

Lawrence Berkeley National Laboratory

LBL Publications

Title

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

Permalink

<https://escholarship.org/uc/item/18v6h17b>

Journal

Nature Biotechnology, 37(6)

ISSN

1087-0156

Authors

Bin Jang, Ho
Bolduc, Benjamin
Zablocki, Olivier
et al.

Publication Date

2019-06-01

DOI

10.1038/s41587-019-0100-8

Peer reviewed

1 **Gene sharing networks to automate genome-based prokaryotic viral taxonomy**

2 Ho Bin Jang^{1*}, Benjamin Bolduc^{1*}, Olivier Zablocki¹, Jens H. Kuhn², Simon Roux³, Evelien M.
3 Adriaenssens⁴, J. Rodney Brister⁵, Andrew M Kropinski^{6,7}, Mart Krupovic⁸, Rob Lavigne⁹, Dann
4 Turner¹⁰, & Matthew B. Sullivan^{1,11#}

5 ¹ Department of Microbiology, Ohio State University, Columbus, OH, USA

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

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

9 ⁴ Institute of Integrative Biology, University of Liverpool, Liverpool, UK

10 ⁵ National Center for Biotechnology Information, National Library of Medicine, National Institutes of
11 Health, Bethesda, MD, USA

12 ⁶ Department of Pathobiology, Ontario Veterinary College, University of Guelph, Guelph, ON, Canada
13 N1G 2W1

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

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

17 ⁹ Laboratory of Gene Technology, Department of Biosystems, Faculty of BioScience Engineering, KU
18 Leuven, Leuven, Belgium

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

21

22 ¹¹ Department of Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH,
23 USA

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

25 * These authors contributed equally to this work.

26

27 **ABSTRACT**

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

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

41 **Main text**

42 Bacteria and archaea modulate the nutrient and energy cycles that drive ocean and soil ecosystems¹⁻⁴,
43 and impact humans by producing metabolites that affect health, behavior, and susceptibility to disease⁵.
44 Viruses that infect these microbes modulate these ‘ecosystem roles’ via killing, metabolic reprogramming
45 and gene transfer^{6,7}, with substantial impacts predicted in the ocean⁸⁻¹⁰, soil^{11,12} and human
46 microbiomes^{13,14}. However, ecosystem-scale understanding is hampered by the lack of universal genes or
47 methods that could facilitate a formalized taxonomy and comparative surveys. In fact, viruses do not have
48 a single, universal marker gene¹⁵, and, thus, no analog to the microbial 16S rRNA-based phylogenies and
49 operational taxonomic units (OTUs) are possible¹⁶.

50 Another potential challenge is that many viruses exhibit ‘rampant genome mosaicism’¹⁷, due to high
51 rates of horizontal gene transfer between viruses and their hosts. If broadly true, this would hamper the
52 creation of a genome-based prokaryotic virus taxonomy¹⁸. Fortunately, explorations of viral sequence
53 space are revealing structure^{19,20} and population genetic support for a biological species definition, driven
54 primarily by vertical evolution²¹, and new hypotheses to explain variable evolution among prokaryotic
55 viruses²². Such findings, alongside rapidly expanding viral genome databases, allowed the International
56 Committee on Taxonomy of Viruses (ICTV) to present a consensus statement suggesting a shift from the
57 ‘traditional’ classification criteria²³ (e.g. virion morphology, single/multiple gene phylogenies) towards a
58 genome-centered, and perhaps one-day, largely automated, viral taxonomy²⁴. This shift is particularly

59 critical given the modern pace of viral discovery in which, as of November 2018, hundreds of thousands
60 of metagenome-derived viral genomes and large genome fragments (735,112 at IMG/VR²⁵, with a total
61 of 110,384 ‘species’ or viral OTUs) now dwarf the 34,091 available from prokaryotic virus sequences in
62 the NCBI GenBank database²⁶. Together with the recently proposed ‘minimum information about
63 uncultivated virus genomes’ (MIUViGs) community guidelines²⁷, evaluation of approaches to establish a
64 scalable, genome-based viral taxonomy is needed to enable a universal classification framework. The
65 ability to classify thousands of microbial UViGs is invaluable for foundational hypothesis testing
66 anywhere microbes, and their viruses, might matter. With microbiome and phage therapy research
67 exploding, this includes ecosystems such as industrial bioreactors and the human gut and lung, as well as
68 the oceans, the deep-sea floor, and soils.

69 Multiple genome-based strategies have been proposed to develop such a unified taxonomic
70 framework for viruses infecting bacterial^{15,28–33}, archaeal³⁴ and eukaryotic³⁵ hosts. For bacterial viruses
71 (“phages”), an early approach to a universal taxonomy targeted phage relationships only by using
72 complete genome pairwise protein sequence comparisons in a phylogenetic framework (the “phage
73 proteomic tree”) and was broadly concordant with ICTV-endorsed virus groupings of the time¹⁵. Such
74 efforts were not widely adopted, presumably because (i) the demand was limited (few metagenomics
75 studies existed), and (ii) the paradigm was that “rampant mosaicism” would blur taxonomic boundaries
76 and violate the assumptions of the underlying phylogenetic algorithms used in the analyses¹⁷. Other
77 efforts sought to establish percent of genes shared and percent identity of shared genes cut-offs to define
78 genera and sub-family affiliations^{36,37}, but lacked taxonomic resolution for several virus groups. This lack
79 of resolution was due to the likelihood that the mode and tempo of prokaryotic virus evolution could vary
80 significantly across the viral sequence space²². Building upon a prokaryotic classification algorithm, the
81 Genome Blast Distance Phylogeny (GBDP)³⁸, a freely accessible online tool (VICTOR) now provides for
82 classification of phage genomes via combined phylogenetic and clustering methods from nucleotide and

83 protein sequences³¹. Although a key advance, this method suffers from limited scalability (100 genomes
84 limit) and taxonomic assignment challenges for viruses that lack closely related reference genomes.

85 Alternatively, several groups reasoned that the highly variable evolutionary rates across phage
86 sequence space could be examined through gene sharing networks^{29,30,39} to determine whether a
87 meaningful structure, and therefore taxonomic signal, occurs in this space. These networks, based on
88 shared protein clusters (PCs) between viral genomes, were largely concordant with ICTV-endorsed taxa
89 independent of whether monopartite²⁹ (a single node type, i.e., viral genomes) or bipartite networks^{34,39}
90 (two node types, i.e., viral genomes and genes) were used. Given these successes, we revisited the
91 monopartite gene sharing network approach to establish an iVirus⁴⁰ app (vConTACT v1.0, hereafter v1.0)
92 to automate a network-based classification pipeline for prokaryotic virus genomes. Performance tests
93 indicated that the network analytics used by v1.0 produced viral clusters (‘VCs’) that were ~75%
94 concordant with recently established ICTV prokaryotic viral genera, even with seven times more genomes
95 now available³⁰. The capacity to incorporate these genomes and accuracy of the network-based analytics
96 have resulted in viral taxonomy applications across large-scale studies of ocean^{41,42}, freshwater⁴³ and
97 soil⁴⁴, and studies of single-virus amplified genomes (vSAGs)^{45,46} – all environments where the viruses
98 observed were largely so novel as to not be classifiable outside of these gene-sharing network approaches.
99 These v1.0 advances were an important step forward, but they could not be used for automated tentative
100 taxonomic assignments. This is because v1.0 created artifactual VCs of both undersampled genomes and
101 highly overlapped regions of viral sequence space³⁰, and also lacked several key, community-desired
102 features including per-VC confidence metrics, a metric for establishing hierarchical taxonomy, and
103 scalability.

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

108

109 RESULTS AND DISCUSSION

110

111 *vConTACT v2.0 key features and updates*

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

123 In a network-based genome taxonomy framework (**Fig. 1a**), related genomes emerge as a group of
124 nodes strongly connected through multiple edges, here termed a Viral Cluster, or ‘VC’. In a taxonomic
125 context and based on the clustering of viral reference genomes, we have previously demonstrated that the
126 network parameters can be tuned such that the VCs best represent genus-level grouping of viral
127 genomes³⁰. In the original v1.0, ~75% of VCs corresponded to established ICTV genera³⁰ (‘concordant
128 VCs’), while the remaining mismatched VCs were termed ‘discordant VCs’. These mismatches in the
129 discordant VCs can occur from any of three scenarios (**Fig. 1b**): (i) VCs that wrongly clustered genomes
130 with no close relatives (termed ‘outlier genomes’ from ‘undersampled VCs’), (ii) a given VC that
131 contained multiple ICTV genera represented by virus genomes that shared many genes and/or modules
132 with other VCs (termed ‘overlapping VCs’), and (iii) VCs that encompassed multiple ICTV genera

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

137 To address these issues in v2.0, we evaluated and ultimately implemented a new clustering algorithm,
138 established confidence scores and distance-based taxon separation for hierarchical taxonomy, and
139 optimized and evaluated scalability and robustness to a large-scale viral metagenomic dataset. Briefly,
140 after the MCL-clustered protein clusters are generated, we optimized the protein-cluster-based gene-
141 sharing information to establish an automated two-step process whereby VCs are defined using
142 ClusterONE⁵⁰ (CL1), instead of MCL used in v1.0, and then subdivided using hierarchical clustering to
143 disentangle problematic regions of the networks (**Fig. 1b**, see Online Methods). This approach considers
144 edge weight (i.e., degree of connection between genomes) to identify outlier genomes that are weakly
145 connected with members of their VC compared to neighbour genomes, detect and separate genomes that
146 ‘bridge’ overlapping VCs, and break down structured VCs into concordant VCs through distance-based
147 hierarchical clustering (**Fig. 1b**).

