Gene sharing networks to automate genome-based prokaryotic viral taxonomy

Ho Bin Jang[1*], Benjamin Bolduc[1*], Olivier Zablocki[1], Jens H. Kuhn[2], Simon Roux[3], Evelien M. Adriaenssens[4], J. Rodney Brister[5], Andrew M Kropinski[6, 7], Mart Krupovic[8], Rob Lavigne[9], Dann Turner[10], & Matthew B. Sullivan[1,11#]

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

[2] Integrated Research Facility at Fort Detrick, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Fort Detrick, Frederick, MD, USA

[3] U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA, USA

[4] Institute of Integrative Biology, University of Liverpool, Liverpool, UK

[5] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

[6] Department of Pathobiology, Ontario Veterinary College, University of Guelph, Guelph, ON, Canada N1G 2W1

[7] Department of Food Science, University of Guelph, Guelph, ON, Canada, N1G 2W1

[8] Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, Department of Microbiology, Paris 75015, France

[9] Laboratory of Gene Technology, Department of Biosystems, Faculty of BioScience Engineering, KU Leuven, Leuven, Belgium

[10] Centre for Research in Biosciences, Department of Applied Sciences, Faculty of Health and Applied Sciences, University of the West of England, Coldharbour Lane, Bristol, UK

[11] Department of Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH, USA

[#] correspondence to: Matthew Sullivan, sullivan.948@osu.edu

[*] These authors contributed equally to this work.

ABSTRACT

Viruses of bacteria and archaea impact natural, engineered and human ecosystems, but their study is hampered by the lack of a universal or scalable taxonomic framework. Here we introduce vConTACT v2.0, a network-based application to establish prokaryotic virus taxonomy that scales to thousands of uncultivated virus genomes, and integrates distance-based hierarchical clustering and confidence scores for all taxonomic predictions. Performance tests demonstrated significant improvements over the original tool and near-identical (96%) correspondence to current International Committee on Taxonomy of

Viruses (ICTV) viral taxonomy where genus-level assignments are available. Beyond these "known viruses", vConTACT v2.0 suggested automatic genus assignments for 1,364 previously unclassified reference viruses, with perfectly scoring assignments submitted as new taxonomic proposals to ICTV. Scaling experiments with 15,280 global ocean large viral genome fragments demonstrated that the reference network was rapidly scalable and robust to adding large-scale viral metagenomic datasets. Together these efforts provide a critically-needed, systematically classified reference network and an accurate, scalable, and automatable taxonomic analysis tool.

Main text

Bacteria and archaea modulate the nutrient and energy cycles that drive ocean and soil ecosystems[1–4], and impact humans by producing metabolites that affect health, behavior, and susceptibility to disease[5]. Viruses that infect these microbes modulate these 'ecosystem roles' via killing, metabolic reprogramming and gene transfer[6,7], with substantial impacts predicted in the ocean[8–10], soil[11,12] and human microbiomes[13,14]. However, ecosystem-scale understanding is hampered by the lack of universal genes or methods that could facilitate a formalized taxonomy and comparative surveys. In fact, viruses do not have a single, universal marker gene[15], and, thus, no analog to the microbial 16S rRNA-based phylogenies and operational taxonomic units (OTUs) are possible[16].

Another potential challenge is that many viruses exhibit 'rampant genome mosaicism'[17], due to high rates of horizontal gene transfer between viruses and their hosts. If broadly true, this would hamper the creation of a genome-based prokaryotic virus taxonomy[18]. Fortunately, explorations of viral sequence space are revealing structure[19,20] and population genetic support for a biological species definition, driven primarily by vertical evolution[21], and new hypotheses to explain variable evolution among prokaryotic viruses[22]. Such findings, alongside rapidly expanding viral genome databases, allowed the International Committee on Taxonomy of Viruses (ICTV) to present a consensus statement suggesting a shift from the 'traditional' classification criteria[23] (e.g. virion morphology, single/multiple gene phylogenies) towards a genome-centered, and perhaps one-day, largely automated, viral taxonomy[24]. This shift is particularly

critical given the modern pace of viral discovery in which, as of November 2018, hundreds of thousands of metagenome-derived viral genomes and large genome fragments (735,112 at IMG/VR[25], with a total of 110,384 'species' or viral OTUs) now dwarf the 34,091 available from prokaryotic virus sequences in the NCBI GenBank database[26]. Together with the recently proposed 'minimum information about uncultivated virus genomes' (MIUViGs) community guidelines[27], evaluation of approaches to establish a scalable, genome-based viral taxonomy is needed to enable a universal classification framework. The ability to classify thousands of microbial UViGs is invaluable for foundational hypothesis testing anywhere microbes, and their viruses, might matter. With microbiome and phage therapy research exploding, this includes ecosystems such as industrial bioreactors and the human gut and lung, as well as the oceans, the deep-sea floor, and soils.

Multiple genome-based strategies have been proposed to develop such a unified taxonomic framework for viruses infecting bacterial[15,28–33], archaeal[34] and eukaryotic[35] hosts. For bacterial viruses ("phages"), an early approach to a universal taxonomy targeted phage relationships only by using complete genome pairwise protein sequence comparisons in a phylogenetic framework (the "phage proteomic tree") and was broadly concordant with ICTV-endorsed virus groupings of the time[15]. Such efforts were not widely adopted, presumably because (i) the demand was limited (few metagenomics studies existed), and (ii) the paradigm was that "rampant mosaicism" would blur taxonomic boundaries and violate the assumptions of the underlying phylogenetic algorithms used in the analyses[17]. Other efforts sought to establish percent of genes shared and percent identity of shared genes cut-offs to define genera and sub-family affiliations[36,37], but lacked taxonomic resolution for several virus groups. This lack of resolution was due to the likelihood that the mode and tempo of prokaryotic virus evolution could vary significantly across the viral sequence space[22]. Building upon a prokaryotic classification algorithm, the Genome Blast Distance Phylogeny (GBDP)[38], a freely accessible online tool (VICTOR) now provides for classification of phage genomes via combined phylogenetic and clustering methods from nucleotide and

protein sequences[31]. Although a key advance, this method suffers from limited scalability (100 genomes limit) and taxonomic assignment challenges for viruses that lack closely related reference genomes.

Alternatively, several groups reasoned that the highly variable evolutionary rates across phage sequence space could be examined through gene sharing networks[29,30,39] to determine whether a meaningful structure, and therefore taxonomic signal, occurs in this space. These networks, based on shared protein clusters (PCs) between viral genomes, were largely concordant with ICTV-endorsed taxa independent of whether monopartite[29] (a single node type, i.e., viral genomes) or bipartite networks[34,39] (two node types, i.e., viral genomes and genes) were used. Given these successes, we revisited the monopartite gene sharing network approach to establish an iVirus[40] app (vConTACT v1.0, hereafter v1.0) to automate a network-based classification pipeline for prokaryotic virus genomes. Performance tests indicated that the network analytics used by v1.0 produced viral clusters ('VCs') that were ~75% concordant with recently established ICTV prokaryotic viral genera, even with seven times more genomes now available[30]. The capacity to incorporate these genomes and accuracy of the network-based analytics have resulted in viral taxonomy applications across large-scale studies of ocean[41,42], freshwater[43] and soil[44], and studies of single-virus amplified genomes (vSAGs)[45,46] – all environments where the viruses observed were largely so novel as to not be classifiable outside of these gene-sharing network approaches. These v1.0 advances were an important step forward, but they could not be used for automated tentative taxonomic assignments. This is because v1.0 created artifactual VCs of both undersampled genomes and highly overlapped regions of viral sequence space[30], and also lacked several key, community-desired features including per-VC confidence metrics, a metric for establishing hierarchical taxonomy, and scalability.

Here we introduce and evaluate vConTACT v2.0 (hereafter v2.0), which significantly updates the network analytics and feature set of the original program. We apply this program to (i) establish a centralized, 'living' taxonomic reference network as a foundational community resource and (ii) demonstrate that v2.0 is robust and scalable to modern datasets.

RESULTS AND DISCUSSION

vConTACT v2.0 key features and updates

The underlying goal of vConTACT is to automatically assign viral genomes into relevant established or tentatively new taxa, with performance assessed relative to ICTV-assigned, manually-curated taxa (see **Fig. 1** for conceptual overview). However, in the current ICTV taxonomy for prokaryotic viruses, taxonomic classifications above the genus level are only sporadically available. For example, of the 2,304 prokaryotic virus genomes available at RefSeq, 84.2% are unclassified at the subfamily level, and 61.6% are unclassified at the order level with virtually all of the remaining 38% lumped into a single "*Caudovirales*" order. Moreover, among the *Caudovirales*, the three phenotypically recognized and dominant bacterial virus family level designations – *Podoviridae*, *Myoviridae* and *Siphoviridae* – are being called into question by genome-based taxonomy methods[47–49] and are thus in flux. Therefore, we focused specifically on assigning viruses at the genus level, as it is the most 'stable' level and constitutes the principal taxon on which molecular classification is based on in the ICTV taxonomy.

In a network-based genome taxonomy framework (**Fig. 1a**), related genomes emerge as a group of nodes strongly connected through multiple edges, here termed a Viral Cluster, or 'VC'. In a taxonomic context and based on the clustering of viral reference genomes, we have previously demonstrated that the network parameters can be tuned such that the VCs best represent genus-level grouping of viral genomes[30]. In the original v1.0, ~75% of VCs corresponded to established ICTV genera[30] ('concordant VCs'), while the remaining mismatched VCs were termed 'discordant VCs'. These mismatches in the discordant VCs can occur from any of three scenarios (**Fig. 1b**): (i) VCs that wrongly clustered genomes with no close relatives (termed 'outlier genomes' from 'undersampled VCs'), (ii) a given VC that contained multiple ICTV genera represented by virus genomes that shared many genes and/or modules with other VCs (termed 'overlapping VCs'), and (iii) VCs that encompassed multiple ICTV genera

sharing many genes and/or gene modules across genomes within the VC, and within subsets of the genomes in the VC (termed 'structured VCs'). While a significant advance, v1.0 struggled to adequately differentiate between these issues. Further, v1.0 lacked several key features to enable broader adoption and utility as described above.

To address these issues in v2.0, we evaluated and ultimately implemented a new clustering algorithm, established confidence scores and distance-based taxon separation for hierarchical taxonomy, and optimized and evaluated scalability and robustness to a large-scale viral metagenomic dataset. Briefly, after the MCL-clustered protein clusters are generated, we optimized the protein-cluster-based gene-sharing information to establish an automated two-step process whereby VCs are defined using ClusterONE[50] (CL1), instead of MCL used in v1.0, and then subdivided using hierarchical clustering to disentangle problematic regions of the networks (**Fig. 1b**, see Online Methods). This approach considers edge weight (i.e., degree of connection between genomes) to identify outlier genomes that are weakly connected with members of their VC compared to neighbour genomes, detect and separate genomes that 'bridge' overlapping VCs, and break down structured VCs into concordant VCs through distance-based hierarchical clustering **(Fig. 1b)**.

Additionally, v2.0 incorporates confidence scores for each VC to help differentiate between meaningful taxonomic assignments and those that might be artifacts. Briefly, each VC receives two types of confidence scores: a topology-based score (value range 0-1), which aggregates information about network topological properties, and a taxonomy-based score (value range 0-1), which estimates the likelihood of predicted VCs to be equivalent to a single ICTV genus (Online Methods). Higher values indicate either more confident linkages within the VC or better taxonomic agreement for the topology and taxonomy-based scores, respectively, and the taxonomy-based score is used to automatically optimize the hierarchical clustering of structured VCs into ICTV-concordant 'subclusters'.

