Title: Quantitative assessment of NCLDV–host interactions predicted by co-occurrence analyses

Running title: NCLDV–host prediction based on co-occurrence analysis

Lingjie Meng[a], Hisashi Endo[a], Romain Blanc-Mathieu[a,b], Samuel Chaffron[c,d], Rodrigo Hernández-Velázquez[a,e], Hiroto Kaneko[a], Hiroyuki Ogata[a,#]

Affiliations:

a. Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011, Japan
b. Laboratoire de Physiologie Cellulaire & Végétale, CEA, Univ. Grenoble Alpes, CNRS, INRA, IRIG, Grenoble, France
c. Université de Nantes, CNRS UMR 6004, LS2N, F-44000 Nantes, France
d. Research Federation (FR2022) Tara Océan GO-SEE, Paris, France
e. Max Planck Institute for Marine Microbiology, Celsiusstraße 1, Bremen, Germany

[#]***Corresponding author***:
Hiroyuki Ogata (E-mail: ogata@kuicr.kyoto-u.ac.jp, Phone: +81-774-38-3270)

Abstract

Nucleocytoplasmic DNA viruses (NCLDVs) are highly diverse and abundant in marine environments. However, knowledge of their hosts is limited because only a few NCLDVs have been isolated so far. Taking advantage of the recent large-scale marine metagenomics census, *in silico* host prediction approaches are expected to fill the gap and further expand our knowledge of virus–host relationships for unknown NCLDVs. In this study, we built co-occurrence networks of NCLDVs and eukaryotic taxa to predict virus–host interactions using *Tara* Oceans sequencing data. Using the positive likelihood ratio to assess the performance of host prediction for NCLDVs, we benchmarked several co-occurrence approaches and demonstrated an increase in the odds ratio of predicting true positive relationships four-fold compared with random host predictions. To further refine host predictions from high-dimensional co-occurrence networks, we developed a phylogeny-informed filtering method, Taxon Interaction Mapper, and showed it further improved the

prediction performance by twelve-fold. Finally, we inferred virophage – NCLDV networks to corroborate that co-occurrence approaches are effective for predicting interacting partners of NCLDVs in marine environments.

Importance

NCLDVs can infect a wide range of eukaryotes although their life cycle is less dependent on hosts compared with other viruses. However, our understanding of NCLDV–host systems is highly limited because few of these viruses have been isolated so far. Co-occurrence information has been assumed to be useful to predict virus–host interactions. In this study, we quantitatively show the effectiveness of co-occurrence inference for NCLDV host prediction. We also improve the prediction performance with a phylogeny-guided method, which leads to a concise list of candidate host lineages for three NCLDV families. Our results underpin the usage of co-occurrence approach for metagenomic exploration of the ecology of this diverse group of viruses.

Introduction

Nucleocytoplasmic large DNA viruses (NCLDVs) represent a group of double-stranded DNA viruses that belong to the viral phylum *Nucleocytoviricota* (Virus Taxonomy: 2019 Release), which was previously referred to as Megavirales (1, 2). NCLDVs usually possess diverse gene repertoires (74 to more than 2,000 proteins), large genomes (45 kb to 2.5 Mb), and outsized virions (80 nm to 1.5 µm) (3–5). NCLDVs have high functional autonomy and encode components of replication, transcription, and translation systems (3). Recently, a virus that belongs to a new family of NCLDVs called “Medusaviridae” was found to encode five types of histones (6). The existence of metabolically active viral factories and infectious virophages also indicates that the life cycle of NCLDVs is less dependent on host cells than other viruses (7, 8). To further understand what makes these giant viruses more independent than other viruses, a first crucial step is to identify their hosts — “Who infects whom?”.

NCLDVs are known to infect a broad range of eukaryotes, from unicellular eukaryotes and macroalgae to animals (9). Amoebae are frequently used hosts in co-culture to isolate large NCLDVs (10). However, there is growing evidence, especially in marine systems, that NCLDVs can infect many phytoplankton groups, such as Pelagophyceae, Mamiellophyceae, Dinophyceae, and Haptophyte (11–13). Several other non-photosynthetic eukaryotic lineages, such as Bicoecea and Choanoflagellatea, were also reported as

experimentally identified NCLDV hosts in marine environments (14, 15). Small to large marine organisms, including invertebrates and vertebrates, are infected by viruses that belong to the NCLDV family *Iridoviridae* (16, 17). Together these studies indicate ubiquitous infectious relationships between NCLDVs and a wide range of marine eukaryotes. However, our understanding of NCLDV–host systems is very limited because few viruses have been isolated so far.

The number of viruses and hosts isolated in the laboratory represents a very small fraction of existing interactions in the ocean. Indeed, NCLDVs have been found to be highly diverse and abundant based on omics data (18, 19). In only a few liters of coastal seawater, more than 5,000 *Mimiviridae* species were detected; by comparison, only 20 *Mimiviridae* with known hosts have been well investigated (20). Global marine metagenomic data have revealed that the richness and phylogenetic diversity of NCLDVs are even higher than those of an entire prokaryotic domain (21). From biogeographical evidence, it is clear that these viruses are prevalent in the marine environment but have a heterogeneous community structure across sizes, depths, and biomes (22). Marine metatranscriptomic data have also shown that NCLDVs are active everywhere in sunlit oceans and may infect hosts from small piconanoplankton (0.8–5 μm) to large mesoplankton (180–2000 μm) (23).

Previous studies also demonstrated that NCLDVs have the potential to infect a greater diversity of hosts than known to date through gene transfer analyses (24, 25). NCLDVs might have started coevolving with eukaryotes even before the last eukaryotic common ancestor (LECA) (26). A recent study supported this hypothesis by showing that some NCLDVs encode viractins (actin-related genes in viruses), which could have been acquired from proto-eukaryotes and possibly reintroduced in the pre-LECA eukaryotic lineage (27). Together, these findings underline a lack of knowledge about NCLDV biology and host diversity. Therefore, more effort is needed to identify hosts to elucidate the poorly known virus–host relationships and the largely unknown NCLDV world.

Substantial effort has been made to reveal interactions between NCLDVs and their putative hosts. Apart from the co-culture method, other alternative methods, including high-throughput cell sorting, are also being used (10, 15). Metagenomics, which is particularly useful to assess a large fraction of ecosystem diversity, has been increasingly used to investigate NCLDVs host range. Comparative genomics analyses, such as identification of horizontal gene transfer (HGT) predictions, have also performed well for NCLDV host prediction (24, 25).

102 Abundance-based co-occurrence analyses have been used for host prediction and are
103 supposed to be effective because viruses can only thrive in an environment where their hosts
104 exist (18, 28). In addition to virus–host relationships, co-occurrence networks have been used
105 to predict the association between NCLDVs and their "parasites" (virophages) (29).
106 However, the co-occurrence-based prediction is also controversial for viral host prediction
107 since the abundance dynamics of viruses and their hosts (e.g., *Emiliania huxleyi* and
108 *Heterosigma akashiwo* viruses) are sometimes not concordant (30, 31). Usually, validation
109 with known virus–host relationships or corroboration with genomic evidence (e.g., HGT) is
110 used to assess network-based predictions (18, 28). However, the effectiveness of previous
111 and novel co-occurrence network methods has never been quantitatively tested for NCLDV
112 host prediction. The current lack of quantitative assessment hinders the widespread use of
113 this approach. Therefore, dedicated methods are needed to test the accuracy of NCLDV host
114 prediction with co-occurrence networks and to improve the performance of co-occurrence-
115 based predictions.
116 The *Tara* Oceans expedition is a global-scale survey on marine ecosystems that
117 expands our knowledge of microbial diversity, organismal interactions, and ecological
118 drivers of community structure (32). The present study used *Tara* Oceans metagenomic and
119 metabarcoding datasets to predict virus–host relationships between NCLDVs and eukaryotes
120 by constructing co-occurrence networks using different methods. To quantitatively assess the
121 performance of network-based host prediction, we employed the positive likelihood ratio
122 (LR+) using reference data for known NCLDV–host relationships. We developed a
123 phylogeny-based enrichment analysis approach, Taxon Interaction Mapper (TIM), to enhance
124 the performance in detecting positive signals in the intricate inferred networks. TIM has
125 previously been used in host predictions for DNA and RNA viruses (33), but without a
126 quantitative assessment on its effectiveness. In this study, we assessed the performance of
127 TIM as a filter of co-occurrence networks. We examined NCLDV–virophage networks,
128 which further justify the use of co-occurrence and filtering approaches to identify NCLDV
129 interaction partners.
130
131
132 # Results
133 **NCLDV–eukaryote co-occurrence networks**
134 From five datasets that corresponded to five size fractions (Fig. S1), we generated five
135 co-occurrence networks on a global scale (Fig. 1, S2A). Altogether, these networks were