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

156 Finally, although v2.0 is presented here as a monopartite (one type of node) network tool, it produces
157 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,

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

194

195 *Performance assessment of vConTACT v2.0 for complex genomic relationships*

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

208 Second, we evaluated the ability of v2.0 to handle overlapping VCs (**Fig. 1b**), which are those VCs
209 that share a greater fraction of genes with members of other VCs than typically expected, presumably due
210 to gene exchange that could erode structure of the network. While overlapping VCs cannot be
211 automatically identified in v1.0, we can detect them in v2.0 through a ‘match coefficient’ that measures
212 the connection within- and between- other VCs (see Online Methods). Sensitivity analyses established a
213 cluster overlap value of 0.8 as diagnostic (Online Methods). This approach identified nine overlapping
214 VCs (ICTV-classified genera only) containing 30 viruses across 11 ICTV genera. These included viruses
215 with known mosaic genomes⁴⁸ (e.g., lambdoid or mu-like phages of the *P22virus*, *Lambdavirus*,
216 *N15virus*, and *Bcepmyovirus* genera), recombinogenic temperate phages^{51,52} (i.e., *Mycobacterium* phages
217 of the *Bignuzvirus*, *Phayoncevirus*, and *Fishburnevirus* genera and *Gordonia* phages of the genus
218 *Wizardvirus*), and three newly-established genera (i.e., *Cd119virus*, *P100virus* and archaeal
219 *Alphapleolipovirus*), all bearing low topology-based confidence scores (averages of 0.32 for these VCs
220 versus 0.52 for concordant VCs; P-value = 6.12e-09, Mann-Whitney U test) (**Supplementary Fig. 3a**).
221 Overlapping VCs are also linked to high horizontal gene flow, since most viruses in these VCs were
222 classified as having high gene content variation (HGCF, **Fig. 2e**, **Supplementary Fig. 3b**) as assigned by
223 a recently proposed framework of phage evolutionary lifestyles²². Though unresolvable in v1.0, v2.0
224 could assign eight of the 11 ICTV genera (24 viruses) into 8 ICTV-concordant VCs (**Supplementary**
225 **Table 2**). The remaining 3 ICTV genera, all comprised of *Mycobacterium* phages⁵³ (6 genomes), could
226 not be resolved (**Supplementary Table 2**), and may not be amenable to automated taxonomy. While
227 these highly recombinogenic genomes represent a minor portion of the viruses classified by ICTV, their
228 prevalence across environments remains to be determined.

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

233 through hierarchical clustering²⁷, in v2.0, we formalized an optimized, quantitative hierarchical
234 decomposition distance measure for this process (Online Methods and **Supplementary Fig. 4**). In the
235 v2.0 network, 23 of the 31 discordant VCs (74%) were structured VCs, spanning 86 genera (**Fig. 3a,b** and
236 **Supplementary Table 2**). The automatic v2.0 approach resolved 30% (26 of 86) of these ICTV genera
237 from 6 of the 23 structured VCs (**Fig. 3c**).

238 Given such strong performance, even with challenging regions of viral sequence space, we suggest
239 that this gene sharing network already offers significant new taxonomic insights. As described earlier,
240 only 41% of the 2,304 reference virus genomes are classified by ICTV at the genus rank, which leaves
241 1,364 reference viruses that are currently not assigned to a genus. In our networks, these 1,364 reference
242 viruses organized into 404 well-supported VCs (**Supplementary Table 2**). Of these 1364, 544 were
243 assorted into 104 VCs with genomes from known ICTV taxa, whereas 820 formed 200 separate VCs. We
244 propose that the 820 can be classified and the 200 VCs represent *bona fide* novel virus genera, to be
245 evaluated as formal ICTV taxa. If adopted, this would immediately double the number of known
246 prokaryotic viral genera, as there are only 264 currently recognized. Beyond providing a starting point for
247 identifying areas of viral sequence space needing revision, the manual curation process itself would
248 provide feedback for improving v2.0.

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

258 result that years of manual curation by experts has resulted in specialized demarcation cut-offs across
259 viral sequence space⁵⁴. However, just has recently been advanced for microbes⁵⁵ any steps towards this,
260 such as distances in well-sampled and well-resolved portions of the reference networks, will be invaluable
261 for automating virus taxonomy. The v2.0 VCs and subclusters provided here provide a case study and a
262 reference baseline for working with the ICTV to translate such network-derived cut-offs into systematic
263 taxonomic demarcation criteria.

264 Finally, there will be some taxon assignments that are not amenable to being resolved by gene-
265 sharing networks. For example, cases where genera are defined based on phenotypic or evolutionary
266 evidence, e.g., archaeal fuselloviruses⁵⁶ (VC42) and bacterial microviruses⁵⁷ (VCs 30 and 49), a gene-
267 sharing network approach will not be appropriate (see **Fig. 3c** and **Supplementary Table 2**). However, an
268 automated vConTACT-based approach can help systematically identify such problematic taxa and
269 drastically speed up these critical revisions to our taxonomy as new data become available.

270

271 *vConTACT v2.0 is scalable to modern virome datasets*

272 Beyond accurate classification, a major bottleneck regarding automated taxonomic assignments is the
273 ability to scalably and robustly integrate large sets of newly discovered virus genomes. To evaluate this,
274 we added 15,280 curated viral genomes and large genome fragments (≥ 10 kb) from the Global Ocean
275 Virome (GOV) dataset⁴¹ to our reference network in successive 10% increments (i.e., 0%, 10%, ..., 100%
276 of the total dataset), totalling 16,960 sequences (**Fig. 4a**). Through this process, we evaluated 2 types of
277 network changes to assess robustness. First, we checked whether the incremental addition of GOV data to
278 the network would lead to changes in node connections, as estimated by the ‘change centrality’ metrics
279 (CC, values range from 0-1 with 0 indicating no change and 1 indicating complete change, see **Fig. 4b**).
280 Second, we evaluated the concordance between v2.0 clustering and ICTV genera using the same
281 performance metrics as above (*Sn*, *Acc* & *PPV*, see **Fig. 4c**). As seen in Fig. 4b, a large fraction of the
282 data initially experiences a moderate change (CC = 0.4), but the whole dataset eventually stabilized, as

283 CC values for most of the data ranged from 0 to 0.1. A similar trend was observed for accuracy (*Acc*, **Fig.**
284 **4c**). This indicated that, not only can v2.0 scalably handle thousands of new input sequences, but our
285 original reference network clustering is robust to large-scale data additions such that the underlying
286 reference network remains consistent with established ICTV genera.

287 Outside of better sampled genera or VCs, we also examined whether GOV data may partially resolve
288 ICTV outlier and singleton genomes as a proxy for assessing the taxonomic ramifications of so much new
289 data. We reasoned that more data might create new connections to singletons such that outliers may
290 become better connected to new or existing VCs. We found that, of 38 single-member VCs of singleton
291 and outlier genomes (**Supplementary Fig. 6**), three *Mycobacterium* phage VCs were improved, while
292 two other *Mycobacterium* virus genomes were merged into larger heterogeneous VCs now composed of
293 six ICTV genera, which did not constitute an improvement. This implies that even though the GOV data
294 derive from an environment very different to that of most of the reference taxa (oceans vs soils and
295 humans), these environmental data could help improve some taxon assignments of isolate genomes.
296 Separately, looking at the overall taxonomic impact, we observed that 919 new VCs were created with the
297 full GOV dataset (15, 280 total contigs). Given the strong concordance of the ‘known’ VCs to ICTV
298 genera, we posit that these new VCs represent 919 new viral genera that are not represented among the
299 264 ICTV genera already known from RefSeq genomes. If true, this demonstrates the power of
300 integrating viral sequence datasets from under-explored environments to better map viral sequence space.
301 While enticing, we agree with the recent consensus statement that any taxonomic reference network be
302 constrained to complete genomes²⁴, and that large genome fragments commonly derived from
303 metagenome-based studies be utilized in a relevant manner to address questions specific to that study.

304

305 *Community availability and future needs*

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

308 optimized tool available in two ways. First, the source code is available through Bitbucket
309 (<https://bitbucket.org/MAVERICLab/vcontact2>) as a downloadable python package. V2.0 is a highly
310 scalable tool, with memory and CPU requirements concomitant with the amount of sequences processed.
311 The largest memory requirements are by the all-versus-all protein comparisons used to build the protein
312 clusters. Overall, there is a strong linear ($R^2 = 0.99$, see **Supplementary Fig. 7**) correlation between
313 number of sequences and runtimes. For example, running the full virus dataset RefSeq with Diamond (for
314 the protein comparisons) would take ~10 minutes on a regular laptop, while a GOV-sized dataset would
315 run for several hours. Second, for ease of use, v2.0 is also available as an app through iVirus⁴⁰, the viral
316 ecology apps and data resource embedded in the CyVerse Cyberinfrastructure, with detailed usage
317 protocols available through Protocol Exchange (<https://www.nature.com/protocolexchange/>) and
318 protocols.io (<https://www.protocols.io/>). Finally, the curated reference network is available at each of
319 these sites, and will be updated approximately bi-yearly with as complete genomes become available and
320 resources exist to support this effort.

321 Although v2.0 performance metrics are strong and provide a critically needed, systematic reference
322 viral taxonomic network, limitations remain. First, the complete reference network needs to be rebuilt
323 each time new data are added. Avoiding this reconstruction step will require the development of
324 approximation methods and/or a placement algorithm (akin to PPlacer for 16S phylogenies⁵⁸) to
325 incorporate new data. Second, CL1-based VC generation may require manual parameter optimization if
326 datasets of an extreme, and probably unlikely, nature are encountered. Such datasets would have to be
327 dominated by overlapping genomes (i.e., those that share a large proportion of genes), and though such
328 genomes exist (e.g., the HGCF *Mycobacterium* phages), they are currently rare among cultured isolates
329 and their frequency in nature remains unknown. Notably, at least for the oceans, such highly
330 recombinogenic viruses are likely uncommon as the incremental addition of GOV data resulted in a stable
331 network and separate analyses of a more recent version of this GOV dataset found ocean viral populations
332 to be quite structured⁵⁹. However, in case such datasets dominated by overlapping genomes are

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

335 Third, solving other future needs will require improved understanding of the broader viral sequence
336 landscape and evolutionary processes. For example, although v2.0 handles reference prokaryotic virus
337 genomes (including ssDNA or dsDNA phages) and large GOV genome fragments, this framework has
338 not been designed, tested or validated for eukaryotic viruses. These viruses will require new solutions as
339 they have broader genomic configurations, such as genome segmentation, overlapping genes, ambisense
340 transcriptional gene configurations, that pose unique computational challenges^{35,60}. However, even these
341 viruses can be classified using genome-based metrics (e.g., hidden Markov models protein profiles and
342 genomic organization models) and network analytics, at least at the family-level³⁵. Separately, shorter
343 complete prokaryotic virus genomes and small fragments of larger genomes (e.g., ≤ 3 PCs or ≤ 5 genes)
344 are of low statistical power in gene-sharing networks, and will require new solutions to establish higher
345 confidence VCs or remain taxonomically inaccessible via these approaches. Finally, genomes identified
346 as singletons, outliers or overlapping are currently excluded from the gene-sharing network, which leaves
347 a large fraction of viral sequence space unclassified. Although singletons and outliers can be resolved by
348 the addition of new data, overlapping VCs can remain challenging to resolve, particularly for the HGCF
349 phages²² that are highly recombinogenic. Such mosaic virus genomes are challenging for viral taxonomy.
350 However, they are identifiable in the networks and, at least to date, represent a small fraction of known
351 viral sequence space. Given increased gene flow among temperate phages²², it will be valuable to explore
352 viral sequence space in environments where temperate phages are thought to predominate (e.g., soils⁶¹,
353 human gut⁶²).

