Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks

Ho Bin Jang[1*], Benjamin Bolduc[1*], Olivier Zablocki[1], Jens H. Kuhn[2], Simon Roux[3], Evelien M. Adriaenssens[4,5], J. Rodney Brister[6], Andrew M Kropinski[7, 8], Mart Krupovic[9], Rob Lavigne[10], Dann Turner[11], & Matthew B. Sullivan[1,12#]

[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] Quadram Institute Bioscience, Norwich Research Park, Norwich, UK

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

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

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

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

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

[11] 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

[12] 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

Microbiomes from every environment contain a myriad of uncultivated archaeal and bacterial viruses, but studying these viruses is hampered by the lack of a universal, scalable, taxonomic framework. We present vConTACT v2.0, a network-based application utilizing whole genome gene-sharing profiles for virus taxonomy that integrates distance-based hierarchical clustering and confidence scores for all taxonomic predictions. We report near-identical (96%) replication of existing genus-level viral taxonomy assignments from the International Committee on Taxonomy of Viruses (ICTV) for NCBI virus Refseq. Application of vConTACT v2.0 to 1,364 previously unclassified viruses deposited in virus

RefSeq as reference genomes produced automatic, high confidence genus assignments for 820/1364. We applied vconTACTv2.0 to analyse 15,280 Global Ocean Virome genome fragments and were able to provide taxonomic assignments for 31% of these data, which shows that our algorithm is scalable to very large metagenomic datasets. Our taxonomy tool can be automated and applied to metagenomes from any environment for virus classification.

Editors summary

Classification of archaeal and bacterial viruses can be automated with an algorithm that identifies relationships based on shared gene content.

Bacteria and archaea have roles in nutrient and energy cycles in ocean and soil ecosystems[1–4], as well as playing a vital part in human and health[5]. Viruses that infect bacteria and archaea modulate these 'ecosystem roles' by killing, metabolic reprogramming or gene transfer[6,7], with substantial effects of viral predation predicted in ocean[8–10], soil[11,12] and human microbiomes[13,14]. However, ecosystem-scale understanding of virus dynamics is hampered by the lack of universal viral genes, or methods that enable a formalized taxonomy or comparative surveys. For example, viruses do not have a single, universal marker gene[15] so microbial-style 16S rRNA-based phylogenies and operational taxonomic units (OTUs) are impossible[16].

Virus sequencing has revealed structure[17,18] and population genetic support for a species definition[19], and hypotheses have been put forward to explain variable evolution among prokaryotic viruses[20]. Together with rapidly expanding viral genome databases, these advances have led the International Committee on Taxonomy of Viruses (ICTV) to present a consensus

64 statement suggesting a shift from the 'traditional' classification criteria[21] e.g. virion morphology,
65 single/multiple gene phylogenies, towards a genome-centered, and perhaps one-day, largely
66 automated, viral taxonomy[22].

67 Given the pace of viral discovery a virus taxonomy is urgently needed. Hundreds of
68 thousands of metagenome-derived viral genomes and large genome fragments (more than
69 700,000 at IMG/VR[23]) dwarf the 34,091 prokaryotic virus genomes present in the NCBI
70 GenBank database[24]. Together with the recently proposed 'minimum information about
71 uncultivated virus genomes' (MIUViGs) community guidelines[25], evaluation of approaches to
72 establish a scalable, genome-based viral taxonomy is needed to enable a universal classification
73 framework.

74 Multiple genome-based strategies have been proposed to develop a taxonomic framework for
75 viruses of bacteria[15,26–31], archaea[32] or eukaryotes[33]. For bacterial viruses ("phages"), one early
76 approach used complete genome pairwise protein sequence comparisons in a phylogenetic
77 framework (the "phage proteomic tree") and was broadly concordant with ICTV-endorsed virus
78 groupings at the time[15]. However, this approach was not widely adopted as it was thought that
79 "rampant mosaicism" might blur taxonomic boundaries and violate the assumptions of the
80 underlying phylogenetic algorithms used in the analyses[34]. Other approaches estimated the
81 fraction of genes shared, and percent identity of shared gene cut-offs, to define genera and sub-
82 family affiliations[35,36], but this approach failed to define taxonomic classification for several
83 known virus groups due to the likelihood that the mode and tempo of prokaryotic virus evolution
84 is highly variable[20]. Building on a prokaryotic classification algorithm, the Genome Blast
85 Distance Phylogeny (GBDP)[37], which comes with a freely accessible online tool (VICTOR),
86 classifies phage genomes by combining phylogenetic and clustering methods[29]. This method has

insufficient scalability (100 genomes limit) and limited taxonomic assignment for viruses that lack reference genomes.

Gene sharing networks, based on shared protein clusters (PCs) between viral genomes, have been shown to be largely concordant with ICTV-endorsed taxa, independent of whether monopartite[27,28,38] (a single node type, i.e., viral genomes) or bipartite networks[32,38] (two node types, i.e., viral genomes and genes) were used. We used a monopartite gene sharing network to build an iVirus[39] app (vConTACT v1.0, hereafter v1.0) to automate network-based classification of prokaryotic viruses. v1.0 produced viral clusters ('VCs') that were ~75% concordant with ICTV prokaryotic viral genera[28]. Network-based analytics have been applied to viral taxonomy in large-scale studies of ocean[40,41], freshwater[42] and soil[43] and studies of single-virus amplified genomes (vSAGs)[44,45]. In all of these environments the viruses could only be classified upon application of a gene sharing network method. v1.0 cannot, however, make tentative taxonomic assignments. This is because v1.0 creates artifactual VCs of both undersampled genomes and highly overlapped regions of viral sequence space[28], and lacks per-VC confidence metrics, necessary for establishing hierarchical taxonomy.

Here we present vConTACT v2.0 (hereafter v2.0), which has a new clustering algorithm, confidence scoring of clusters and network analytics that together enable automation, improved taxonomy assignments and scalability to much larger datasets. We apply v2.0 to establish a centralized, 'living' taxonomic reference network as a community resource and show that v2.0 is robust and scalable to large metagenomic datasets.

