruses into order, family, subfamily, genus and species. Historically, this organization derives from numerous viral features, such as morphology, genome composition, segmentation, replication strategies and amino- and nucleic-acid similarities – all of which is thought to roughly organize viruses according to their evolutionary histories (Simmonds, 2015). As of 2015, the latest report issued, the ICTV has classified 7 orders, 111 families, 27 subfamilies, 609 genera and 3704 species (http://ictvonline.org/virusTaxInfo.asp).

Problematically, however, current ICTV classification procedures cannot keep pace with viral discovery and may need revision where viruses are not brought into culture. For example, of the 4400 viral isolate genomes deposited into National Center for Biotechnology information (NCBI) viral RefSeq, only 43% had been ICTV-classified by 2015. This is because the lengthy 'proposal' processes lags deposition of new viral genomes, in some cases for years (Fauquet & Fargette, 2005). Concurrently, new computational approaches are providing access to viral genomes and large genome fragments at unprecedented rates. One approach mines microbial genomic datasets to provide virus sequences where the host is known – already adding 12,498 new prophages from publicly available bacterial and archaeal microbial genomes (Roux et al., 2015a) and 89 (69 and 20, respectively) new virus sequences from single cell amplified genome sequencing projects (Roux et al., 2014; Labonté et al., 2015). A second approach assembles viral genomes and large genome fragments from metagenomics datasets. These studies have added 15,222 and 125,842 new virus sequences from oceanic viral metagenomes (Tara Oceans Virome, or TOV (Brum et al., 2015b) and Global Oceans Virome, or GOV (Roux et al., 2016)), and from microbial and viral metagenomes from a diversity of ecosystems (Paez-Espino et al., 2016), respectively. A third approach leverages large-insert cloning and sequencing approaches to identify new viral sequences – with 208 from a single seawater sample (Mizuno et al., 2013). Such new virus genomes and large genome fragments will keep coming for the foreseeable future and represent an incredible resource for viral ecology, but they also represent a daunting challenge for taxonomy.

Currently such rapidly expanding genomic databases of the virosphere remain unclassified and challenging to integrate into a systematic framework for three reasons. First, viruses lack a universal marker gene, which prevents the taxonomic starting place that is so valuable for microbes (Woese, Kandler & Wheelis, 1990). Second, though genomes and large genome fragments are now much more readily available, researchers are reticent to use genomes as a basis for taxonomy as a paradigm has emerged whereby viruses are rampantly mosaic and therefore must exist as part of a genomic continuum such that any clustering in 'sequence space' is an artifact of sampling. This is most well studied in the many genomes of mycobacteriophages (Pope et al., 2015), but is contrasted by observations in cyanophages where efforts have been made to more deeply sample variability in a single site with findings suggesting clear population structure for naturally-occurring cyanophages (Deng et al., 2014) and that cyanophage populations appear to fit a population genetics-based species definition (Marston & Amrich, 2009; Gregory et al., 2016). It is possible that gene flow in fast evolving RNA and ssDNA viruses is rampant, but that slower evolving dsDNA viruses, particularly if obligately lytic rather than temperate, could evolve into relatively stable populations that could be the basis of taxonomy. Thus, it remains unclear whether viral genomes can serve as the sole basis for taxonomy, or whether exploration of available data could help identify areas of viral genome sequence space that are amenable to taxonomic 'rules' and others that are not.

Despite these challenges, numerous reference-independent, automated, genome-based classification schemes have been proposed. An early effort recognized that more genes are shared within related virus groups than between them (Lawrence, Hatfull & Hendrix, 2002), which led to virologists grappling with natural diversity to use translated genomes as the basis of whole genome phylogenomic tree classifications – e.g., the Phage Proteomic Tree (Edwards & Rohwer 2002). Simulations showed this method to be very accurate for assigning fragmented reads to the correct genomes (Edwards & Rohwer, 2005) but it suffers from the availability of phage genomes. A second approach that has emerged for relatively well-studied virus groups, is to use pairwise distances between aligned sequences to identify discontinuities that can indicate classification thresholds. Two tools – Pairwise Sequence Comparison (PASC; Bao 2014) and DEmARC (Lauber & Gorbalenya, 2012a) – are available to align the sequences either in context of previous knowledge of the taxonomic affiliations (PASC) or naively using pairwise distributions (DEmARC). These tools have worked well for several virus families, such as the *Picornaviridae* (Lauber & Gorbalenya, 2012a) and *Filoviridae* (Lauber & Gorbalenya, 2012b) for DEmARC, and *Bornaviridae* and arenavirus for PASC (Kuhn et al., 2014; Radoshitzky et al., 2015). However, DEmARC and PASC suffer from several issues: (i) they are not generalizable to the coming deluge of environmental viral genome sequences as they require *a priori* expert knowledge to impose similarity thresholds at each level, (ii) ICTV subcommittees have established varied sequence similarity thresholds across viral groups (Simmonds, 2015), which would require a sliding threshold, and (iii) the methods can only classify sequences that are similar to database references (Zanotto et al., 1996), which for the oceans at least represents <1% of the viral genomes recovered (Brum et al., 2015a).

Complementarily, two network-based approaches have been utilized to organize virus genome sequence space in a manner that enables classification without *a priori* knowledge. The first, a gene sharing network (Lima-Mendez et al., 2008), predicts viral genes in all the genomes, translates them into proteins, organizes these proteins into MCL-based protein families (protein clusters, “PCs”), evaluates the number of shared protein clusters pairwise throughout the dataset to establish a protein profile, and then represents this information as a weighted graph, with nodes representing viral genomes and edges the similarity score of their shared protein content. Given the 306 bacterial viruses (phages) known at the time, this method was precise as it correctly placed 92% and 95% of these phages into their correct ICTV genus or family, respectively (Lima-Mendez et al., 2008). The second, a bipartite network (Iranzo, Krupovic & Koonin, 2016), incorporates both gene sharing as above and genomic similarity. In this work, all dsDNA viruses along with mobile genetic elements were analyzed, which revealed a module-based structure to the dsDNA virosphere. These two studies imply that even very distantly related viruses can be organized into discrete populations by genomes alone and that there may be hope for automated, genome-based viral taxonomy, at least for dsDNA viruses.

Here we optimize gene sharing networks and re-evaluate their efficacy for recapitulating ICTV-based classifications using an expanded dataset of 2,010 bacterial and archaeal virus genomes (available as of RefSeq v75), while also deeply exploring where network-based methods have lower resolution and/or yield discontinuities with currently established taxonomies. Further, we make these approaches accessible to researchers by developing a tool, vConTACT (Viral CONTigs Automatic Clustering and Taxonomy), and deploy it as part of the iVirus ecosystem of apps that leverages the CyVerse cyberinfrastructure (Bolduc et al., 2016).

Materials and Methods.

Terminology. Network topological parameters, their definitions and abbreviations are available in Table 1.

Reference datasets. To test this methodology, we downloaded the entire NCBI viral reference dataset (“ViralRefSeq”, version 75, containing 5539 viruses) and removed eukaryotic viruses by filtering against tables downloaded on NCBI’s ViralRefSeq viral genome page (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239). The resulting file (“Bacterial and Archaeal viruses”; BAV) contained 2010 total viruses; 1905 dsDNA, 88 ssDNA, 5 dsRNA and 12 ssRNA. All viruses contained taxonomic affiliation information, though not all viruses had affiliations associated with each level of the taxonomy (i.e. not all viruses have a “sub-family” designation). To improve taxonomic assignments, the ICTV taxonomy was also retrieved (https://talk.ictvonline.org/files/master-species-lists/) and the ICTV affiliations were used to supplement the NCBI version.

Building protein cluster profiles. To generate sequence profiles – with information about the presence or absence of a sequence within one or more protein clusters (described previously as protein *families* (Lima-Mendez et al., 2008)), proteins from each sequence were first extracted from the ViralRefSeq proteins file. BLASTP (Altschul et al., 1997) was used to compare all proteins (198,102) from the sequences in an all-verses-all pairwise comparison. Protein clusters were subsequently identified using the Markov clustering algorithm (MCL) with an inflation value of 2, resulting in 23,022 protein clusters (“PCs”). Finally, we generated protein cluster profiles for each genome such that the presence of a gene within a protein cluster of a viral genome was given a value of “1” and the absence “0”. This resulted in a large 2,010 x 23,022 matrix.

Generating the similarity network. The similarity network is a graph where the nodes (i.e. reference sequences) are linked by edges when the similarity between their pc-profiles is considered significant. In other words, the network represents the overall similarity between sequences based on the number of shared protein clusters. To calculate the similarity between the profiles of two sequences (sequence A and sequence B), the hypergeometric formula was used to estimate the probability that at least *c* protein clusters would be in common:

$$P(X \geq c) = \sum_{i=c}^{\min(a,b)} \frac{C_a^i C_{n-a}^{b-i}}{C_n^b} \tag{1}$$

Simply stated, the hypergeometric formula is the probability that genomes A and B would have *c* protein clusters in common by chance. The probability can be converted to an expectation value (E; for false positives) by multiplying the probability (P) by the total number of comparisons (T). The expectation value can then be converted into a significance score:

$$S(A,B) =- \log(E) =- \log(P \times T) \tag{2}$$

Genome pairs with significance scores greater than 1 (i.e. E-value < 0.1) are considered sufficiently similar (see permutation test, below) and were joined by an edge in the similarity network with a weight equal to their significance score. We refer to sequences within the network as *nodes*, the relationships connecting them, *edges* and the strength of that relationship, edge *weight*.

After generating the similarity network, groups of similar sequences (referred to as viral clusters, “VCs”) were clustered by applying MCL with an inflation of 2.

Measuring the proportion of shared genes between genomes. Given that genome sizes between pairs can differ greatly, this can lead to large differences in the proportion of the shared genes (Ågren et al., 2012). To counter this, we characterized the proportion of shared PCs between two genomes using the geometric index (G) as a symmetric index:

$$\mathrm{G}_{AB} = \frac{|N(A) \cap N(B)|}{|N(A)| \times |N(B)|} \tag{3}$$

where *N(A)* and *N(B)* indicate the numbers of protein clusters (PCs) in the genomes of A and B, respectively.