Finally, although v2.0 is presented here as a monopartite (one type of node) network tool, it produces the necessary output to also be visualized as a bipartite network (**Supplementary Fig. 1**). In these

158 bipartite visualizations, two types of nodes are used to display genomes and their connecting, shared
159 protein clusters (PCs). This information about which PCs link a given set a viruses together are also
160 provided (**Supplementary Table 1**; see Online Methods), as it enables researchers to identify specific
161 core virus group genes that may be of value for establishing novel gene markers and other additional
162 downstream analyses.
163
164 *General performance comparison of vConTACT versions 1.0 and 2.0*
165 To assess clustering performance of v1.0 and v2.0, we quantified ICTV correspondence for the 940
166 archaeal and bacterial virus genomes that had ICTV genus-level classification (accessed January 2018,
167 see Online Methods and **Supplementary Table 2**). Clustering performance was evaluated through a
168 composite performance score of Accuracy (*Acc*) and Separation (*Sep*). Both *Acc* and *Sep* are aggregate
169 measures themselves (see Online Methods), which indicate clustering precision, and how resulting
170 clusters (or VCs) correspond to a single ICTV genus, respectively (**Fig. 2a**). Each of these metrics has a
171 value range from 0 to 1 with 1 indicating perfect clustering accuracy and/or coverage.
172 Compared to the original v1.0's performance (which used MCL at an inflation factor, or IF, of 1.4,
173 see Online Methods), v2.0's CL1, combined with hierarchical clustering, resulted in an overall
174 performance improvement of 28.8% (**Fig. 2a**). This increase indicated the overall improved ability of the
175 tool to correctly group viruses into their appropriate VC, and how each VC corresponded to its ICTV
176 genus counterpart. Though further feature enhancements also advanced v2.0, we wondered which aspect
177 of our changes most improved performance. To assess this, we further optimized v1.0's MCL-based VC
178 clustering and found that, at an IF of 7, we could achieve nearly equivalent performance (**Fig. 2a,**
179 **Supplementary Table 3**) and more VCs predicted by the optimized MCL-based configuration as it
180 organized the 940 viral genomes into 180 VCs, whereas v2.0's CL1 identified 157 VCs. However, higher
181 values in *Sep* for CL1 indicate better performance for assigning single genera into single VCs, even
182 though MCL at its optimal IF value (i.e., 7) generated more VCs (**Supplementary Table 1**). Thus,

although more VCs were assigned to ICTV genera by the optimized MCL configuration, they were largely discordant VCs of either lumped or split ICTV genera, or both; whereas this behavior was ~50% reduced using CL1 (see **Supplementary Fig. 2a and b**). Among these 22 lumped or split VCs from the optimized MCL configuration, the virus genomes shared very few proteins (average = 17% range: 1-30%; **Supplementary Fig. 1b**) similarities, which modern cut-offs would suggest should have been separated as separate genera, here outliers in the network. To better resolve these issues, we added a post-processing, Euclidean distance-based hierarchical clustering step to split mismatched VCs in v2.0. This step accurately and automatically classified 36 additional genera from the problematic structured VCs (**Supplementary Table 2**), which increased v2.0's *Sep* value by 7%. Together, these findings suggested that both upgrading the clustering algorithm and adding hierarchical clustering were critical to improve automatic VC assignments.

Performance assessment of vConTACT v2.0 for complex genomic relationships

Next, we assessed how v2.0 specifically handled areas of the reference network that represent the three broad scenarios that yield discordant VCs (**Fig. 1b**). First, 55% of ICTV genera are undersampled (**Supplementary Table 2**), which in a gene-sharing network manifests as weakly connected, small VCs prone to artifactual clustering (**Fig. 1b**, top row) due to outlier genomes only weakly connected to any given VC. In v1.0, undersampled VCs accounted for 64% (28/44) of all discordant VCs, and could not be resolved by increasing IF values (**Fig. 2b and d** and **Supplementary Table 2**). In contrast, v2.0 automatically and accurately handled these same 28 undersampled VCs (comprising 60 genomes) by splitting the 37 problematic genera into 22 outliers (i.e., genera with only one member) and correctly placing the remaining 38 genomes from 15 genera into 15 now concordant VCs (**Fig. 2c and d** and **Supplementary Table 2**). Thus, v1.0 performed poorly on under-sampled VCs, whereas v2.0 was able to automatically resolve all under-sampled VCs in this reference network into their appropriate ICTV genera.

Second, we evaluated the ability of v2.0 to handle overlapping VCs (**Fig. 1b**), which are those VCs that share a greater fraction of genes with members of other VCs than typically expected, presumably due to gene exchange that could erode structure of the network. While overlapping VCs cannot be automatically identified in v1.0, we can detect them in v2.0 through a 'match coefficient' that measures the connection within- and between- other VCs (see Online Methods). Sensitivity analyses established a cluster overlap value of 0.8 as diagnostic (Online Methods). This approach identified nine overlapping VCs (ICTV-classified genera only) containing 30 viruses across 11 ICTV genera. These included viruses with known mosaic genomes[48] (e.g., lambdoid or mu-like phages of the *P22virus*, *Lambdavirus*, *N15virus*, and *Bcepmuvirus* genera), recombinogenic temperate phages[51,52] (i.e., *Mycobacterium* phages of the *Bignuzvirus, Phayoncevirus,* and *Fishburnevirus* genera and *Gordonia* phages of the genus *Wizardvirus*), and three newly-established genera (i.e., *Cd119virus*, *P100virus* and archaeal *Alphapleolipovirus*), all bearing low topology-based confidence scores (averages of 0.32 for these VCs versus 0.52 for concordant VCs; P-value = 6.12e-09, Mann-Whitney U test) (**Supplementary Fig. 3a**). Overlapping VCs are also linked to high horizontal gene flow, since most viruses in these VCs were classified as having high gene content variation (HGCF, **Fig. 2e, Supplementary Fig. 3b**) as assigned by a recently proposed framework of phage evolutionary lifestyles[22]. Though unresolvable in v1.0, v2.0 could assign eight of the 11 ICTV genera (24 viruses) into 8 ICTV-concordant VCs (**Supplementary Table 2**). The remaining 3 ICTV genera, all comprised of *Mycobacterium* phages[53] (6 genomes), could not be resolved (**Supplementary Table 2**), and may not be amenable to automated taxonomy. While these highly recombinogenic genomes represent a minor portion of the viruses classified by ICTV, their prevalence across environments remains to be determined.

Third, structured VCs (**Fig. 1b**, bottom row) contained genomes that our gene sharing networks placed into a single VC due to many shared genes and/or gene modules across all the member genomes, but distributed into several ICTV genera due to subsets of the genomes also sharing additional genes (see **Supplementary Note 1**). While we showed in v1.0 that these structured VCs could be decomposed

through hierarchical clustering[27], in v2.0, we formalized an optimized, quantitative hierarchical decomposition distance measure for this process (Online Methods and **Supplementary Fig. 4**). In the v2.0 network, 23 of the 31 discordant VCs (74%) were structured VCs, spanning 86 genera (**Fig. 3a,b** and **Supplementary Table 2**). The automatic v2.0 approach resolved 30% (26 of 86) of these ICTV genera from 6 of the 23 structured VCs (**Fig. 3c**).

Given such strong performance, even with challenging regions of viral sequence space, we suggest that this gene sharing network already offers significant new taxonomic insights. As described earlier, only 41% of the 2,304 reference virus genomes are classified by ICTV at the genus rank, which leaves 1,364 reference viruses that are currently not assigned to a genus. In our networks, these 1,364 reference viruses organized into 404 well-supported VCs (**Supplementary Table 2**). Of these 1364, 544 were assorted into 104 VCs with genomes from known ICTV taxa, whereas 820 formed 200 separate VCs. We propose that the 820 can be classified and the 200 VCs represent *bona fide* novel virus genera, to be evaluated as formal ICTV taxa. If adopted, this would immediately double the number of known prokaryotic viral genera, as there are only 264 currently recognized. Beyond providing a starting point for identifying areas of viral sequence space needing revision, the manual curation process itself would provide feedback for improving v2.0.

As first evidence of the value of such an iterative process, we note that v2.0 clustering suggested an updated taxonomy among ten currently established ICTV genera: *Barnyardvirus*, *Bcep78virus*, *Bpp1virus*, *Che8virus*, *Jerseyvirus*, *P68virus*, *Pbunavirus*, *Phietavirus*, *Phikmvvirus*, and *Yuavirus* (**Supplementary Fig. 5** and **Supplementary Note 2**), and manual inspection had already recommended some of these ICTV genera be revised (e.g. *Phikmvvirus* viruses, ICTV proposal 2015.007a-Db). Hierarchical decomposition of structured VCs into subclusters indicated that the gene content-based distance correctly recapitulated the ICTV taxonomy, but that the cut-offs used to define subclusters are different from the ones currently used to delineate established genera (**Fig. 3c** and **Supplementary Fig. 4**). It is long thought that universal cut-offs may not be appropriate across all of viral sequence space with the

result that years of manual curation by experts has resulted in specialized demarcation cut-offs across viral sequence space[54]. However, just has recently been advanced for microbes[55] any steps towards this, such as distances in well-sampled and well-resolved portions of the reference networks, will be invaluable for automating virus taxonomy. The v2.0 VCs and subclusters provided here provide a case study and a reference baseline for working with the ICTV to translate such network-derived cut-offs into systematic taxonomic demarcation criteria.

Finally, there will be some taxon assignments that are not amenable to being resolved by gene-sharing networks. For example, cases where genera are defined based on phenotypic or evolutionary evidence, e.g., archaeal fuselloviruses[56] (VC42) and bacterial microviruses[57] (VCs 30 and 49), a gene-sharing network approach will not be appropriate (see **Fig. 3c** and **Supplementary Table 2**). However, an automated vConTACT-based approach can help systematically identify such problematic taxa and drastically speed up these critical revisions to our taxonomy as new data become available.

vConTACT v2.0 is scalable to modern virome datasets

Beyond accurate classification, a major bottleneck regarding automated taxonomic assignments is the ability to scalably and robustly integrate large sets of newly discovered virus genomes. To evaluate this, we added 15,280 curated viral genomes and large genome fragments (≥ 10 kb) from the Global Ocean Virome (GOV) dataset[41] to our reference network in successive 10% increments (i.e., 0%, 10%, …, 100% of the total dataset), totalling 16,960 sequences (**Fig. 4a**). Through this process, we evaluated 2 types of network changes to assess robustness. First, we checked whether the incremental addition of GOV data to the network would lead to changes in node connections, as estimated by the 'change centrality' metrics (CC, values range from 0-1 with 0 indicating no change and 1 indicating complete change, see **Fig. 4b**). Second, we evaluated the concordance between v2.0 clustering and ICTV genera using the same performance metrics as above (*Sn*, *Acc* & *PPV*, see **Fig. 4c**). As seen in Fig. 4b, a large fraction of the data initially experiences a moderate change (CC = 0.4), but the whole dataset eventually stabilized, as

CC values for most of the data ranged from 0 to 0.1. A similar trend was observed for accuracy (*Acc*, **Fig. 4c**). This indicated that, not only can v2.0 scalably handle thousands of new input sequences, but our original reference network clustering is robust to large-scale data additions such that the underlying reference network remains consistent with established ICTV genera.