RESULTS

Description of vConTACTv2.0

The aim of vConTACT is to automatically assign viral genomes into established or new taxa, with performance assessed relative to ICTV-assigned, manually-curated taxa (**Fig. 1**). However, in the current ICTV taxonomy for prokaryotic viruses, taxonomic classifications above the genus level are only sporadically available for sub-family and order ranks. For example, of the 2,304 prokaryotic virus genomes available in 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[46–48] and are thus in flux. Therefore, we focused specifically on assigning viruses at the genus level, as it constitutes the principal taxon of molecular classification 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[28]. In v1.0, ~75% of VCs corresponded to established ICTV genera[28] ('concordant VCs'), but ~25% 'discordant VCs' were present. Discordant VCs can occur by production of outlier cluster genomes with no close relatives from 'undersampled VCs', or by incorrect overlapping of multiple ICTV genera that share many genes or by misassignment of multiple ICTV genera into a structured VC (**Fig. 1b**).

To address these problems we developed a new clustering algorithm, established confidence scores and distance-based taxon separation for hierarchical taxonomy, and optimized and evaluated scalability and robustness using 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[49] (CL1), rather than MCL which is used in v1.0, and then subdivided using hierarchical clustering to disentangle problematic regions of the networks (**Fig. 1b**, Online Methods). This approach considers edge weight (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 artefacts. 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 we present v2.0 as a monopartite (one type of node) network tool, it produces the necessary output to be visualized as a bipartite network (**Supplementary Fig. 1**). In bipartite visualizations, two types of nodes are used to display genomes and their connecting, shared protein clusters (PCs). Information about which PCs link a given set a viruses together is also provided (**Supplementary Table 1**; Online Methods), as it can enable identification of core virus group genes that might be useful for downstream analyses.

Comparison of vConTACT 1.0 and 2.0

To assess clustering performance of v1.0 and v2.0, we quantified concordance with the set of 940 prokaryotic virus genomes that have ICTV genus-level classification (accessed January 2018, Online Methods and **Supplementary Table 2**). Clustering performance was evaluated by a composite performance score of Accuracy (*Acc*) and Separation (*Sep*). Both *Acc* and *Sep* are aggregate measures themselves (Online Methods), and report clustering precision, and how resulting clusters (or VCs) correspond to a single ICTV genus, respectively (**Fig. 2a**). Each metric has a value between 0 and 1, with 1 indicating perfect clustering accuracy and/or coverage.

v2.0's CL1, combined with hierarchical clustering, resulted in an overall performance improvement of 28.8% (**Fig. 2a**). To assess which changes in v2.0 contributed to improved performance we further optimized v1.0's MCL-based VC clustering and found that, at an IF of 7, we could achieve nearly equivalent performance (**Fig. 2a, Supplementary Table 3**) and more VCs predicted by the optimized MCL-based configuration as it organized the 940 viral genomes into 180 VCs, whereas v2.0's CL1 identified 157 VCs. However, higher values in *Sep* for CL1 indicate better performance for assigning single genera into single VCs, even 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 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.

vConTACT v2.0 can analyse genomic relationships

Next, we tested whether v2.0 could resolve 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**). v2.0 correctly places 38 genomes from 15 genera into 15 now concordant VCs using the same input data (**Fig. 2c and d** and **Supplementary Table 2**). Second, we evaluated the ability of v2.0 to resolve overlapping VCs (**Fig. 1b**). We detected overlapping VCs using a 'match coefficient' that measures the connection within- and between-other VCs (Online Methods). This approach identified nine overlapping VCs (ICTV-classified genera only) containing 30 viruses in 11 ICTV genera. These included viruses with known mosaic genomes[47] (lambdoid or mu-like phages of the *P22virus*, *Lambdavirus*, *N15virus*, and *Bcepmuvirus* genera), recombinogenic temperate phages[50,51] (*Mycobacterium* phages of the *Bignuzvirus, Phayoncevirus,* and *Fishburnevirus* genera and *Gordonia* phages of the genus *Wizardvirus*), and three newly-established genera (*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 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[20]. 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[52] (6 genomes), could not be resolved (**Supplementary Table 2**), and may not be amenable to automated taxonomy.

Third, structured VCs (**Fig. 1b**, bottom row) contained genomes that both 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 (**Supplementary Note 1**). For v1.0 we previously reported that these structured VCs could be decomposed through hierarchical clustering[27], but 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**). Automated v2.0 resolved 30% (26 of 86) of these ICTV genera from 6 of the 23 structured VCs (**Fig. 3c**).

Of the 2,304 reference virus genomes classified by ICTV at the genus rank, 1,364 are currently unassigned to a genus. This set of 1,364 reference viruses was organized into 404 well-supported VCs with v2.0 (**Supplementary Table 2**). 544/1364 were placed in 104 VCs with genomes from known ICTV taxa, whereas 820/1364 formed 200 separate VCs. We propose that

these 820 genomes can be as 200 *bona fide* novel virus genera and have submitted these to the ICTV for consideration. If ratified, application of vContact2.0 will double the number of prokaryotic viral genera (which is currently 264).

v2.0 clustering changed the taxonomy of ten established ICTV genera: *Barnyardvirus*, *Bcep78virus*, *Bpp1virus*, *Che8virus*, *Jerseyvirus*, *P68virus*, *Pbunavirus*, *Phietavirus*, *Phikmvvirus*, and *Yuavirus* (**Supplementary Fig. 5** and **Supplementary Note 2**), and manual inspection by ICTV members involved in this study has recommended revision of *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 the cut-offs used to define subclusters are different from those currently used to delineate established genera (**Fig. 3c** and **Supplementary Fig. 4**). Universal cut-offs are known to be of limited use. Manual curation by experts has resulted in different cut-offs across viral sequence space[53]. A standardized taxonomy has been proposed for bacteria and archaea[54] and for viruses standardization would be invaluable for automating virus taxonomy. v2.0 VCs and subclusters will provide a reference baseline for the ICTV to translate network-derived cut-offs into systematic taxonomic demarcation criteria.