136 composed of 20,148 V9 and 5,234 *polB* OTUs (nodes) and 47,978 *polB*–V9 associations
137 (edges). Out of these associations, 47,296 had positive weights, and 682 had negative weights
138 (Fig. 2A). The associations that involved the family *Mimiviridae* were numerically dominant
139 (n = 36,830) among the different NCLDV families. The second largest family was
140 *Phycodnaviridae*, with 5,521 edges involving eukaryotes. No other family had more than 2,000
141 associations with eukaryotes. *Marseilleviridae*, forming the least associations in the networks,
142 had 132 edges with eukaryotes. Taxonomic annotation of eukaryotic OTUs indicated that
143 Alveolata, Opisthokonta, Rhizaria, and Stramenopiles were the major four eukaryotic groups
144 connected to NCLDVs (with 21,167, 9,179, 6,521, and 5,327 edges, respectively). Three of
145 these eukaryotic groups belong to the SAR supergroup (i.e., Stramenopiles, Alveolata, and
146 Rhizaria), which represented 68.81% of the total associations. Regarding the pairs between
147 viral families and eukaryotic lineages, *Mimiviridae* and Alveolata showed the largest number
148 of edges (n = 16,548). Besides NCLDV–eukaryote associations, we detected 57,495 *polB–polB*
149 associations and 234,448 V9–V9 associations (Fig. S2B). We also included environmental
150 parameters in the network inference and identified 25 pairs of associations between
151 environmental parameters and *polB* OTUs (Table S1).
152 The number of NCLDV–eukaryote associations generally decreased with enlarging
153 size fraction (Fig. S2A). The largest number of *polB*–V9 associations were found in the 0.8–
154 5-μm fraction (n = 10,647). Correspondingly, the eukaryotic community in 0.8–5-μm fraction
155 had the greatest diversity (Fig. S3). However, the 0.8–inf-μm size fraction network was the
156 largest (n = 10,477) for edges with positive weights. With the annotation of major lineages,
157 the eukaryotic community compositions in the networks varied across different size fractions
158 (Fig. 2B). In the smallest size fraction (0.8–5 μm) and the large range size fraction (0.8–inf
159 μm), Marine Alveolate Group II was the eukaryotic lineage with the largest number of
160 associations with NCLDVs (21.39% and 19.98%, respectively). Dinophyceae was the second
161 largest group connected to NCLDVs in these two size fractions and showed the largest
162 number of connections with NCLDVs in the 5–20-μm size fraction network (22.22% of total
163 interactions). The viral associations with Metazoa and Collodaria increased with increasing
164 size fractions. In the largest 180–2000-μm size fraction network, Metazoa contributed
165 39.31% of the total *polB*–V9 edges.
166 We calculated the degree of nodes (number of connected edges) for each NCLDV
167 *polB* OTU (Fig. 3A, B). Naturally, the average degree of positive associations per *polB* was
168 higher than negative edges in all size fractions and decreased along with increasing size
169 fractions (2.69, 2.40, 2.25, and 2.10 from 0.8–5 μm to 180–2000 μm, and 2.76 for 0.8–inf).

170 Most of the *polB* nodes had more than one positive association (Fig. 3A). Together with the
171 taxonomic annotation of nodes, *polB*–V9 associations in the networks generated with the
172 *Tara* Oceans data revealed their high dimensionality and complexity.
173
174 **Network validation**
175 We quantitatively assessed the performance of predicting *polB*–V9 associations using
176 the positive likelihood ratio (LR+) (Fig. 1). By defining groups of metagenomic PolBs as
177 described in the Materials and Methods, 932 OTUs were recruited in the validation, and these
178 sequences contributed 6191 *polB*–V9 associations in the FlashWeave networks (Fig. S4). To
179 obtain an overall performance, we assessed the pooled associations (by removing
180 redundancy) from the five co-occurrence networks. LR+ was separately calculated for edges
181 with positive and negative weights because they can represent different infectious patterns.
182 As shown in Fig. 4A, the LR+ of host prediction for positive associations was higher than 1
183 (LR+ = 1 indicates no change in the likelihood of the condition). The LR+ generally
184 increased with the cut-off for FlashWeave weights, which indicated that condition positive
185 cases are enriched in the edges with higher weights. This result demonstrated that the co-
186 occurrence-based host prediction of NCLDVs outperformed random prediction (i.e., random
187 inference of virus–host pairs). In high-weight regions: 1) weight > 0.6, the LR+ of
188 associations was higher than 10; 2) weight > 0.4, the LR+ was roughly higher than 4.
189 Nonetheless, the false discovery rate (FDR) was high (Fig. S5A), which indicated that the
190 predictions contained numerous virus–host edges that were not considered condition positive.
191 FDR was 91.67% and 96.34% when weight is greater than 0.6 and 0.4, respectively. An
192 assessment of the host prediction for negative weight associations was also carried out. There
193 were no known NCLDV–host pairs found in the negative networks (Fig. S5B). The analysis
194 of the remaining part of our study was thus conducted for positive associations.
195 Comparing the performance between different size fractions indicated that the
196 networks of small size fractions (including the 0.8–inf-μm size fraction) performed better in
197 predicting the NCLDV–host relationships (Fig. 4B, S6). The 0.8–inf-μm size fraction had the
198 highest average LR+ out of the five size fractions (LR+ = 4.97). The LR+ of small size
199 fractions was generally higher than that of large size fractions, but there were exceptions
200 between 180–2000 and 20–180 μm. The LR+ of the associations in the 0.8–inf-μm, 0.8–5-μm
201 and 5–20-μm was greater than 1. Different from the average results, when the weight is
202 greater than 0.8, the associations of 5–20-μm size fraction had the best performance in terms
203 of both LR+ and FDR (Fig. S6 A, B).

204 We also compared abundance filtration strategies using Flashweave-S (sensitive
205 model) and FlashWeave-HE (heterogeneous model) but did not find a consistent pattern in
206 prediction performance (Fig. S7). The networks from the Q1 filtration strategy performed
207 best using Flashweave-S, but Q1 (lower quartile) filtration was not better than Q2 (middle
208 quartile) for Flashweave-HE inferred networks. Flashweave-S had a better performance than
209 HE model with any filtration strategy. Finally, we compared the performance of networks
210 inferred by all three methods: FlashWeave-S, FastSpar and Spearman. Three methods
211 generated a comparable number of positive associations, but FlashWeave-S made the largest
212 number of true positive predictions (Fig. S8A). Noteworthy, the LR+ of these three methods
213 were all larger than 1, however, FlashWeave-S and Spearman performed better than FastSpar
214 (Fig. S8B).
215
216 **Assessment of host prediction improvement**
217 Then we used the newly developed phylogeny-guided host prediction tool, TIM, to
218 filter *polB*–V9 associations, which is based on the assumption that evolutionarily related
219 viruses tend to infect evolutionarily related hosts (see Materials and Methods). We identified
220 24 eukaryotic taxonomic groups specifically associated with NCLDVs (Fig. S9). To compare
221 the performance of the TIM results with the above raw FlashWeave results, we converted the
222 three primary eukaryotic taxonomic ranks to their associated major lineages (Table S2), and
223 the associations were plotted as a network (Fig. 5A). This network showed that three out of
224 nine NCLDV families (*Mimiviridae, Phycodnaviridae*, and *Iridoviridae*) had enriched
225 connections in specific eukaryotic lineages. Among the network edges, known virus–host
226 pairs were found, such as Haptophyta–*Mimiviridae*, Mamiellophyceae–*Phycodnaviridae*, and
227 Metazoa–*Iridoviridae*. The associations in the TIM-filtered results showed a sharp
228 improvement in performance from the original result with and without an edge weight cut-
229 off. The average LR+ of TIM-enriched associations was 42.22, which was higher than the
230 raw FlashWeave associations without a weight cut-off (3.43), with a weight cut-off of 0.4
231 (5.20), and with a cut-off at 0.668 (14.23) (Fig. 5B, S9A). The FDR dropped from 0.97 (no
232 cut-off) and 0.95 (weight cut-off of 0.4) to 0.74 (Fig. 5C).
233 From the network, diverse putative hosts (13 lineages) emerged for *Mimiviridae,*
234 including algae, protozoans, and metazoans. Metazoa had the most enriched nodes connected
235 to *Mimiviridae*; additionally, MAST-3,12, Cryptophyta, Foraminifera, and Ciliophora had
236 strong relationships with *Mimiviridae*. For *Phycodnaviridae*, there were six eukaryotic
237 lineages retained after TIM filtration. Among these, Bacillariophyta, "other filosan (part of