Outside of better sampled genera or VCs, we also examined whether GOV data may partially resolve ICTV outlier and singleton genomes as a proxy for assessing the taxonomic ramifications of so much new data. We reasoned that more data might create new connections to singletons such that outliers may become better connected to new or existing VCs. We found that, of 38 single-member VCs of singleton and outlier genomes (**Supplementary Fig. 6**), three *Mycobacterium* phage VCs were improved, while two other *Mycobacterium* virus genomes were merged into larger heterogeneous VCs now composed of six ICTV genera, which did not constitute an improvement. This implies that even though the GOV data derive from an environment very different to that of most of the reference taxa (oceans vs soils and humans), these environmental data could help improve some taxon assignments of isolate genomes. Separately, looking at the overall taxonomic impact, we observed that 919 new VCs were created with the full GOV dataset (15, 280 total contigs). Given the strong concordance of the 'known' VCs to ICTV genera, we posit that these new VCs represent 919 new viral genera that are not represented among the 264 ICTV genera already known from RefSeq genomes. If true, this demonstrates the power of integrating viral sequence datasets from under-explored environments to better map viral sequence space. While enticing, we agree with the recent consensus statement that any taxonomic reference network be constrained to complete genomes[24], and that large genome fragments commonly derived from metagenome-based studies be utilized in a relevant manner to address questions specific to that study.

Community availability and future needs

The utility of v2.0 depends upon its expert evaluation and community availability. In close discussion with members of the ICTV Bacterial and Archaeal Viruses Subcommittee, we made the resulting

optimized tool available in two ways. First, the source code is available through Bitbucket (https://bitbucket.org/MAVERICLab/vcontact2) as a downloadable python package. V2.0 is a highly scalable tool, with memory and CPU requirements concomitant with the amount of sequences processed. The largest memory requirements are by the all-versus-all protein comparisons used to build the protein clusters. Overall, there is a strong linear (R^2 = 0.99, see **Supplementary Fig. 7**) correlation between number of sequences and runtimes. For example, running the full virus dataset RefSeq with Diamond (for the protein comparisons) would take ~10 minutes on a regular laptop, while a GOV-sized dataset would run for several hours. Second, for ease of use, v2.0 is also available as an app through iVirus[40], the viral ecology apps and data resource embedded in the CyVerse Cyberinfrastructure, with detailed usage protocols available through Protocol Exchange (https://www.nature.com/protocolexchange/) and protocols.io (https://www.protocols.io/). Finally, the curated reference network is available at each of these sites, and will be updated approximately bi-yearly with as complete genomes become available and resources exist to support this effort.

Although v2.0 performance metrics are strong and provide a critically needed, systematic reference viral taxonomic network, limitations remain. First, the complete reference network needs to be rebuilt each time new data are added. Avoiding this reconstruction step will require the development of approximation methods and/or a placement algorithm (akin to PPlacer for 16S phylogenies[58]) to incorporate new data. Second, CL1-based VC generation may require manual parameter optimization if datasets of an extreme, and probably unlikely, nature are encountered. Such datasets would have to be dominated by overlapping genomes (i.e., those that share a large proportion of genes), and though such genomes exist (e.g., the HGCF *Mycobacterium* phages), they are currently rare among cultured isolates and their frequency in nature remains unknown. Notably, at least for the oceans, such highly recombinogenic viruses are likely uncommon as the incremental addition of GOV data resulted in a stable network and separate analyses of a more recent version of this GOV dataset found ocean viral populations to be quite structured[59]. However, in case such datasets dominated by overlapping genomes are

encountered, we have added an auto-optimization option for determining the optimal distance for hierarchical decomposition of structured VCs in v2.0.

Third, solving other future needs will require improved understanding of the broader viral sequence landscape and evolutionary processes. For example, although v2.0 handles reference prokaryotic virus genomes (including ssDNA or dsDNA phages) and large GOV genome fragments, this framework has not been designed, tested or validated for eukaryotic viruses. These viruses will require new solutions as they have broader genomic configurations, such as genome segmentation, overlapping genes, ambisense transcriptional gene configurations, that pose unique computational challenges[35,60]. However, even these viruses can be classified using genome-based metrics (e.g., hidden Markov models protein profiles and genomic organization models) and network analytics, at least at the family-level[35]. Separately, shorter complete prokaryotic virus genomes and small fragments of larger genomes (e.g., $\leq$ 3 PCs or $\leq$ 5 genes) are of low statistical power in gene-sharing networks, and will require new solutions to establish higher confidence VCs or remain taxonomically inaccessible via these approaches. Finally, genomes identified as singletons, outliers or overlapping are currently excluded from the gene-sharing network, which leaves a large fraction of viral sequence space unclassified. Although singletons and outliers can be resolved by the addition of new data, overlapping VCs can remain challenging to resolve, particularly for the HGCF phages[22] that are highly recombinogenic. Such mosaic virus genomes are challenging for viral taxonomy. However, they are identifiable in the networks and, at least to date, represent a small fraction of known viral sequence space. Given increased gene flow among temperate phages[22], it will be valuable to explore viral sequence space in environments where temperate phages are thought to predominate (e.g., soils[61], human gut[62]).

In spite of these limitations, vConTACT v2.0 already provides upgrades in performance and its feature set (per-VC confidence scores, user-desired outputs and usability) such that it offers a scalable, robust, systematic and automated means to classify large swaths of bacterial and archaeal virus sequence space. Its limitations are largely surmountable through future research, and coordinated efforts with the

research community and the ICTV will only make these gene-sharing network approaches more scalable and broadly applicable. Assuming broad acceptance of these efforts, and parallel efforts with eukaryotic viruses[35], we might finally have the foundation needed to realize the consensus statement goals[24,27] of establishing a genome-based viral taxonomy to better capture the broader viral sequence landscape emerging from environmental surveys.

METHODS

Data sets. Full-length viral genomes were obtained from the National Center for Biotechnology Information (NCBI) viral reference dataset[26,63] ('ViralRefSeq', version 85, as of January, 2018), downloaded from NCBI's viral genome page (https://www.ncbi.nlm.nih.gov/genome/viruses/) and eukaryotic viruses were removed. The resulting file contained a total of 2,304 RefSeq viral genomes including 2,213 bacterial viruses and 91 archaeal viruses (**Supplementary Table 2**). In parallel, the ICTV taxonomy (ICTV Master Species List v1.3, as of February, 2018) was retrieved from the ICTV homepage (https://talk.ictvonline.org/files/master-species-lists/). ICTV-classifications were available for a subset of genomes at each taxonomic rank, and the final dataset included: 884 viruses from two orders, 974 viruses from 23 families, 363 viruses from 28 subfamilies, and 940 viruses from 264 genera. To maintain hierarchical ranks of taxonomy, we manually incorporated 2016 and 2017 ICTV updates[49,64,65] to NCBI taxonomy when ICTV taxonomy was absent.

Generation of viral protein clusters. Both version 1 and 2 of vConTACT share an identical protein clustering initial step, in which viral proteins are grouped in protein clusters (PCs) through MCL, followed by the formation of viral clusters (VCs) using either MCL (version 1) or ClusterOne (version 2). First, a total of 231,166 protein sequences were extracted from the 2,304 viral genomes (above). Second, to group protein sequences into homologous protein clusters (PCs)[30], all proteins were subjected to all-

versus-all BLASTP[66] searches (default parameters, cut-offs of $1E^{-5}$ on e-value and 50 on bit score). Third, PCs were generated by applying MCL (inflation factor of 2.0), and resulted into all the proteins being organized into 25,513 PCs, with a fraction of proteins (26,625 or 11.5%) as singletons (i.e. isolated protein with no relatives).

Calculating genome similarity between viruses. The resulting output was parsed in the form of a matrix comprised of genomes, PCs and singleton proteins (i.e., 2,304 × 52,138 matrix) (**Supplementary Table 1**). We then determined the similarities between genomes by calculating a one-tailed *P* value of observing at least *c* PCs in common between each pair of genomes, based on the following hypergeometric equation as per Lima-Mendez et al[29]:

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

in which *c* is the number of PCs in common; *a* and *b* are the numbers of PCs and singletons in genomes A and B, respectively; and *n* is the total number of PCs and singletons in the dataset. The hypergeometric formula calculates the probability of sharing a number of common PCs between two genomes at or above the number (*c*) under the null hypothesis that the observed result is likely to occur by chance. A score of similarity between genomes was obtained by taking the negative logarithm (base 10) of the hypergeometric P-value multiplied by the total number of pairwise genome comparisons (i.e., $(2{,}304 \times 2{,}303)/2$). Genome pairs with a similarity score ≥1 were previously shown to be significantly similar through permutation test where PCs and singleton proteins with genome pairs having a similarity score below the given threshold (negative control) were randomly rearranged. None of the genome pairs in this negative control produced similarity score >1, indicating values above this threshold did not occur by chance[30].

Network visualization. The gene (protein)-sharing network was constructed, in which nodes are genomes and edges connect significantly similar genomes. This network was visualized with Cytoscape software (version 3.6.0; http://cytoscape.org/), using an edge-weighted spring embedded model, which places the genomes sharing more PCs closer to each other.

Parameter optimization for viral cluster formation of vConTACT v1.0 and 2.0. Due to different criteria for parameter optimization between the clustering methods, different number and size of the clusters are often generated, which can make objective performance comparisons difficult[67]. Thus, to more comprehensively compare performance, v1.0's MCL-based VCs were generated at inflation factors (IFs) of 2.0 to 7.0 by 1.0 increments, with an optimal IF of 1.4 showing the highest intra-cluster clustering coefficient (ICCC)[29] (**Supplementary Table 2** and **Supplementary Fig. 8**). Unlike MCL, which uses a single parameter[29] (i.e., the inflation factor), VC formation with CL1 (used in vConTACT v2.0), involves multiple parameters that can detect complex network relationships[50]. The three main parameters of CL1, minimum density/node penalty, haircut, and overlap, automatically quantify (i) the cohesiveness of a cluster, (ii) the boundaries of the clusters (i.e. outlier genomes), and (iii) the size of overlap between clusters , respectively[50]. Of these parameters, the first one is used to detect the coherent groups of VCs as follows:

$$C = \frac{W_{in}(V)}{W_{in}(V) + W_{out}(V) + p|C|} \quad (2)$$

in which $W_{in}(V)$ and $W_{out}(V)$ are the total weight of edges that lie within cluster V and that connect the cluster V and the rest of the network, respectively, $|C|$ is the size of the cluster, p is a penalty that counts the possibility of uncharted connections for each node.

The second parameter, the haircut, can find loosely connected regions of the network (outliers) by measuring the ratio of connectivity of the node g within the cluster c to that of its neighbouring node h as:

$$\Delta_{out} = k \sum_{j=1}^{l} W_{h,j} / \sum_{i=1}^{k} W_{g,i} \quad (3)$$

in which k is the number of edges of the node g, and $W$is the total weight of edges of the respective nodes g and h. If the total weight of edges from a node (h) to the rest of the cluster (c) is less than x times that we specified the average weight of nodes (g) within the given cluster, CL1 will remove the node (h) from a given VC and consider it an outlier.