Some taxon assignments are not amenable to being resolved by gene-sharing networks. For example, when genera are defined on phenotypic or evolutionary evidence, e.g., archaeal fuselloviruses[55] (VC42) or bacterial microviruses[56] (VCs 30 and 49), a gene-sharing network approach will not be suitable (see **Fig. 3c** and **Supplementary Table 2**). An automated vConTACT-based approach can however identify problematic taxa and speed up revisions to the taxonomy.

vConTACT v2.0 can scale to large virome datasets

To evaluate scalability of our algorithm, we added 15,280 curated viral genomes and large genome fragments (≥ 10 kb) from the Global Ocean Virome (GOV) dataset[40] to our reference network in 10% increments (i.e., 0%, 10%, …, 100% of the total dataset). The final network comprised 16,960 sequences (**Fig. 4a**). We evaluated whether the incremental addition of GOV data to the network led 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; **Fig. 4b**). We also evaluated concordance between v2.0 clustering and ICTV genera using the Sn, Acc & PPV performance metrics (**Fig. 4c**). A large fraction of added data initially experiences a moderate change (CC = 0.4), but the entire 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 v2.0 can scale to thousands of input sequences, and that our reference network clustering is robust to large-scale data additions.

We assessed whether GOV data can resolve ICTV outlier and singleton genomes as a proxy for assessing taxonomic ramifications of adding data. We reasoned that more data might connect outliers to new or existing VCs. Of 38 single-member VCs (**Supplementary Fig. 6**) three *Mycobacterium* phage VCs were improved, while two *Mycobacterium* virus genomes were merged into larger heterogeneous VCs composed of six ICTV genera, which did not constitute an improvement. We observed that 919 new VCs were created with the full GOV dataset (15, 280 total contigs). We propose that these new VCs represent 919 viral genera that are not represented in the existing 264 ICTV genera. According to a recent consensus statement, any taxonomic reference network must be constrained to complete genomes[22], and large genome

fragments commonly derived from metagenome-based studies must be utilized in a relevant manner to address questions specific to that study, so these results remain preliminary.

Discussion

vConTACT v2.0 offers a scalable, robust, systematic and automated means to classify bacterial and archaeal virus sequences. V2.0 is a highly scalable tool. 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 would take ~10 minutes on a regular laptop, while a GOV-sized dataset would run for several hours.

There are limitation of v2.0. 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[57]) to incorporate new data. Second, CL1-based VC generation may require manual parameter optimization if datasets with overlapping genomes are included. We have added an auto-optimization option for determining the optimal distance for hierarchical decomposition of structured VCs in v2.0. v2.0 can run with prokaryotic viruses but has not been designed, tested or validated for eukaryotic viruses. These viruses will require new algorithms for classification as they have more diverse genomic configurations (segmentation, overlapping genes and ambisense transcriptional gene configurations) that pose unique computational challenges[33,58]. Short, complete prokaryotic virus genomes and small fragments of larger genomes (e.g., ≤ 3 PCs or ≤ 5 genes) have low statistical power in gene-sharing networks, and will require new solutions to establish higher confidence VCs, and remain taxonomically inaccessible using v2.0. 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.

Assuming broad acceptance of vConTACT v2.0, and parallel efforts with eukaryotic viruses[33], we may finally have the foundation to realize the consensus statement goals[22,25] of establishing a genome-based viral taxonomy to better capture the broader viral sequence landscape emerging from environmental surveys.

ACKNOWLEDGEMENTS.
We thank Laura Bollinger, G. Trubl, and I. Tolstoy for their comments on improving the manuscript, as well as Z-Q. 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; NSF Biological Oceanography awards (OCE#1536989 and OCE#1756314), 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 publication does not necessarily reflect the views or policies of the US Department of Health and Human Services or of the institutions and companies affiliated with the authors.

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

COMPETING INTERESTS.
The authors declare no competing interests.

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

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. Deng, L. *et al.* Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature* **513,** 242–245 (2014).
18. Gregory, A. C. *et al.* Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics* **17,** (2016).
19. Bobay, L. & Ochman, H. Biological species in the viral world. **115,** (2018).
20. Mavrich, T. N. & Hatfull, G. F. Bacteriophage evolution differs by host, lifestyle and genome. *Nat. Microbiol.* **2,** (2017).
21. 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
22. Simmonds, P. *et al.* Consensus statement: Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* **15,** 161–168 (2017).
23. 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).
24. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral Genomes resource.

Nucleic Acids Res. **43,** D571–D577 (2015).
25. 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).
26. Nishimura, Y. *et al.* ViPTree: The viral proteomic tree server. *Bioinformatics* **33,** 2379–2380 (2017).
27. 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).
28. Bolduc, B. *et al.* vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. *PeerJ* **5,** e3243 (2017).
29. Meier-Kolthoff, J. P. & Göker, M. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx440
30. Yu, C. *et al.* Real Time Classification of Viruses in 12 Dimensions. *PLoS One* **8,** (2013).
31. Gao, Y. & Luo, L. Genome-based phylogeny of dsDNA viruses by a novel alignment-free method. *Gene* **492,** 309–314 (2012).
32. 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).
33. 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).
34. 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).
35. Lavigne, R. *et al.* Classification of myoviridae bacteriophages using protein sequence similarity. *BMC Microbiol.* **9,** (2009).
36. 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).
37. Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K. & Schuster, S. C. Whole-genome prokaryotic phylogeny. *Bioinformatics* **21,** 2329–2335 (2005).
38. Iranzo, J., Krupovic, M. & Koonin, E. V. The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. *MBio* **7,** (2016).
39. 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).
40. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537,** 689–693 (2016).
41. Vik, D. R. *et al.* Putative archaeal viruses from the mesopelagic ocean. *PeerJ* **5,** e3428 (2017).
42. Roux, S. *et al.* Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nat. Commun.* **8,** (2017).
43. 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
44. Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat. Commun.* **8,** (2017).