238 filosan Cercozoa)", and Mamiellophyceae had comparatively strong associations. Moreover,
239 Rhodophyta, Ciliophora, and Dictyochophyceae had links to both *Mimiviridae* and
240 *Phycodnaviridae*. There was also a connection between *Iridoviridae* and Metazoa.
241
242 **Associations between virophages and NCLDVs**
243 Using 6,818 NCLDV *polB* OTUs and 195 virophage major capsid proteins (MCPs),
244 we identified 535 FlashWeave associations (196 and 339 for pico- and femto-size fractions,
245 respectively) (Fig. 6A). Most of the associations had positive weights (n = 490), whereas
246 some had negative weights (n = 45). The average number of associations per virophage MCP
247 was different in two size fractions: 3.2 in femto- and 5.6 in pico-size fractions. The network
248 revealed that *Mimiviridae* had the largest number of virophage associations in both size
249 fractions. We also detected 84 positive associations between virophages and
250 *Phycodnaviridae*.
251 The phylogenetic tree defined three main virophage clades, and they were all found to
252 have many connections to NCLDVs. To investigate significant relationships, Fisher's exact
253 test was performed between virophage clades and NCLDV families. Families other than
254 *Phycodnaviridae* and *Mimiviridae* did not show significant associations. Therefore, we made
255 a group, "Other NCLDVs," to include all families except *Phycodnaviridae* and *Mimiviridae*.
256 First, we only used FlashWeave results with a weight > 0.4, as previous results showed that a
257 FlashWeave weight of 0.4 is a suitable cut-off that produced moderate performance (Fig.
258 3A). From the femto-size fraction network, we found two significantly enriched connections
259 (Fig. 6B): one was between virophage group C and *Mimiviridae* (p = 0.0022) and the other
260 was between group A and "Other NCLDVs" (p = 0.0439). Another significantly enriched
261 relationship between virophage group B and *Phycodnaviridae* (p = 0.0410) was found in
262 pico-size fractions when we used all associations without edge weight cut-off.
263 Finally, we examined HGTs of virophage MCPs in NCLDV genomes. We found two
264 HGTs of virophage MCPs; both showed links between clade A and *Iridoviridae* (Table S3).
265 This result was consistent with the Fisher's exact test result, which revealed a connection
266 between virophage clade A and "Other NCLDVs" including *Iridoviridae*.
267
268 # Discussion
269 NCLDVs can infect a wide range of eukaryotes, from unicellular to multicellular
270 organisms (34). However, we are still far from a comprehensive knowledge of their hosts
271 because few have been isolated so far. Therefore, better host prediction algorithms are needed

272 to understand the ecological functions and evolutionary significance of NCLDVs. To make
273 these predictions, we constructed global ocean co-occurrence networks based on the marine
274 metagenome and metabarcoding datasets from 85 stations of the *Tara* Oceans expedition,
275 which cover all major oceanic provinces across an extensive latitudinal gradient from pole to
276 pole. The edges (associations) between *polB* and V9 nodes (OTUs) in the networks were
277 generated using FlashWeave. The networks were particularly dense (Fig. 2A, S2A), thus
278 suggesting that NCLDVs interact with numerous eukaryotes in the ocean. This was expected
279 given the high abundance and diversity of NCLDVs in marine environments (18, 21) and the
280 identification of HGT between these viruses and diverse eukaryotic lineages (24). The
281 networks were dominated by the *Mimiviridae* nodes, which is consistent with previous
282 reports that *Mimiviridae* is the most abundant and has the widest array of transcribed genes
283 out of NCLDV families in marine environments (22, 23). *Mimiviridae* was known to infect
284 amoebae, algae, and stramenopiles (3). In our study, these three eukaryotic groups were all
285 found to have numerous associations with *Mimiviridae*. *Phycodnaviridae* has been known to
286 infect many species of aquatic organisms, such as *Emiliania huxleyi* (Haptophyta),
287 *Ectocarpus siliculosus* (Phaeophyceae), *Chlorella heliozoae* (Trebouxiophyceae), and
288 *Ostreococcus tauri* (Mamiellophyceae) (35–37). Correspondingly, plenty of associations of
289 *Phycodnaviridae* were found in the co-occurrence networks. For the eukaryotic nodes, all
290 high taxonomic rank groups, including the SAR supergroup (i.e., Stramenopiles, Alveolata,
291 Rhizaria), Opisthokonta, Archaeplastida, Amoebozoa, Excavata, and other eukaryotes, have
292 associations with NCLDVs. Among these groups, the SAR supergroup contributed the most
293 (~68%) *polB*–V9 associations. However, this is still lower than in other microbial co-
294 occurrence analyses; for example, a previous study showed SAR supergroup dominated
295 ~92% of the total aquatic microbial associations (38). A substantial proportion (~32%) of
296 NCLDV–eukaryote interactions were from non-SAR groups, which covered the known
297 NCLDV host range, such as Archaeplastida and Haptophyta.
298 However, it is difficult to accurately predict NCLDV hosts from constructed networks
299 because of the high degree of associations per *polB* OTU (Fig. 3A). One node connected to
300 multiple edges was expected in the co-occurrence analysis. In previous NCLDV host
301 prediction studies, additional processing was performed to filter the high dimensional
302 associations to predict the meaningful interactions, such as weight cut-off or a combination of
303 different co-occurrence network inference methods (18, 29). Moreover, no previous study
304 quantitively assessed the performance of co-occurrence networks when predicting NCLDV–
305 host relationships. Qualitatively identifying known pairs and detecting HGT (without

306 validation) have been commonly used to assess prediction reliability (18, 28). Therefore, we
307 aimed to 1) quantitatively assess the performance of co-occurrence-based host prediction for
308 NCLDVs and 2) improve the prediction results using filtering methods.
309 In a previous study of bacteriophage host prediction, ROC curves were used as an
310 assessment metric to compare different prediction methods (39). However, the number of
311 known virus–host pairs of NCLDVs is not sufficient to generate a dataset for ROC
312 assessment. Therefore, in this study, we carried out an alternative method, the LR+, to assess
313 the performance. LR+ is calculated with two relative values, sensitivity and specificity (Fig.
314 1). The LR+ of co-occurrence-based host predictions for positive associations was higher
315 than 1 and increased along with increasing cut-off values for the edge weights (Fig. 4A),
316 which demonstrated that positive prediction results were more likely to be true positives than
317 those based on random prediction. In high-weight regions (> 0.6 and > 0.4), LR+ values were
318 larger than 10 and 4, respectively. These LR+ values indicate that FlashWeave can increase
319 the probability of predicting true positives (40). However, both the true positive rate
320 (sensitivity, < 18.9%) and false positive rate (< 6.37%) were very low (Fig. S5C). These low
321 rates were from FlashWeave with a cut-off of alpha < 0.01, which excluded a large
322 proportion of the *polB*–V9 pairs from the results. So only about 4000 predictions could be
323 validated from a set of 6191 *polB*–V9 FlashWeave associations in this study (Fig. S5D, as
324 described in the Materials and Methods). The FDR of co-occurrence was even higher than
325 90% (Fig. S5A). Such a high FDR in co-occurrence networks demonstrates that condition
326 positive connections (i.e., known interactions) are embedded in an immense pool of condition
327 negative connections. However, these negative signals can correspond to either unidentified
328 (i.e., currently unknown true interactions), indirect, or false relationships.
329 We also found that true positive predictions only existed in positive weight
330 associations, whereas negative weight associations did not contribute to NCLDV–host
331 detection (Fig. S5B). This result indicates that the abundance dynamics of NCLDVs and their
332 potential hosts were positively correlated with each other in the analyzed samples, which
333 were collected at a global scale; this might be because NCLDVs detected in the dataset were
334 active viruses that replicate locally in their hosts. Similar results were obtained in other co-
335 occurrence-based host prediction studies (28, 41). However, several experimental studies
336 showed that the abundance dynamics of NCLDVs and hosts showed a delay in time (30, 31).
337 It is possible that the global-scale samples did not have sufficiently high resolution to detect
338 negative correlations (or correlations with a time delay) due to lack of time-resolution (42).
339 Therefore, further studies, especially those that focus on a high temporal resolution, are

340 needed to better understand the detailed dynamics of virus–host associations and the capacity
341 of co-occurrence-based methods for host prediction.
342 The networks of different size fractions showed different performance patterns in
343 predicting NCLDV–host relationships (Fig. 4B). This pattern is not dependent on the
344 diversity of eukaryotic communities (Fig. S3A, B). Generally, small-sized fractions (0.8–5-
345 µm and 5–20-µm) networks performed better than large-sized fractions (20-180-µm and 180-
346 2000-µm) networks. This result is not dependent on the diversity of eukaryotic communities.
347 To date, most of the known NCLDV hosts are small, such as the genera *Micromonas*,
348 *Aureococcus*, and *Ostreococcus* are within the range of 0.8–5-µm, and *Prymnesium*,
349 *Heterosigma*, and *Heterocapsa* are within the range of 5–20-µm. Because of this, our
350 assessment method might be biased toward small size fractions as smaller organisms tend to
351 be more abundant in the environment (43). However, it is also possible that NCLDV
352 infections are more prevalent in smaller size fractions. Notably, the 0.8–inf-µm size fraction
353 network, which covered all four individual size fractions, performed best. This might be
354 because NCLDVs can infect not only small hosts but also hosts from a broad size range.
355 Trimming of low-abundance OTUs was recommended to improve the prediction of
356 true interactions and was often used in co-occurrence studies (44, 45). In our study, however,
357 we did not achieve such performance improvement by treating input abundance data with a
358 rigorous filtration (upper quartile) (Fig. S7). This result might be because the true positive
359 and false positive rates defined in this study were too low; therefore, the validation may not
360 be sufficiently sensitive to reflect the change between different abundance trimming
361 strategies. However, it is also possible that low-abundance NCLDV OTUs are indeed
362 network participants, as was demonstrated in a study showing that rare cyanobacterial species
363 might play fundamental roles in blooming (46). Our result also revealed that FlashWeave-S
364 was better than FlashWeave-HE at predicting NCLDV–host interactions (Fig. S7). The
365 difference between FlashWeave-HE and FlashWeave-S is that HE mode can remove
366 structural zeros during network inference. Structural zero is a typical property of
367 heterogeneous datasets, like *Tara* Oceans datasets, and may lead to false-positive edges (47).
368 Conversely, our results suggested that retaining structural zeros did not negatively influence
369 the result, which indicates that the "presence–absence" pattern is as informative as the
370 "more–less" pattern when identifying NCLDV–host relationships. This result is consistent
371 with a previous "K-r-strategist" hypothesis: some NCLDVs, like mimiviruses, are K-
372 strategists that decay slowly and can form stable associations with their hosts (48, 49). A
373 recent report supported these non-"boom and bust" dynamics of prasinoviruses and their

374 hosts with an experiment-based mathematical model (50). Overall, our results support co-
375 occurrence networks as a useful method for predicting NCLDV–host interactions in marine
376 metagenomes, and likelihood ratios as useful quantitative metrics for assessing the
377 performance of co-occurrence analysis for viral host predictions.