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

358 research community and the ICTV will only make these gene-sharing network approaches more scalable
359 and broadly applicable. Assuming broad acceptance of these efforts, and parallel efforts with eukaryotic
360 viruses³⁵, we might finally have the foundation needed to realize the consensus statement goals^{24,27} of
361 establishing a genome-based viral taxonomy to better capture the broader viral sequence landscape
362 emerging from environmental surveys.

363

364 **METHODS**

365

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

377

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

383 versus-all BLASTP⁶⁶ searches (default parameters, cut-offs of 1E⁻⁵ on e-value and 50 on bit score). Third,
384 PCs were generated by applying MCL (inflation factor of 2.0), and resulted into all the proteins being
385 organized into 25,513 PCs, with a fraction of proteins (26,625 or 11.5%) as singletons (i.e. isolated
386 protein with no relatives).

387

388 **Calculating genome similarity between viruses.** The resulting output was parsed in the form of a matrix
389 comprised of genomes, PCs and singleton proteins (i.e., 2,304 × 52,138 matrix) (**Supplementary Table**
390 **1**). We then determined the similarities between genomes by calculating a one-tailed *P* value of observing
391 at least *c* PCs in common between each pair of genomes, based on the following hypergeometric equation
392 as per Lima-Mendez et al²⁹:

393

$$394 \quad P(X \geq c) = \sum_{i=c}^{\min(a,b)} \frac{c_a^i c_n^{b-i}}{c_n^b} \quad (1)$$

395

396 in which *c* is the number of PCs in common; *a* and *b* are the numbers of PCs and singletons in genomes A
397 and B, respectively; and *n* is the total number of PCs and singletons in the dataset. The hypergeometric
398 formula calculates the probability of sharing a number of common PCs between two genomes at or above
399 the number (*c*) under the null hypothesis that the observed result is likely to occur by chance. A score of
400 similarity between genomes was obtained by taking the negative logarithm (base 10) of the
401 hypergeometric P-value multiplied by the total number of pairwise genome comparisons (i.e., (2,304 ×
402 2,303)/2). Genome pairs with a similarity score ≥1 were previously shown to be significantly similar
403 through permutation test where PCs and singleton proteins with genome pairs having a similarity score
404 below the given threshold (negative control) were randomly rearranged. None of the genome pairs in this
405 negative control produced similarity score >1, indicating values above this threshold did not occur by
406 chance³⁰.

407

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

412
413 **Parameter optimization for viral cluster formation of vConTACT v1.0 and 2.0.** Due to different
414 criteria for parameter optimization between the clustering methods, different number and size of the
415 clusters are often generated, which can make objective performance comparisons difficult⁶⁷. Thus, to
416 more comprehensively compare performance, v1.0's MCL-based VCs were generated at inflation factors
417 (IFs) of 2.0 to 7.0 by 1.0 increments, with an optimal IF of 1.4 showing the highest intra-cluster clustering
418 coefficient (ICCC)²⁹ (**Supplementary Table 2** and **Supplementary Fig. 8**). Unlike MCL, which uses a
419 single parameter²⁹ (i.e., the inflation factor), VC formation with CL1 (used in vConTACT v2.0), involves
420 multiple parameters that can detect complex network relationships⁵⁰. The three main parameters of CL1,
421 minimum density/node penalty, haircut, and overlap, automatically quantify (i) the cohesiveness of a
422 cluster, (ii) the boundaries of the clusters (i.e. outlier genomes), and (iii) the size of overlap between
423 clusters, respectively⁵⁰. Of these parameters, the first one is used to detect the coherent groups of VCs as
424 follows:

425
426
$$C = \frac{W_{in}(V)}{W_{in}(V) + W_{out}(V) + p|C|} \quad (2)$$

427
428 in which $W_{in}(V)$ and $W_{out}(V)$ are the total weight of edges that lie within cluster V and that connect the
429 cluster V and the rest of the network, respectively, $|C|$ is the size of the cluster, p is a penalty that counts
430 the possibility of uncharted connections for each node.

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

433

434

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

435

436 in which k is the number of edges of the node g , and W is the total weight of edges of the respective
437 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
438 that we specified the average weight of nodes (g) within the given cluster, CL1 will remove the node (h)
439 from a given VC and consider it an outlier.

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

442

443

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

444

445 in which i is the size of overlap, which is divided by the product of the sizes of the two clusters under
446 consideration (a and b). Since CL1 identifies overlap between VCs, it can find both hierarchical and
447 overlapping structures within viral groups. This ability is a significant improvement over v1.0, as v1.0's
448 MCL cannot handle modules with overlaps⁷. Specifically, for each pair of clusters, CL1 calculates the
449 overlap score between them (above) and merges these clusters if the overlap is larger than a given
450 threshold. Thus, in the resulting output file, viral groups (or clusters) having the identical member viruses
451 can be found in multiple clusters, called 'overlapping viral clusters' (**Supplementary Table 2 and Fig.**
452 **1b, middle row**).

453 To determine the best parameter combination to use for CL1, we tested a wide range of values for
454 the three aforementioned parameters: minimum density ranging from 0 to 1 by 0.1 increments; node
455 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
456 for the other parameters: 2 as minimum cluster size, weighted as edge weight, single-pass as merging,
457 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 described⁶⁸. 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 CL1⁶⁸ (**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 .

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

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:

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

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:

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

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:

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

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

$$509 \quad PPV_{cl_j} = \max_{i=1}^n PPV_{i,j} \quad (9)$$

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

$$511 \quad PPV = \frac{\sum_{j=1}^m T_j PPV_{cl_j}}{\sum_{j=1}^m T_j} \quad (10)$$

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

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

$$515 \quad 516 \quad Acc = \sqrt{Sn \times PPV} \quad (11)$$

517
 518 **Complex- and Cluster-wise separations:** With the same contingency table used for Sn , PPV , and Acc , we
 519 calculated the relative frequencies with respect to the marginal sums for each row ($F_{row_{i,j}}$) and each
 520 column ($F_{col_{i,j}}$), respectively:

$$521 \quad F_{row_{i,j}} = T_{i,j} / \sum_{j=1}^m T_{i,j} \quad (12)$$

$$522 \quad F_{col_{i,j}} = T_{i,j} / \sum_{i=1}^n T_{i,j} \quad (13)$$

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

$$524 \quad Sep_{i,j} = F_{col_{i,j}} \times F_{row_{i,j}} \quad (14)$$

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

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

530

$$531 \quad Sep_{co} = \frac{\sum_{i=1}^n Sep_{co_i}}{n} \quad (15)$$

532

$$533 \quad Sep_{cl} = \frac{\sum_{j=1}^m Sep_{cl_j}}{m} \quad (16)$$

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

536

$$537 \quad Sep = \sqrt{Sep_{co} \times Sep_{cl}} \quad (17)$$

538 High clustering-wise separation values indicate a bidirectional correspondence between a sub-/cluster and
539 each ICTV taxon: a score of 1.0 indicates that a cluster corresponds perfectly to each taxon. For overall
540 comparison, we used a composite score⁵⁰, calculated by multiplying Acc by Sep .

541 As an internal measure, the fraction of PCs³⁰ between two genomes (i.e., proteome similarity) was
542 computed by using the geometric index (G). The proteome similarity was estimated as:

543

$$544 \quad G_{AB} = \frac{|N(A) \cap N(B)|}{|N(A)| \times |N(B)|} \quad (18)$$

545

546 in which $N(A)$ and $N(B)$ indicate the number of PCs in the genomes of A and B, respectively. A total of
547 400,234 pairs of genomes with >1% proteome similarity are shown in **Supplementary Table 4**.

548

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

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

554

$$555 \quad Q_c = W_{in}/(W_{in} + W_{out}) \quad (19)$$

556

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

566

$$567 \quad prediction = \sum l_{i,j} / l_c \quad (20)$$

568

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

575

576 **Measuring effect of GOV on network structural changes.** GOV contigs (14,656 sequences) were
577 added in 10% increments (randomly selected at each iteration) to NCBI Viral RefSeq and processed using
578 vConTACT v2.0 with one difference – Diamond⁷¹ instead of BLASTp was used to construct the all-
579 versus-all protein comparison underlying the PC generation. For running this large number of sequences,
580 high-memory computer nodes from the Ohio State supercomputer Center⁷² were used. Once generated,
581 vConTACT v2.0 networks were post-processed using a combination of the Scipy⁷³, Numpy, Pandas⁷⁴ and
582 Scikit-learn⁷⁵ python 3.6 packages. Networks were rendered using iGraph⁷⁶. The method to calculate
583 change centrality was calculated as described previously⁷⁷. CCs were calculated in a successive way, in
584 which each addition was compared to Viral RefSeq 85 independently of other additions (0% versus 10%,
585 0% vs 20%, [...], 0% vs 100%).