45. de la Cruz Peña, M. J. *et al.* Deciphering the Human Virome with Single-Virus Genomics and Metagenomics. *Viruses* **10,** 113 (2018).
46. 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).
47. 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).
48. 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
49. Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* **9,** 471–472 (2012).
50. Doyle, E. L. *et al.* Genome Sequences of Four Cluster P Mycobacteriophages. *Genome Announc.* **6,** e01101-17 (2018).
51. Pope, W. H. *et al.* Bacteriophages of Gordonia spp. Display a spectrum of diversity and genetic relationships. *MBio* **8,** (2017).
52. 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).
53. Nelson, D. Phage taxonomy: We agree to disagree. *Journal of Bacteriology* **186,** 7029–7031 (2004).
54. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36,** 996 (2018).
55. 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).
56. 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).
57. 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).
58. Marz, M. *et al.* Challenges in RNA virus bioinformatics. *Bioinformatics* **30,** 1793–1799 (2014).

Figure Legends

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 of vConTACT 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. 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. All box plots (n=61) were defined in terms of the minima, center, maxima, percentiles and sample size (**Supplementary Table 5**). (**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). All box plots (n=60) were defined in terms of the

minima, center, maxima, percentiles and sample size (**Supplementary Table 6**). **(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**) Change centralities on a per-genome (grey) 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. Numbers in parentheses indicate the number of genomes corresponding the GOV % above and numbers in bracket indicate the number of corresponding VCs. 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**) 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. 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).

ONLINE METHODS

Data sets. Full-length viral genomes were obtained from the National Center for Biotechnology Information (NCBI) viral reference dataset[24,59] ('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[48,60,61] 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)[28], all proteins were subjected to all-versus-all BLASTP[62] 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[27]:

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

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[28].

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[63]. 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)[27] (**Supplementary Table 2** and **Supplementary Fig. 8**). Unlike MCL, which uses a single parameter[27] (i.e., the inflation factor), VC formation with CL1 (used in vConTACT v2.0), involves multiple parameters that can detect complex network relationships[49]. 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[49]. Of these parameters, the first one is used to detect the coherent groups of VCs as follows:

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

623 in which $W_{in}(V)$ and $W_{out}\ (V)$ are the total weight of edges that lie within cluster V and that
624 connect the cluster V and the rest of the network, respectively, $|C|$ is the size of the cluster, p is a
625 penalty that counts the possibility of uncharted connections for each node.
626 The second parameter, the haircut, can find loosely connected regions of the network (outliers)
627 by measuring the ratio of connectivity of the node g within the cluster c to that of its
628 neighbouring node h as:
629
630 $$\Delta_{out} = \ k \sum_{j=1}^{l} W_{h,j} / \sum_{i=1}^{k} W_{g,i}$$
631 (3)
632
633 in which k is the number of edges of the node g, and W is the total weight of edges of the
634 respective nodes g and h. If the total weight of edges from a node (h) to the rest of the cluster (c)
635 is less than x times that we specified the average weight of nodes (g) within the given cluster,
636 CL1 will remove the node (h) from a given VC and consider it an outlier.
637 The third CL1 parameter, the overlap size, determines the maximum allowed overlap (ω)
638 between two clusters, measured by the match coefficient, as follows:
639
640 $$\omega \ = \ i^2 / a * b$$
641 (4)
642
643 in which i is the size of overlap, which is divided by the product of the sizes of the two clusters
644 under consideration (a and b). Since CL1 identifies overlap between VCs, it can find both
645 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 determine the highest performance on our genome data set (above)To identify the best parameter combination, we used the geometric mean value of prediction accuracy (*Acc*) and clustering-wise separation (*Sep*, see next section), as previously described[64]. The final, optimized CL1 parameters were a 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 280 VCs (**Supplementary Table 2**).

Next, to further decompose 'discordant VCs', we added as a post-clustering step in v2.0, which allows additional hierarchical separation of such VCs into sub-clusters using the unweighted pair group method with arithmetic mean (UPGMA) with pairwise Euclidean distances (implemented in Scipy). To determine the optimal distance for sub-clustering of VCs, we assessed the distances of sub-clusters across all the VCs in the network. We tested the effect of these distances (ranging from 1 to 20 in 0.5 increments) and picked as optimal distance the one which maximized the composite score by multiplying the prediction accuracy (*Acc*) and

clustering-wise separation (*Sep*) at the ICTV genus rank (see next section). A distance of 9.0 yielded the highest composite score of *Acc* and *Sep* (**Supplementary Fig. 4**). Notably, vConTACT v2.0 was designed to help users optimize these parameters for grouping of genomes/contigs into VCs and distance for post-decomposition of VCs into sub-clusters. This tool automatically evaluates the robustness of each VCs and sub-clusters, based on the external performance evaluation statistics (below).

Performance comparison between vConTACT v1.0 and v2.0. Six external quality metrics were used to compare clustering performance between MCL and CL1[64] (**Fig. 2a**) . Specifically, the performance of v1.0 (MCL) and v2.0 (CL1 alone and CL1 + hierarchical sub-clustering) were evaluated based on : (i) cluster-wise sensitivity, *Sn* (ii) positive predictive value, *PPV* (iii) geometric mean of *Sn* and *PPV*, *Acc* (iv) cluster-wise separation, Sep_{cl} (v) complex (ICTV taxon)-wise separation Sep_{co}, and (vi) geometric mean of Sep_{cl} and Sep_{co}, *Sep*. As an internal parameter, we computed the intra- and inter-cluster proteome similarities (fraction of shared genes between genome that are within the same VCs and different VCs, respectively). For vConTACT v1.0, we only included clustering results which had been determined to yield the highest clustering accuracy value (i.e., inflation factor of 7.0), and this configuration was used for comparison to v2.0's clustering. Therefore, testing each parameter combination (6 performance metrics, for one taxon rank, for 10 clustering results, all cross-compared; i.e., 6 x 1 x 45) resulted in 270 comparisons.