Permutation test. The stringency of the significant score was evaluated through randomization of the original matrix where rows present viral genomes and columns PCs or singletons that are not shared with any other protein sequences (Leplae et al., 2004). Briefly, with an in-house R script, 1,000 matrices were generated by randomly rearranging PCs and/or singletons within pairs of genomes having the significant score ≤ 1 (a negative control) and calculated a new significant score each time. None of the genome pairs in this negative control produced significant scores >1, indicating values above this significance threshold did not occur by chance (Lima-Mendez et al., 2008).

Affiliating sequence clusters with taxonomic groups. To assign (in the case of unknown sequences) or compare nodes (genomes) within clusters to their reference counterparts, we first defined *membership* of a node *c* to a cluster *k* $B(c,k)$ according to two methods, conservative and permissive. The conservative method 4) directly takes the result from the MCL clustering to assign a node to a cluster:

$$B(c,k) = \begin{cases} 1 \text{ if Contig } c \in \text{Cluster } k, \\ 0 \quad \text{otherwise} \end{cases} \tag{4}$$

while the permissive method takes the sum of all edge weights *w* linking the node to nodes of the cluster, with the node becoming a member of its maximal membership cluster (5):

$$B^{'}(c,k) = \frac{\sum_{i \in k} w_{c,i}}{\sum_{p \in \{\text{Clusters}\}} \sum_{j \in p} w_{g,j}} \tag{5}$$

The precision $P(k,t)$ of the taxonomic class *t* with respect to a cluster *k* as its membership of nodes of the class *t* in the memberships of reference nodes in the cluster *k*.

$$P(k,t) = \frac{\sum_{\forall i \in \{\text{sequence of class } t\}} B(i,k)}{\sum_{\forall j \in \{\text{reference sequence}\}} B(j,k)} \qquad (6)$$

A cluster and all its node members are then affiliated with its maximal precision class. For the conservative method, the cluster is affiliated with the taxonomic class associated with the majority of its members (i.e. ≥ 50%). In cases where clusters do not contain at least half reference sequences, the entire cluster will be unaffiliated.

Measuring the connectivity of genomes to clusters. The connection strength of a node *g* to cluster *c* was calculated as the average edge weight linking it to nodes of cluster *c*:

$$W_{g,c} = \frac{1}{k}\sum_{i=1}^{k} wg,i \qquad (7)$$

where *k* and *w* are the number and total weight of edges of the node *g* in the cluster *c*, respectively. We refer to the average edge weight for node *g* to the cluster it belongs to as its in-VC average weight, and to other clusters within the network as out-VC average weight.

Identifying sub-clusters. To further subdivide heterogeneous clusters (those comprising ≥ 2 taxa), cluster-wise module profiles (i.e. a module profile only including viruses previously identified as belonging to the same viral cluster) were hierarchically clustered using UPGMA with pairwise distances calculated using Euclidean distance implemented in Scipy.

Statistical calculations. All calculations, statistics, network statistical analyses were performed using in-house python scripts, with the Numpy, Scipy, Biopython and Pandas python-packages. vConTACT is implemented in python with the same dependencies. The tool is available at https://bitbucket.org/MAVERICLab/vcontact. Scripts used in the generational and calculations of data are available at https://bitbucket.org/MAVERICLab/vcontact-SI.

Network visualization and analysis. The network was visualized with Cytoscape (version 3.1.1; http://cytoscape.org/), using an edge-weighted spring embedded model, which places the genomes or fragments sharing more PCs closer to each other. Topological properties were estimated using a combination of python and the Network Analyzer 2.7 Cytoscape plug-in (Assenov et al., 2008).

Results and Discussion.

vConTACT analytical workflow and terminology**:** The vConTACT analyses are based on previously established gene sharing network methods (Lima-Mendez 2008). Briefly, PCs are established across all genomes in the dataset; with vConTACT doing this by default using MCL clustering from all-verses-all BLASTP comparisons (though user-specified clusters can also be used). PC *profiles* of genomes or genome fragments (herein ‘genome’) are then calculated, where the presence and absence of PCs (from the entire PC dataset) along a genome are established and then compared pairwise between genomes (Fig. 1). The pairwise genome comparisons are then mathematically adjusted (using the hypergeometric similarity formula) to establish a probability that any genome pair would share *n* PCs, given the total number of all PCs. This probability is log-transformed (in similar fashion to BLAST E-values) into a significance score and applied as a *weight* to an edge between the two paired genomes in a similarity network. High significance scores represent a low probability that two genomes would share *n* PCs by chance, which can be interpreted as evidence of gene-sharing and presumably evolutionary relatedness between the paired genomes. After evaluating all pairings in the dataset, significance scores ≥1 are retained, and a network of the remaining genome pairs is constructed. MCL is subsequently applied to identify structure in the gene sharing network, but now the clusters represent groups or related genomes and are termed *viral* clusters ("VCs"). MCL is also applied against the network of PCs, whose members can be similar to members of other PCs. This effectively organizes the PCs into a higher-order structure known as a protein module. The relationship information identified from the genomes (organized into VCs) and PCs (organized into protein modules) are used to create a *module profile*, which can then be mined for taxonomic identification, functional profiling, etc.

Benchmarking network-based taxonomy: To benchmark the ability of network-based taxonomy to capture 'known' viral relationships, we evaluated how vConTACT "re-classified" viral sequences at various taxonomic levels using 2,010 bacterial and archaeal viral genomes from VirRefSeq (v75). Of these reference genomes, ICTV-classifications were only available for a subset; 654 viruses from 2 orders, 738 viruses from 19 families, 152 viruses from 11 subfamilies, and 562 viruses from 158 genera. The network was then decomposed into VCs (described above) and a permutation test was used to establish significance score thresholds to prevent random relationships from entering the network. This analysis used the initial network's edge information to construct a matrix between genome pairs, and then permuted the edges 1,000 times. No edges were found to be significant during these tests, suggesting that relationships seen within the network did not arise by chance and could be confidently used to establish taxonomic groupings (see Materials and Methods, Table S1).

The resulting network, consisting of 1,964 viruses (nodes) and 65,393 relationships (edges) between them (Fig. 2A), was then used as a basis for comparison to the ICTV-based classifications. A total of 211 VCs were identified, spread among 46 components (unconnected subnetworks), which more than doubles the 17 connected components identified previously (Lima-Mendez et al., 2008). Of the 46 components, 38 included 1,891 phages representing 194

VCs (left, Fig. 2A), and 8 components included 73 archaeal viruses representing 17 VCs (right, Fig. 2A). Most (87%) of the 1,891 phages belonged to the order *Caudovirales*, and comprised the largest connected component (LCC) in the analysis (top left, Fig. 2A). At the VC level, the network clustering performed well with average (across each taxonomic level) recall / precision percentages of 100% / 100%, 90% / 86%, and 80% / 80% at the order, family and genus levels, respectively (Fig. 2B). Of the 211 VCs resolved by the network, 76.4% contained a single ICTV-accepted genus, suggesting concordance between the network VCs and accepted taxonomy, whereas 15.1% and 8.5% of the VCs contained two and 3 or more genera, respectively (Fig. 2A and C). Thus, roughly 4 out of 5 of the VCs correspond to ICTV genera.

Mechanistically, these discrepancies between network clustering and the ICTV classification could derive from: (*i*) undersampling such that VCs with fewer members may not represent the naturally-occurring diversity of that viral group, or (*ii*) gene flow between viral genomes that ameliorates taxonomic boundaries by providing excessive access to genes outside the VC's gene pool. While much of abundant viral genome sequence space is recently being mapped (Paez-Espino et al., 2016; Roux et al., 2016), there remains contrasting paradigms about the role of gene flow in structuring mycobacteriophage ("mycophages") vs cyanophage populations (Gregory et al., 2016).

To discriminate between these possibilities, we identified these "ICTV-discordant" areas of the network containing 2 or more ICTV genera (which we define as *heterogeneous VCs*), focusing on three of the more well-populated (many member genomes) heterogeneous VCs, and the archaeal virus heterogeneous VCs, which are among the least well-sampled taxa. Of the well-sampled VCs, VCs containing the 2nd, 3rd, and 4th most members (i.e. genomes), included the following: (i) VC1 contains the 8 genera belonging to the *Tevenvirinae* subfamily (*T4virus*, *Cc31virus*, *Js98virus*, *Rb49virus*, *Rb69virus*, *S16virus*, *Sp18virus*, and *Schizot4virus*) and a genus of the *Eucamyvirinae* (*Cp8virus*), as well as the *Tg1virus* and *Secunda5virus* that are not assigned to a particular subfamily, (ii) VC2 contains three genera (*Biseptimavirus*, *Phietavirus*, and *Triavirus*), and (iii) VC3 contains four genera (*Kayvirus*, *Silviavirus*, *Twortvirus*, and *P100virus*) belonging to the *Spounavirinae* and the six *Bacillus* virus genera (*Agatevirus*, *B4virus*, *Bc431virus*, *Bastillevirus*, *Nit1virus*, and *Wphvirus*). Finally, among the 73 archaeal viruses, only the *Fuselloviridae* were accurately classified at the genus level, while most (63%) archaeal viruses were incorrectly classified at the genus level.

***Gene content analyses suggest ICTV classifications, not VC-based classifications should be revised*:** A total of 23.6% of the VCs contained genomes from more than one ICTV-recognized genus, which suggests ‘lumping’ by the network analyses (via MCL) or ‘splitting’ during ICTV classification. To assess this, we computed the fraction of PCs that were shared both within an ICTV genus and between the multiple ICTV genera found in each heterogeneous VC and represented them as the percentage of intragenus similarity and intergenera similarity, respectively. Of the 25 VCs, intragenus similarities of all but one (VC9) suggested they shared more than 40% of their PCs (Fig. 3A, Table S2), which is consistent with the threshold commonly used to define a new dsDNA viral genus (Lavigne et al., 2009). In contrast, the intergenera similarities varied widely – some VCs (VCs 1-3, 9-11, 17, 20, 25, 33, 58, 91, 95) shared 20-40% of their PCs (subfamily level), whereas others shared more than ~40% (VCs 12, 14, 24, 26, 37, 44, and 51) or less than ~20% (VCs 39, 55, 63, 74, and 77) of their PCs. Where intergenera similarities are high (>40% of the PCs are shared), there may be a case to be made for merging the currently recognized ICTV genera. Consistent with this, all 6 of these highly (>40%) similar VCs (12, 14, 24, 26, 37 and 51) are suggested to be in need of revision, as these include *G7Cvirus*, *N4virus*, *T1virus*, *HP34virus*, and *PhiKMVvirus* (Wittmann et al., 2015; Eriksson et al., 2015; Niu et al., 2014; Krupovic et al., 2016). Additionally, we found that in VC44, the phage CAjan, belonging to the *Seuratvirus*, shared 41.6-42.7% of its genes with three phages (JenP1 and 2 and JenK1 of the *Nongaviru*s (Table S2)). Where intergenera similarities are lower (<20%, or 20-40% of the PCs are shared), the appropriate taxonomic assignment may require deeper sampling of viral genome sequence space and/or further network analytic development.