378 Although the results generated by FlashWeave already improved the accuracy of
379 predictions, the condition positive interactions were still embedded in many noise edges, as
380 shown by a very high FDR (Fig. S5A, S6B). To overcome this situation, we developed TIM
381 to reduce the high dimension of associations and improve NCLDV host prediction (33). The
382 results showed that NCLDVs had enriched connections with 15 major eukaryotic lineages,
383 which included 24 taxonomic groups in three different ranks (order, class, and phylum)
384 (Table S2) (Fig. 5A, S7). Using the LR+ as a prediction diagnostic metric, NCLDV host
385 prediction improved 12-fold with TIM filtration (Fig. 5B). FDR dropped below 23% after
386 TIM treatment (Fig. 5C). In TIM-enriched connections, some are known NCLDV–host pairs,
387 such as *Phycodnaviridae* and Mamiellophyceae, *Mimiviridae* and Haptophyta, and
388 *Iridoviridae* and Metazoa. Some other studies revealed that *Mimiviridae* could exclusively
389 infect diverse putative hosts (24, 51). Our results support the assumption that *Mimiviridae* has
390 connections with 13 eukaryotic lineages out of 15 total lineages. Among these lineages,
391 *Mimiviridae* had the most numerous links to Metazoa. Some mimiviruses (namaoviruses) are
392 known to infect freshwater sturgeon, *Acipenser fulvescens* (52). Metazoans are presumed to
393 be susceptible to mimiviruses, because the choanoflagellates, a group of eukaryotes that is
394 phylogenetically close to metazoans, were recently identified to be the host of a species of
395 *Mimiviridae* (15). Moreover, the TIM result revealed that *Phycodnaviridae* is closely
396 connected to Bacillariophyta, which consists of three NCBI taxonomic groups:
397 Thalassiophysales, Cymbellales, and Bacillariophyceae. Thalassiophysales was shown to
398 have many HGT candidates with a large range of NCLDVs, and Bacillariophyceae also has a
399 significant HGT candidate with phaeoviruses (24). Although Dictyochophyceae itself has not
400 been proven to be a phycodnavirus host, its sister group Pelagophyceae was experimentally
401 identified as an AaV host (53). Additionally, it is interesting to note the connection between
402 Metazoa (Calanoida) and *Iridoviridae*. Calanoida is an order of arthropods commonly found
403 as zooplankton; most of the sizes are 500–2000 μm. The viruses of the family *Iridoviridae*
404 infect many Arthropod species, including insects and crustaceans (17).

405 Furthermore, we also inferred associations between virophages and NCLDVs. To
406 date, all isolated virophages are only known to infect *Mimiviridae* (54). As expected,
407 *Mimiviridae* was the family with dominant connections to virophages (Fig. 6A). Recently, *in*

408 *silico* evidence demonstrated that virophages can infect *Phycodnaviridae*, which indicated
409 that the virophage host range might be larger than we know (55). In support of this
410 hypothesis, a relatively large number of virophage OTUs were found to be associated with
411 *Phycodnaviridae* in our study. The enrichment analysis also revealed significant connections
412 between three virophage clades and NCLDV families (Fig. 6B). To support the enrichment
413 analysis, we conducted an HGT analysis because gene transfers have previously been found
414 between Sputnik virophages and giant viruses (56). Our HGT analysis indicated a previously
415 undescribed infectious relationship between virophage clade A and *Iridoviridae*. Overall, the
416 results of virophage–NCLDV associations support our previous statement that co-occurrence
417 networks inference and analysis are appropriate for investigating NCLDV interactions in
418 marine metagenomic data.
419

420 # Materials and Methods

421 **Metagenomic and metabarcoding data**

422 The microbial metagenomic and eukaryotic metabarcoding data used in this study
423 were previously generated from plankton samples collected by the *Tara* Oceans expedition
424 from 2008 to 2013 (57, 58). Because our research requires paired metagenomic and
425 metabarcoding datasets, we used data derived from the euphotic zone samples, namely those
426 from the surface (SRF) and Deep Chlorophyll Maximum (DCM) layers (59). Type B DNA
427 polymerase (*polB*) was used as the marker gene for NCLDVs. A total of 6818 NCLDV *polB*
428 OTUs were extracted from the metagenomic datasets (i.e., the second version of the Ocean
429 Microbial Reference Gene Catalog, OM-RGC.v2) using the pplacer phylogenetic placement
430 method (ML tree) (22, 60, 61). These *polB* sequences were classified into seven NCLDV
431 families (*Mimiviridae*, *Phycodnaviridae*, *Marseilleviridae*, *Ascoviridae*, *Iridoviridae*,
432 *Asfarviridae*, and *Poxviridae*) and two other giant virus groups ("Medusaviridae" and
433 "Pithoviridae"). For eukaryotes, we employed used the metabarcoding data for eukaryotes,
434 which targeting the 18S ribosomal RNA gene hypervariable V9 region (V9) (62). Taxonomic
435 annotation of the eukaryotic metabarcoding data was previously performed by the *Tara*
436 Oceans consortium using an extensive V9_PR2 reference database (59), which was derived
437 from the original Protist Ribosomal Reference (PR2) database (63). The diversity index of
438 eukaryotic communities was calculated using the package "vegan" (64). Processed frequency
439 data are available from GenomeNet
440 (ftp://ftp.genome.jp/pub/db/community/tara/Cooccurrence).
441

442 **Data processing**

443 A relative abundance matrix for the NCLDV *polB* OTUs was extracted from OM-
444 RGC.v2 for the samples derived from the pico-size fractions (0.22–1.6 or 0.22–3.0 µm). We
445 converted the relative abundances of *polB* OTUs back to absolute read counts based on gene
446 length and read length (assumed to be 100 nt). This process was required because small
447 decimal numbers cannot be used by FlashWeave and because relative abundance data suffer
448 from apparent correlations, which reduce the specificity of co-occurrence networks in
449 revealing microbial interactions (44). To build comprehensive interaction networks involving
450 eukaryotes of different sizes, we extracted the V9 read count matrices from the
451 metabarcoding dataset for the following five size fractions: 0.8–5 µm; 5–20 µm and 3–20 µm
452 (hereafter referred to as "5–20 µm" for simplicity); 20–180 µm; 180–2000 µm; and > 0.8 µm
453 (hereafter referred to as 0.8–inf µm). To create the input files for network inference, the *polB*
454 matrix was combined with each of the V9 matrices (corresponding to different size fractions),
455 and only the samples represented by both *polB* and V9 files were placed in new files. In total,
456 samples from 84 *Tara* Oceans stations (a total of 560 samples for two depths and five size
457 fractions) widely distributed across oceans were used in this study (Fig. S1A). Depending on
458 the individual size fractions, 84–127 samples were retained and included in the co-occurrence
459 analysis (Fig. S1B). Read counts in the newly generated matrices were normalized using
460 centered log-ratio (*clr*) transformation after adding a pseudo count of one to all matrix
461 elements because zero cannot be transformed in *clr.* Following *clr* normalization, we filtered
462 out low-abundance OTUs with a lower quartile (Q1) filtering approach. Specifically, OTUs
463 were retained in the matrices when their *clr*-normalized abundance was higher than Q1
464 (among the non-zero counts in the original count matrix prior to the addition of a pseudo
465 count of one) in at least five samples. Normalization and filtering were separately applied to
466 *polB* and V9. The numbers of OTUs in the final matrices are provided in Fig. S1C.
467

468 **Co-occurrence-based network inference**

469 Network inference was performed using FlashWeave [v0.15.0 (47)]. FlashWeave is a
470 fast and compositionally robust tool for ecological network inference based on the local-to-
471 global learning framework. Meta-variables (such as environmental parameters) can be
472 included in the FlashWeave network to remove potential indirect associations. We used
473 temperature, salinity, nitrate, phosphate, and silicate concentrations as meta-variables in our
474 network inferences to determine their correlations with *polB* OTUs. FlashWeave provides a
475 heterogeneous mode (FlashWeave-HE), which helps overcome sample heterogeneity.