To generate six external measures, we first built a contingency table *T*, in which row *i* corresponds to the i^{th} annotated reference complex (i.e., ICTV-recognized order, family, subfamily, or genus), and column *j* corresponds to the j^{th} predicted complex (i.e., sub-/clusters).

692 The value of a cell T_{ij} denotes the number of member viruses in common between the i^{th}

693 reference complex and j^{th} predicted complex.

694 ***Sensitivity***: The sensitivity can be defined as the fraction of member viruses of complex ***i*** which

695 are found in sub-/cluster *j*.

696 $$Sn_{i,j} = T_{i,j}/N_i$$

697 (5)

698 In the formula above, N_i is the number of member viruses of complex *i*. We then calculated the

699 coverage of complex *i* by its best-matching cluster Sn_{co_i} , as the maximal fraction of member

700 viruses of complex *i* assigned to the same sub-/cluster by the formula below:

701 $$Sn_{co_i} = max_{j=1}^{m} Sn_{i,j}$$

702 (6)

703 The clustering-wise sensitivity was computed as the weighted average of Sn_{co_i} over all

704 complexes. Higher *Sn* values indicate a better coverage of the member viruses in the real

705 complexes as:

706 $$Sn = \frac{\sum_{i=1}^{n} NiSn_{co_i}}{\sum_{i=1}^{n} N_i}$$

707 (7)

708 ***Positive predictive value***: The positive predictive value (*PPV*) indicates the proportion of

709 member viruses of the sub-/cluster *j* which belong to complex *i*, relative to the total number of

710 member viruses of the sub-/cluster assigned to all complexes by:

711 $$PPV_{i,j} = T_{i,j}/\sum_{i=1}^{n} T_{i,j} = T_{i,j}/T_{.j}$$

712 (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[49], calculated by multiplying *Acc* by *Sep*.

As an internal measure, the fraction of PCs[28] 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[65,66], 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:

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

779

780 in which W_{in} and W_{out} are the total weight of edges that lie within sub-cluster c and across
781 others, respectively. For the second method, we evaluated the P-value of a one-sided Mann-
782 Whitney U test for in-weights and out-weights of sub-clusters. The rationale behind this test is
783 that sub-clusters with a lower P-value contains significantly higher in-weights than out-weights,
784 thus indicative that a formed sub-cluster is valid, and not a random fluctuation. These two
785 independent values, weight of cluster quality and the probability of cluster quality are then
786 multiplied to derive a topology-based confidence score for each cluster. Along with this
787 confidence score, we quantified the likelihood that each sub-cluster corresponds to an ICTV-
788 approved genus (or equivalent) by using distance threshold that are specified at the ICTV genus
789 rank, which we refer to as "taxon predictive score". This score can be calculated as:
790

791 $$prediction = \sum l_{i,j} / l_c \quad (20)$$

792

793 Specifically, for a sub-cluster (c) having the genus-level assignment, vConTACT v2.0
794 automatically measures the maximum distance between taxonomically-known member viruses
795 and calculate the scores by dividing the sum of links having less than the given maximum
796 distance threshold between nodes (i and j) by the total number of links (l_c) between all nodes.
797 For a sub-cluster that does not have the genus-level assignment, v2.0 uses Euclidean distance of
798 9.0 that can maximize the prediction accuracy and clustering-wise separation (see above) as
799 distance threshold.
800

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[67] 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[68] were used. Once generated, vConTACT v2.0 networks were post-processed using a combination of the Scipy[69], Numpy, Pandas[70] and Scikit-learn[71] python 3.6 packages. Networks were rendered using iGraph[72]. The method to calculate change centrality was calculated as described previously[73]. 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%).

Data and code availability statement

The set of reference genomes used to evaluate vConTACT were retrieved from https://www.ncbi.nlm.nih.gov/genome/viruses/. The Global Ocean Virome (GOV) contigs were retrieved from the publically available CyVerse data commons repository, accessible at http://datacommons.cyverse.org/browse/iplant/home/shared/iVirus/GOV.The utility of v2.0 depends upon its expert evaluation and community availability. The tool is available through Bitbucket (https://bitbucket.org/MAVERICLab/vcontact2) as a downloadable python package, and usable as an app through iVirus[39], 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

824 approximately bi-yearly with as complete genomes become available and resources exist to
825 support this effort.