To further assess these cases, we next examined four VCs (1-3, 14) that contained more than 4 ICTV-recognized genera using hierarchical clustering of PC presence-absence data for each genome (Fig. 3B). In parallel, we computed the actual connectivity of the genomes within these heterogeneous VCs according to the average weight of edges that (i) are between genomes of the same VC (in-VC avg. weight) and (ii) between the genomes of other VCs (out-VC avg. weight) (Table S3; Materials and Methods). For example, within VC1, 8 genera of the *Tevenvirinae* (*S16virus*, *Cc31virus*, *T4virus*, *Rb69virus*, *Sp18virus*, *Js98virus*, *Rb49virus* and *Schizotvirus*) and their relatives (*Tg1virus* and *Secunda5virus*) share, on average, 61% and 38% of their total PCs, respectively, and 39% between all 10 genera (Table S2). Outside VC1, they share ~11.2% of genes with other viral groups (Table S2). We found that the 10 genera within VC1 are more tightly interconnected than those of the 210 VCs overall, with average in-cluster values of 223.7 and 131.9 and average out-cluster values of 13.1 and 9.0, respectively (Table S3). These observations indicate that higher cross-similarities of 10 genera can be attributed to a large fraction of their shared genes, whereas only a small fraction of gene shared by other groups can hold them together.

Upon closer inspection, some of this ‘lumping’ appeared to be due to poorly sampled regions of sequence space. For example, VC1 also contained the *Cp8virus* of the subfamily *Eucampyvirinae*, which is odd to be placed alongside the *Tevevirinae*, given that other ICTV-recognized genus (*Cp220virus*) of the *Eucampyvrinae* is grouped into the separate cluster (VC 87). Since both genera (*Cp8virus* and *Cp220virus*) are distantly related to the *Tevenvirinae* (Javed et al., 2014), containing only ~11% the shared genes to (an average weight of 18.5) and ~6% (11.8) , respectively (Tables S2 and S3), these groupings might be driven by the lack of diversity among the *Cp220virus* where only 2 reference genomes (i.e., *Campylobacter* phages CPX and NCTC12673) available in our ViralRefSeq dataset. To test this, we computationally doubled the number of the genomes for this group while holding the number and weight of their connections constant, finding that the *Cp220virus* genomes clearly separated from VC1 and instead were correctly placed alongside VC 87 (Table S4). Consistently, among the heterogeneous VCs 39, 55, 63, 74, and 77 showing < ~20% intergenera similarities (Figs. 3A and S1), increasing the genome numbers of poorly-sampled ICTV genera led to clustering of members of those genera into their correct VCs (Table S4). Together these findings suggest that additional sampling in poorly sampled areas of viral sequence space will be required to most accurately establish genome-based taxonomy – issues that parallel those presented by long branch attraction for phylogenies (Bergsten, 2005).

Similar structure emerged from hierarchical clustering of PC presence / absence data from the 3 other well-represented heterogeneous VCs. In VC2, the three known subgroups of the *Phietavirus* (Gutiérrez et al., 2014) were resolved, sharing 44.9% of their PCs, and separate from two other subgroups – the *Biseptimavirus* and *Triavirus*, which shared 22.3% of their PCs (Fig. 3B, Table S2). In VC3, containing the *Spounavirinae* (Krupovic et al., 2016), each sub-cluster has a corresponding ICTV genus with largely overlapping sets of genes while also showing a clearly distinct set(s) of genes. Of these, the six *Bacillus* virus genera (*Wphvirus*, *Bastillevirus*, *B4virus*, *Bc431virus*, *Agatevirus*, and *Nit1virus*) appear to be closely related to the *Spounavirinae*, with ~20% of total PCs in common (Fig. 3B, Table S2). Finally, VC14 produced a clear division of the *Tunavirinae* (Krupovic et al., 2016), in which the *Escherichia* virus Jk06 is placed in a separate branch due to its less shared common genes (~56%) to the other *Rogue1virus* members (~82%); their highly-overlapped genes between genera above the genus boundary (40%) are associated with "taxonomic lumping" as described above (Niu et al., 2014; Krupovic et al., 2016).

We next evaluated three phage groups which were poorly represented in the S277 network (Lima-Mendez et al 2008) and also represent some of the most abundant, widespread, and/or extensively studied phage groups (Grose & Casjens, 2014; Pope et al., 2015; Roux et al., 2015b)) – the mycobacteriophages, *Teveniviriae*, *Autographivirinae* and the archaeal viruses.

Mycobacterium phages. The largest viral group covering 16.1% of the total population of the LCC (mostly *Caudovirales*, top left Fig. 1A) includes phages infecting *Mycobacteria*. The 318 mycophage genomes were assigned to 14 VCs (Fig. 4A), 13 of which were composed of reference genomes belonging to a single ICTV-recognized genus for each VC. The 14th mycophage VC, VC25, contained three ICTV-recognized genera – the *Bignuzvirus*, *Charlievirus*, and *Che9cvirus*. Although the module-based approach discerned the structure in this VC, which would group them into the known genera (Fig. S1), this "lumping" into a single VC reflects (*i*) their undersampling (i.e., each genus has 1 to at most 3 viruses) and/or (*ii*) highly-overlapped genes between genera. Indeed, of the 3 phages belonging to the *Che9cvirus*, phages Babsiella and Che9c shared 45% of their genes, but also shared 35% and 36% of their genes with the *Bignuzvirus* and 28% and 32% with the *Charlievirus*, respectively (Table S2), which results in higher connectivity between three genera than other viral groups to which they linked (Table S3). These findings contrast those in the rest of the network, and suggest that some phage groups (e.g., mycophages) may more frequently exchange genes than others.

To quantify this, we next examined features of the network. For example, many VC59 mycophages were broadly linked to nine VCs that contain other mycophages and phages from diverse hosts (Fig. 4A). To characterize this further, we analyzed the topological properties using the betweenness centrality (BC), as it can identify the node residing in the shortest path between other nodes (Halary et al., 2009). Specifically, in the shared-gene network, high-betweenness nodes (phages) can act as bridges between phages that would remain disconnected, due to their mosaic content of genes (Lima-Mendez et al., 2008). Indeed, these eight VC 59 phages had 42-fold higher average BC than those of other mycophages and their relatives (0.04 vs. 9.45E-04) (Fig. S2), strongly indicating they may be prone to increased gene flow and thus exceptionally 'mosaic' genomes (Halary et al., 2009; Pope et al., 2015a).

The Tevenvirinae As the second-largest group, containing 94 viruses in the heterogeneous VC1, which were further connected to 74 distant relatives and taxonomically unclassified myo-/siphoviruse(s), appeared to be restricted to a densely interconnected region (Fig. 4). A subsequent hierarchical clustering within VC1 grouped these 168 viral genomes into 5 subgroups (Fig. S3). Interestingly, three phages infecting cyanobacteria (P-SSM2, P-SSM4, and S-PM2) and T4-like phages that were initially found in a single cluster (Lima-Mendez et al., 2008) are separated into two clusters of the Exo T-evens (VC_8) and T-evens/Pseudo/Schizo T-evens (VC_1), respectively (Filee, 2006) (upper in Fig. 4B; Fig. S3). This network grouping can identify the Exo T-evens including cyano- and pelagiphages, which the literature suggests to be only distantly related to other T4 superfamily viruses (Comeau & Krisch, 2008; Roux et al., 2015b).

The Autographivirinae We further identified 8 VCs associated with the *Autographivirinae*. Of four genera defined by the NCBI and/or ICTV, the *T7virus*, *SP6virus*, *Kp34virus* were found in VCs 4, 28, and 37, respectively, whereas the *PhiKMVvirus* were spread across VCs 13 and 37 (Fig. 4B; also Fig. S4). Notably, a previous phylogenetic study based on three conserved proteins (i.e., RNA polymerase, head-tail connector and the DNA maturase B) showed considerable diversity of the *phiKMVvirus* (Eriksson et al., 2015). We also observed distinct patterns of PC sharing between the PhiKMV-related genome(s) and other viruses in each cluster (Fig. S4), suggesting that the *PhiKMVvirus* should likely be divided into two new subgroups.

In addition, as the recently emerged groups, nine *Acinetobacter* phages (Huang et al., 2013), as well as phage vB_CsaP_GAP227 (Abbasifar et al., 2013) and its close relatives were found in VCs 54 and 93, respectively (Fig. S4); all of them encode T7-specific RNA polymerase (Lavigne et al., 2009), which suggest that they fall within the *Autographivirnae* subfamily.

Finally, many viruses are now thought to co-opt host genes to improve viral fitness; these stolen ‘auxiliary metabolic genes’ are known from cyanophage genomes (photosynthesis genes; (Sullivan et al., 2006; Millard et al., 2009; Labrie et al., 2013), but also from ocean viral metagenomes where viruses are now shown to contain genes involved in central carbon metabolism (Hurwitz, Hallam & Sullivan, 2013) and nitrogen and sulfur cycling (Roux et al., 2016) in ways that likely drive niche differentiation (Hurwitz, Brum & Sullivan, 2014). Thus it is striking that VC22 in our network, which contains 19 cyanopodoviruses, had many linkages to taxonomically disparate *Tevenvirinae*, which turned out the be driven by photosynthesis genes shared across these viral taxa (Fig. 4B). Such “host” genes in viruses can bring taxonomically disparate viral groups closer together, and the network can thus help identify such niche defining viral genes for viruses infecting well studied hosts.