476 However, FlashWeave-HE may not be appropriate for the *Tara* Oceans data because it was
477 shown to predict an insufficient number of known planktonic interactions (47). Therefore, we
478 mainly used FlashWeave-S with default settings except for the FlashWeave normalization
479 step and comparison between FlashWeave-S and FlashWeave-HE. A threshold to determine
480 the statistical significance was set to alpha < 0.01. All detected pairwise associations were
481 assigned a value called "weight" that ranged between −1 and +1. Edges with weights > 0 or <
482 0 were referred to as positive and negative associations, respectively. To compare the
483 performance of FlashWeave-S to other co-occurrence methods, we used FlashWeave-HE,
484 Spearman, and FastSpar (65). The FlashWeave-HE settings were the same as FlashWeave-S
485 but with a command "heterogeneous". For Spearman, we used stats.spearmanr in package
486 "Scipy" (66). In FastSpar, we used 50 iterations, 20 excluded iterations, and a threshold of
487 0.1 to generate associations. To reduce the high dimensionality of the datasets, upper quartile
488 (Q3) filtered matrices were used for comparison among FlashWeave-S, Spearman, and
489 FastSpar.
490
491 **Network validation**
492 We validated the virus–host associations in inferred networks based on a confusion
493 matrix defined by the known NCLDV–host information (Fig. 1). Briefly, we manually
494 compiled 69 known virus–host relationships for NCLDVs (Table S4). In the validation
495 process, eukaryotic taxonomic groups were annotated at the level of the "Major lineages" in
496 the extensive PR2 database (updated after publication) (62). The "Major lineages" were used
497 in the present study because 1) the deficiency of known virus–host relationships limited the
498 use of lower eukaryotic taxonomy ranks, such as genus, for assessment, and 2) these lineages
499 adequately represented marine eukaryotes by covering the full spectrum of cataloged
500 eukaryotic V9 diversity at a comparable phylogenetic depth (62). Then, we performed
501 BLASTp [2.10.1 (67)] searches from the *Tara* Oceans PolB sequences against the NCLDV
502 reference database to define groups of metagenomic PolBs with a threshold of 65% sequence
503 identity by retaining only the best hit for each environmental PolB sequence. This threshold
504 was determined because, by using reference PolB sequences and RefSeq protein sequence
505 databases, we found that 60–70% of sequence identity could distinguish whether the
506 NCLDVs infected hosts of the same major lineages; this was mainly tested for
507 *Phycodnaviridae* because of the lack of host information for closely related viruses in other
508 NCLDV families (Table S5). Then, 65% was chosen because it could provide a better LR+
509 (as described below) than 60% and 70%.

510 The positive likelihood ratio was used in for assessment to estimate the predictions
511 accuracy. This approach is commonly used in diagnostic testing to assess if a test (host
512 prediction in this study) usefully changes the probability of the existence of condition
513 positive (true positive). In this study, the LR+ was used because host prediction is a test to
514 discover condition positive states (68). LR+ is calculated by dividing the true-positive rate
515 (sensitivity) by the false-positive rate (1 –specificity). If LR+ is close to 1, the performance of
516 the prediction is comparable to a random prediction. If LR+ >> 1, a positive prediction result
517 is more likely to be a true positive than that based on random prediction. From the set of
518 detected associations between a given *polB* OTU and V9 OTUs that belong to a given major
519 eukaryotic lineage, we only kept the best positive and negative associations (i.e., the edges
520 with the highest absolute weights) to simplify the prediction scheme. As an auxiliary
521 assessment, the FDR was also calculated by dividing the number of false positives by the
522 number of positive predictions (Fig. 1). For the comparison among five size fractions, we
523 only used the abundance in the overlap samples of 0.8–5 µm, 5–20 µm, 20–180 µm, and
524 180–2000 µm sizes. So the number of samples in five size fractions is comparable (84, 88,
525 88, 88, 88), which could reduce the bias that may influence the topology of networks (69).
526
527 **Phylogeny-guided filtering of host predictions and its assessment**
528 We developed Taxon Interaction Mapper (TIM) to improve host predictions by co-
529 occurrence approaches (33). TIM assumes that evolutionarily related viruses tend to infect
530 evolutionarily related hosts (17, 70), and extract the most likely virus-host associations from
531 the co-occurrence networks. TIM requires a phylogenetic tree of viruses (based on marker
532 genes) and a set of connections between viruses and eukaryotes (co-occurrence edges), and
533 then tests whether leaves (i.e., viral OTUs) under a node of the virus tree is enriched with a
534 specific predicted host group compared with the rest of the tree using Fisher's exact test and
535 Benjamini–Hochberg adjustment (Fig. S9A) (33). TIM is available from
536 https://github.com/RomainBlancMathieu/TIM.
537 We pooled network associations using FlashWeave analysis for five size fractions. To
538 build a concise and credible viral phylogenetic tree, we removed all of the PolB sequences
539 that were absent in the FlashWeave network associations, and the remaining sequences were
540 filtered by the amino acid sequence length (≥ 500 aa). Protein alignment was conducted using
541 MAFFT-*linsi* [version 7.471 (71)], and 18 sequences were manually removed because they
542 were not well aligned with other PolB sequences. A total of 501 PolB sequences were used to
543 make a maximum likelihood phylogenetic tree with FastTree [version 2.1.11 (72)]. Then, the

544 PolB–V9 associations were mapped on the tree to calculate the significance of the enrichment
545 of specific associations using TIM. TIM provides a list of nodes in the viral tree and
546 associated NCBI taxonomies (order, class, and phylum) of eukaryotes that show significant
547 enrichment in the leaves under the nodes. The TIM result was visualized with iTOL [version
548 5 (73)]. The TIM result was converted to a network, in which nodes correspond to the major
549 eukaryotic lineages. The network edge weight was defined by the number of tree nodes in
550 each viral family subtree enriched with a specific major eukaryotic lineage. The network was
551 visualized with Cytoscape [version 3.7.1] using prefuse force directed layout (74). To assess
552 the effectiveness of TIM in improving prediction, we extracted all the associations predicted
553 by TIM and compared their performance with the raw and weight cut-off results.
554
555 **Virophage–NCLDV associations**
556 We inferred the networks between NCLDVs and virophages using. *mcp* was used as
557 the marker gene for virophages. First, 47 reference MCP amino acid sequences were
558 collected from public databases and used to build an HMM profile. The HMM profile was
559 used to search against the amino acid sequences of OM-RGC v2 using HMMER hmmsearch
560 [version 3.3.1] with the threshold of E-value < 1E−90 (75). This threshold was determined by
561 searching reference sequences against the GenomeNet nr-aa database. The search detected
562 195 *Tara* Oceans virophage MCP sequences in the OM-RGC database. Together with 47
563 reference MCPs, a phylogenetic tree of MCP amino acid sequences was built using MAFFT
564 and FastTree.
565 We extracted the abundance profiles for the 195 MCP sequences from the pico-
566 (0.22–1.6 or 0.22–3.0 μm) and femto-size (< 0.22 μm) fractions. We used samples from the
567 SRF and DCM depths. PolB and MCP abundance profiles were merged into two matrices
568 corresponding to the two virophage size fractions. Then, network inference was conducted
569 using the FlashWeave default settings after Q1 filtration. In the MCP phylogenetic tree, three
570 virophage clades contributed most of the NCLDV connections. Thus, an NCLDV enrichment
571 analysis for the three clades was carried out using Fisher's exact test, and the *p*-value was
572 adjusted by the Benjamin–Hochberg method. This approach was the same as TIM, but we did
573 not use the TIM software because the current version of TIM requires inputs of eukaryotic
574 nodes with NCBI taxonomy annotations.
575 We used another approach, HGT, to predict the virophage-NCLDV interactions. First,
576 we generated an NCLDV genome database, which includes 56 reference NCLDV genomes
577 corresponding to our *polB* dataset and 2,074 metagenome-assembled genomes from a

previous study (24). A total of 827,548 coding sequences were included in this database. Then, 195 virophage MCPs from the metagenomic data were BLASTp searched against this database using an E-value cut-off of 1E−10 (with a minimum query coverage of 50% and a minimum sequence identity of 50%). If a virophage MCP obtained a hit in the NCLDV genome database with a lower E-value compared with hits in the MCP database (the hit to itself was removed), the hit in the NCLDV genome database was considered an HGT candidate.

Acknowledgments

This work was supported by JSPS/KAKENHI (Nos. 18H02279 and 19H05667 to H.O. and Nos. 19K15895 and 19H04263 to H.E.), Kyoto University Research Coordination Alliance (funding to H.E.), and the Collaborative Research Program of the Institute for Chemical Research, Kyoto University (Nos. 2019-30 and 2020-27). Computational time was provided by the SuperComputer System, Institute for Chemical Research, Kyoto University. We further thank the *Tara* Oceans consortium, and the people and sponsors who supported *Tara* Oceans. *Tara* Oceans (including both the *Tara* Oceans and *Tara* Oceans Polar Circle expeditions) would not exist without the leadership of the *Tara* Expeditions Foundation and the continuous support of 23 institutes (https://oceans.taraexpeditions.org). This article is contribution number XXX of *Tara* Oceans. We thank Mallory Eckstut, PhD, from Edanz Group (https://en-author-services.edanzgroup.com/ac) for editing a draft of this manuscript.