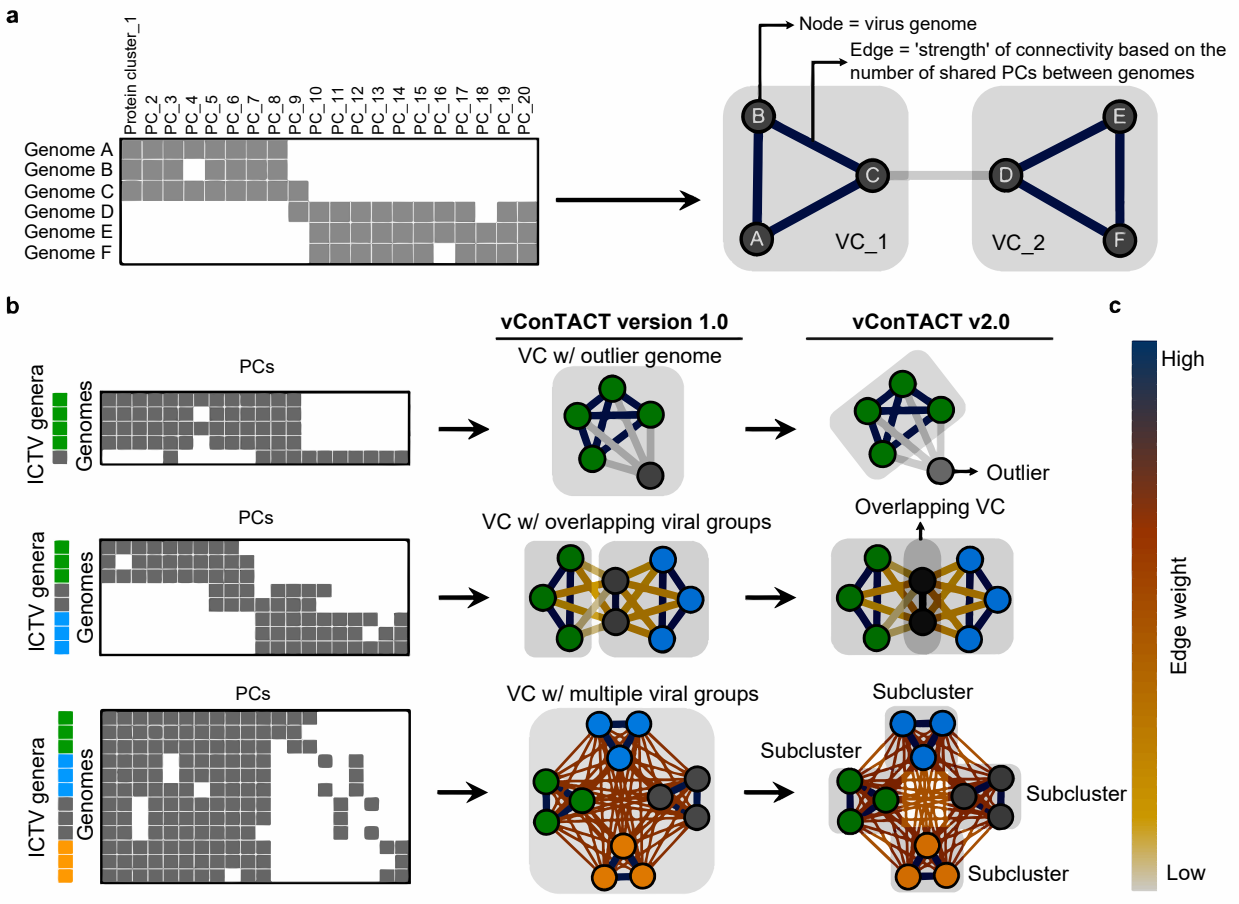
826 Methods-only References

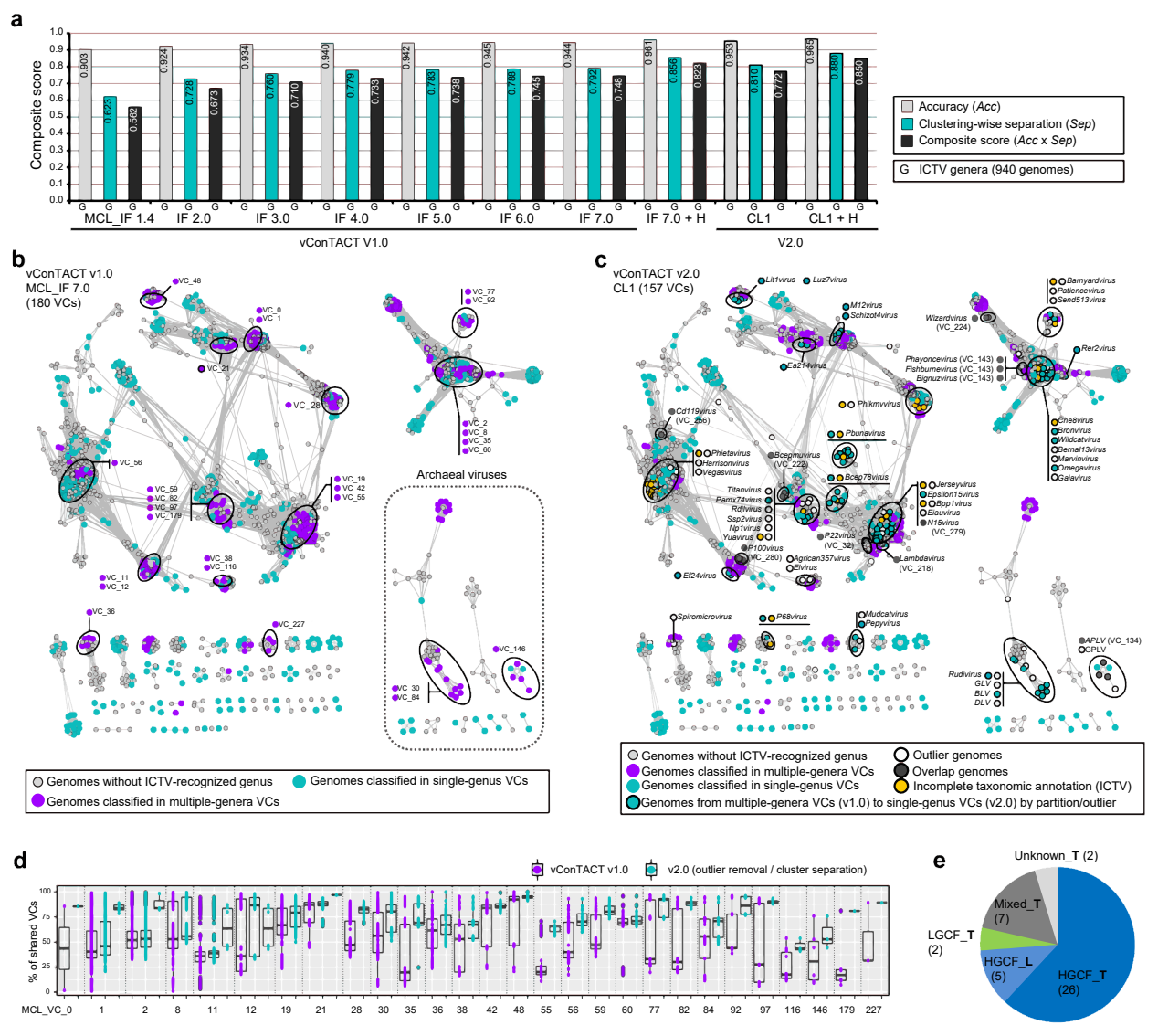
827 59. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status,
828 taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44,** D733–D745
829 (2016).
830 60. Krupovic, M. *et al.* Taxonomy of prokaryotic viruses: update from the ICTV bacterial and
831 archaeal viruses subcommittee. *Arch. Virol.* **161,** 1095–1099 (2016).
832 61. Adams, M. J. *et al.* Changes to taxonomy and the International Code of Virus
833 Classification and Nomenclature ratified by the International Committee on Taxonomy of
834 Viruses (2017). *Arch. Virol.* **162,** 2505–2538 (2017).
835 62. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: A new generation of protein
836 database search programs. *Nucleic Acids Research* **25,** 3389–3402 (1997).
837 63. Wiwie, C., Baumbach, J. & Röttger, R. Comparing the performance of biomedical
838 clustering methods. *Nat. Methods* **12,** 1033–1038 (2015).
839 64. Brohée, S. & van Helden, J. Evaluation of clustering algorithms for protein-protein
840 interaction networks. *BMC Bioinformatics* **7,** (2006).