The third CL1 parameter, the overlap size, determines the maximum allowed overlap (ω) between two clusters, measured by the match coefficient, as follows:

$$\omega = i^2 / a * b \quad (4)$$

in which i is the size of overlap, which is divided by the product of the sizes of the two clusters under consideration (a and b). Since CL1 identifies overlap between VCs, it can find both hierarchical and overlapping structures within viral groups. This ability is a significant improvement over v1.0, as v1.0's MCL cannot handle modules with overlaps[7]. Specifically, for each pair of clusters, CL1 calculates the overlap score between them (above) and merges these clusters if the overlap is larger than a given threshold. Thus, in the resulting output file, viral groups (or clusters) having the identical member viruses can be found in multiple clusters, called 'overlapping viral clusters' (**Supplementary Table 2 and Fig. 1b, middle row**).

To determine the best parameter combination to use for CL1, we tested a wide range of values for the three aforementioned parameters: minimum density ranging from 0 to 1 by 0.1 increments; node penalty from 1 to 10 by 1.0; haircut from 0 to 1 by 0.05; overlap from 0 to 1 by 0.05) and default settings for the other parameters: 2 as minimum cluster size, weighted as edge weight, single-pass as merging, unused nodes as seeding. This resulted in 53,361 clustering results, which we evaluated individually to

458 determine the highest performance on our genome data set (above)To identify the best parameter
459 combination, we used the geometric mean value of prediction accuracy (*Acc*) and clustering-wise
460 separation (*Sep*, see next section), as previously described[68]. The final, optimized CL1 parameters were a
461 minimum density of 0.3, a node penalty of 2.0, a haircut of 0.65, and an overlap of 0.8, which resulted in
462 280 VCs (**Supplementary Table 2**).
463 Next, to further decompose 'discordant VCs', we added as a post-clustering step in v2.0, which
464 allows additional hierarchical separation of such VCs into sub-clusters using the unweighted pair group
465 method with arithmetic mean (UPGMA) with pairwise Euclidean distances (implemented in Scipy). To
466 determine the optimal distance for sub-clustering of VCs, we assessed the distances of sub-clusters across
467 all the VCs in the network. We tested the effect of these distances (ranging from 1 to 20 in 0.5
468 increments) and picked as optimal distance the one which maximized the composite score by multiplying
469 the prediction accuracy (*Acc*) and clustering-wise separation (*Sep*) at the ICTV genus rank (see next
470 section). A distance of 9.0 yielded the highest composite score of *Acc* and *Sep* (**Supplementary Fig. 4**).
471 Notably, vConTACT v2.0 was designed to help users optimize these parameters for grouping of
472 genomes/contigs into VCs and distance for post-decomposition of VCs into sub-clusters. This tool
473 automatically evaluates the robustness of each VCs and sub-clusters, , based on the external performance
474 evaluation statistics (below).
475
476 **Performance comparison between vConTACT v1.0 and v2.0.** Six external quality metrics were used
477 to compare clustering performance between MCL and CL1[68] (**Fig. 2a**) . Specifically, the performance of
478 v1.0 (MCL) and v2.0 (CL1 alone and CL1 + hierarchical sub-clustering) were evaluated based on : (i)
479 cluster-wise sensitivity, *Sn* (ii) positive predictive value, *PPV* (iii) geometric mean of *Sn* and *PPV*, *Acc*
480 (iv) cluster-wise separation, Sep_{cl} (v) complex (ICTV taxon)-wise separation Sep_{co}, and (vi) geometric
481 mean of Sep_{cl} and Sep_{co}, *Sep*. As an internal parameter, we computed the intra- and inter-cluster proteome
482 similarities (fraction of shared genes between genome that are within the same VCs and different VCs,

483 respectively). For vConTACT v1.0, we only included clustering results which had been determined to
484 yield the highest clustering accuracy value (i.e., inflation factor of 7.0), and this configuration was used
485 for comparison to v2.0's clustering. Therefore, testing each parameter combination (6 performance
486 metrics, for one taxon rank, for 10 clustering results, all cross-compared; i.e., 6 x 1 x 45) resulted in 270
487 comparisons.
488 To generate six external measures, we first built a contingency table *T*, in which row *i* corresponds to the
489 i^{th} annotated reference complex (i.e., ICTV-recognized order, family, subfamily, or genus), and column *j*
490 corresponds to the j^{th} predicted complex (i.e., sub-/clusters). The value of a cell T_{ij} denotes the number of
491 member viruses in common between the i^{th} reference complex and j^{th} predicted complex.
492 ***Sensitivity***: The sensitivity can be defined as the fraction of member viruses of complex *i* which are found
493 in sub-/cluster *j*.

494 $$Sn_{i,j} = T_{i,j}/N_i \quad (5)$$

495 In the formula above, N_i is the number of member viruses of complex *i*. We then calculated the coverage
496 of complex *i* by its best-matching cluster Sn_{co_i} , as the maximal fraction of member viruses of complex *i*
497 assigned to the same sub-/cluster by the formula below:

498 $$Sn_{co_i} = max_{j=1}^{m} Sn_{i,j} \quad (6)$$

499 The clustering-wise sensitivity was computed as the weighted average of Sn_{co_i} over all complexes.
500 Higher *Sn* values indicate a better coverage of the member viruses in the real complexes as:

501 $$Sn = \frac{\sum_{i=1}^{n} NiSn_{co_i}}{\sum_{i=1}^{n} N_i} \quad (7)$$

502 ***Positive predictive value***: The positive predictive value (*PPV*) indicates the proportion of member viruses
503 of the sub-/cluster *j* which belong to complex *i*, relative to the total number of member viruses of the sub-
504 /cluster assigned to all complexes by:

505 $$PPV_{i,j} = T_{i,j}/\sum_{i=1}^{n} T_{i,j} = T_{i,j}/T_{.j} \quad (8)$$

where $T_{.j}$ is the marginal sum of a column *j*. We calculated the maximal fraction of member viruses of sub-/cluster *j* found in the same annotated complex PPV_{cl_j}, as the prediction reliability of sub-/cluster *j* to belong to its best-matching complex as:

$$PPV_{cl_j} = max_{i=1}^{n} PPV_{i,j} \quad (9)$$

The clustering-wise PPV was then computed as the weighted average of PPV_{cl_j} over all sub/clusters by:

$$PPV = \frac{\sum_{j=1}^{m} T_{.j} PPV_{cl_j}}{\sum_{j=1}^{m} T_{.j}} \quad (10)$$

Higher *PPV* values indicate that the predicted sub-/clusters are likely to be true positives.

***Accuracy*:** As a summary metric, the *Acc* can be obtained by computing the geometrical mean of the *Sn* and *PPV* values as:

$$Acc = \sqrt{Sn \times PPV} \quad (11)$$

***Complex- and Cluster-wise separations*:** With the same contingency table used for *Sn*, *PPV*, and *Acc*, we calculated the relative frequencies with respect to the marginal sums for each row ($F_{row_{i,j}}$) and each column ($F_{col_{i,j}}$), respectively:

$$F_{row_{i,j}} = T_{i,j} / \sum_{j=1}^{m} T_{i,j} \quad (12)$$

$$F_{col_{i,j}} = T_{i,j} / \sum_{i=1}^{n} T_{i,j} \quad (13)$$

Then the separation is computed as the product of column-wise and row-wise frequencies as:.

$$Sep_{i,j} = F_{col_{i,j}} \times F_{row_{i,j}} \quad (14)$$

The separation values range from 0 to 1, with 1 indicating a perfect correspondence between complex *j* and sub-/cluster *i* (i.e., the cluster contains all the members of the complex and only them). Additionally, the separation penalizes the case when member viruses of a given complex are split into multiple sub-

/clusters. The complex-wise Sep_{co} and cluster-wise Sep_{cl} values are calculated as the average of Sep_{co_i} over all complexes, and of Sep_{cl_j} over all sub-/cluster, respectively:

$$Sep_{co} = \frac{\sum_{i=1}^{n} Sep_{co_i}}{n} \quad (15)$$

$$Sep_{cl} = \frac{\sum_{j=1}^{m} Sep_{cl_j}}{m} \quad (16)$$

To estimate these separation results as a whole, the geometric mean (clustering-wise separation; *Sep*) of Sep_{co} and Sep_{cl} was computed:

$$Sep = \sqrt{Sep_{co} \times Sep_{cl}} \quad (17)$$

High clustering-wise separation values indicate a bidirectional correspondence between a sub-/cluster and each ICTV taxon: a score of 1.0 indicates that a cluster corresponds perfectly to each taxon. For overall comparison, we used a composite score[50], calculated by multiplying *Acc* by *Sep*.

As an internal measure, the fraction of PCs[30] between two genomes (i.e., proteome similarity) was computed by using the geometric index (G). The proteome similarity was estimated as:

$$G_{AB} = \frac{|N(A) \cap N(B)|}{|N(A)| \times |N(B)|} \quad (18)$$

in which *N(A)* and *N(B)* indicate the number of PCs in the genomes of A and B, respectively. A total of 400,234 pairs of genomes with >1% proteome similarity are shown in **Supplementary Table 4**.

Clustering-based confidence score. To generate confidence scores for each viral cluster prediction, we used three previously described confidence scoring methods[69,70], with some modifications. Two of them exploit the network topology properties by assessing the weight of cluster quality and the probability of

cluster quality. We then combined these two values as an aggregate topology-based confidence score per VC. For the first scoring method, we computed the quality (Q) of sub-cluster (c) as:

$$Q_c = W_{in}/(W_{in} + W_{out}) \quad (19)$$

in which W_{in} and W_{out} are the total weight of edges that lie within sub-cluster c and across others, respectively. For the second method, we evaluated the P-value of a one-sided Mann-Whitney U test for in-weights and out-weights of sub-clusters. The rationale behind this test is that sub-clusters with a lower P-value contains significantly higher in-weights than out-weights, thus indicative that a formed sub-cluster is valid, and not a random fluctuation. These two independent values, weight of cluster quality and the probability of cluster quality are then multiplied to derive a topology-based confidence score for each cluster. Along with this confidence score, we quantified the likelihood that each sub-cluster corresponds to an ICTV-approved genus (or equivalent) by using distance threshold that are specified at the ICTV genus rank, which we refer to as "taxon predictive score". This score can be calculated as:

$$prediction = \sum l_{i,j} / l_c \quad (20)$$

Specifically, for a sub-cluster (c) having the genus-level assignment, vConTACT v2.0 automatically measures the maximum distance between taxonomically-known member viruses and calculate the scores by dividing the sum of links having less than the given maximum distance threshold between nodes (i and j) by the total number of links (l_c) between all nodes. For a sub-cluster that does not have the genus-level assignment, v2.0 uses Euclidean distance of 9.0 that can maximize the prediction accuracy and clustering-wise separation (see above) as distance threshold.

Measuring effect of GOV on network structural changes. GOV contigs (14,656 sequences) were added in 10% increments (randomly selected at each iteration) to NCBI Viral RefSeq and processed using vConTACT v2.0 with one difference – Diamond[71] instead of BLASTp was used to construct the all-versus-all protein comparison underlying the PC generation. For running this large number of sequences, high-memory computer nodes from the Ohio State supercomputer Center[72] were used. Once generated, vConTACT v2.0 networks were post-processed using a combination of the Scipy[73], Numpy, Pandas[74] and Scikit-learn[75] python 3.6 packages. Networks were rendered using iGraph[76]. The method to calculate change centrality was calculated as described previously[77]. CCs were calculated in a successive way, in which each addition was compared to Viral RefSeq 85 independently of other additions (0% versus 10%, 0% vs 20%, […], 0% vs 100%).