586

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

589

590 REFERENCES

- 591 1. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive earth's
592 biogeochemical cycles. *Science* **320**, 1034–1039 (2008).
- 593 2. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* (80-.). **348**,
594 (2015).
- 595 3. Moran, M. A. The global ocean microbiome. *Science* **350**, (2015).
- 596 4. Zhao, M. *et al.* Microbial mediation of biogeochemical cycles revealed by simulation of global
597 changes with soil transplant and cropping. *ISME J.* **8**, 2045–2055 (2014).
- 598 5. Cho, I. & Blaser, M. J. The human microbiome: At the interface of health and disease. *Nature*
599 *Reviews Genetics* **13**, 260–270 (2012).
- 600 6. Fernández, L., Rodríguez, A. & García, P. Phage or foe: an insight into the impact of viral
601 predation on microbial communities. *ISME Journal* 1–9 (2018). doi:10.1038/s41396-018-0049-5
- 602 7. Hurwitz, B. L. & U'Ren, J. M. Viral metabolic reprogramming in marine ecosystems. *Current*
603 *Opinion in Microbiology* **31**, 161–168 (2016).
- 604 8. Suttle, C. a. Marine viruses-major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–
605 812 (2007).
- 606 9. Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* (80-.). **348**,

- 607 (2015).
- 608 10. Danovaro, R. *et al.* Virus-mediated archaeal hecatomb in the deep seafloor. *Sci. Adv.* **2**, (2016).
- 609 11. Pratama, A. A. & van Elsas, J. D. The ‘Neglected’ Soil Virome - Potential Role and Impact.
610 *Trends in Microbiology* (2018). doi:10.1016/j.tim.2017.12.004
- 611 12. Gómez, P. & Buckling, A. Bacteria-phage antagonistic coevolution in soil. *Science (80-.)*. **332**,
612 106–109 (2011).
- 613 13. Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F. & Gordon, J. I. Going viral: Next-
614 generation sequencing applied to phage populations in the human gut. *Nature Reviews*
615 *Microbiology* **10**, 607–617 (2012).
- 616 14. Abeles, S. R. & Pride, D. T. Molecular bases and role of viruses in the human microbiome.
617 *Journal of Molecular Biology* **426**, 3892–3906 (2014).
- 618 15. Rohwer, F. & Edwards, R. The phage proteomic tree: A genome-based taxonomy for phage. *J.*
619 *Bacteriol.* **184**, 4529–4535 (2002).
- 620 16. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using
621 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635–645 (2014).
- 622 17. Lawrence, J. G., Hatfull, G. F. & Hendrix, R. W. Imbroglios of viral taxonomy: Genetic exchange
623 and failings of phenetic approaches. *J. Bacteriol.* **184**, 4891–4905 (2002).
- 624 18. Sullivan, M. B. Viromes, Not Gene Markers, for Studying Double-Stranded DNA Virus
625 Communities. *J. Virol.* **89**, 2459–2461 (2015).
- 626 19. Deng, L. *et al.* Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence
627 space. *Nature* **513**, 242–245 (2014).
- 628 20. Gregory, A. C. *et al.* Genomic differentiation among wild cyanophages despite widespread
629 horizontal gene transfer. *BMC Genomics* **17**, (2016).
- 630 21. Bobay, L. & Ochman, H. Biological species in the viral world. **115**, (2018).
- 631 22. Mavrich, T. N. & Hatfull, G. F. Bacteriophage evolution differs by host, lifestyle and genome.
632 *Nat. Microbiol.* **2**, (2017).
- 633 23. Ackermann, H.-W. Phage Classification and Characterization BT - Bacteriophages: Methods and
634 Protocols, Volume 1: Isolation, Characterization, and Interactions. in (eds. Clokie, M. R. J. &
635 Kropinski, A. M.) 127–140 (Humana Press, 2009). doi:10.1007/978-1-60327-164-6_13
- 636 24. Simmonds, P. *et al.* Consensus statement: Virus taxonomy in the age of metagenomics. *Nat. Rev.*
637 *Microbiol.* **15**, 161–168 (2017).
- 638 25. Paez-Espino, D. *et al.* IMG/VR v.2.0: an integrated data management and analysis system for
639 cultivated and environmental viral genomes. *Nucleic Acids Res.* gky1127-gky1127 (2018).
- 640 26. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral Genomes resource. *Nucleic*
641 *Acids Res.* **43**, D571–D577 (2015).
- 642 27. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG): a
643 community consensus on standards and best practices for describing genome sequences from
644 uncultivated viruses. *Nat. Biotechnol.* (2018).

- 645 28. Nishimura, Y. *et al.* ViPTree: The viral proteomic tree server. *Bioinformatics* **33**, 2379–2380
646 (2017).
- 647 29. Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. Reticulate representation of
648 evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* **25**, 762–777
649 (2008).
- 650 30. Bolduc, B. *et al.* vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect
651 *Archaea* and *Bacteria*. *PeerJ* **5**, e3243 (2017).
- 652 31. Meier-Kolthoff, J. P. & Göker, M. VICTOR: genome-based phylogeny and classification of
653 prokaryotic viruses. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx440
- 654 32. Yu, C. *et al.* Real Time Classification of Viruses in 12 Dimensions. *PLoS One* **8**, (2013).
- 655 33. Gao, Y. & Luo, L. Genome-based phylogeny of dsDNA viruses by a novel alignment-free method.
656 *Gene* **492**, 309–314 (2012).
- 657 34. Iranzo, J., Koonin, E. V., Prangishvili, D. & Krupovic, M. Bipartite Network Analysis of the
658 Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile
659 Elements. *J. Virol.* **90**, 11043–11055 (2016).
- 660 35. Aiewsakun, P. & Simmonds, P. The genomic underpinnings of eukaryotic virus taxonomy:
661 creating a sequence-based framework for family-level virus classification. *Microbiome* **6**, 38
662 (2018).
- 663 36. Lavigne, R. *et al.* Classification of myoviridae bacteriophages using protein sequence similarity.
664 *BMC Microbiol.* **9**, (2009).
- 665 37. Lavigne, R., Seto, D., Mahadevan, P., Ackermann, H. W. & Kropinski, A. M. Unifying classical
666 and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools.
667 *Res. Microbiol.* **159**, 406–414 (2008).
- 668 38. Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K. & Schuster, S. C. Whole-genome
669 prokaryotic phylogeny. *Bioinformatics* **21**, 2329–2335 (2005).
- 670 39. Iranzo, J., Krupovic, M. & Koonin, E. V. The double-stranded DNA virosphere as a modular
671 hierarchical network of gene sharing. *MBio* **7**, (2016).
- 672 40. Bolduc, B., Youens-Clark, K., Roux, S., Hurwitz, B. L. & Sullivan, M. B. iVirus: Facilitating new
673 insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure.
674 *ISME J.* **11**, 7–14 (2017).
- 675 41. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean
676 viruses. *Nature* **537**, 689–693 (2016).
- 677 42. Vik, D. R. *et al.* Putative archaeal viruses from the mesopelagic ocean. *PeerJ* **5**, e3428 (2017).
- 678 43. Roux, S. *et al.* Ecogenomics of virophages and their giant virus hosts assessed through time series
679 metagenomics. *Nat. Commun.* **8**, (2017).
- 680 44. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat.*
681 *Microbiol.* (2018). doi:10.1038/s41564-018-0190-y
- 682 45. Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and abundant
683 viruses. *Nat. Commun.* **8**, (2017).

- 684 46. de la Cruz Peña, M. J. *et al.* Deciphering the Human Virome with Single-Virus Genomics and
685 Metagenomics. *Viruses* **10**, 113 (2018).
- 686 47. Aiewsakun, P., Adriaenssens, E. M., Lavigne, R., Kropinski, A. M. & Simmonds, P. Evaluation of
687 the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common
688 bioinformatic platform: Steps towards a unified taxonomy. *J. Gen. Virol.* **99**, 1331–1343 (2018).
- 689 48. Hulo, C., Masson, P., Le Mercier, P. & Toussaint, A. A structured annotation frame for the
690 transposable phages: A new proposed family ‘Saltoviridae’ within the Caudovirales. *Virology* **477**,
691 155–163 (2015).
- 692 49. Adriaenssens, E. M. *et al.* Taxonomy of prokaryotic viruses: 2017 update from the ICTV Bacterial
693 and Archaeal Viruses Subcommittee. *Archives of Virology* 1–5 (2018). doi:10.1007/s00705-018-
694 3723-z
- 695 50. Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-protein
696 interaction networks. *Nat. Methods* **9**, 471–472 (2012).
- 697 51. Doyle, E. L. *et al.* Genome Sequences of Four Cluster P Mycobacteriophages. *Genome Announc.*
698 **6**, e01101-17 (2018).
- 699 52. Pope, W. H. *et al.* Bacteriophages of *Gordonia* spp. Display a spectrum of diversity and genetic
700 relationships. *MBio* **8**, (2017).
- 701 53. Pope, W. H. *et al.* Whole genome comparison of a large collection of mycobacteriophages reveals
702 a continuum of phage genetic diversity. *Elife* **4**, e06416 (2015).
- 703 54. Nelson, D. Phage taxonomy: We agree to disagree. *Journal of Bacteriology* **186**, 7029–7031
704 (2004).
- 705 55. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially
706 revises the tree of life. *Nat. Biotechnol.* **36**, 996 (2018).
- 707 56. Krupovic, M., Quemin, E. R. J., Bamford, D. H., Forterre, P. & Prangishvili, D. Unification of the
708 Globally Distributed Spindle-Shaped Viruses of the Archaea. *J. Virol.* **88**, 2354–2358 (2014).
- 709 57. Rokyta, D. R., Burch, C. L., Caudle, S. B. & Wichman, H. A. Horizontal gene transfer and the
710 evolution of microvirid coliphage genomes. *J. Bacteriol.* **188**, 1134–1142 (2006).
- 711 58. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and
712 Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*
713 **11**, 538 (2010).
- 714 59. Gregory, A. *et al.* Marine viral macro- and micro-diversity from pole to pole.
- 715 60. Marz, M. *et al.* Challenges in RNA virus bioinformatics. *Bioinformatics* **30**, 1793–1799 (2014).
- 716 61. Howard-Varona, C., Hargreaves, K. R., Abedon, S. T. & Sullivan, M. B. Lysogeny in nature:
717 Mechanisms, impact and ecology of temperate phages. *ISME Journal* **11**, 1511–1520 (2017).
- 718 62. Mirzaei, M. K. & Maurice, C. F. Ménage à trois in the human gut: Interactions between host,
719 bacteria and phages. *Nature Reviews Microbiology* **15**, 397–408 (2017).
- 720 63. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic
721 expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
- 722 64. Krupovic, M. *et al.* Taxonomy of prokaryotic viruses: update from the ICTV bacterial and