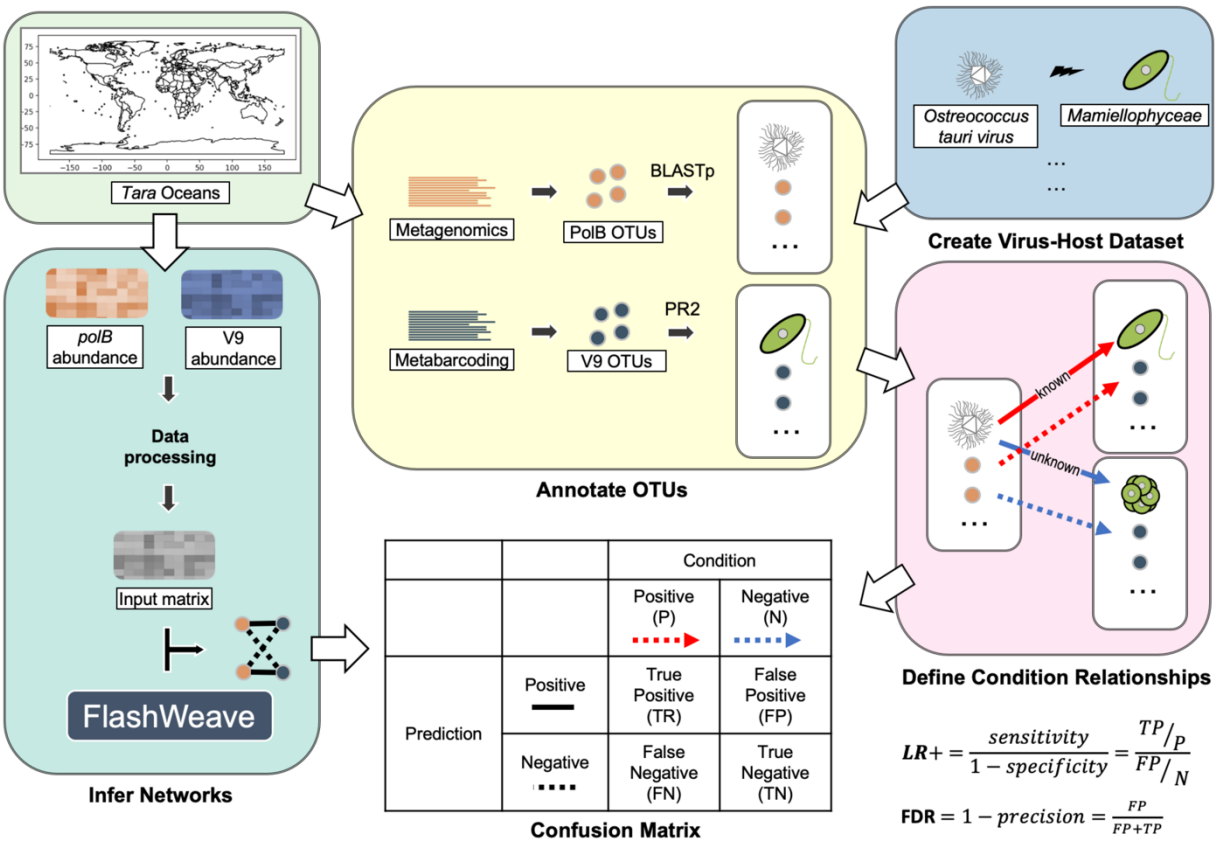

Figure 1. Overall workflow for inferring co-occurrence networks and quantitative assessment. This figure shows how the input data (*Tara* Oceans metagenomics and metabarcoding data) were used in this study. The definition of the confusion matrix for quantitative assessment is shown in the table. The LR+ and FDR equations are given at the lower right corner of the plot.

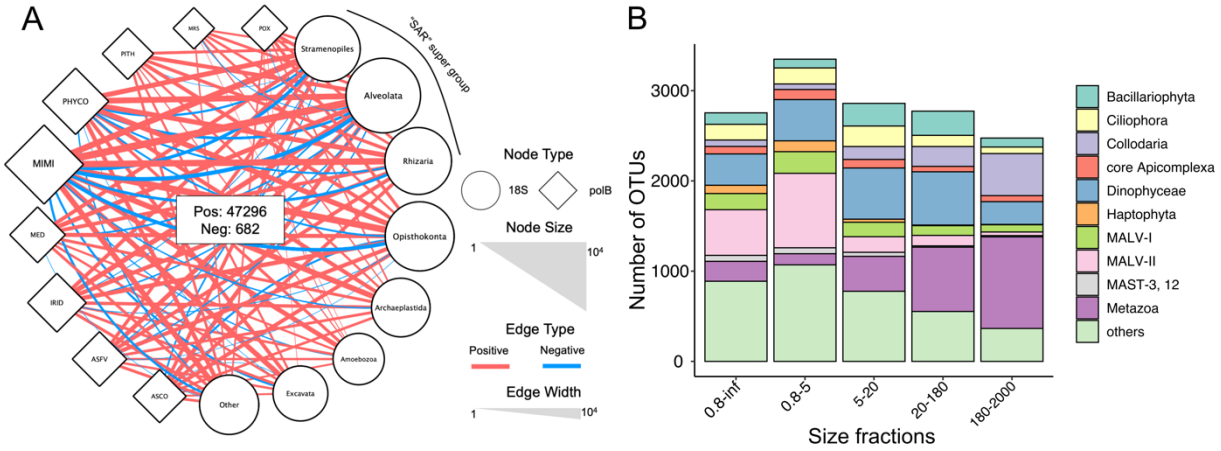

Figure 2. *polB*–V9 co-occurrence network. We performed co-occurrence analysis at the OTU level and constructed the network with pooled *polB* –V9 associations from five size fraction networks (A). To better display co-occurrence patterns, PolB OTUs were grouped at the family or family-like level, and V9 OTUs were grouped using annotation at high taxonomic ranks. The size of each node indicates the number of OTUs that belong to the group, and the width of each edge indicates the number of associations between two connected groups. Associations with positive weight are shown in red and negative associations are shown in blue. (B) Number of associations connected to NCLDVs for each major eukaryotic lineage in five size fractions. The top 10 lineages were retained, and other lineages were omitted and shown as "others." Size fractions are presented in µm.

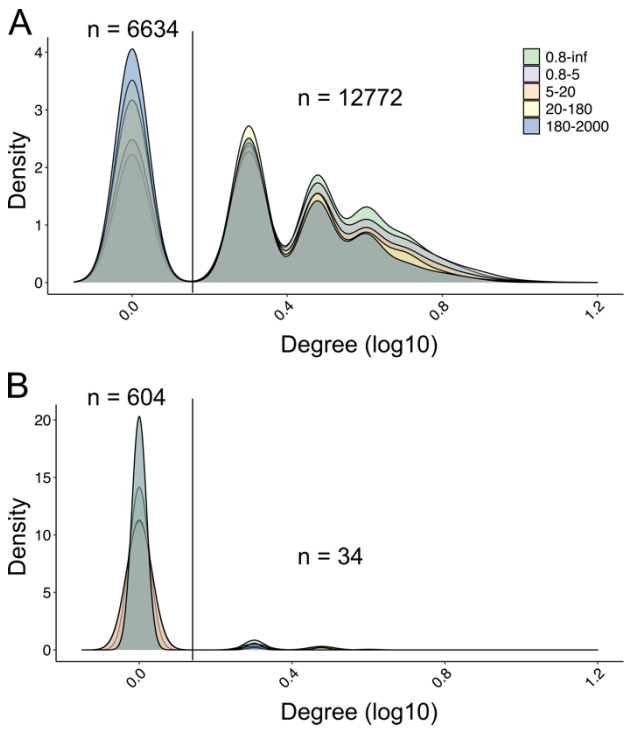

Figure 3. Density plots for the degree of NCLDV nodes in co-occurrence networks. The degree of an NCLDV node is given by the associations between this node and eukaryotes in the networks. The amount of NCLDV nodes are given on the top of the density values. (A) Positive degree (number of positive associations per node) for NCLDV nodes in five size fraction networks. (B) Negative degree (number of negative associations per node) for NCLDV nodes in five size fraction networks. Size fractions are presented in µm. NCLDV nodes with degree = 1 and degree > 1 are separated using a vertical line, and the number of nodes is given.