Code availability. The vConTACT v2.0 package is freely distributed through Bit Bucket as a python package (https://bitbucket.org/MAVERICLab/vcontact2).

REFERENCES

1. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive earth's biogeochemical cycles. *Science* **320,** 1034–1039 (2008).
2. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science (80-.).* **348,** (2015).
3. Moran, M. A. The global ocean microbiome. *Science* **350,** (2015).
4. Zhao, M. *et al.* Microbial mediation of biogeochemical cycles revealed by simulation of global changes with soil transplant and cropping. *ISME J.* **8,** 2045–2055 (2014).
5. Cho, I. & Blaser, M. J. The human microbiome: At the interface of health and disease. *Nature Reviews Genetics* **13,** 260–270 (2012).
6. Fernández, L., Rodríguez, A. & García, P. Phage or foe: an insight into the impact of viral predation on microbial communities. *ISME Journal* 1–9 (2018). doi:10.1038/s41396-018-0049-5
7. Hurwitz, B. L. & U'Ren, J. M. Viral metabolic reprogramming in marine ecosystems. *Current Opinion in Microbiology* **31,** 161–168 (2016).
8. Suttle, C. a. Marine viruses-major players in the global ecosystem. *Nat. Rev. Microbiol.* **5,** 801–812 (2007).
9. Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science (80-.).* **348,**

(2015).

10. Danovaro, R. *et al.* Virus-mediated archaeal hecatomb in the deep seafloor. *Sci. Adv.* **2,** (2016).

11. Pratama, A. A. & van Elsas, J. D. The 'Neglected' Soil Virome - Potential Role and Impact. *Trends in Microbiology* (2018). doi:10.1016/j.tim.2017.12.004

12. Gómez, P. & Buckling, A. Bacteria-phage antagonistic coevolution in soil. *Science (80-.).* **332,** 106–109 (2011).

13. Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F. & Gordon, J. I. Going viral: Next-generation sequencing applied to phage populations in the human gut. *Nature Reviews Microbiology* **10,** 607–617 (2012).

14. Abeles, S. R. & Pride, D. T. Molecular bases and role of viruses in the human microbiome. *Journal of Molecular Biology* **426,** 3892–3906 (2014).

15. Rohwer, F. & Edwards, R. The phage proteomic tree: A genome-based taxonomy for phage. *J. Bacteriol.* **184,** 4529–4535 (2002).

16. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12,** 635–645 (2014).

17. Lawrence, J. G., Hatfull, G. F. & Hendrix, R. W. Imbroglios of viral taxonomy: Genetic exchange and failings of phenetic approaches. *J. Bacteriol.* **184,** 4891–4905 (2002).

18. Sullivan, M. B. Viromes, Not Gene Markers, for Studying Double-Stranded DNA Virus Communities. *J. Virol.* **89,** 2459–2461 (2015).

19. Deng, L. *et al.* Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature* **513,** 242–245 (2014).

20. Gregory, A. C. *et al.* Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics* **17,** (2016).

21. Bobay, L. & Ochman, H. Biological species in the viral world. **115,** (2018).

22. Mavrich, T. N. & Hatfull, G. F. Bacteriophage evolution differs by host, lifestyle and genome. *Nat. Microbiol.* **2,** (2017).

23. Ackermann, H.-W. Phage Classification and Characterization BT - Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions. in (eds. Clokie, M. R. J. & Kropinski, A. M.) 127–140 (Humana Press, 2009). doi:10.1007/978-1-60327-164-6_13

24. Simmonds, P. *et al.* Consensus statement: Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* **15,** 161–168 (2017).

25. Paez-Espino, D. *et al.* IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res.* gky1127-gky1127 (2018).

26. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral Genomes resource. *Nucleic Acids Res.* **43,** D571–D577 (2015).

27. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG): a community consensus on standards and best practices for describing genome sequences from uncultivated viruses. *Nat. Biotechnol.* (2018).

28. Nishimura, Y. *et al.* ViPTree: The viral proteomic tree server. *Bioinformatics* **33,** 2379–2380 (2017).

29. Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* **25,** 762–777 (2008).

30. Bolduc, B. *et al.* vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. *PeerJ* **5,** e3243 (2017).

31. Meier-Kolthoff, J. P. & Göker, M. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx440

32. Yu, C. *et al.* Real Time Classification of Viruses in 12 Dimensions. *PLoS One* **8,** (2013).

33. Gao, Y. & Luo, L. Genome-based phylogeny of dsDNA viruses by a novel alignment-free method. *Gene* **492,** 309–314 (2012).

34. Iranzo, J., Koonin, E. V., Prangishvili, D. & Krupovic, M. Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements. *J. Virol.* **90,** 11043–11055 (2016).

35. Aiewsakun, P. & Simmonds, P. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. *Microbiome* **6,** 38 (2018).

36. Lavigne, R. *et al.* Classification of myoviridae bacteriophages using protein sequence similarity. *BMC Microbiol.* **9,** (2009).

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

38. Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K. & Schuster, S. C. Whole-genome prokaryotic phylogeny. *Bioinformatics* **21,** 2329–2335 (2005).

39. Iranzo, J., Krupovic, M. & Koonin, E. V. The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. *MBio* **7,** (2016).

40. Bolduc, B., Youens-Clark, K., Roux, S., Hurwitz, B. L. & Sullivan, M. B. IVirus: Facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. *ISME J.* **11,** 7–14 (2017).

41. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537,** 689–693 (2016).

42. Vik, D. R. *et al.* Putative archaeal viruses from the mesopelagic ocean. *PeerJ* **5,** e3428 (2017).

43. Roux, S. *et al.* Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nat. Commun.* **8,** (2017).

44. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* (2018). doi:10.1038/s41564-018-0190-y

45. Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat. Commun.* **8,** (2017).

46. de la Cruz Peña, M. J. *et al.* Deciphering the Human Virome with Single-Virus Genomics and Metagenomics. *Viruses* **10,** 113 (2018).

47. Aiewsakun, P., Adriaenssens, E. M., Lavigne, R., Kropinski, A. M. & Simmonds, P. Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: Steps towards a unified taxonomy. *J. Gen. Virol.* **99,** 1331–1343 (2018).

48. Hulo, C., Masson, P., Le Mercier, P. & Toussaint, A. A structured annotation frame for the transposable phages: A new proposed family 'Saltoviridae' within the Caudovirales. *Virology* **477,** 155–163 (2015).

49. Adriaenssens, E. M. *et al.* Taxonomy of prokaryotic viruses: 2017 update from the ICTV Bacterial and Archaeal Viruses Subcommittee. *Archives of Virology* 1–5 (2018). doi:10.1007/s00705-018-3723-z

50. Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* **9,** 471–472 (2012).

51. Doyle, E. L. *et al.* Genome Sequences of Four Cluster P Mycobacteriophages. *Genome Announc.* **6,** e01101-17 (2018).

52. Pope, W. H. *et al.* Bacteriophages of Gordonia spp. Display a spectrum of diversity and genetic relationships. *MBio* **8,** (2017).

53. Pope, W. H. *et al.* Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *Elife* **4,** e06416 (2015).

54. Nelson, D. Phage taxonomy: We agree to disagree. *Journal of Bacteriology* **186,** 7029–7031 (2004).

55. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36,** 996 (2018).

56. Krupovic, M., Quemin, E. R. J., Bamford, D. H., Forterre, P. & Prangishvili, D. Unification of the Globally Distributed Spindle-Shaped Viruses of the Archaea. *J. Virol.* **88,** 2354–2358 (2014).

57. Rokyta, D. R., Burch, C. L., Caudle, S. B. & Wichman, H. A. Horizontal gene transfer and the evolution of microvirid coliphage genomes. *J. Bacteriol.* **188,** 1134–1142 (2006).

58. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11,** 538 (2010).

59. Gregory, A. *et al.* Marine viral macro- and micro-diversity from pole to pole.

60. Marz, M. *et al.* Challenges in RNA virus bioinformatics. *Bioinformatics* **30,** 1793–1799 (2014).

61. Howard-Varona, C., Hargreaves, K. R., Abedon, S. T. & Sullivan, M. B. Lysogeny in nature: Mechanisms, impact and ecology of temperate phages. *ISME Journal* **11,** 1511–1520 (2017).

62. Mirzaei, M. K. & Maurice, C. F. Ménage à trois in the human gut: Interactions between host, bacteria and phages. *Nature Reviews Microbiology* **15,** 397–408 (2017).

63. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44,** D733–D745 (2016).

64. Krupovic, M. *et al.* Taxonomy of prokaryotic viruses: update from the ICTV bacterial and

archaeal viruses subcommittee. *Arch. Virol.* **161,** 1095–1099 (2016).

65. Adams, M. J. *et al.* Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2017). *Arch. Virol.* **162,** 2505–2538 (2017).

66. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* **25,** 3389–3402 (1997).

67. Wiwie, C., Baumbach, J. & Röttger, R. Comparing the performance of biomedical clustering methods. *Nat. Methods* **12,** 1033–1038 (2015).

68. Brohée, S. & van Helden, J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **7,** (2006).

69. Kamburov, A., Stelzl, U. & Herwig, R. IntScore: A web tool for confidence scoring of biological interactions. *Nucleic Acids Res.* **40,** (2012).

70. Goldberg, D. S. & Roth, F. P. Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci.* **100,** 4372–4376 (2003).

71. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12,** 59–60 (2015).

72. Ohio Supercomputer Center . (1987).

73. Oliphant, T. E. SciPy: Open source scientific tools for Python. *Comput. Sci. Eng.* **9,** 10–20 (2007).

74. McKinney, W. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.* **1697900,** 51–56 (2010).

75. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

76. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* **1695,** 1–9 (2006).

77. Federico, P., Pfeffer, J., Aigner, W., Miksch, S. & Zenk, L. Visual Analysis of Dynamic Networks Using Change Centrality. in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* 179–183 (2012). doi:10.1109/ASONAM.2012.39

ACKNOWLEDGEMENTS. We thank Laura Bollinger, Gareth Trubl, and Igor Tolstoy for their comments on improving the manuscript, as well as Wesley Zhi-Qiang You for helping push the network analytics. High performance computational support was provided as an award from the Ohio Supercomputer Center to MBS. Funding was provided in part by the Department of Energy's Genome Sciences Program Soil Microbiome Scientific Focus Area award (#SCW1632) to Lawrence Livermore National Laboratory; an NSF Biological Oceanography award (OCE#1536989), and a Gordon and Betty Moore Foundation Investigator Award (#3790) to MBS. Funding was provided to JRB by the Intramural Research Program of the NIH, National Library of Medicine. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract DE-AC02-05CH11231 to SR. This work was funded in part through Battelle Memorial Institute's prime contract with the US National Institute of Allergy and Infectious Diseases (NIAID) under Contract No. HHSN272200700016I to JHK. The content of this

763 publication does not necessarily reflect the views or policies of the US Department of Health and Human
764 Services or of the institutions and companies affiliated with the authors.

765 **AUTHOR CONTRIBUTIONS.** HBJ, BB and MBS designed the study. OZ and MBS wrote the
766 manuscript with significant contributions from all co-authors. HBJ and BB performed the statistical and
767 network analyses.