The Archaeal Viruses. Of the 72 archaeal viruses, 66 were associated with 18 VCs, while 6 viruses (Haloviruses HHTV-1 and VNH-1, Hyperthermophilic Archaeal Virus 1 & 2, Pyrococcous abyssi virus 1, and His 1 virus) were not included in the network, due to lack of statistically significant similarity to any other virus. Of the 25 heterogeneous VCs, archaeal viruses comprise 3 of them (VCs 51, 74 and 77), likely owing to their gene products showing little similarity to published viruses outside of other archaeal viruses (Prangishvili, Garrett & Koonin, 2006). All 3 VCs show considerable sharing of PCs within each VC (61.3 %, 50.2 % and 67.6 %, respectively). VCs 74 and 77, each consisting of 2 genera (*Gammalipothrixvirus*/R*udivirus* and *Betalipothrixvirus*/*Deltalipothrixvirus*) unify the entire *Ligamenvirales* order (2 families). Though the genera are distinguished mainly by their virion morphology (Prangishvili & Krupovič, 2012), it can be argued that some lipothrixviruses share as much similarity within the *Lipothrixviridae* family as to the rudiviruses, exemplified by the 10 genes shared between AFV-1 (a lipothrixvirus) and SIRV1 (a rudivirus) (Prangishvili & Krupovič, 2012) and that they likely derive from a common ancestor (Goulet et al., 2009). In addition to the number of PCs shared between AFV-1 and the rudivirus in VC74 (Fig. S1), the more “distal” position between AFV-2 (*Deltalipothrixvirus*) and the other VC77 members (*Betalipothrixvirus*) (Fig. S1), the order-level separation is easily seen in the overall network (Fig. 2). VC55 (*Alphafusellovirus*/*Betafusellovirus*) consists of all known *Fuselloviridae* members. Like VCs 74 and 77, their genera are separated mainly through virion morphology, with Alphafusellovirus lemon-shaped and Betafusellovirus pleomorphic, and also through their attachment structures (Redder et al., 2009). The large number of "core" genes (13) shared among all family members argues for frequent recombination events, with even distant fuselloviruses potentially capable of recombination during integration. Furthermore, some fuselloviruses exhibit regions >70% pairwise identity on the nucleotide level, including ASV-1 (*Betafusellovirus*) and SSV-K1 (*Alphafusellovirus*) (Redder et al., 2009). Despite shared non-core regions between the *fusselloviridae*, the high similarity between the two genera is also revealed in the network through unification into a single VC. The most recently identified member of the *Fuselloviridae*, *Sulfolobales* Mexican fusellovirus 1 (SMF1) has no official ICTV classification between family, though clustering within the VC shows clear association with the *Betafusellovirus*. It is remarkable that nearly all known archaeal viruses not only fall within the network, but that most of their VCs follow a genus-level affiliation.

vConTACT, ***an iVirus tool for network-based viral taxonomy***: Given the strong and robust performance of these network classification methods (Lima-Mendez et al., 2008) to largely capture known viral taxonomy from genomes alone, we sought to democratize the analytical capability. To this end, we developed a tool named "vContact" (overview of its logic in Fig. 1) and integrated it into iVirus, a virus ecology-focused set of tools also known as "apps" and databases (Bolduc et al., 2016). Such implementation at iVirus enables any user to run the application simply by inputting viral sequences with all compute, storage and data repository happening via the CyVerse cyberinfrastructure (formerly the iPlant Collaborative (Goff et al., 2011).

Conclusions

Network-based approaches have been widely used to explore mathematical, statistical, biological, and structural properties of a set of entities (nodes) and the connections between them (edges) in a variety of biological and social systems (Dagan, 2011; Barberán et al., 2012). Such approaches are invaluable for developing a quantitative framework to evaluate if and where taxonomically meaningful classifications can be made in viral sequence space (Simmonds et al., 2016). By expanding upon prior large-scale analyses (Lima-Mendez et al., 2008; Koonin, Krupovic & Yutin, 2015; Roux et al., 2015a) we sought here to quantitatively evaluate when and where such network-based classifications will perform poorly. These findings suggest that under sampled viral sequence space and some phage groups with exceptionally high gene flow (e.g., mosaic genomes of the mycophages) are currently challenging and represent about 1 in 4 publicly-available, dsDNA viral genomes. While these problematic viral genomes await improved representation of viral sequence space and/or improvements in network analytics to best resolve their taxonomy, the remaining ¾ of viral genomes appear ready for gene sharing network-based viral taxonomy. To this end, we present vConTACT as a publicly-available tool for researchers to effectively enable large-scale, automated virus classification. Given the scale of thousands of new virus genomes and genome fragments discovered through increasingly used metagenomics approaches (Roux 2016 and Paez et al 2016), such step-wise progress towards an automated taxonomic classifier will be foundational to most rapidly integrate viruses into models that seek to make predictions in ecosystems ranging from the oceans and soils to bioreactors and humans.

Acknowledgements.

We thank Kate Hargreaves, Consuelo Gazitua, Gareth Trubl, and Dean Vik for testing out beta versions of the vConTACT app, Bonnie Hurwitz and Ken Youens-Clark and CyVerse for help implementing the app, Sullivan Lab for critical review through the years and comments on the manuscript.

References

Abbasifar R., Kropinski AM., Sabour PM., Ackermann H-W., Alanis Villa A., Abbasifar A., Griffiths MW. 2013. The Genome of Cronobacter sakazakii Bacteriophage vB_CsaP_GAP227 Suggests a New Genus within the Autographivirinae. *Genome Announcements* 1:e00122-12-e00122-12. DOI: 10.1128/genomeA.00122-12.

Ågren J., Sundström A., Håfström T., Segerman B. 2012. Gegenees: Fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PLoS ONE* 7. DOI: 10.1371/journal.pone.0039107.

Altschul SF., Madden TL., Schäffer AA., Zhang J., Zhang Z., Miller W., Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389–3402.

Assenov Y., Ramirez F., Schelhorn S-E., Lengauer T., Albrecht M. 2008. Computing topological parameters of biological networks. *Bioinformatics* 24:282–284. DOI: 10.1093/bioinformatics/btm554.

Barberán A., Bates ST., Casamayor EO., Fierer N. 2012. Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME Journal* 6:343–351.

Bergsten J. 2005. A review of long-branch attraction. *Cladistics* 21:163–193. DOI: 10.1111/j.1096-0031.2005.00059.x.

Bolduc B., Youens-Clark K., Roux S., Hurwitz BL., Sullivan MB. 2016. iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. *The ISME Journal*:1–8. DOI: 10.1038/ismej.2016.89.

Brum JR., Ignacio-Espinoza JC., Roux S., Doulcier G., Acinas SG., Alberti A., Chaffron S., Cruaud C., de Vargas C., Gasol JM., Gorsky G., Gregory AC., Guidi L., Hingamp P., Iudicone D., Not F., Ogata H., Pesant S., Poulos BT., Schwenck SM., Speich S., Dimier C., Kandels-Lewis S., Picheral M., Searson S., Tara Oceans Coordinators., Bork P., Bowler C., Sunagawa S., Wincker P., Karsenti E., Sullivan MB. 2015a. Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science (New York, N.Y.)* 348:1261498. DOI: 10.1126/science.1261498.

Brum JR., Ignacio-Espinoza JC., Roux S., Doulcier G., Acinas SG., Alberti A., Chaffron S., Cruaud C., de Vargas C., Gasol JM., Gorsky G., Gregory AC., Guidi L., Hingamp P., Iudicone D., Not F., Ogata H., Pesant S., Poulos BT., Schwenck SM., Speich S., Dimier C., Kandels-Lewis S., Picheral M., Searson S., Bork P., Bowler C., Sunagawa S., Wincker P., Karsenti E., Sullivan MB. 2015b. Patterns and ecological drivers of ocean viral communities. *Science* 348:1261498–1261498. DOI: 10.1126/science.1261498.

Comeau AM., Krisch HM. 2008. The Capsid of the T4 Phage Superfamily: The Evolution, Diversity, and Structure of Some of the Most Prevalent Proteins in the Biosphere. *Molecular Biology and Evolution* 25:1321–1332. DOI: 10.1093/molbev/msn080.

Dagan T. 2011. Phylogenomic networks. *Trends in Microbiology* 19:483–491. DOI: 10.1016/j.tim.2011.07.001.

Deng L., Ignacio-Espinoza JC., Gregory AC., Poulos BT., Weitz JS., Hugenholtz P., Sullivan MB. 2014. Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature* advance on. DOI: 10.1038/nature13459.

Edwards RA., Rohwer F. 2005. Viral metagenomics. *Nature Reviews Microbiology* 3:504–510.

Eriksson H., Maciejewska B., Latka A., Majkowska-Skrobek G., Hellstrand M., Melefors Ö., Wang JT., Kropinski AM., Drulis-Kawa Z., Nilsson AS. 2015. A suggested new bacteriophage genus, “Kp34likevirus”, within the Autographivirinae subfamily of podoviridae. *Viruses* 7:1804–1822. DOI: 10.3390/v7041804.

Fauquet CM., Fargette D. 2005. International Committee on Taxonomy of Viruses and the 3,142 unassigned species. *Virology journal* 2:64. DOI: 10.1186/1743-422X-2-64.

Filee J. 2006. A Selective Barrier to Horizontal Gene Transfer in the T4-Type Bacteriophages That Has Preserved a Core Genome with the Viral Replication and Structural Genes. *Molecular Biology and Evolution* 23:1688–1696. DOI: 10.1093/molbev/msl036.

Goff S a., Vaughn M., McKay S., Lyons E., Stapleton AE., Gessler D., Matasci N., Wang L., Hanlon M., Lenards A., Muir A., Merchant N., Lowry S., Mock S., Helmke M., Kubach A., Narro M., Hopkins N., Micklos D., Hilgert U., Gonzales M., Jordan C., Skidmore E., Dooley R., Cazes J., McLay R., Lu Z., Pasternak S., Koesterke L., Piel WH., Grene R., Noutsos C., Gendler K., Feng X., Tang C., Lent M., Kim S-J., Kvilekval K., Manjunath BS., Tannen V., Stamatakis A., Sanderson M., Welch SM., Cranston K a., Soltis P., Soltis D., O'Meara B., Ane C., Brutnell T., Kleibenstein DJ., White JW., Leebens-Mack J., Donoghue MJ., Spalding EP., Vision TJ., Myers CR., Lowenthal D., Enquist BJ., Boyle B., Akoglu A., Andrews G., Ram S., Ware D., Stein L., Stanzione D. 2011. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Frontiers in Plant Science* 2:1–16. DOI: 10.3389/fpls.2011.00034.