- 723 archaeal viruses subcommittee. *Arch. Virol.* **161**, 1095–1099 (2016).
- 724 65. Adams, M. J. *et al.* Changes to taxonomy and the International Code of Virus Classification and
725 Nomenclature ratified by the International Committee on Taxonomy of Viruses (2017). *Arch.*
726 *Virol.* **162**, 2505–2538 (2017).
- 727 66. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: A new generation of protein database
728 search programs. *Nucleic Acids Research* **25**, 3389–3402 (1997).
- 729 67. Wiwie, C., Baumbach, J. & Röttger, R. Comparing the performance of biomedical clustering
730 methods. *Nat. Methods* **12**, 1033–1038 (2015).
- 731 68. Brohée, S. & van Helden, J. Evaluation of clustering algorithms for protein-protein interaction
732 networks. *BMC Bioinformatics* **7**, (2006).
- 733 69. Kamburov, A., Stelzl, U. & Herwig, R. IntScore: A web tool for confidence scoring of biological
734 interactions. *Nucleic Acids Res.* **40**, (2012).
- 735 70. Goldberg, D. S. & Roth, F. P. Assessing experimentally derived interactions in a small world.
736 *Proc. Natl. Acad. Sci.* **100**, 4372–4376 (2003).
- 737 71. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat.*
738 *Methods* **12**, 59–60 (2015).
- 739 72. Ohio Supercomputer Center . (1987).
- 740 73. Oliphant, T. E. SciPy: Open source scientific tools for Python. *Comput. Sci. Eng.* **9**, 10–20 (2007).
- 741 74. McKinney, W. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.*
742 **1697900**, 51–56 (2010).
- 743 75. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830
744 (2011).
- 745 76. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal*
746 *Complex Syst.* **1695**, 1–9 (2006).
- 747 77. Federico, P., Pfeffer, J., Aigner, W., Miksch, S. & Zenk, L. Visual Analysis of Dynamic Networks
748 Using Change Centrality. in *2012 IEEE/ACM International Conference on Advances in Social*
749 *Networks Analysis and Mining* 179–183 (2012). doi:10.1109/ASONAM.2012.39

750

751 **ACKNOWLEDGEMENTS.** We thank Laura Bollinger, Gareth Trubl, and Igor Tolstoy for their
752 comments on improving the manuscript, as well as Wesley Zhi-Qiang You for helping push the network
753 analytics. High performance computational support was provided as an award from the Ohio
754 Supercomputer Center to MBS. Funding was provided in part by the Department of Energy’s Genome
755 Sciences Program Soil Microbiome Scientific Focus Area award (#SCW1632) to Lawrence Livermore
756 National Laboratory; an NSF Biological Oceanography award (OCE#1536989), and a Gordon and Betty
757 Moore Foundation Investigator Award (#3790) to MBS. Funding was provided to JRB by the Intramural
758 Research Program of the NIH, National Library of Medicine. The work conducted by the U.S.
759 Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S.
760 Department of Energy under Contract DE-AC02-05CH11231 to SR. This work was funded in part
761 through Battelle Memorial Institute’s prime contract with the US National Institute of Allergy and
762 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

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

16
17 **Figure 2. Performance evaluation of vConTACT versions 1.0 and 2.0 on prokaryotic virus genomes.**
18 (a) The same colors denote individual performance metrics for the ICTV genera (G) including 940 viral
19 genomes, which are achieved by the Markov clustering (MCL) algorithm at each inflation factor (v1.0)
20 as well as ClusterONE (CL1) and CL1 followed by distance-based hierarchical clustering (CL1 + H)
21 (v2.0), respectively. For more objective comparisons, MCL at an inflation factor of 7.0 followed by
22 hierarchical clustering (IF 7.0 + H) with the same distance (i.e., 9.0) used for v2.0 was included. The total
23 height and number of each bar indicate the composite score for overall performance comparison (Fig. 2a).
24 For details, see Online Methods. (b, c) Gene-sharing networks were built using 2,304 archaeal and
25 bacterial virus genomes retrieved from Viral RefSeq v85. Viral clusters (VCs) were obtained by
26 vConTACT v1.0 (b) and v2.0 (c) that used MCL with inflation factor (IF) of 7.0 and CL1, respectively
27 (see Online Methods). For both networks, genomes (nodes) are color-coded according to their taxonomic
28 assignments. For example, genomes (only members of the ICTV-recognized genus) that are classified in
29 VCs containing a single ICTV genus are colored in cyan, while genomes found in VCs containing more
30 than two genera are colored in pink. Genomes without ICTV genus affiliation are in grey. Nodes with
31 bold borders indicate those that were correctly identified either as outlier, overlap genomes or separate
32 VCs through v1.0 (b), compared to v2.0 (c). Genomes whose taxonomic assignments and/or annotation
33 are incomplete are colored in yellow, identified through v2.0 (c). For details, see **Supplementary Figs 3**
34 **and 5** and **Supplementary Table 2**. (d) Box plots of the percentage of shared protein clusters (PCs)
35 **between member viruses within 28 v1.0-generated undersampled VCs having**
36 **≥2 genera before (pink), and after (cyan) removal of outlier and/or**
37 **separation into individual clusters by v2.0.** (e) Pie charts depicting the number of
38 overlapping genomes that belong to the high (HGCF) or low (LGCF) gene content flux evolutionary
39 modes or mixed and lytic or temperate phages. Data on the lifestyle and evolutionary modes of 74 viruses
40 were collected from Mavrich and Hatfull²². For details, see **Supplementary Fig. 3**.

41
42 **Figure 3. Application of the hierarchical decomposition to discordant VCs.** (a) Distribution of all 31
43 discordant VCs across the archaeal and bacterial virus gene sharing network, where genomes (nodes) of
44 the given VCs are highlighted in pink and others in grey. (b) Box plots show the fraction (%) of protein
45 clusters (PCs) that were shared within an ICTV genus (i.e., intra-genus proteome similarity) and between
46 multiple genera (i.e., inter-genera similarity) found in each discordant VC including structured clusters
47 whose member genera have similar inter-genera and intra-genus similarities (black dot). (c) Left, A full
48 link dendrogram is represented. Note that the Euclidean distance of nine yielded the highest composite
49 score of accuracy (*Acc*) and clustering-wise separation (*Sep*) for sub-clusters from all v2.0-generated
50 VCs, which was used to split the discordant clusters (**Online Methods** and **Supplementary Fig. 4**).

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

57
58 **Figure 4. Adding the Global Ocean Virome to NCBI Viral RefSeq.** (a) Selected network images from
59 the largest connected component of GOV-additions. Red nodes are virus RefSeq genomes, and grey
60 nodes are GOV. Despite adding 15,280 new genomes, the network maintains its overall structure. (b)
61 Pairwise heatmap comparison at all GOV incremental additions using normalized mutual information
62 (NMI) values. NMI measures VC similarity to other VCs by comparing genome content changes across
63 incremental additions of data. Darker blue hues correspond to more similar information content (i.e.
64 genomes maintaining the same VC membership). (c) Boxplots depicting the average Euclidean distance
65 within VCs across GOV data increments. Grey boxes are samples prior to hierarchical trimming, while
66 blue boxes are post-trimming. Points represent discordant VCs, with darker hues representing increasing
67 discordance (i.e., more genera per VC). (d) Change centralities on a per-genome (gray) and per-VC
68 (aqua) basis through successive, 10% increments of GOV data. A value of zero in change centrality (Y-
69 axis) represent no change in any of the nodes connected to the origin node (or that the node was
70 removed), while a value of one represents origin node creation. High change centrality scores imply that
71 nodes are being created adjacent to the origin node, with the further a node's creation is from the origin
72 node, the less of an impact it has on the origin node's centrality. Dotted lines in each violin represent
73 quartiles, whereas the width of each violin plot is scaled to be equal between GOV % (X-axis), such that
74 distributions can be compared between datasets. (e) GOV network performance through successive data
75 accumulations. As GOV sequences are added (X-axis), individual performance score (ranging from 0 to
76 1, Y-axis; calculated from the clustering-wise positive predictive value (PPV), clustering-wise sensitivity
77 and accuracy) across genus- and family-level predictions (represented by circular and square data points,
78 respectively) generally trend towards stabilization.

79
80
81 **Supplementary Figure 1. Bipartite network of reference virus genomes generated by vConTACT 2.**
82 (a) The full network is represented (b) Close-up of 3 viral clusters displayed as bipartite network. In this
83 configuration, pink nodes represent individual genomes, while dark grey nodes depict proteins clusters
84 that are shared between viral clusters.

85
86 **Supplementary Figure 2. Clustering comparisons between vConTACT v1.0 and v2.0** (a) Number of
87 total viral clusters (VCs), and genus-assigned VCs, concordant and discordant VCs, as detected by
88 vConTACT v1.0 (left) and v2.0 (right). Discordant VCs are (i) those that have a mix of the different
89 genera (i.e., lumped genera), (ii) those that have different member virus(es) of the same genus by splitting
90 them into multiple clusters (i.e., split genera), or (iii) mix of (i) and (ii). (b) Proteome similarities of
91 viruses within 22 concordant VCs of vConTACT version 1.0. For all plots, the x-axis is the individual
92 pairwise comparisons and y-axis is the proteome similarity (i.e., percentage fraction of shared protein
93 cluster between genomes). In the case of VCs 6, 26, 66, and 130 (highlighted by bold borders), these all
94 contains taxonomically-misplaced member virus(es) of the *Che8virus*, *Pbunavirus*, *PI00virus*, and
95 *Bcep78virus*, respectively, all of which were correctly captured by v2.0 and ratified by the ICTV (see
96 **Supplementary Fig. 5**). Like these four VCs, the remaining 18 VCs contained distant relatives, with only
97 1-30% of similarities to the rest of the given clusters or discrete viral group(s) displaying a number of
98 discontinuous similarities, which were identified as outliers or separated VCs by v2.0, respectively (see
99 **Supplementary Table 4**).