768 **COMPETING INTERESTS.** The authors declare no competing interests.

769 **MATERIALS & CORRESPONDENCE.** Correspondence and material requests should be addressed to
770 Matthew B. Sullivan at sullivan.948@osu.edu.

771

Figure 1. Virus genome classification visualized as networks. (a) Left side panel: matrix of shared protein clusters (PCs, grey blocks) between a set of virus genomes can be visualized as a network of interconnected nodes, as shown on the right-side of the panel. Each node in this sample 6-node network represents a virus genome that may be connected to other nodes through edges. The edge value represents the strength of connectivity between nodes. If a set of nodes have considerably higher edge weights than the rest of the network they are linked to, these are grouped together to form a viral cluster, or 'VC'. **(b)** Each row depicts a node clustering scenario in which vConTACT v2.0 has improved upon. On the left side, each scenario is first depicted as a genome-PC matrix highlighting how shared protein clusters between certain genomes may induce erroneous virus groupings due to outlier genomes, overlapping viral groups or VCs containing multiple viral groups. On the right side of the matrices, the topology of each clustering scenario is depicted as small networks of nodes (color-coded according to the ICTV genera colors next to the matrices), and shows how vConTACT version 1 and 2 handled clustering of problematic genomes and/or VCs. **(c)** Heatmap key corresponding to the various values related to edge weight in **(a)** and **(b)**, which serve to connect the nodes in the networks and shows how closely related each connected node is to other nodes based on the number of common PCs between genomes.

Figure 2. Performance evaluation of vConTACT versions 1.0 and 2.0 on prokaryotic virus genomes. (**a)** The same colors denote individual performance metrics for the ICTV genera (G) including 940 viral genomes , which are achieved by the Markov clustering (MCL) algorithm at each inflation factor (v1.0) as well as ClusterONE (CL1) and CL1 followed by distance-based hierarchical clustering (CL1 + H) (v2.0), respectively. For more objective comparisons, MCL at an inflation factor of 7.0 followed by hierarchical clustering (IF 7.0 + H) with the same distance (i.e., 9.0) used for v2.0 was included. The total height and number of each bar indicate the composite score for overall performance comparison (**Fig. 2a**). For details, see Online Methods. (**b, c)** Gene-sharing networks were built using 2,304 archaeal and bacterial virus genomes retrieved from Viral RefSeq v85. Viral clusters (VCs) were obtained by vConTACT v1.0 (**b**) and v2.0 (**c**) that used MCL with inflation factor (IF) of 7.0 and CL1, respectively (see Online Methods). For both networks, genomes (nodes) are color-coded according to their taxonomic assignments. For example, genomes (only members of the ICTV-recognized genus) that are classified in VCs containing a single ICTV genus are colored in cyan, while genomes found in VCs containing more than two genera are colored in pink. Genomes without ICTV genus affiliation are in grey. Nodes with bold borders indicate those that were correctly identified either as outlier, overlap genomes or separate VCs through v1.0 (**b**), compared to v2.0 (**c**). Genomes whose taxonomic assignments and/or annotation are incomplete are colored in yellow, identified through v2.0 (**c**). For details, see **Supplementary Figs 3** and **5** and **Supplementary Table 2**. **(d)** Box plots of the percentage of shared protein clusters (PCs) between member viruses within 28 v1.0-generated undersampled VCs having ≥2 genera before (pink), and after (cyan) removal of outlier and/or separation into individual clusters by v2.0. **(e)** Pie charts depicting the number of overlapping genomes that belong to the high (HGCF) or low (LGCF) gene content flux evolutionary modes or mixed and lytic or temperate phages. Data on the lifestyle and evolutionary modes of 74 viruses were collected from Mavrich and Hatfull[22]. For details, see **Supplementary Fig. 3**.

Figure 3. Application of the hierarchical decomposition to discordant VCs. (a) Distribution of all 31 discordant VCs across the archaeal and bacterial virus gene sharing network, where genomes (nodes) of the given VCs are highlighted in pink and others in grey. **(b)** Box plots show the fraction (%) of protein clusters (PCs) that were shared within an ICTV genus (i.e., intra-genus proteome similarity) and between multiple genera (i.e., inter-genera similarity) found in each discordant VC including structured clusters whose member genera have similar inter-genera and intra-genus similarities (black dot). **(c)** Left, A full link dendrogram is represented. Note that the Euclidean distance of nine yielded the highest composite score of accuracy (*Acc*) and clustering-wise separation (*Sep*) for sub-clusters from all v2.0-generated VCs, which was used to split the discordant clusters (**Online Methods** and **Supplementary Fig. 4**).

Right, module profiles showing the presence and absence of 7,662 total protein clusters (PCs) across 362 genomes. Each row represents a phage and each column represents a PC, with a unique color (left of the module) representing the genome's VC and ICTV genus, respectively. Sub-clusters, which are generated by distance-based hierarchical grouping, are represented across all discordant VCs on the right side of the heat map. From the 12 discordant VCs, 37 sub-clusters (corresponding to a single ICTV genus), are highlighted as green boxes. For details, see **Supplementary Table 2**.

Figure 4. Adding the Global Ocean Virome to NCBI Viral RefSeq. (**a**) Selected network images from the largest connected component of GOV-additions. Red nodes are virus RefSeq genomes, and grey nodes are GOV. Despite adding 15,280 new genomes, the network maintains its overall structure. (**b**) Pairwise heatmap comparison at all GOV incremental additions using normalized mutual information (NMI) values. NMI measures VC similarity to other VCs by comparing genome content changes across incremental additions of data. Darker blue hues correspond to more similar information content (i.e. genomes maintaining the same VC membership). (**c**) Boxplots depicting the average Euclidean distance within VCs across GOV data increments. Grey boxes are samples prior to hierarchical trimming, while blue boxes are post-trimming. Points represent discordant VCs, with darker hues representing increasing discordance (i.e., more genera per VC). (**d**) Change centralities on a per-genome (gray) and per-VC (aqua) basis through successive, 10% increments of GOV data. A value of zero in change centrality (Y-axis) represent no change in any of the nodes connected to the origin node (or that the node was removed), while a value of one represents origin node creation. High change centrality scores imply that nodes are being created adjacent to the origin node, with the further a node's creation is from the origin node, the less of an impact it has on the origin node's centrality. Dotted lines in each violin represent quartiles, whereas the width of each violin plot is scaled to be equal between GOV % (X-axis), such that distributions can be compared between datasets. **(e)** GOV network performance through successive data accumulations. As GOV sequences are added (X-axis), individual performance score (ranging from 0 to 1, Y-axis; calculated from the clustering-wise positive predictive value (PPV), clustering-wise sensitivity and accuracy) across genus- and family-level predictions (represented by circular and square data points, respectively) generally trend towards stabilization.

Supplementary Figure 1. Bipartite network of reference virus genomes generated by vConTACT 2. (a) The full network is represented **(b)** Close-up of 3 viral clusters displayed as bipartite network. In this configuration, pink nodes represent individual genomes, while dark grey nodes depict proteins clusters that are shared between viral clusters.

Supplementary Figure 2. Clustering comparisons between vConTACT v1.0 and v2.0 (a) Number of total viral clusters (VCs), and genus-assigned VCs, concordant and discordant VCs, as detected by vConTACT v1.0 (left) and v2.0 (right). Discordant VCs are (i) those that have a mix of the different genera (i.e., lumped genera), (ii) those that have different member virus(es) of the same genus by splitting them into multiple clusters (i.e., split genera), or (iii) mix of (i) and (ii). **(b)** Proteome similarities of viruses within 22 concordant VCs of vConTACT version 1.0. For all plots, the x-axis is the individual pairwise comparisons and y-axis is the proteome similarity (i.e., percentage fraction of shared protein cluster between genomes). In the case of VCs 6, 26, 66, and 130 (highlighted by bold borders), these all contains taxonomically-misplaced member virus(es) of the *Che8virus*, *Pbunavirus*, *P100virus*, and *Bcep78virus*, respectively, all of which were correctly captured by v2.0 and ratified by the ICTV (see **Supplementary Fig. 5**). Like these four VCs, the remaining 18 VCs contained distant relatives, with only 1-30% of similarities to the rest of the given clusters or discrete viral group(s) displaying a number of discontinuous similarities, which were identified as outliers or separated VCs by v2.0, respectively (see **Supplementary Table 4**).

Supplementary Figure 3. vConTACT v2.0-based detection and characterization of overlapping viral genomes. (a) Box plots depicting the distribution of the topology-based confidence scores between viruses identified as overlaps and non-overlaps, which vConTACT v2.0 placed into two clusters (see panel b) and single clusters, respectively. For details, see Methods. **(b)** List of phages and archaeal viruses identified as overlaps and their ICTV genus. Details on the lifestyle and evolutionary modes of 74 viruses were collected from Mavrich and Hatfull[22] and from the Actinobacteriophage Database website (http://phagesdb.org/). The high (HGCF) and low (LGCF) gene content flux evolutionary modes indicate the predicted lifestyle based on the gene content dissimilarity between viral genomes. Bioinformatically-predicted temperate phages indicate those that contain the integrase (for integrating temperate phage genomes into host) or *parA* (partitioning gene found in extrachromosomal temperate phages) genes.

Supplementary Figure 4. Evaluation of optimal distance thresholds for hierarchical clustering of VCs. The X-axis denotes distance threshold increments from dist=1 to dist=20 in 0.5 intervals. The Y-axis denotes composite scores by multiplying Accuracy (*Acc*, cyan) and clustering-wise separation (*Sep*, pink) when trying to recapitulate ICTV genera, which are geometric means of Sensitivity and Positive predictive value as well as Complex-wise separation and Cluster-wise separation, respectively. From these data, a distance of 9.0 yielded the highest composite score for the sub-clusters partitioned from all vConTACT v2.0-generated viral clusters. For details, see Online Methods.

Supplementary Figure 5. vConTACT v2.0-based detection and characterization of boundary genome(s) within the ICTV-recognized genera. (a) Box plots show the percentage of shared protein clusters (PCs) between members of an ICTV genus (red), and the same metrics after excluding viruses recognized as outlier(s) by vConTACT v2.0 (cyan). The proteome similarities for the *Barnyardvirus* are shown between (1) member viruses of the *Barnyardvirus*, (2) *Mycobacterium* virus Barnyard and the remaining members of the *Barnyardvirus*, (3) the *Barnyardvirus* and *Patiencevirus*, (4) *Mycobacterium* virus Barnyard and the *Patiencevirus*, and (5) the *Barnyardvirus* without outliers. For the *Phikmvvirus*, the proteome similarities between (1) member viruses of the *Phikmvvirus*, (2) *Pseudomonas virus phiKMV* and the remaining members of the *Phikmvvirus*, and (3) *Pseudomonas virus phiKMV* and members of VC_33 are shown, respectively. (**b)** Module profiles show the presence (dark) and absence (light) of homologous PCs across genomes. Each row represents a virus and each column a PC. The genomes were hierarchically clustered based on pairwise Euclidean distance. The ICTV and vConTACT v2.0 classifications are indicated next to each virus.

Supplementary Figure 6. Evaluation of singletons/outlier genomes over GOV increments. Thirty-eight ICTV recognized singleton and outlier genomes (one per row) were observed to evaluate whether the addition of GOV sequences would improve their classification. The coloring gradient on the left indicates the numbers of genera per VC. Clearly improved clustering was observed in 3 of the 38 genomes (black bolded), with 2 genomes clustering to 6-genera discordant VCs (red), though the majority saw only minor, if any, change in their clustering assignment.