Goulet A., Blangy S., Redder P., Prangishvili D., Felisberto-Rodrigues C., Forterre P., Campanacci V., Cambillau C. 2009. Acidianus filamentous virus 1 coat proteins display a helical fold spanning the filamentous archaeal viruses lineage. *Proceedings of the National Academy of Sciences* 106:21155–21160.

Gregory AC., Solonenko SA., Ignacio-Espinoza JC., LaButti K., Copeland A., Sudek S., Maitland A., Chittick L., dos Santos F., Weitz JS., Worden AZ., Woyke T., Sullivan MB. 2016. Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics* 17:930. DOI: 10.1186/s12864-016-3286-x.

Grose JH., Casjens SR. 2014. Understanding the enormous diversity of bacteriophages: The tailed phages that infect the bacterial family Enterobacteriaceae. *Virology* 468:421–443. DOI: 10.1016/j.virol.2014.08.024.

Halary S., Leigh JW., Cheaib B., Lopez P., Bapteste E. 2009. Network analyses structure genetic diversity in independent genetic worlds. *Proceedings of the National Academy of Sciences* 107:127–132. DOI: 10.1073/pnas.0908978107.

Huang G., Le S., Peng Y., Zhao Y., Yin S., Zhang L., Yao X., Tan Y., Li M., Hu F. 2013.

Characterization and genome sequencing of phage Abp1, a new phiKMV-like virus infecting multidrug-resistant acinetobacter baumannii. *Current Microbiology* 66:535–543. DOI: 10.1007/s00284-013-0308-7.

Hurwitz BL., Brum JR., Sullivan MB. 2014. Depth-stratified functional and taxonomic niche specialization in the “core” and “flexible” Pacific Ocean Virome. *The ISME journal*:1–13. DOI: 10.1038/ismej.2014.143.

Hurwitz BL., Hallam SJ., Sullivan MB. 2013. Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biology* 14:R123. DOI: 10.1186/gb-2013-14-11-r123.

Iranzo J., Krupovic M., Koonin E V. 2016. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *mBio* 7:e00978-16. DOI: 10.1128/mBio.00978-16.

Koonin E V., Krupovic M., Yutin N. 2015. Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses. *Annals of the New York Academy of Sciences*:n/a-n/a. DOI: 10.1111/nyas.12728.

Krupovic M., Dutilh BE., Adriaenssens EM., Wittmann J., Vogensen FK., Sullivan MB., Rumnieks J., Prangishvili D., Lavigne R., Kropinski AM., Klumpp J., Gillis A., Enault F., Edwards RA., Duffy S., Clokie MRC., Barylski J., Ackermann H-W., Kuhn JH. 2016. Taxonomy of prokaryotic viruses: update from the ICTV bacterial and archaeal viruses subcommittee. *Archives of Virology* 161:1095–1099. DOI: 10.1007/s00705-015-2728-0.

Kuhn JH., Dürrwald R., Bào Y., Briese T., Carbone K., Clawson AN., deRisi JL., Garten W., Jahrling PB., Kolodziejek J., Rubbenstroth D., Schwemmle M., Stenglein M., Tomonaga K., Weissenböck H., Nowotny N. 2014. Taxonomic reorganization of the family Bornaviridae. *Archives of Virology* 160:621–632. DOI: 10.1007/s00705-014-2276-z.

Labonté JM., Swan BK., Poulos B., Luo H., Koren S., Hallam SJ., Sullivan MB., Woyke T., Eric Wommack K., Stepanauskas R. 2015. Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *The ISME Journal* 9:2386–2399. DOI: 10.1038/ismej.2015.48.

Labrie SJ., Frois-Moniz K., Osburne MS., Kelly L., Roggensack SE., Sullivan MB., Gearin G., Zeng Q., Fitzgerald M., Henn MR., Chisholm SW. 2013. Genomes of marine cyanopodoviruses reveal multiple origins of diversity. *Environmental Microbiology* 15:1356–1376. DOI: 10.1111/1462-2920.12053.

Lauber C., Gorbalenya AE. 2012a. Partitioning the Genetic Diversity of a Virus Family: Approach and Evaluation through a Case Study of Picornaviruses. *Journal of virology* 86:3890–3904.

Lauber C., Gorbalenya AE. 2012b. Genetics-based classification of filoviruses calls for expanded sampling of genomic sequences. *Viruses* 4:1425–1437. DOI: 10.3390/v4091425.

Lavigne R., Darius P., Summer EJ., Seto D., Mahadevan P., Nilsson AS., Ackermann HW., Kropinski AM. 2009. Classification of Myoviridae bacteriophages using protein sequence similarity. *BMC Microbiology* 9:224. DOI: 10.1186/1471-2180-9-224.

Lawrence JG., Hatfull GF., Hendrix RW. 2002. Imbroglios of Viral Taxonomy: Genetic Exchange and Failings of Phenetic Approaches. *Journal of Bacteriology* 184:4891–4905. DOI: 10.1128/JB.184.17.4891-4905.2002.

Leplae R., Hebrant A., Wodak SJ., Toussaint A. 2004. ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic acids research* 32:D45-9. DOI: 10.1093/nar/gkh084.

Lima-Mendez G., Van Helden J., Toussaint A., Leplae R. 2008. Reticulate representation of evolutionary and functional relationships between phage genomes. *Molecular Biology and Evolution* 25:762–777. DOI: 10.1093/molbev/msn023.

Marston MF., Amrich CG. 2009. Recombination and microdiversity in coastal marine cyanophages. *Environmental Microbiology* 11:2893–2903. DOI: 10.1111/j.1462-2920.2009.02037.x.

Millard AD., Zwirglmaier K., Downey MJ., Mann NH., Scanlan DJ. 2009. Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of Synechococcus host genes localized to a hyperplastic region: Implications for mechanisms of cyanophage evolution. *Environmental Microbiology* 11:2370–2387. DOI: 10.1111/j.1462-2920.2009.01966.x.

Mizuno CM., Rodriguez-Valera F., Kimes NE., Ghai R. 2013. Expanding the Marine Virosphere Using Metagenomics. *PLoS Genetics* 9. DOI: 10.1371/journal.pgen.1003987.

Niu YD., McAllister TA., Nash JHE., Kropinski AM., Stanford K. 2014. Four Escherichia coli O157:H7 Phages: A New Bacteriophage Genus and Taxonomic Classification of T1-Like Phages. *PLoS ONE* 9:e100426. DOI: 10.1371/journal.pone.0100426.

Paez-Espino D., Eloe-Fadrosh EA., Pavlopoulos GA., Thomas AD., Huntemann M., Mikhailova N., Rubin E., Ivanova NN., Kyrpides NC. 2016. Uncovering Earth's virome. *Nature* 536:425–430. DOI: 10.1038/nature19094.

Pope WH., Bowman C a., Russell D a., Jacobs-Sera D., Asai DJ., Cresawn SG., Jacobs WR., Hendrix RW., Lawrence JG., Hatfull GF. 2015. Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *eLife* 4:e06416. DOI: 10.7554/eLife.06416.

Prangishvili D., Garrett RA., Koonin E V. 2006. Evolutionary genomics of archaeal viruses: Unique viral genomes in the third domain of life. *Virus Research* 117:52–67.

Prangishvili D., Krupovič M. 2012. A new proposed taxon for double-stranded DNA viruses, the order "Ligamenvirales." *Archives of Virology* 157:791–795.

Radoshitzky SR., Bào Y., Buchmeier MJ., Charrel RN., Clawson AN., Clegg CS., DeRisi JL., Emonet S., Gonzalez JP., Kuhn JH., Lukashevich IS., Peters CJ., Romanowski V., Salvato MS., Stenglein MD., de la Torre JC arlos. 2015. Past, present, and future of arenavirus taxonomy. *Archives of virology* 160:1851–1874. DOI: 10.1007/s00705-015-2418-y.

Redder P., Peng X., Brugger K., Shah SA., Roesch F., Greve B., She Q., Schleper C., Forterre P., Garrett RA., Prangishvili D. 2009. Four newly isolated fuselloviruses from extreme geothermal environments reveal unusual morphologies and a possible interviral recombination mechanism. *Environmental Microbiology* 11:2849–2862.

Roux S., Hawley AK., Torres Beltran M., Scofield M., Schwientek P., Stepanauskas R., Woyke T., Hallam SJ., Sullivan MB. 2014. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta- genomics. *eLife*:e03125. DOI: 10.7554/eLife.03125.

Roux S., Hallam SJ., Woyke T., Sullivan MB. 2015a. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife* 4:e08490. DOI:

10.7554/eLife.08490.

Roux S., Enault F., Ravet V., Pereira O., Sullivan MB. 2015b. Genomic characteristics and environmental distributions of the uncultivated Far-T4 phages. *Frontiers in microbiology* 6:199. DOI: 10.3389/fmicb.2015.00199.

Roux S., Brum JR., Dutilh BE., Sunagawa S., Duhaime MB., Loy A., Poulos BT., Solonenko N., Lara E., Poulain J., Pesant S., Kandels-Lewis S., Dimier C., Picheral M., Searson S., Cruaud C., Alberti A., Duarte CM., Gasol JM., Vaqué D., Bork P., Acinas SG., Wincker P., Sullivan MB. 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 537:689–693. DOI: 10.1038/nature19366.

Simmonds P. 2015. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *Journal of General Virology* 96:1193–1206. DOI: 10.1099/vir.0.000016.

Sullivan MB., Lindell D., Lee J a., Thompson LR., Bielawski JP., Chisholm SW. 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biology* 4:1344–1357. DOI: 10.1371/journal.pbio.0040234.

Woese CR., Kandler O., Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences* 87:4576–4579.