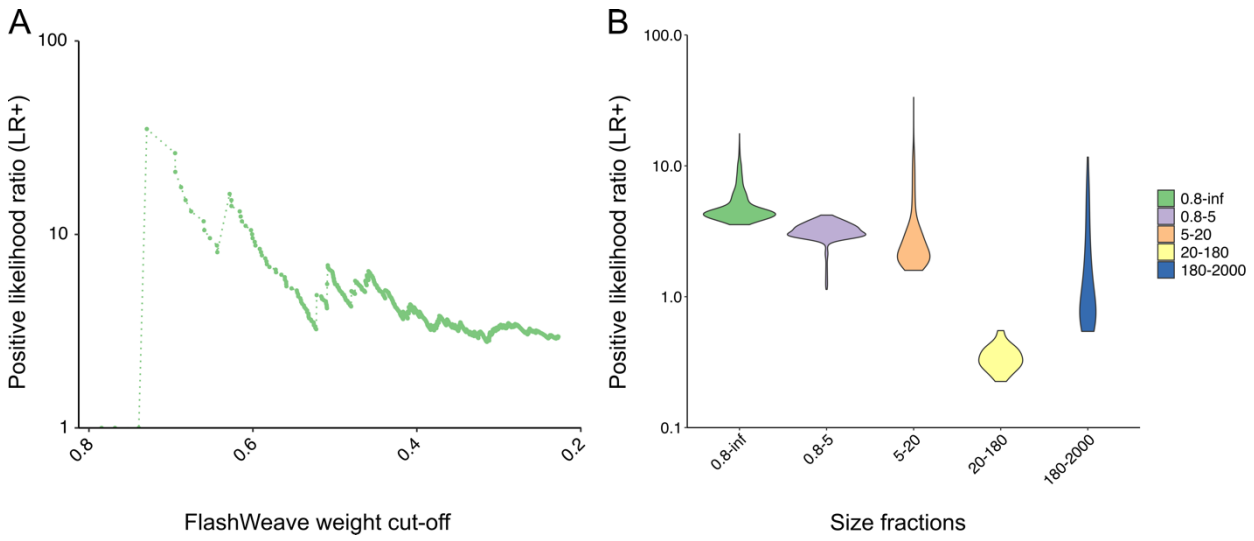

Figure 4. Positive likelihood ratios (LR+) in the NCLDV virus–host validation. (A) General performance of co-occurrence networks is shown with the LR+ calculated with associations pooled from five size fractions networks. To show the relationship between LR+ and FlashWeave association weight, the LR+ values are plotted along with the association weight. (B) Performance of each size fraction network is shown with the violin plot by ggplot2 with a bandwidth of 2. Size fractions are presented in μm.

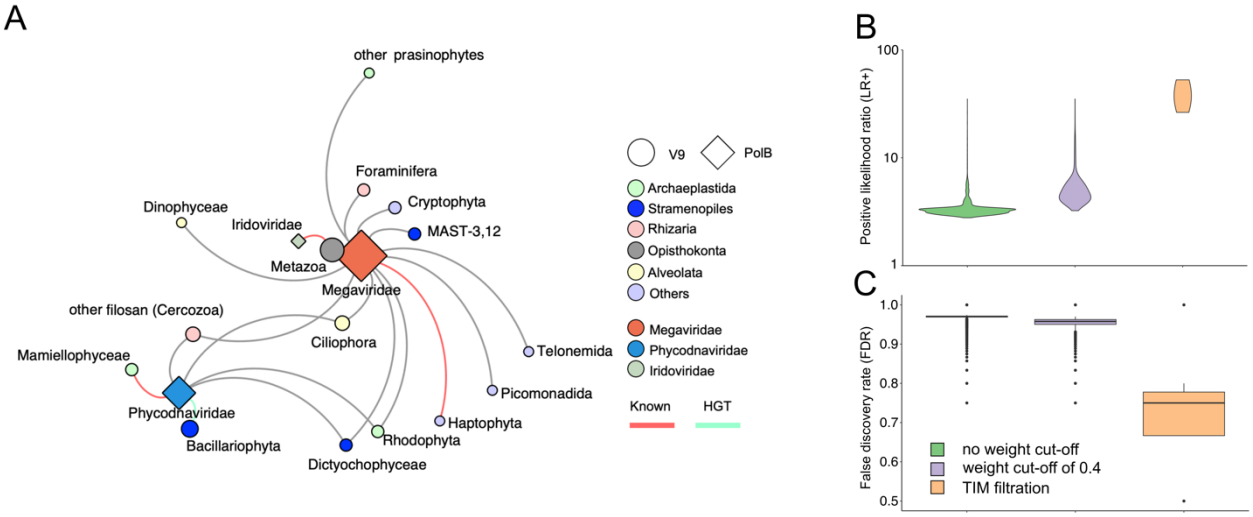

Figure 5. Prediction of NCLDV virus–host relationships with TIM. (A) Undirected network that shows the relationships between NCLDVs and eukaryotes after TIM filtration. The size of each node indicates the number of predicted interactions of this group. The weight of network edges as defined by the number of tree nodes enriched in each viral family subtree to specific eukaryotic major lineages in the TIM analysis. Known virus–host relationships are highlighted in red, and the pairs found to have horizontal gene transfer are highlighted in yellow (1). (B) Performance of networks on NCLDV host prediction for original FlashWeave results without a weight cut-off, weight cut-off > 0.4, and TIM filtration, plotted by ggplot2 with a bandwidth of 2. (C) FDR of networks for NCLDV host prediction with the original FlashWeave results without a weight cut-off, weight cut-off > 0.4, and TIM filtration.

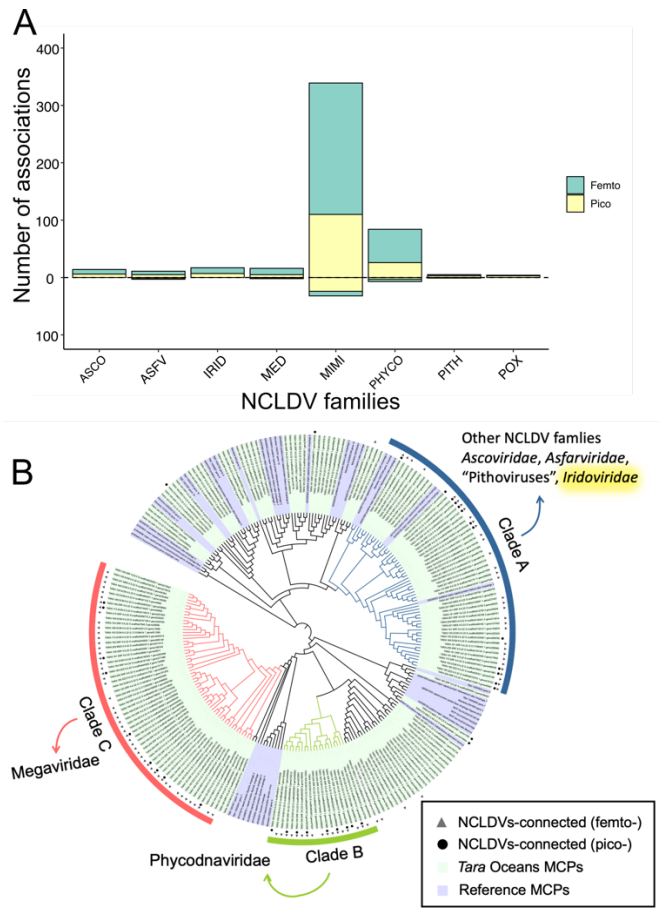

Figure 6. Associations between virophages and NCLDVs. (A) Number of associations with virophages is shown for seven NCLDV families and two unclassified groups, "Medusaviruses" and "Pithoviruses." Associations in the femto-size fraction network are shown in yellow, and in the pico-size fraction network are shown in green. The number of positive associations is above the zero axis, and the number of negative associations is below the zero axis. (B) Phylogenetic tree was constructed from 195 environmental virophages and 47 reference MCP sequences. The outside layer indicates three major virophage clades. The inner two layers indicate that the virophage OTUs have at least one association with NCLDVs in femto- or pico-size fractions networks.

References

1. Colson P, De Lamballerie X, Yutin N, Asgari S, Bigot Y, Bideshi DK, Cheng XW, Federici BA, Van Etten JL, Koonin E V., La Scola B, Raoult D. 2013. "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Arch Virol 158:2517–2521.
2. Koonin E V., Dolja V V., Krupovic M, Varsani A, Wolf YI, Yutin N, Zerbini FM, Kuhn JH. 2020. Global Organization and Proposed Megataxonomy of the Virus World. Microbiol Mol Biol Rev 84:e00061-19.

716 3. Koonin E V., Yutin N. 2019. Evolution of the Large Nucleocytoplasmic DNA Viruses
717 of Eukaryotes and Convergent Origins of Viral Gigantism. Advances in Virus
718 Research 103:167-202.
719 4. Legendre M, Alempic JM, Philippe N, Lartigue A, Jeudy S, Poirot O, Ta NT, Nin S,
720 Couté Y, Abergel C, Claverie JM. 2019. Pandoravirus celtis illustrates the
721 microevolution processes at work in the giant Pandoraviridae genomes. Front
722 Microbiol 10:1–11.
723 5. Boratto PVM, Oliveira GP, Machado TB, Andrade ACSP, Baudoin J-P, Klose T,
724 Schulz F, Azza S, Decloquement P, Chabrière E, Colson P, Levasseur A, La Scola B,
725 Abrahão JS. 2020. Yaravirus: A novel 80-nm virus infecting Acanthamoeba
726 castellanii. Proc Natl Acad Sci 117:16579–16586.
727 6. Yoshikawa G, Blanc-Mathieu R, Song C, Kayama Y, Mochizuki T, Murata K, Ogata
728 H, Takemura M. 2019. Medusavirus, a Novel Large DNA Virus Discovered from Hot
729 Spring Water. J Virol 93:1–3.
730 7. Forterre P. 2011. Manipulation of cellular syntheses and the nature of viruses: The
731 virocell concept. Comptes Rendus Chim 14:392–399.
732 8. Mougari S, Sahmi-Bounsiar D, Levasseur A, Colson P, Scola B La. 2019. Virophages
733 of giant viruses: An update at eleven. Viruses 11:1–28.
734 9. Koonin E V., Yutin N. 2010. Origin and evolution of eukaryotic large nucleo-
735 cytoplasmic DNA viruses. Intervirology 53:284–292.
736 10. La Scola B, Audic S, Robert C, Jungang L, De Lamballerie X, Drancourt M, Birtles R,
737 Claverie JM, Raoult D. 2003. A giant virus in amoebae. Science 299:2033.
738 11. Nagasaki K, Yamaguchi M. 1997. Isolation of a virus infectious to the harmful bloom
739 causing microalga Heterosigma akashiwo (Raphidophyceae). Aquat Microb Ecol
740 13:135–140.
741 12. Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH,
742 Wilhelm SW. 2014. Genome of brown tide virus (AaV), the little giant of the
743 Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution.
744 Virology 466–467:60–70.
745 13. Wilson WH, Schroeder DC, Allen MJ, Holden MTG, Parkhill J, Barrell BG, Churcher
746 C, Hamlin N, Mungall K, Norbertczak H, Quail MA, Price C, Rabbinowitsch E,
747 Walker D, Craigon M, Roy D, Ghazal P. 2005. Complete genome sequence and lytic
748 phase transcription profile of a Coccolithovirus. Science 309:1090–1092.
749 14. Colson P, Gimenez G, Boyer M, Fournous G, Raoult D. 2011. The giant Cafeteria