Supplementary Figure 7. vConTACT v2.0 computational runtimes. In the plot, processing time (in seconds, Y-axis) is plotted against increasing GOV sequence data (GOV % added, X-axis). There is a strong linear correlation between runtime and memory usage with data volume to be processed.

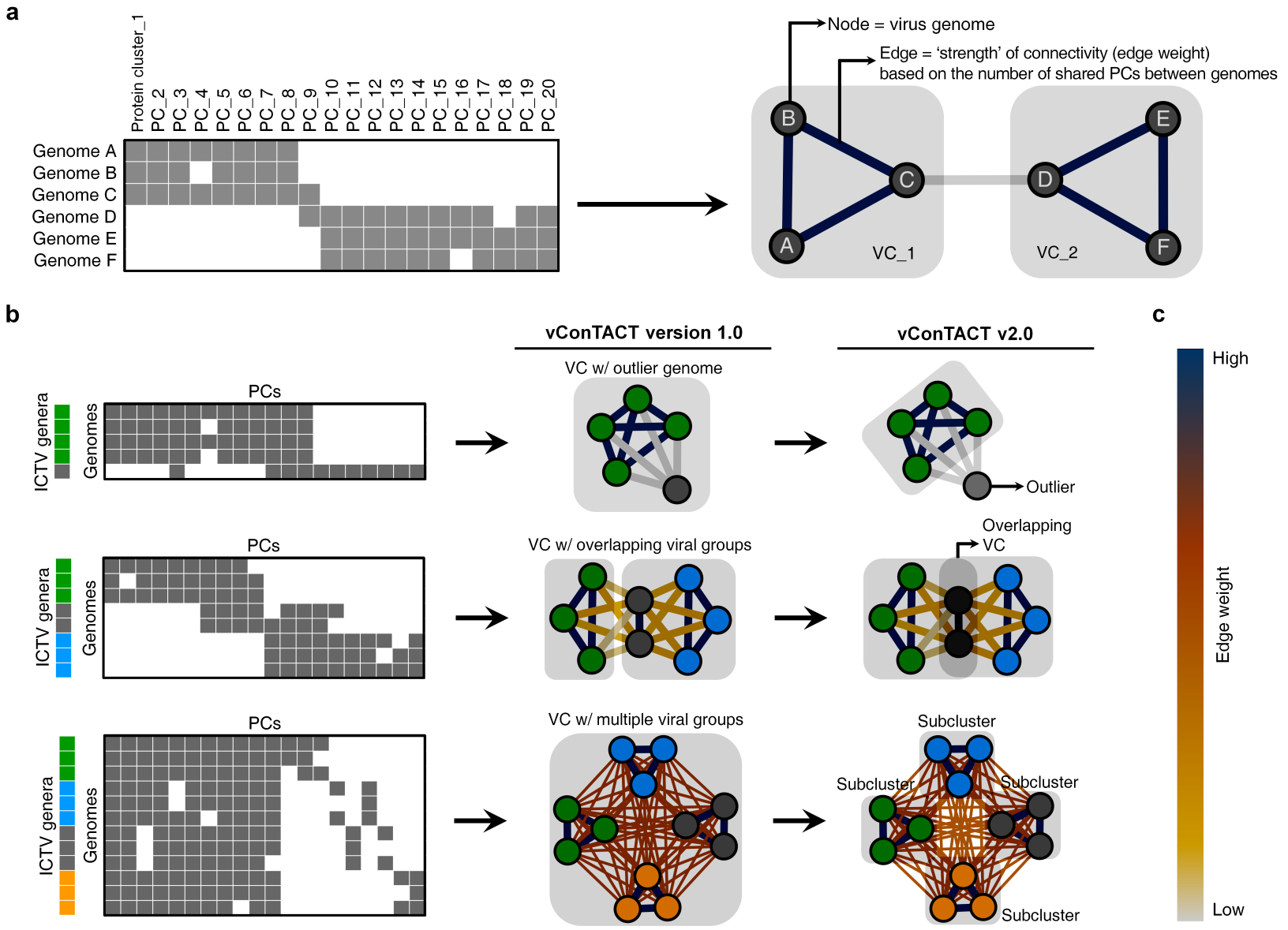
Supplementary Figure 8. Impact of the inflation factor on viral genome clustering based on the Markov clustering (MCL) algorithm. Left panel: average intra-cluster clustering coefficient (ICCC) and number of viral clusters (VCs), which are predicted as a function of the inflation factors ranging from 1.0 to 5.0 with a step of 0.2, are indicated. Right panel: Curve representing the ICCC values for the network containing 2,304 archaeal and bacterial virus genomes.

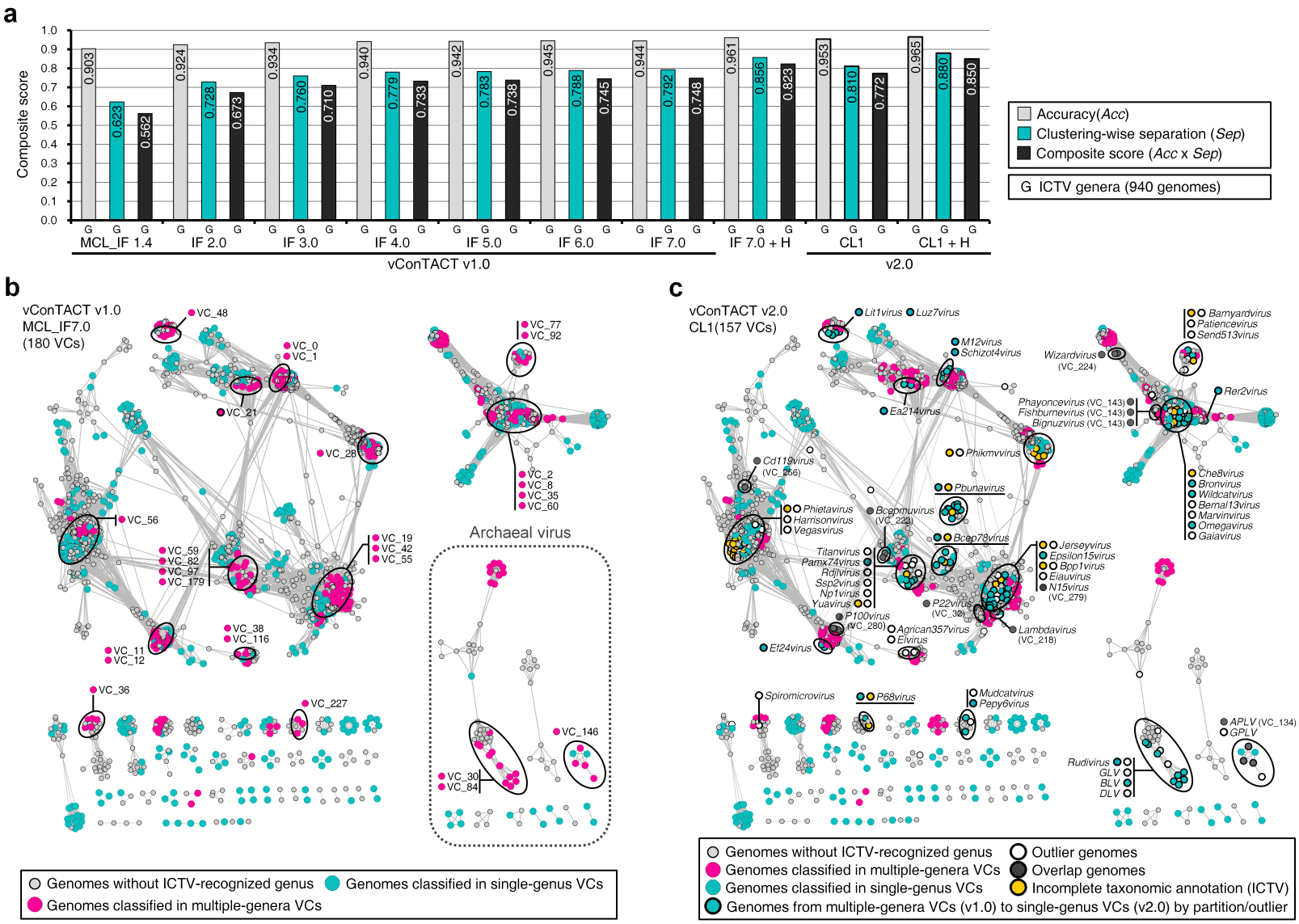
Supplementary Table 1: A genome-gene matrix. In this matrix, a total of 231,166 protein-coding genes from 2,304 bacterial and archaeal virus genomes are shown, along with their annotations, protein clusters and singleton proteins (i.e., isolated protein with no relatives). Each singleton protein was represented as the blank in the column of "cluster".

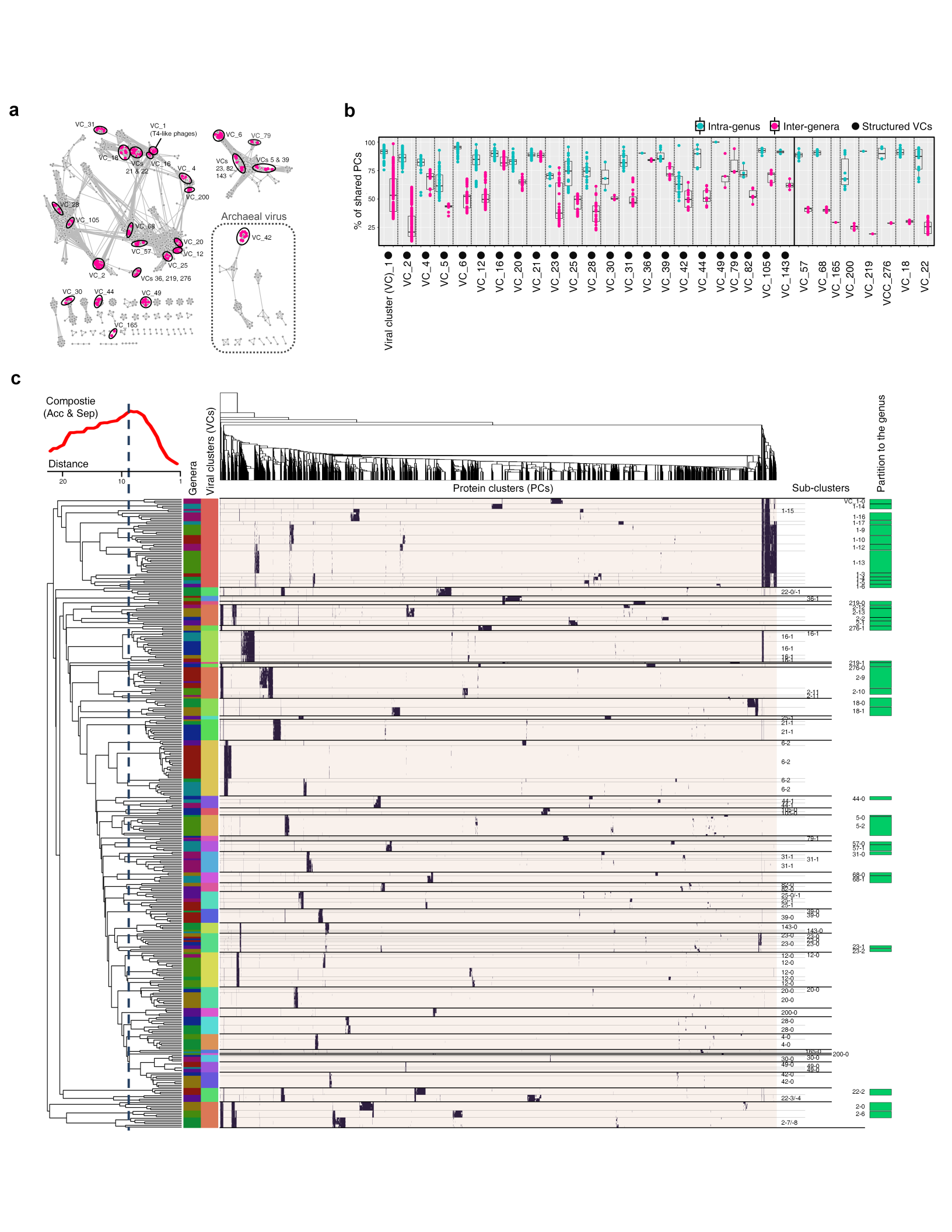
Supplementary Table 2. List of 2,304 archaeal and bacterial virus genomes used to evaluate vConTACT v1.0 and v2.0. Genomes were retrieved from NCBI RefSeq, their main taxonomic features from NCBI and ICTV as well as the clustering outputs from vConTACT v1.0 and v2.0 is indicated. Particularly for v2.0, additional features for viral clusters (VCs) by CL1 and subclusters by CL1 followed, by distance-based hierarchical clustering, as well as quality and p-values, were used to generate the topology-based score. For each subcluster, a taxonomy-based scores is also provided. For the outlier genomes, their parent VCs (i.e., those from which v2.0 separated the given genomes as outliers) are included. For the genomes identified as part of overlapping VCs, dual VC membership is represented by '/'. The former VC, as an overlapping cluster, can be also found in the latter cluster (mix of two VCs). See Online Method for details.