Zanotto PM de., Gibbs MJ., Gould EA., Holmes EC. 1996. A reevaluation of the higher taxonomy of viruses based on RNA polymerases. *Journal of virology* 70:6083–6096.

Table 1(on next page)

Terminology used.

Terminology	Definition
Nodes	Also known as *vertices*, these are points within a network. In this work, they are viral genomes.
Edges	Also known as *arcs*, these lines connect nodes in the network. In this work, edges have a property called *weight*, which represents the strength (as measured by significance score) between two genomes.
Betweenness centrality (BC)	Measure of how influential a node is within a network, measured by the number of shortest paths that pass through the node from all other nodes.
Connected component	A subgraph in which any two nodes are connected to each other directly (to each other) or indirectly (through other nodes).
Largest connected component (LCC)	The connected component with the greatest number of nodes.
Viral cluster (VC)	A group of viral sequences sharing a significant number of genes.
Protein cluster (PC)	A group of highly similar and related proteins, defined in this work using MCL on BLAST E-values between proteins.
Module Profile	A table-like representation of the presence/absence data between groups of protein clusters (modules) and groups of genomes (viral clusters).
Precision (P)	Also known as the *positive predictive value*, is a measure of how many true positives are identified.
Recall (R)	Also known as *sensitivity*, is a measure of how many of the total positives are identified.

Figure 1

Overview of the vContact processing pipeline.

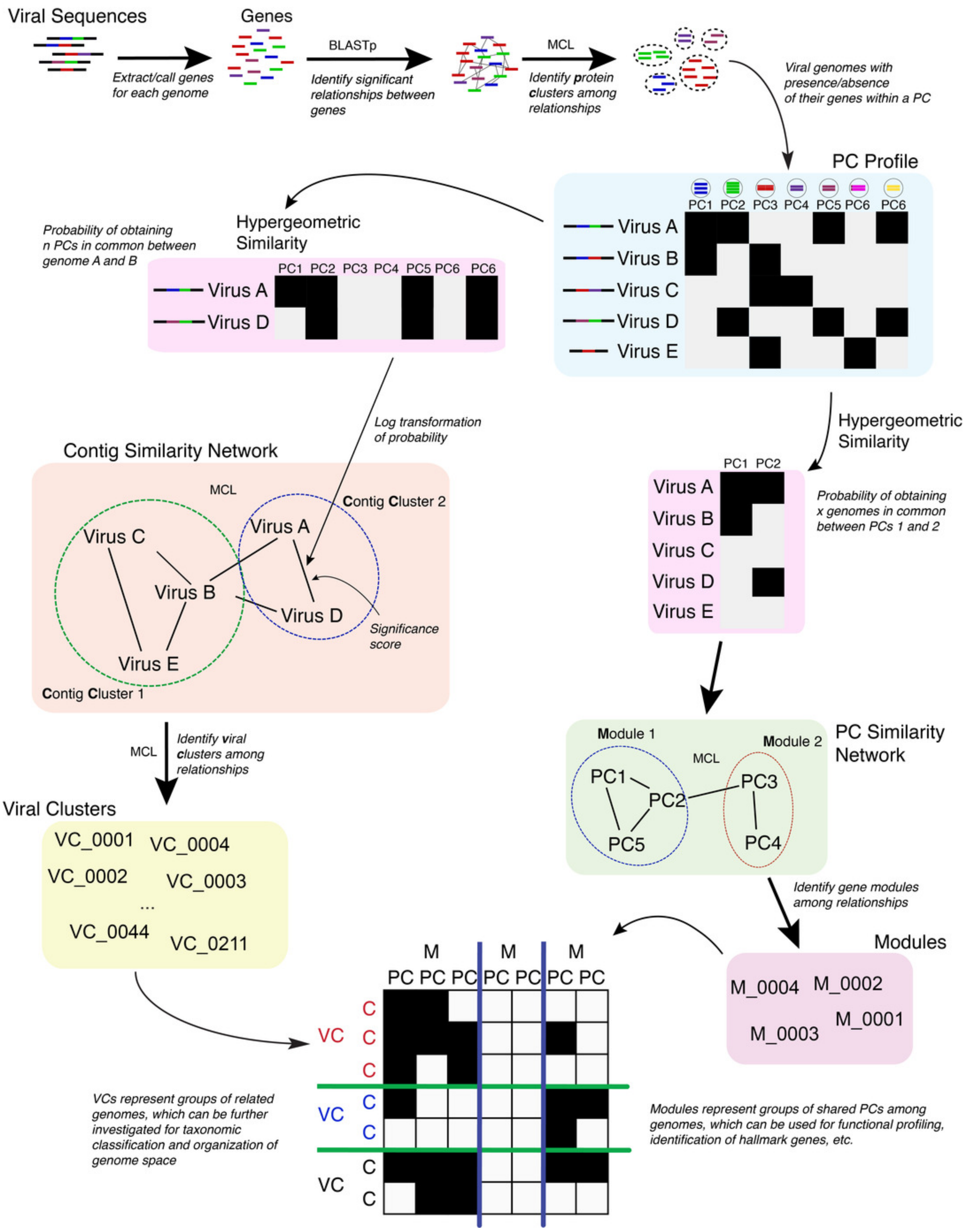

Viral Sequences
Extract/call genes for each genome
Genes
BLASTp
Identify significant relationships between genes
MCL
Identify protein clusters among relationships
Viral genomes with presence/absence of their genes within a PC
PC Profile
PC1 PC2 PC3 PC4 PC5 PC6 PC6
Virus A
Virus B
Virus C
Virus D
Virus E
Probability of obtaining n PCs in common between genome A and B
Hypergeometric Similarity
PC1 PC2 PC3 PC4 PC5 PC6 PC6
Virus A
Virus D
Log transformation of probability
Contig Similarity Network
MCL
Contig Cluster 2
Virus C
Virus A
Virus B
Virus D
Virus E
Significance score
Contig Cluster 1
Hypergeometric Similarity
PC1 PC2
Virus A
Virus B
Virus C
Virus D
Virus E
Probability of obtaining x genomes in common between PCs 1 and 2
Module 1
Module 2
MCL
PC Similarity Network
PC1
PC2
PC3
PC5
PC4
MCL
Identify viral clusters among relationships
Viral Clusters
VC_0001
VC_0004
VC_0002
VC_0003
...
VC_0044
VC_0211
Identify gene modules among relationships
Modules
M_0004
M_0002
M_0001
M_0003
M M M
PC PC PC PC PC PC PC
VC
C
C
C
VC
C
C
VC
C
C
VCs represent groups of related genomes, which can be further investigated for taxonomic classification and organization of genome space
Modules represent groups of shared PCs among genomes, which can be used for functional profiling, identification of hallmark genes, etc.

Figure 2

Protein-sharing network.

Protein-sharing network for 1,964 archaeal and bacterial virus genomes benchmarked against ICTV-accepted viral taxonomy. (A) Each node represents a viral genome from RefSeq, with its shape representing the viral family (as indicated in the legend) and each distinct color the node's viral cluster (VC). Edges between nodes indicate a statistically significant relationship between the protein profiles of their viral genomes, with edge colors (darker = more significant) corresponding to their weighted similarity scores (threshold of ≥1). VCs within the network are discriminated using the MCL algorithm (Materials and Methods) and denoted as separate colors. The position of 26 heterogeneous VCs that contain 2 or more genera is indicated. (B) Precision and recall of network-based assignments as compared to ICTV assignments for each taxonomic level (genus, family, order, and type). (C) Percentage (Y-axis) of VCs that contain the number (X-axis) of each ICTV taxonomic level (genus, family, and order).