101 **Supplementary Figure 3. vConTACT v2.0-based detection and characterization of overlapping**
102 **viral genomes.** (a) Box plots depicting the distribution of the topology-based confidence scores between
103 viruses identified as overlaps and non-overlaps, which vConTACT v2.0 placed into two clusters (see
104 panel b) and single clusters, respectively. For details, see Methods. (b) List of phages and archaeal viruses
105 identified as overlaps and their ICTV genus. Details on the lifestyle and evolutionary modes of 74 viruses
106 were collected from Mavrich and Hatfull²² and from the Actinobacteriophage Database website
107 (<http://phagesdb.org/>). The high (HGCF) and low (LGCF) gene content flux evolutionary modes indicate
108 the predicted lifestyle based on the gene content dissimilarity between viral genomes. Bioinformatically-
109 predicted temperate phages indicate those that contain the integrase (for integrating temperate phage
110 genomes into host) or *parA* (partitioning gene found in extrachromosomal temperate phages) genes.
111

112 **Supplementary Figure 4. Evaluation of optimal distance thresholds for hierarchical clustering of**
113 **VCs.** The X-axis denotes distance threshold increments from dist=1 to dist=20 in 0.5 intervals. The Y-
114 axis denotes composite scores by multiplying Accuracy (*Acc*, cyan) and clustering-wise separation (*Sep*,
115 pink) when trying to recapitulate ICTV genera, which are geometric means of Sensitivity and Positive
116 predictive value as well as Complex-wise separation and Cluster-wise separation, respectively. From
117 these data, a distance of 9.0 yielded the highest composite score for the sub-clusters partitioned from all
118 vConTACT v2.0-generated viral clusters. For details, see Online Methods.
119

120 **Supplementary Figure 5. vConTACT v2.0-based detection and characterization of boundary**
121 **genome(s) within the ICTV-recognized genera.** (a) Box plots show the percentage of shared protein
122 clusters (PCs) between members of an ICTV genus (red), and the same metrics after excluding viruses
123 recognized as outlier(s) by vConTACT v2.0 (cyan). The proteome similarities for the *Barnyardvirus* are
124 shown between (1) member viruses of the *Barnyardvirus*, (2) *Mycobacterium* virus Barnyard and the
125 remaining members of the *Barnyardvirus*, (3) the *Barnyardvirus* and *Patiencevirus*, (4) *Mycobacterium*
126 virus Barnyard and the *Patiencevirus*, and (5) the *Barnyardvirus* without outliers. For the *Phikmvvirus*,
127 the proteome similarities between (1) member viruses of the *Phikmvvirus*, (2) *Pseudomonas* virus
128 *phiKMV* and the remaining members of the *Phikmvvirus*, and (3) *Pseudomonas* virus *phiKMV* and
129 members of VC_33 are shown, respectively. (b) Module profiles show the presence (dark) and absence
130 (light) of homologous PCs across genomes. Each row represents a virus and each column a PC. The
131 genomes were hierarchically clustered based on pairwise Euclidean distance. The ICTV and vConTACT
132 v2.0 classifications are indicated next to each virus.
133

134 **Supplementary Figure 6. Evaluation of singletons/outlier genomes over GOV increments.** Thirty-
135 eight ICTV recognized singleton and outlier genomes (one per row) were observed to evaluate whether
136 the addition of GOV sequences would improve their classification. The coloring gradient on the left
137 indicates the numbers of genera per VC. Clearly improved clustering was observed in 3 of the 38
138 genomes (black bolded), with 2 genomes clustering to 6-genera discordant VCs (red), though the
139 majority saw only minor, if any, change in their clustering assignment.
140

141 **Supplementary Figure 7. vConTACT v2.0 computational runtimes.** In the plot, processing time (in
142 seconds, Y-axis) is plotted against increasing GOV sequence data (GOV % added, X-axis). There is a
143 strong linear correlation between runtime and memory usage with data volume to be processed.
144

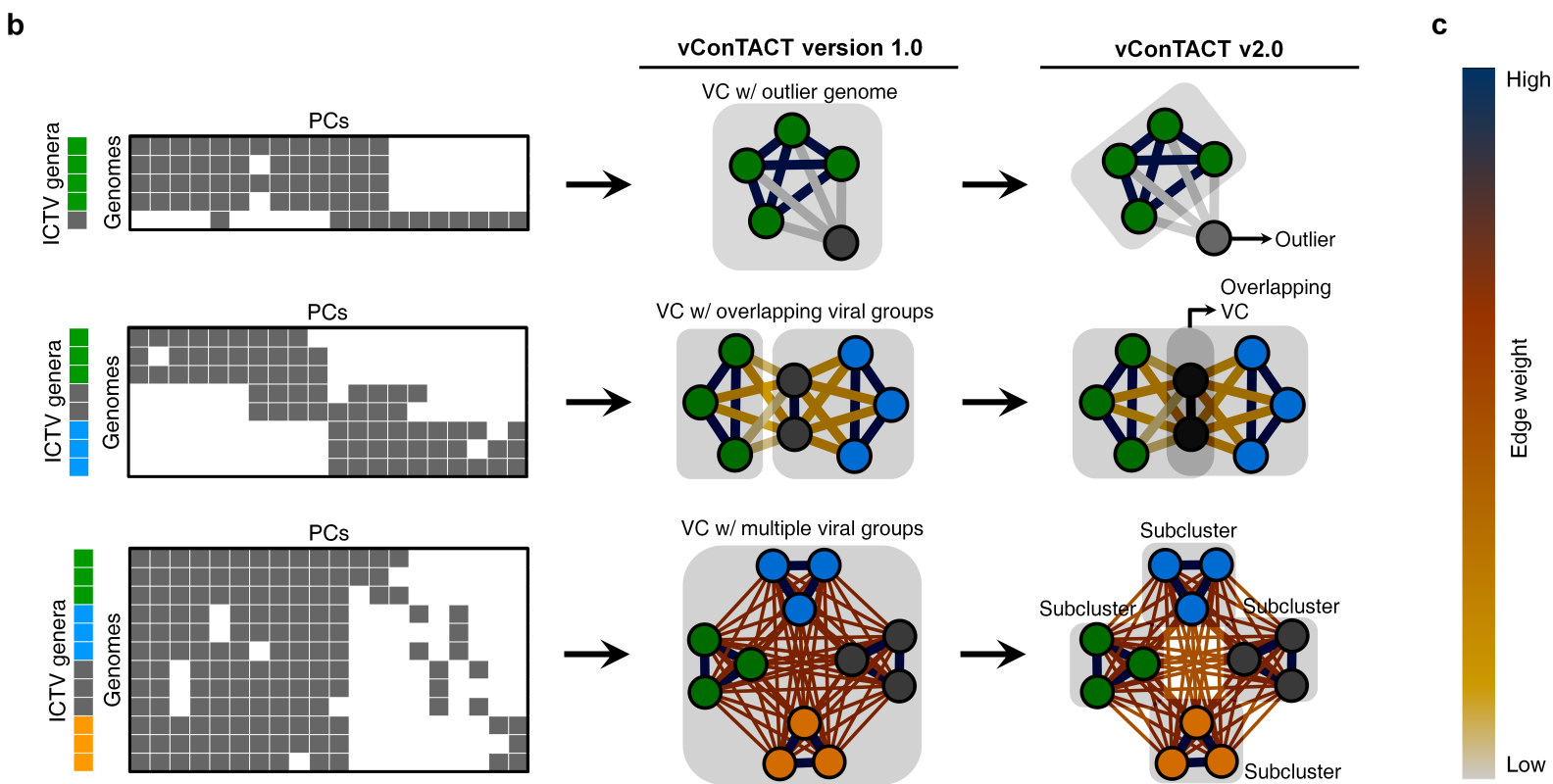
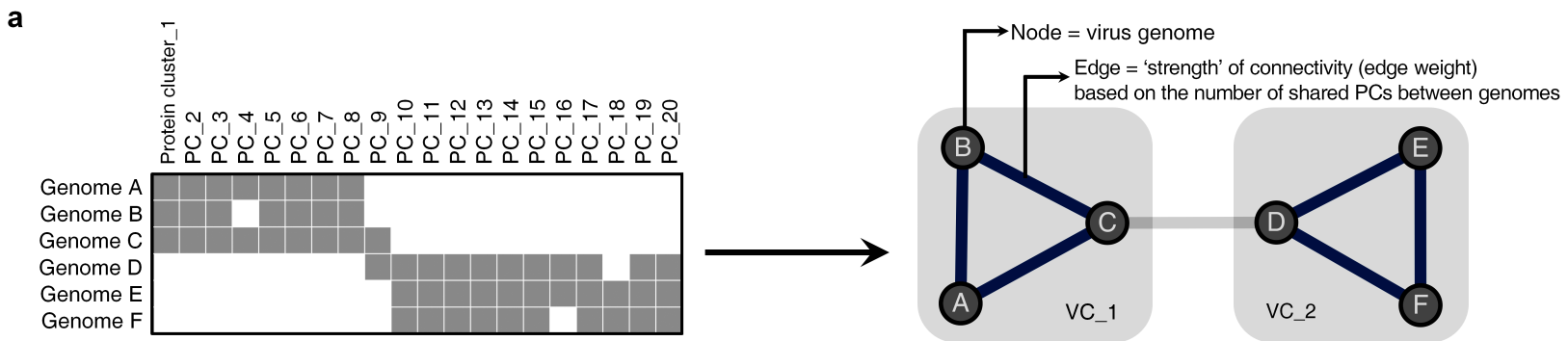
145 **Supplementary Figure 8. Impact of the inflation factor on viral genome clustering based on the**
146 **Markov clustering (MCL) algorithm.** Left panel: average intra-cluster clustering coefficient (ICCC)
147 and number of viral clusters (VCs), which are predicted as a function of the inflation factors ranging from
148 1.0 to 5.0 with a step of 0.2, are indicated. Right panel: Curve representing the ICCC values for the
149 network containing 2,304 archaeal and bacterial virus genomes.
150