841 65. Kamburov, A., Stelzl, U. & Herwig, R. IntScore: A web tool for confidence scoring of
842 biological interactions. *Nucleic Acids Res.* **40,** (2012).
843 66. Goldberg, D. S. & Roth, F. P. Assessing experimentally derived interactions in a small
844 world. *Proc. Natl. Acad. Sci.* **100,** 4372–4376 (2003).
845 67. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
846 DIAMOND. *Nat. Methods* **12,** 59–60 (2015).
847 68. Ohio Supercomputer Center . (1987).
848 69. Oliphant, T. E. SciPy: Open source scientific tools for Python. *Comput. Sci. Eng.* **9,** 10–20
849 (2007).
850 70. McKinney, W. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci.*
851 *Conf.* **1697900,** 51–56 (2010).
852 71. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,**
853 2825–2830 (2011).
854 72. Csárdi, G. & Nepusz, T. The igraph software package for complex network research.
855 *InterJournal Complex Syst.* **1695,** 1–9 (2006).
856 73. Federico, P., Pfeffer, J., Aigner, W., Miksch, S. & Zenk, L. Visual Analysis of Dynamic
857 Networks Using Change Centrality. in *2012 IEEE/ACM International Conference on*
858 *Advances in Social Networks Analysis and Mining* 179–183 (2012).
859 doi:10.1109/ASONAM.2012.39