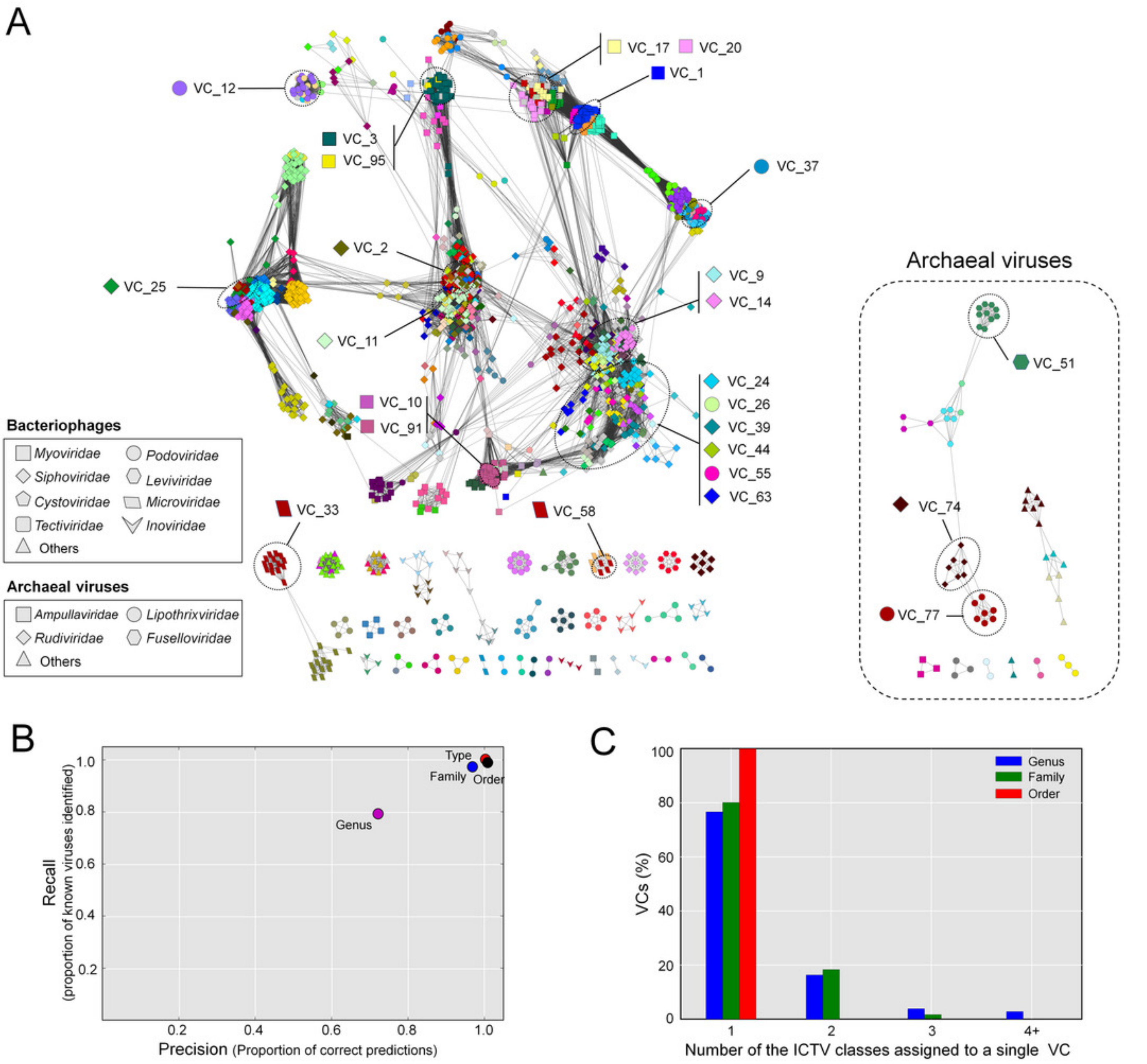

A
VC_12
VC_3
VC_95
VC_17
VC_20
VC_1
VC_37
VC_2
VC_25
VC_11
VC_9
VC_14
VC_24
VC_26
VC_39
VC_44
VC_55
VC_63
VC_10
VC_91
VC_33
VC_58
Archaeal viruses
VC_51
VC_74
VC_77
Bacteriophages
Myoviridae
Podoviridae
Siphoviridae
Leviviridae
Cystoviridae
Microviridae
Tectiviridae
Inoviridae
Others
Archaeal viruses
Ampullaviridae
Lipothrixviridae
Rudiviridae
Fuselloviridae
Others
B
Recall
(proportion of known viruses identified)
1.0
0.8
0.6
0.4
0.2
Type
Family
Order
Genus
0.2
0.4
0.6
0.8
1.0
Precision (Proportion of correct predictions)
C
VCs (%)
100
80
60
40
20
0
Genus
Family
Order
1
2
3
4+
Number of the ICTV classes assigned to a single VC

Figure 3

Detailed view of three major viral groups with relatives.

A detailed view of network regions containing three major viral groups and their relatives. Viruses (nodes) are grouped by the MCL clustering. The different shapes and colors of the nodes represent different viral families (Figure 2) and viral clusters (VCs, legends), respectively. The location of viral groups is indicated for illustrative purpose.

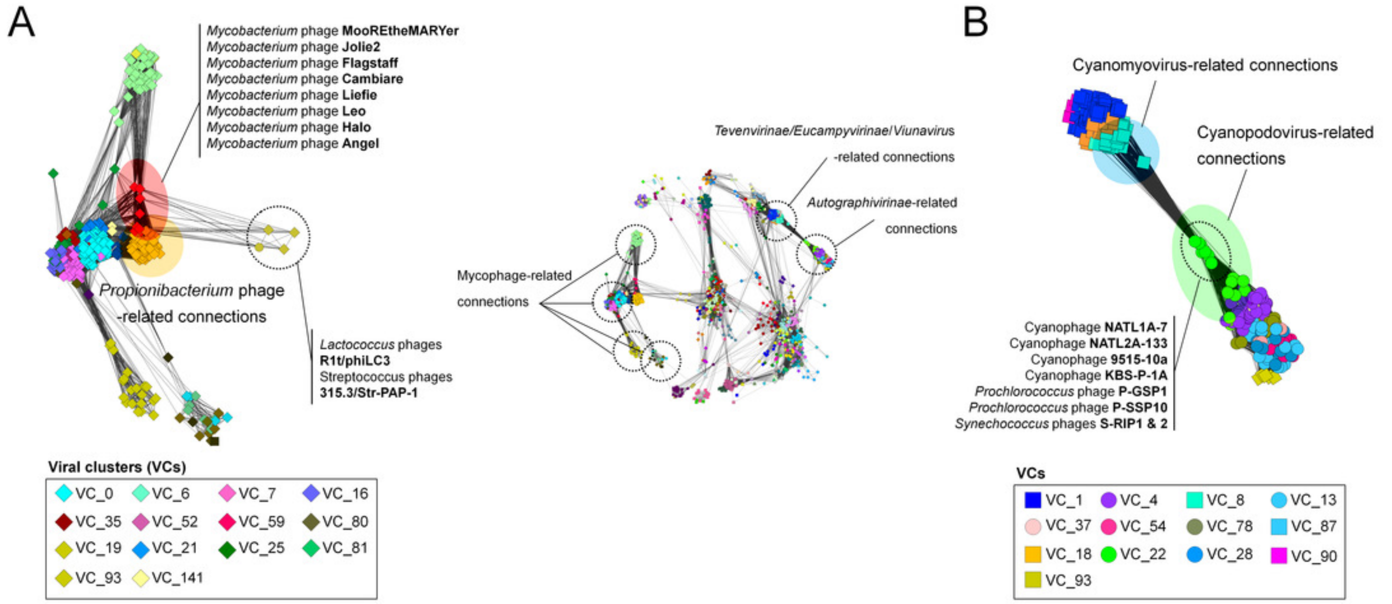
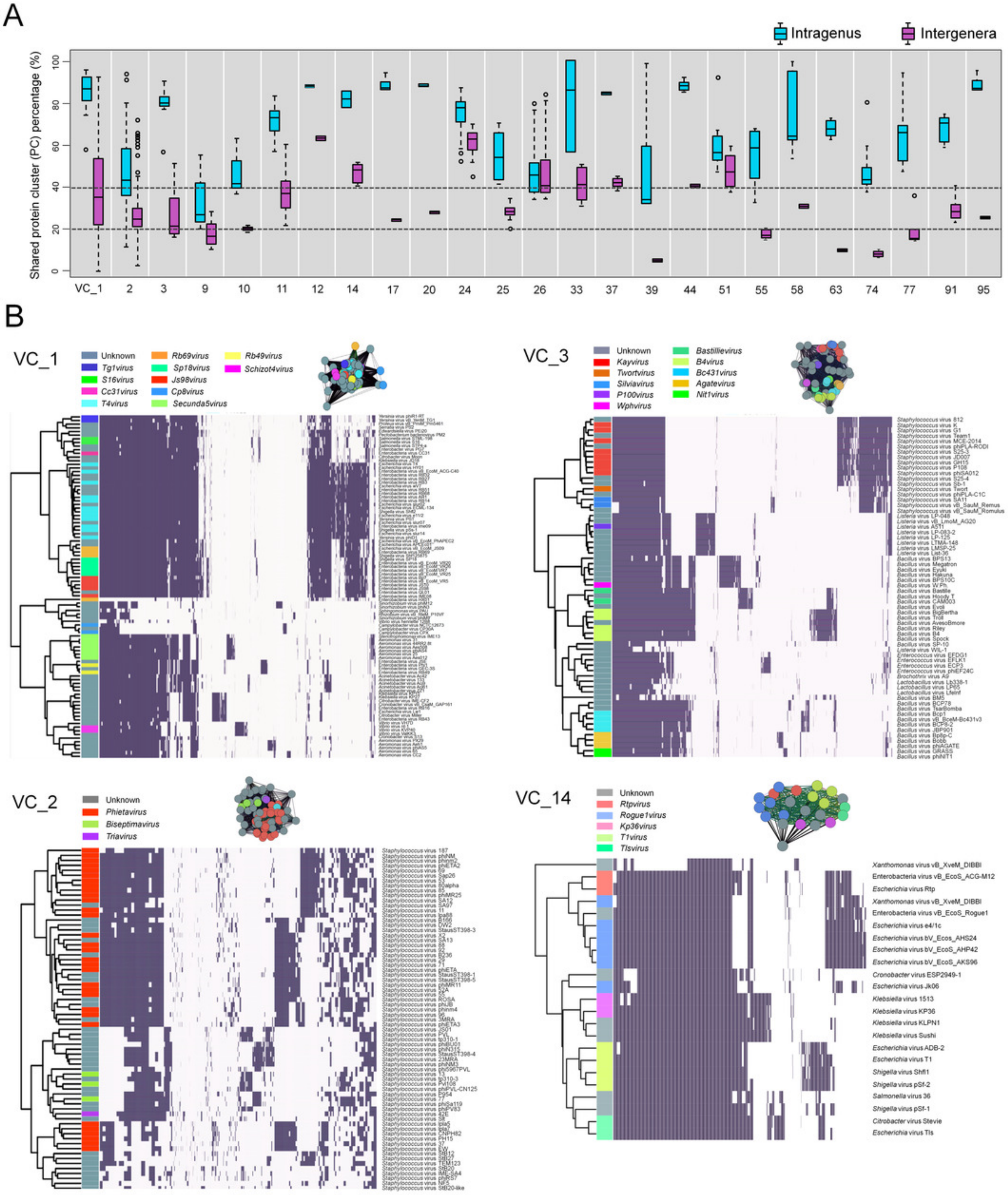

A
Mycobacterium phage **MooREtheMARYer**
Mycobacterium phage **Jolie2**
Mycobacterium phage **Flagstaff**
Mycobacterium phage **Cambiare**
Mycobacterium phage **Liefie**
Mycobacterium phage **Leo**
Mycobacterium phage **Halo**
Mycobacterium phage **Angel**
Propionibacterium phage
-related connections
Lactococcus phages
R1t/phiLC3
Streptococcus phages
315.3/Str-PAP-1
Viral clusters (VCs)
VC_0
VC_6
VC_7
VC_16
VC_35
VC_52
VC_59
VC_80
VC_19
VC_21
VC_25
VC_81
VC_93
VC_141
Tevenvirinae/Eucampyvirinae/Viunavirus
-related connections
Autographivirinae-related
connections
Mycophage-related
connections
B
Cyanomyovirus-related connections
Cyanopodovirus-related
connections
Cyanophage **NATL1A-7**
Cyanophage **NATL2A-133**
Cyanophage **9515-10a**
Cyanophage **KBS-P-1A**
Prochlorococcus phage **P-GSP1**
Prochlorococcus phage **P-SSP10**
Synechococcus phages **S-RIP1 & 2**
VCs
VC_1
VC_4
VC_8
VC_13
VC_37
VC_54
VC_78
VC_87
VC_18
VC_22
VC_28
VC_90
VC_93

Figure 4

Heterogeneous VCs

Evaluation of VCs which contained taxon representatives from more than one ICTV genus. (A) Box plots show the percent inter- and intra-genus proteome similarities in the heterogeneous VCs. Dotted lines indicate the cut-off values of 20% and 40% proteome similarities to define the subfamily and genus, respectively, which have been ratified by the ICTV Bacterial and Archaeal Viruses Subcommittee. (B) Module profiles showing the presence and absence of PCs across genomes. Presence (dark box) denotes a gene that is present within a protein cluster. Genes from related genomes often cluster into the same PC, with alignments of highly related genomes showing large groups of PCs. Genomes are further partitioned using hierarchical clustering (see materials and methods).