Supplementary Table 3. Clustering performance evaluations of the vConTACT v1.0, v2.0, and v2.0 followed by distance-based hierarchical clustering for the genus rank. The numbers represent the values of Sensitivity (*Sn*), Positive prediction value (*PPV*), and Accuracy (geometrical mean of *Sn* and *PPV*, *Acc*) as well as Cluster-wise separation (Sep_{cl}), Complex-wise separation (Sep_{co}), and Clustering-wise separation (*Sep*) across all clustering results, respectively. "Distance" indicates whether a hierarchical clustering was applied to all VCs (either "NoDist" if no clustering was applied, or "9" if a clustering with a cutoff of 9.0 on euclidean distance was applied).

Supplementary Table 4. Fraction of PCs in common between 2,304 genomes. The average shared PC percentage is calculated based on the geometric formula (see Online Methods) and 400,234 pairs of genomes having >1% proteome similarity are shown. For vConTACT v1.0 that uses MCL, the proteome similarities of VCs was further calculated with IFs of 1.4 and 2.0 to 7.0 by a step of 1.0. The VC and ICTV taxa (i.e., subfamily and genera) that each genome belongs to are represented.

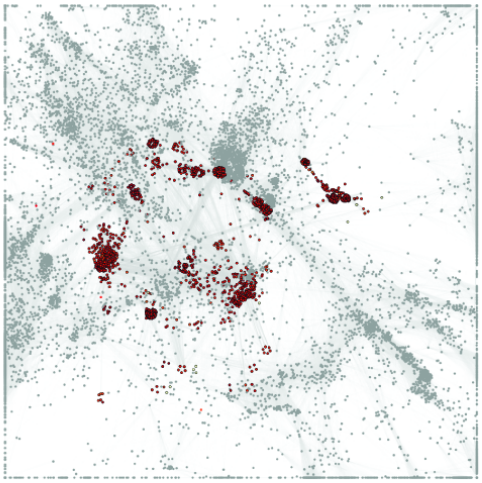

a
Protein cluster_1
PC_2
PC_3
PC_4
PC_5
PC_6
PC_7
PC_8
PC_9
PC_10
PC_11
PC_12
PC_13
PC_14
PC_15
PC_16
PC_17
PC_18
PC_19
PC_20
Genome A
Genome B
Genome C
Genome D
Genome E
Genome F
Node = virus genome
Edge = 'strength' of connectivity (edge weight)
based on the number of shared PCs between genomes
A
B
C
D
E
F
VC_1
VC_2
b
vConTACT version 1.0
vConTACT v2.0
PCs
ICTV genera
Genomes
VC w/ outlier genome
Outlier
VC w/ overlapping viral groups
Overlapping
VC
VC w/ multiple viral groups
Subcluster
Subcluster
Subcluster
Subcluster
c
High
Edge weight
Low


a
Composite score
Accuracy(Acc)
Clustering-wise separation (Sep)
Composite score (Acc x Sep)
G ICTV genera (940 genomes)
MCL_IF 1.4
IF 2.0
IF 3.0
IF 4.0
IF 5.0
IF 6.0
IF 7.0
IF 7.0 + H
CL1
CL1 + H
vConTACT v1.0
v2.0
b
vConTACT v1.0
MCL_IF7.0
(180 VCs)
Archaeal virus
Genomes without ICTV-recognized genus
Genomes classified in single-genus VCs
Genomes classified in multiple-genera VCs
c
vConTACT v2.0
CL1(157 VCs)
Genomes without ICTV-recognized genus
Genomes classified in multiple-genera VCs
Genomes classified in single-genus VCs
Genomes from multiple-genera VCs (v1.0) to single-genus VCs (v2.0) by partition/outlier
Outlier genomes
Overlap genomes
Incomplete taxonomic annotation (ICTV)
d
vConTACT v1.0
v2.0 (outlier removal/cluster separation)
% of shared PCs
MCL_VC_0
e
Unknown_T (2)
Mixed_T (7)
LGCF_T (2)
HGCF_L (5)
HGCF_T (26)

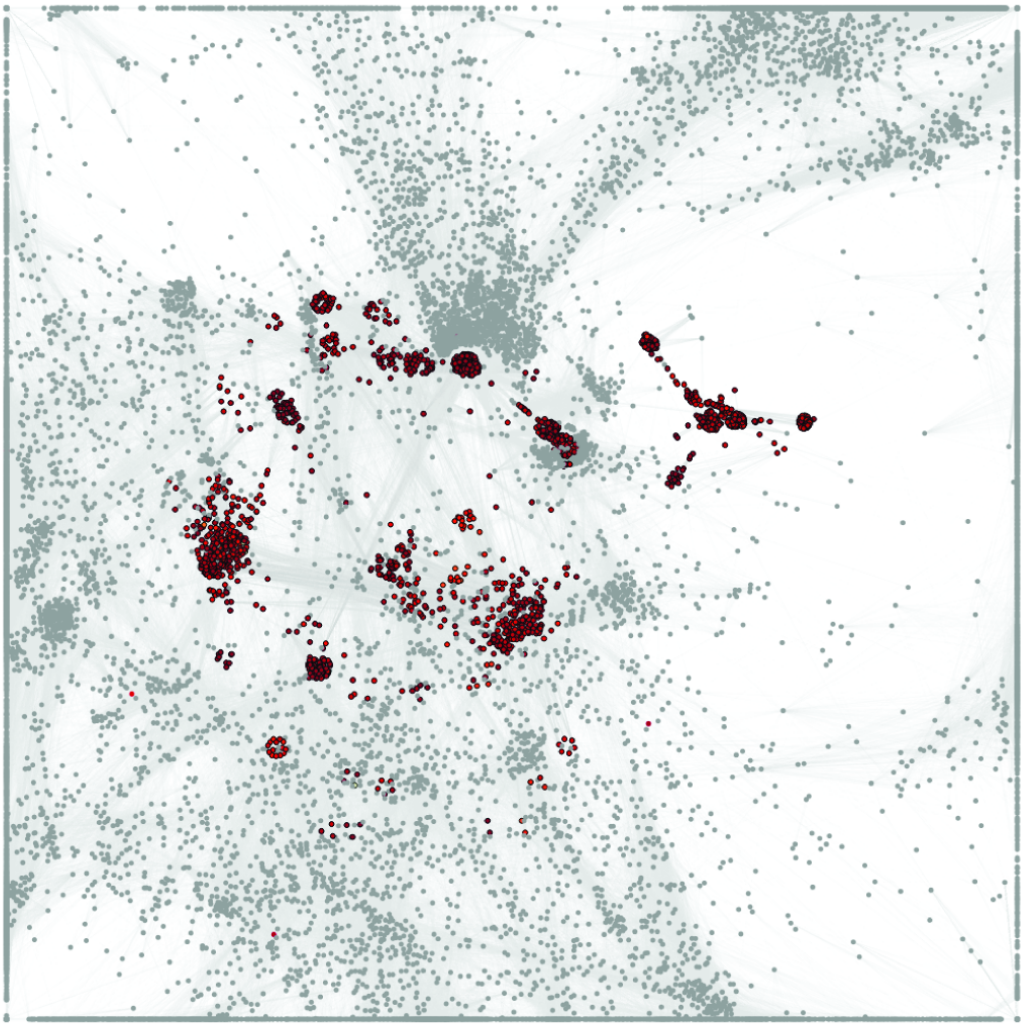

a
VC_31
VC_1
(T4-like phages)
VC_18
VCs
21 & 22
VC_16
VC_ 4
VC_200
VC_6
VC_79
VCs
23, 82,
143
VCs 5 & 39
VC_28
VC_105
VC_68
VC_20
VC_57
VC_12
VC_25
VC_2
VCs 36, 219, 276
Archaeal virus
VC_42
VC_30
VC_44
VC_49
VC_165
b
Intra-genus
Inter-genera
Structured VCs
% of shared PCs
100
75
50
25
Viral cluster (VC)_1
VC_2
VC_4
VC_5
VC_6
VC_12
VC_16
VC_20
VC_21
VC_23
VC_25
VC_28
VC_30
VC_31
VC_36
VC_39
VC_42
VC_44
VC_49
VC_79
VC_82
VC_105
VC_143
VC_57
VC_68
VC_165
VC_200
VC_219
VCC_276
VC_18
VC_22
c
Compostie
(Acc & Sep)
Distance
20
10
1
Genera
Viral clusters (VCs)
Protein clusters (PCs)
Sub-clusters
Partition to the genus

a

30% (5,737)

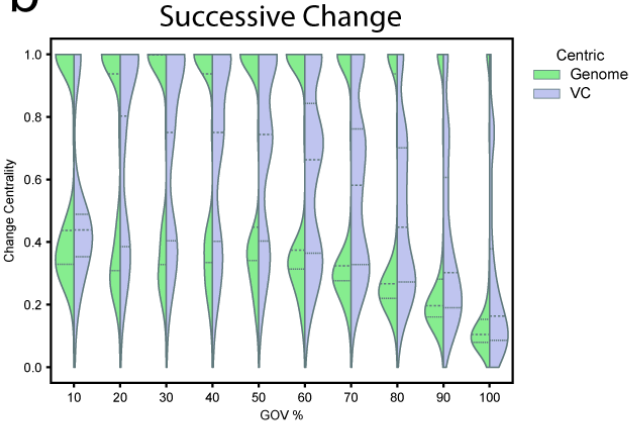


70% (10,638)



100% (14,524)

b



c