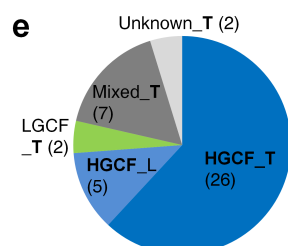
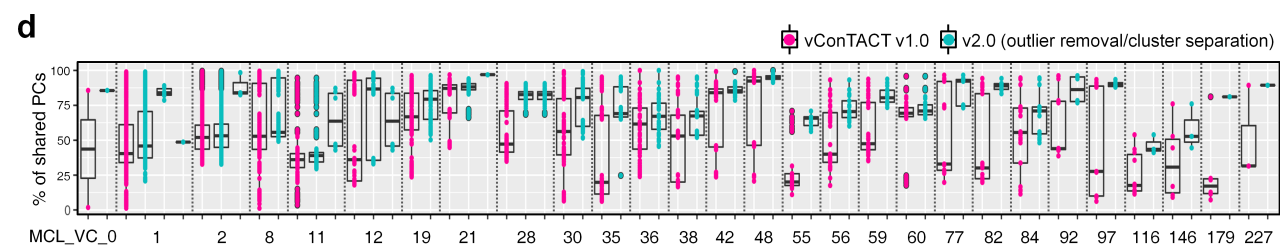
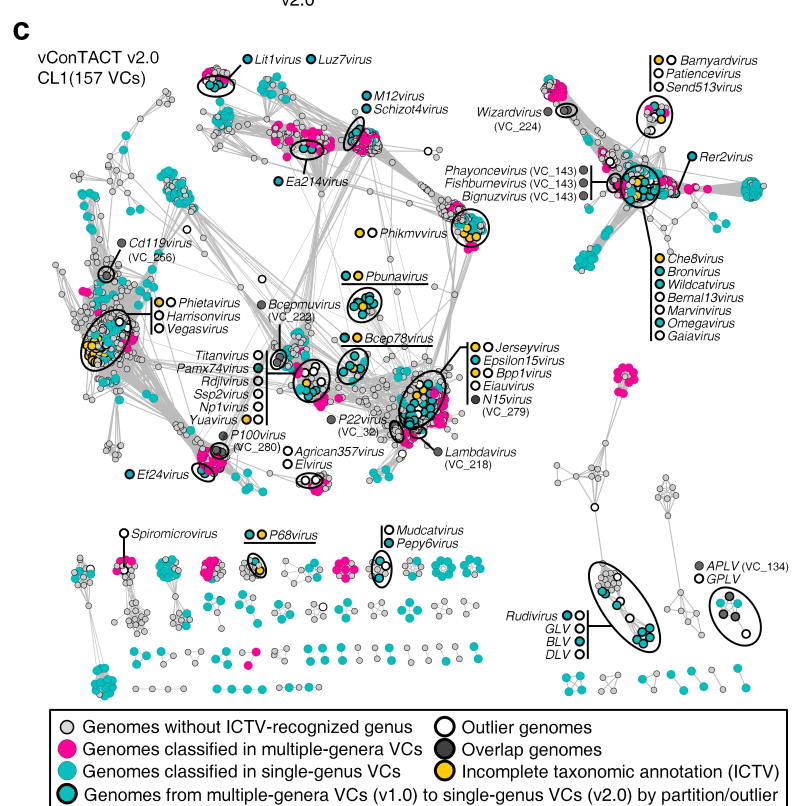
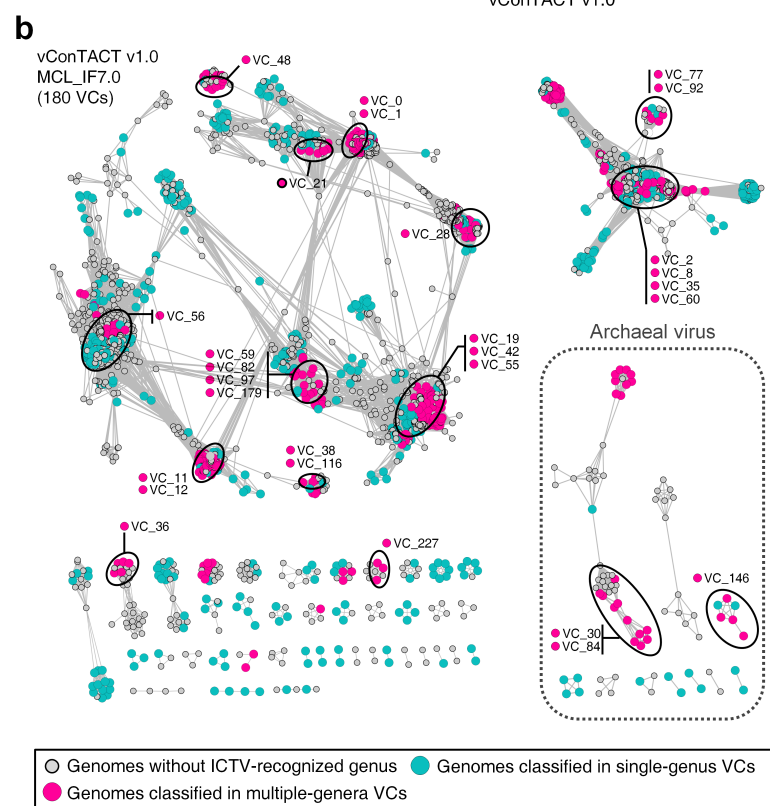
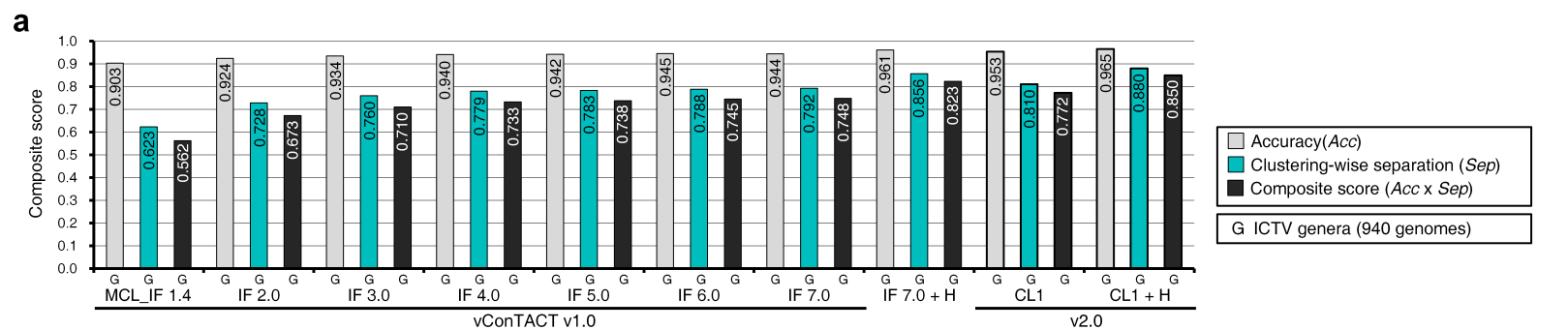
151 **Supplementary Table 1: A genome-gene matrix.** In this matrix, a total of 231,166 protein-coding genes
152 from 2,304 bacterial and archaeal virus genomes are shown, along with their annotations, protein clusters
153 and singleton proteins (i.e., isolated protein with no relatives). Each singleton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