A
Intragenus
Intergenera
Shared protein cluster (PC) percentage (%)
0
20
40
60
80
100
VC_1
2
3
9
10
11
12
14
17
20
24
25
26
33
37
39
44
51
55
58
63
74
77
91
95
B
VC_1
Unknown
Tg1virus
S16virus
Cc31virus
T4virus
Rb69virus
Sp18virus
Js98virus
Cp8virus
Secunda5virus
Rb49virus
Schizot4virus
Yersinia virus phiR1-RT
Yersinia virus vB_YenM_TG1
Proteus virus vB_PmiM_Pm5461
Serratia virus PS2
Edwardsiella virus PEi20
Pectobacterium bacteriophage PM2
Salmonella virus STML-198
Salmonella virus S16
Salmonella virus STP4-a
Enterobacter virus PG7
Enterobacteria virus CC31
Citrobacter virus Moon
Klebsiella virus JD18
Escherichia virus T4
Escherichia virus HY01
Enterobacteria virus vB_EcoM_ACG-C40
Enterobacteria virus RB32
Enterobacteria virus RB27
Enterobacteria virus RB3
Escherichia virus wV7
Enterobacteria virus RB51
Enterobacteria virus RB68
Enterobacteria virus AR1
Enterobacteria virus RB14
Escherichia virus slur02
Escherichia virus ECML-134
Shigella virus Shfl2
Escherichia virus e11/2
Yersinia virus PST
Escherichia virus slur07
Enterobacteria virus ime09
Shigella virus pSs-1
Escherichia virus slur14
Yersinia virus phiD1
Escherichia virus vB_EcoM_PhAPEC2
Escherichia virus APCEc01
Escherichia virus vB_EcoM_JS09
Enterobacteria virus RB69
Shigella virus Shf125875
Shigella virus SP18
Enterobacteria virus vB_EcoM_VR20
Enterobacteria virus vB_EcoM_VR26
Enterobacteria virus vB_EcoM_VR7
Enterobacteria virus vB_EcoM_VR25
Enterobacteria virus Bp7
Enterobacteria virus vB_EcoM_VR5
Enterobacteria virus JS10
Enterobacteria virus JS98
Enterobacteria virus QL01
Enterobacteria virus IME08
Enterobacteria virus HX01
Sinorhizobium virus phiM12
Sinorhizobium virus phiN3
Sphingomonas virus PAU
Rhizobium virus vB_RleM_P10VF
Sinorhizobium virus phiM9
Vibrio virus henriette 12B8
Campylobacter virus NCTC12673
Campylobacter virus CP30A
Campylobacter virus CPX
Stenotrophomonas virus IME13
Aeromonas virus 31
Aeromonas virus 44RR2.8t
Aeromonas virus Aes508
Aeromonas virus phiAS4
Aeromonas virus 25
Aeromonas virus Aes012
Enterobacteria virus JSE
Enterobacteria virus Phi1
Enterobacteria virus GEC-3S
Enterobacteria virus RB49
Acinetobacter virus Ac42
Acinetobacter virus 133
Acinetobacter virus Acj9
Acinetobacter virus Acj61
Acinetobacter virus ZZ1
Klebsiella virus KP15
Klebsiella virus KP27
Citrobacter virus IME-CF2
Cronobacter virus vB_CsaM_GAP161
Enterobacteria virus RB16
Escherichia virus Lw1
Citrobacter virus Miller
Enterobacteria virus RB43
Vibrio virus VH7D
Vibrio virus nt-1
Vibrio virus KVP40
Vibrio virus ValKK3
Cronobacter virus S13
Aeromonas virus PX29
Aeromonas virus Aeh1
Aeromonas virus phiAS5
Aeromonas virus 65
Aeromonas virus CC2
VC_3
Unknown
Kayvirus
Twortvirus
Silviavirus
P100virus
Wphvirus
Bastillievirus
B4virus
Bc431virus
Agatevirus
Nit1virus
Staphylococcus virus 812
Staphylococcus virus K
Staphylococcus virus G1
Staphylococcus virus Team1
Staphylococcus virus MCE-2014
Staphylococcus virus phiPLA-RODI
Staphylococcus virus S25-3
Staphylococcus virus JD007
Staphylococcus virus GH15
Staphylococcus virus P108
Staphylococcus virus phiSA012
Staphylococcus virus S25-4
Staphylococcus virus Sb-1
Staphylococcus virus Twort
Staphylococcus virus phiPLA-C1C
Staphylococcus virus SA11
Staphylococcus virus vB_SauM_Remus
Staphylococcus virus vB_SauM_Romulus
Listeria virus LP-048
Listeria virus vB_LmoM_AG20
Listeria virus A511
Listeria virus LP-083-2
Listeria virus LP-125
Listeria virus LTMA-148
Listeria virus LMSP-25
Listeria virus List-36
Bacillus virus BPS13
Bacillus virus Megatron
Bacillus virus Eyuki
Bacillus virus Hakuna
Bacillus virus BPS10C
Bacillus virus W.Ph.
Bacillus virus Bastille
Bacillus virus Hoody T
Bacillus virus CAM003
Bacillus virus Evoli
Bacillus virus BigBertha
Bacillus virus Troll
Bacillus virus AvesoBmore
Bacillus virus Riley
Bacillus virus B4
Bacillus virus Spock
Bacillus virus SP-10
Listeria virus WIL-1
Enterococcus virus EFDG1
Enterococcus virus EFLK1
Enterococcus virus ECP3
Enterococcus virus phiEF24C
Brochothrix virus A9
Lactobacillus virus Lb338-1
Lactobacillus virus LP65
Lactobacillus virus Lfeinf
Bacillus virus BM5
Bacillus virus BCP78
Bacillus virus TsarBomba
Bacillus virus Bcp1
Bacillus virus vB_BceM-Bc431v3
Bacillus virus BCP8-2
Bacillus virus JBP901
Bacillus virus Bp8p-C
Bacillus virus Bobb
Bacillus virus phiAGATE
Bacillus virus GRASS
Bacillus virus phiNIT1
VC_2
Unknown
Phietavirus
Biseptimavirus
Triavirus
Staphylococcus virus 187
Staphylococcus virus phiNM
Staphylococcus virus phinm2
Staphylococcus virus phiETA2
Staphylococcus virus 69
Staphylococcus virus Sap26
Staphylococcus virus 53
Staphylococcus virus 80alpha
Staphylococcus virus 85
Staphylococcus virus phiMR25
Staphylococcus virus SA12
Staphylococcus virus SA97
Staphylococcus virus 11
Staphylococcus virus Ipla88
Staphylococcus virus B166
Staphylococcus virus DW2
Staphylococcus virus StauST398-3
Staphylococcus virus X2
Staphylococcus virus SA13
Staphylococcus virus 88
Staphylococcus virus 92
Staphylococcus virus B236
Staphylococcus virus 47
Staphylococcus virus 71
Staphylococcus virus phiETA
Staphylococcus virus StauST398-1
Staphylococcus virus StauST398-5
Staphylococcus virus phiMR11
Staphylococcus virus 52A
Staphylococcus virus 55
Staphylococcus virus ROSA
Staphylococcus virus phiJB
Staphylococcus virus phinm4
Staphylococcus virus 96
Staphylococcus virus 3MRA
Staphylococcus virus phiETA3
Staphylococcus virus JS01
Staphylococcus virus PVL
Staphylococcus virus tp310-1
Staphylococcus virus phiBU01
Staphylococcus virus phiN315
Staphylococcus virus StauST398-4
Staphylococcus virus 23MRA
Staphylococcus virus phiNM3
Staphylococcus virus phi5967PVL
Staphylococcus virus 13
Staphylococcus virus tp310-3
Staphylococcus virus Pvl108
Staphylococcus virus phiPVL-CN125
Staphylococcus virus P954
Staphylococcus virus 77
Staphylococcus virus phiSa119
Staphylococcus virus phiPV83
Staphylococcus virus 42E
Staphylococcus virus SH
Staphylococcus virus Ipla5
Staphylococcus virus Ipla7
Staphylococcus virus CNPH82
Staphylococcus virus PH15
Staphylococcus virus 37
Staphylococcus virus EW
Staphylococcus virus StB12
Staphylococcus virus StB27
Staphylococcus virus TEM123
Staphylococcus virus StB20
Staphylococcus virus IME-SA4
Staphylococcus virus phiRS7
Staphylococcus virus NF5
Staphylococcus virus StB20-like
VC_14
Unknown
Rtpvirus
Rogue1virus
Kp36virus
T1virus
Tlsvirus
Xanthomonas virus vB_XveM_DIBBI
Enterobacteria virus vB_EcoS_ACG-M12
Escherichia virus Rtp
Xanthomonas virus vB_XveM_DIBBI
Enterobacteria virus vB_EcoS_Rogue1
Escherichia virus e4/1c
Escherichia virus bV_Ecos_AHS24
Escherichia virus bV_EcoS_AHP42
Escherichia virus bV_EcoS_AKS96
Cronobacter virus ESP2949-1
Escherichia virus Jk06
Klebsiella virus 1513
Klebsiella virus KP36
Klebsiella virus KLPN1
Klebsiella virus Sushi
Escherichia virus ADB-2
Escherichia virus T1
Shigella virus Shfl1
Shigella virus pSf-2
Salmonella virus 36
Shigella virus pSf-1
Citrobacter virus Stevie
Escherichia virus Tls