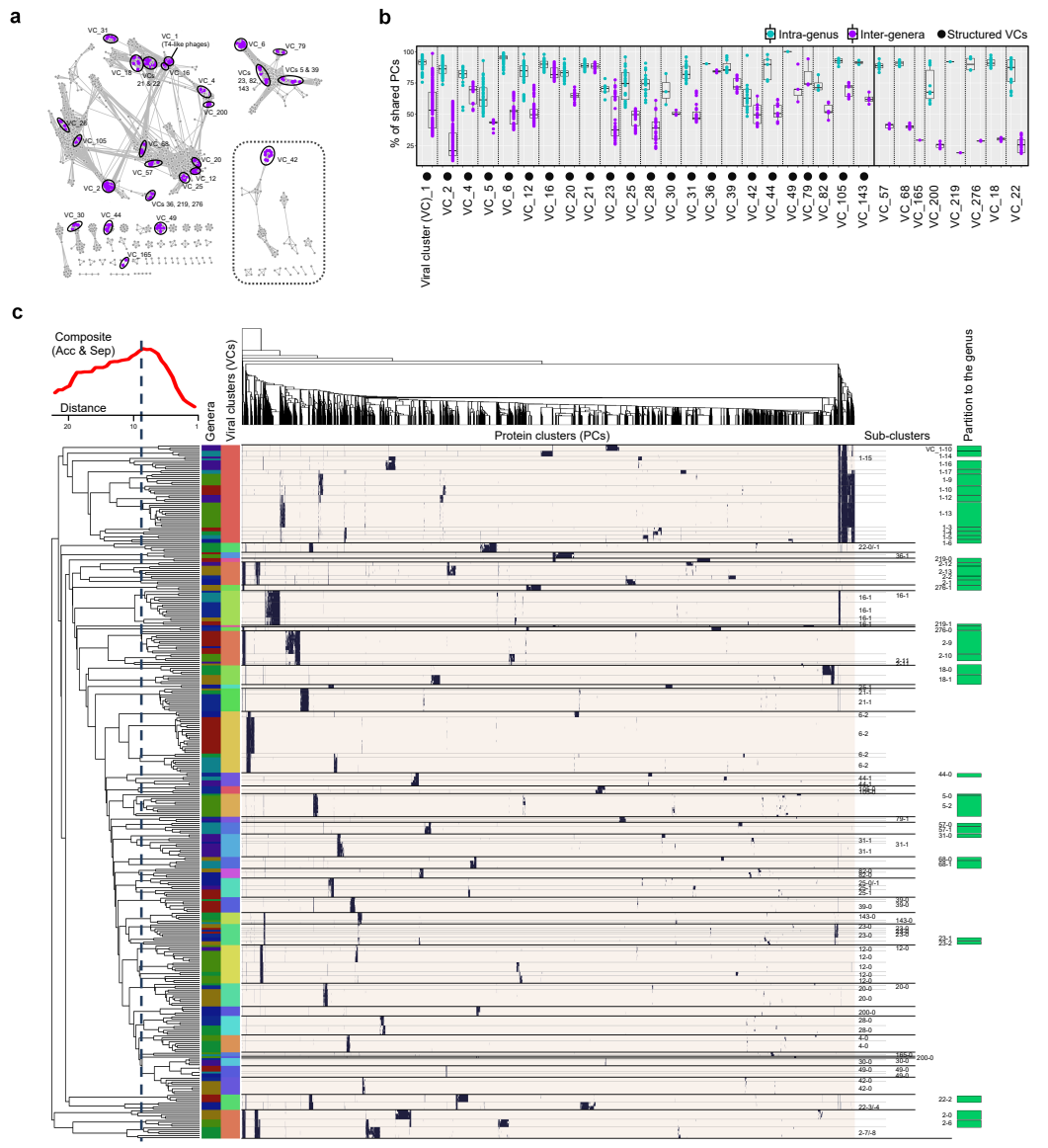
860

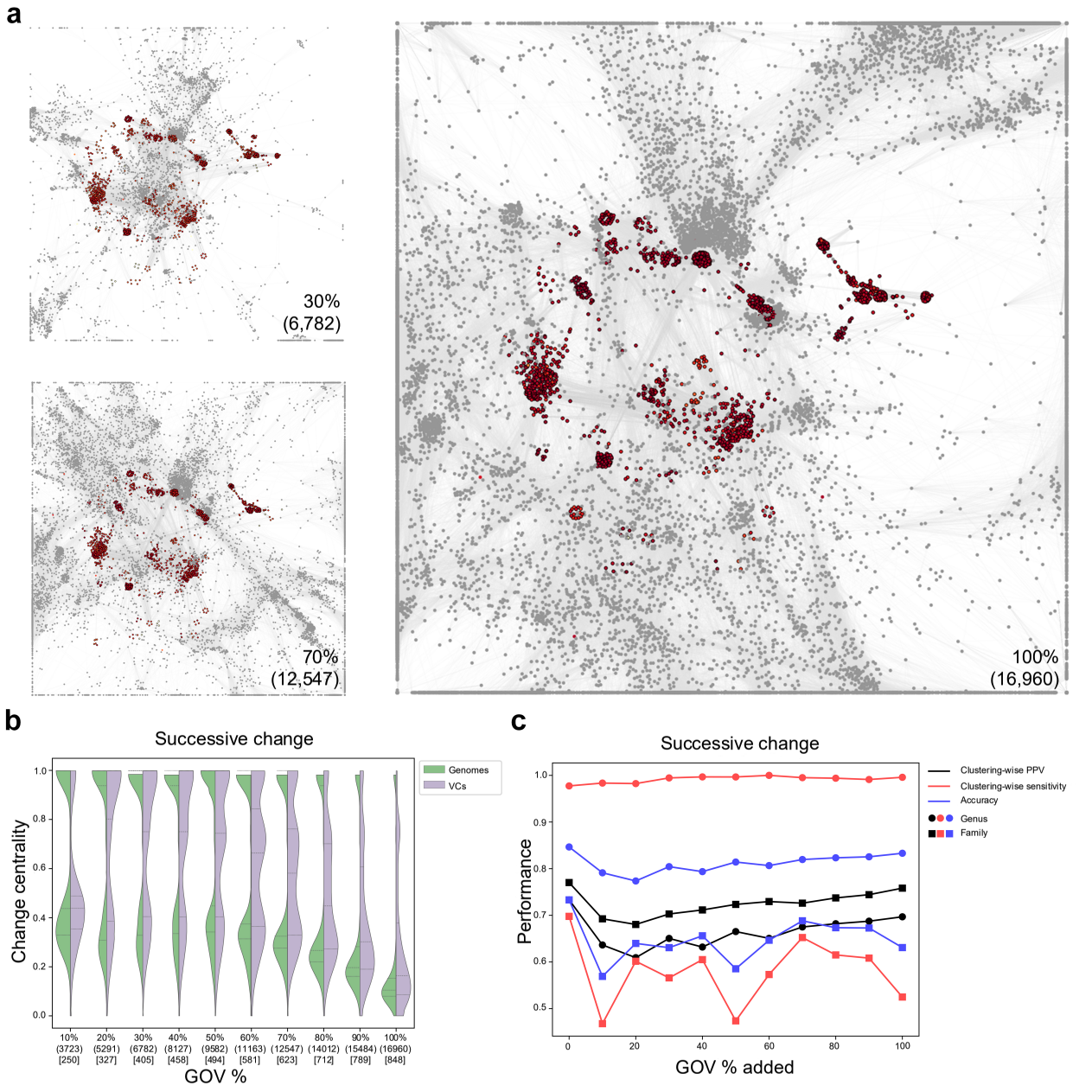
861
862
863
864
865

866
867
868


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 based on the number of shared PCs between genomes
B
C
A
VC_1
E
D
F
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_IF 7.0
(180 VCs)
Archaeal viruses
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 VCs
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_6
VC_79
VC_18
VC_16
VCs 5 & 39
VC_4
VC_200
VC_36
VC_105
VC_68
VC_20
VC_57
VC_12
VC_25
VC_42
VC_7
VCs 36, 219, 276
VC_30
VC_44
VC_49
VC_165
b
Intra-genus
Inter-genera
Structured VCs
% of shared PCs
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
VC_276
VC_18
VC_22
c
Composite
(Acc & Sep)
Distance
Genera
Viral clusters (VCs)
Protein clusters (PCs)
Sub-clusters
Partition to the genus


a
30%
(6,782)
70%
(12,547)
100%
(16,960)
b
Successive change
Change centrality
GOV %
Genomes
VCs
10%
(3723)
[250]
20%
(5291)
[327]
30%
(6782)
[405]
40%
(8127)
[458]
50%
(9582)
[494]
60%
(11163)
[581]
70%
(12547)
[623]
80%
(14012)
[712]
90%
(15484)
[789]
100%
(16960)
[848]
c
Successive change
Performance
GOV % added
Clustering-wise PPV
Clustering-wise sensitivity
Accuracy
Genus
Family