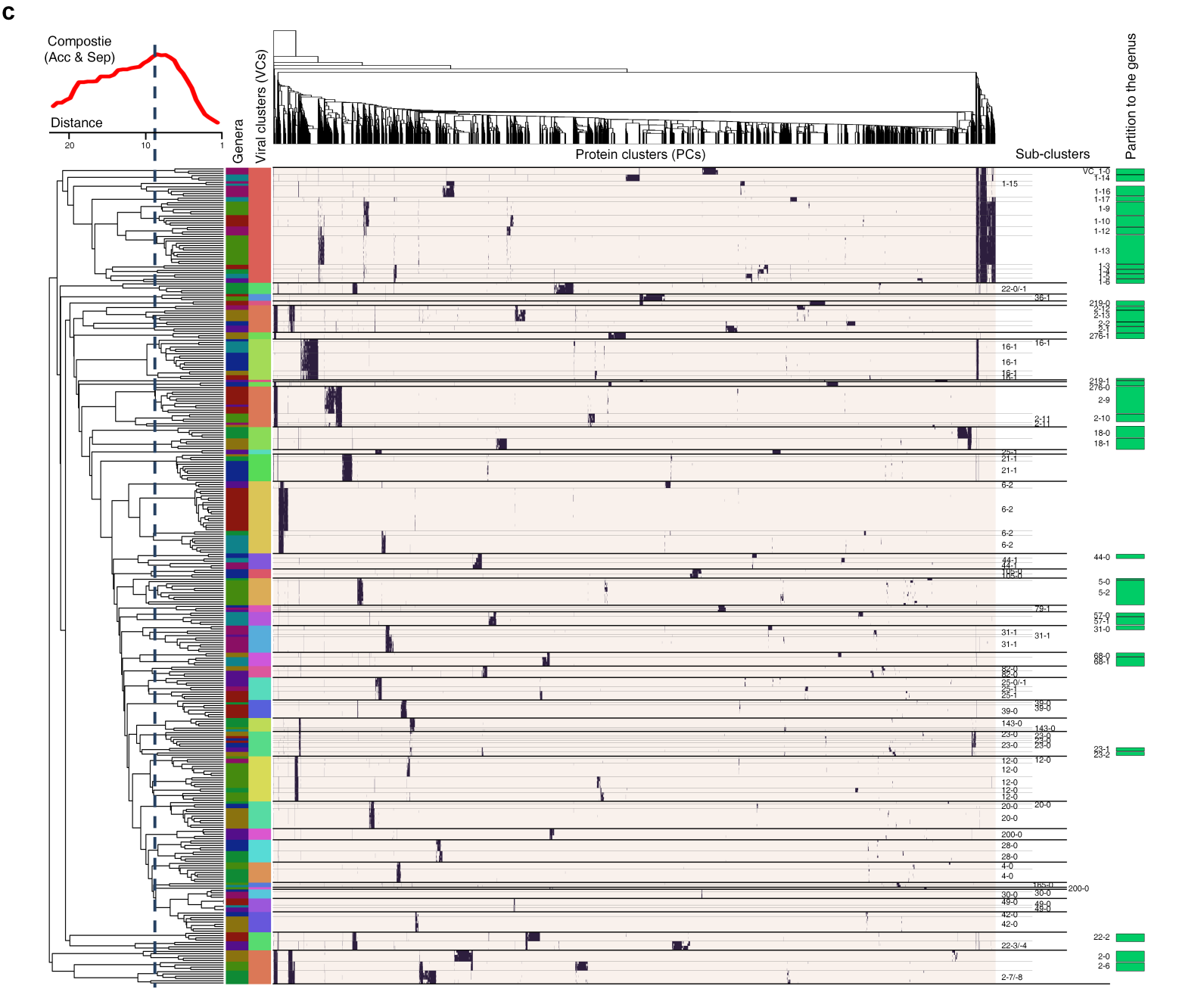
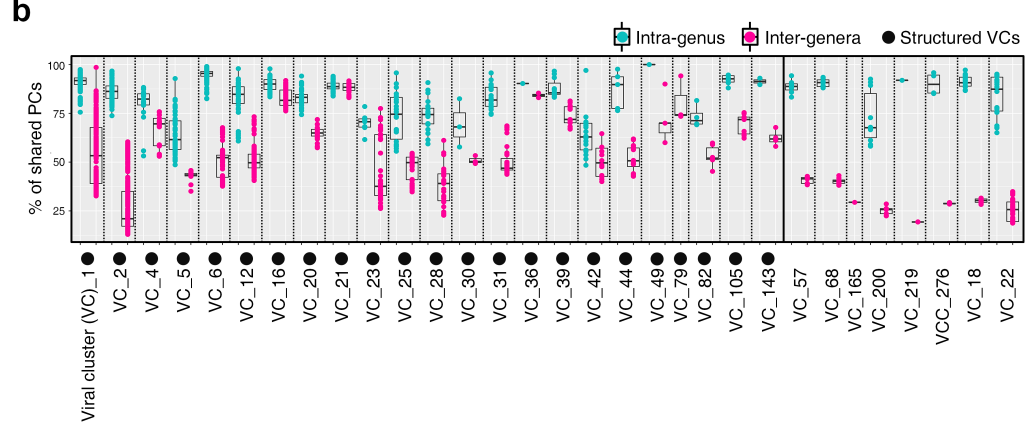
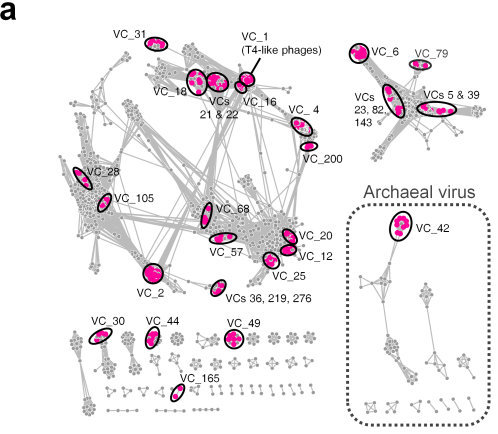
155
156 **Supplementary Table 2. List of 2,304 archaeal and bacterial virus genomes used to evaluate**
157 **vConTACT v1.0 and v2.0.** Genomes were retrieved from NCBI RefSeq, their main taxonomic features
158 from NCBI and ICTV as well as the clustering outputs from vConTACT v1.0 and v2.0 is indicated.
159 Particularly for v2.0, additional features for viral clusters (VCs) by CL1 and subclusters by CL1 followed,
160 by distance-based hierarchical clustering, as well as quality and p-values, were used to generate the
161 topology-based score. For each subcluster, a taxonomy-based scores is also provided. For the outlier
162 genomes, their parent VCs (i.e., those from which v2.0 separated the given genomes as outliers) are
163 included. For the genomes identified as part of overlapping VCs, dual VC membership is represented by
164 ‘/’. The former VC, as an overlapping cluster, can be also found in the latter cluster (mix of two VCs).
165 See Online Method for details.

166
167 **Supplementary Table 3. Clustering performance evaluations of the vConTACT v1.0, v2.0, and v2.0**
168 **followed by distance-based hierarchical clustering for the genus rank.** The numbers represent the
169 values of Sensitivity (Sn), Positive prediction value (PPV), and Accuracy (geometrical mean of Sn and
170 PPV , Acc) as well as Cluster-wise separation (Sep_{cl}), Complex-wise separation (Sep_{co}), and Clustering-
171 wise separation (Sep) across all clustering results, respectively. “Distance” indicates whether a
172 hierarchical clustering was applied to all VCs (either “NoDist” if no clustering was applied, or “9” if a
173 clustering with a cutoff of 9.0 on euclidean distance was applied).

174
175 **Supplementary Table 4. Fraction of PCs in common between 2,304 genomes.** The average shared PC
176 percentage is calculated based on the geometric formula (see Online Methods) and 400,234 pairs of
177 genomes having >1% proteome similarity are shown. For vConTACT v1.0 that uses MCL, the proteome
178 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
179 ICTV taxa (i.e., subfamily and genera) that each genome belongs to are represented.

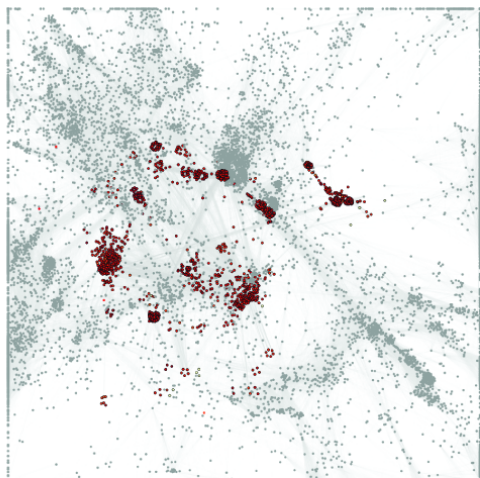




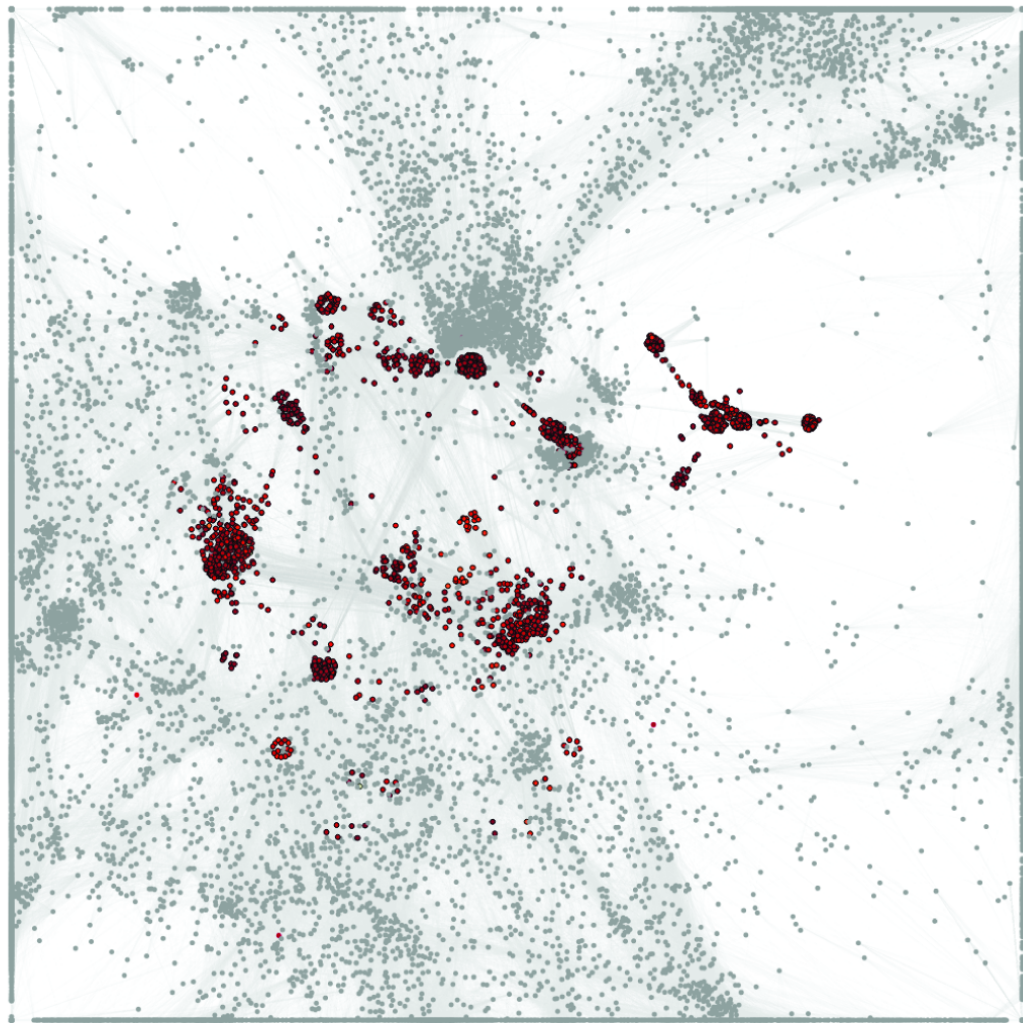


a

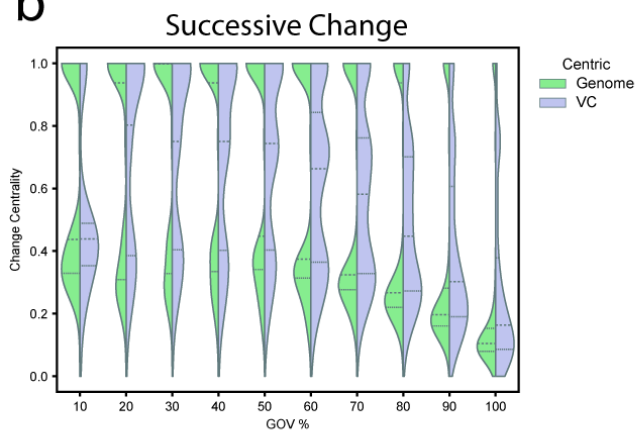
30% (5,737)



70% (10,638)



100% (14,524)

b**c**