750 roenbergensis virus that infects a widespread marine phagocytic protist is a new
751 member of the fourth domain of life. PLoS One 6:13–17.
752 15. Needham DM, Yoshizawa S, Hosaka T, Poirier C, Choi CJ, Hehenberger E, Irwin
753 NAT, Wilken S, Yung CM, Bachy C, Kurihara R, Nakajima Y, Kojima K, Kimura-
754 Someya T, Leonard G, Malmstrom RR, Mende DR, Olson DK, Sudo Y, Sudek S,
755 Richards TA, DeLong EF, Keeling PJ, Santoro AE, Shirouzu M, Iwasaki W, Worden
756 AZ. 2019. A distinct lineage of giant viruses brings a rhodopsin photosystem to
757 unicellular marine predators. Proc Natl Acad Sci U S A 116:20574–20583.
758 16. Williams T. 2008. Natural Invertebrate Hosts of Iridoviruses (Iridoviridae). Neotrop
759 Entomol 37:615–632.
760 17. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, Hingamp
761 P, Goto S, Ogata H. 2016. Linking virus genomes with host taxonomy. Viruses 8:10–
762 15.
763 18. Hingamp P, Grimsley N, Acinas SG, Clerissi C, Subirana L, Poulain J, Ferrera I,
764 Sarmento H, Villar E, Lima-Mendez G, Faust K, Sunagawa S, Claverie JM, Moreau
765 H, Desdevises Y, Bork P, Raes J, De Vargas C, Karsenti E, Kandels-Lewis S, Jaillon
766 O, Not F, Pesant S, Wincker P, Ogata H. 2013. Exploring nucleo-cytoplasmic large
767 DNA viruses in Tara Oceans microbial metagenomes. ISME J 7:1678–1695.
768 19. Monier A, Claverie JM, Ogata H. 2008. Taxonomic distribution of large DNA viruses
769 in the sea. Genome Biol 9:R106.
770 20. Li Y, Hingamp P, Watai H, Endo H, Yoshida T, Ogata H. 2018. Degenerate PCR
771 primers to reveal the diversity of giant viruses in coastal waters. Viruses 10:496.
772 21. Mihara T, Koyano H, Hingamp P, Grimsley N, Goto S, Ogata H. 2018. Taxon richness
773 of "Megaviridae" exceeds those of bacteria and archaea in the ocean. Microbes
774 Environ 33:162–171.
775 22. Endo H, Blanc-Mathieu R, Li Y, Salazar G, Henry N, Labadie K, de Vargas C,
776 Sullivan MB, Bowler C, Wincker P, Karp-Boss L, Sunagawa S, Ogata H. 2020.
777 Biogeography of marine giant viruses reveals their interplay with eukaryotes and
778 ecological functions. Nat Ecol Evol https://doi.org/10.1038/s41559-020-01288-w.
779 23. Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R,
780 Lima-Mendez G, Rocha F, Tirichine L, Labadie K, Kirilovsky A, Bertrand A, Engelen
781 S, Madoui MA, Méheust R, Poulain J, Romac S, Richter DJ, Yoshikawa G, Dimier C,
782 Kandels-Lewis S, Picheral M, Searson S, Acinas SG, Boss E, Follows M, Gorsky G,
783 Grimsley N, Karp-Boss L, Krzic U, Pesant S, Reynaud EG, Sardet C, Sieracki M,

784 Speich S, Stemmann L, Velayoudon D, Weissenbach J, Jaillon O, Aury JM, Karsenti
785 E, Sullivan MB, Sunagawa S, Bork P, Not F, Hingamp P, Raes J, Guidi L, Ogata H,
786 De Vargas C, Iudicone D, Bowler C, Wincker P. 2018. A global ocean atlas of
787 eukaryotic genes. Nat Commun 9:373.
788 24. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, McMahon KD,
789 Konstantinidis KT, Eloe-Fadrosh EA, Kyrpides NC, Woyke T. 2020. Giant virus
790 diversity and host interactions through global metagenomics. Nature 578:432–436.
791 25. Gallot-Lavallée L, Blanc G. 2017. A glimpse of nucleo-cytoplasmic large DNA virus
792 biodiversity through the eukaryotic genomicswindow. Viruses 9:17.
793 26. Guglielmini J, Woo AC, Krupovic M, Forterre P, Gaia M. 2019. Diversification of
794 giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes.
795 Proc Natl Acad Sci U S A 116:19585–19592.
796 27. Cunha V Da, Gaïa M, Ogata H, Jaillon O, Delmont TO, Forterre P. 2020. Giant
797 viruses encode novel types of actins possibly related to the origin of eukaryotic actin:
798 the viractins. bioRxiv doi.org/10.1101/2020.06.16.150565.
799 28. Moniruzzaman M, Wurch LL, Alexander H, Dyhrman ST, Gobler CJ, Wilhelm SW.
800 2017. Virus-host relationships of marine single-celled eukaryotes resolved from
801 metatranscriptomics. Nat Commun 8:16054.
802 29. Roux S, Chan LK, Egan R, Malmstrom RR, McMahon KD, Sullivan MB. 2017.
803 Ecogenomics of virophages and their giant virus hosts assessed through time series
804 metagenomics. Nat Commun 8:858.
805 30. Tomaru Y, Katanozaka N, Nishida K, Shirai Y, Tarutani K, Yamaguchi M, Nagasaki
806 K. 2004. Isolation and characterization of two distinct types of HcRNAV, a single-
807 stranded RNA virus infecting the bivalve-killing microalga Heterocapsa
808 circularisquama. Aquat Microb Ecol 34:207–218.
809 31. Martínez Martínez J, Schroeder DC, Larsen A, Bratbak G, Wilson WH. 2007.
810 Molecular dynamics of Emiliania huxleyi and cooccurring viruses during two separate
811 mesocosm studies. Appl Environ Microbiol 73:554–562.
812 32. Sunagawa S, Acinas SG, Bork P, Bowler C, Eveillard D, Gorsky G, Guidi L, Iudicone
813 D, Karsenti E, Lombard F, Ogata H, Pesant S, Sullivan MB, Wincker P, de Vargas C.
814 2020. Tara Oceans: towards global ocean ecosystems biology. Nat Rev Microbiol
815 18:428–445.
816 33. Kaneko H, Blanc-Mathieu R, Endo H, Chaffron S, Delmont TO, Gaia M, Henry N,
817 Hernández-Velázquez R, Nguyen CH, Mamitsuka H, Forterre P, Jaillon O, De Vargas

818 C, Sullivan MB, Suttle CA, Guidi L, Ogata H. 2019. Eukaryotic virus composition can
819 predict the efficiency of carbon export in the global ocean. bioRxiv
820 https://doi.org/doi.org/10.1101/710228.
821 34. Fischer MG. 2016. Giant viruses come of age. Curr Opin Microbiol 31:50–57.
822 35. Delaroque N, Wolf S, Muller DG, Knippers R. 2000. The brown algal virus EsV-1
823 particle contains a putative hybrid histidine kinase. Virology 273:383–390.
824 36. Fitzgerald LA, Graves M V., Li X, Hartigan J, Pfitzner AJP, Hoffart E, Van Etten JL.
825 2007. Sequence and annotation of the 288-kb ATCV-1 virus that infects an
826 endosymbiotic chlorella strain of the heliozoon Acanthocystis turfacea. Virology
827 362:350–361.
828 37. Clerissi C, Desdevises Y, Grimsley N. 2012. Prasinoviruses of the Marine Green Alga
829 Ostreococcus tauri Are Mainly Species Specific. J Virol 86:4611–4619.
830 38. Bjorbækmo MFM, Evenstad A, Røsæg LL, Krabberød AK, Logares R. 2020. The
831 planktonic protist interactome: where do we stand after a century of research? ISME J
832 14:544–559.
833 39. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. 2016. Computational approaches
834 to predict bacteriophage-host relationships. FEMS Microbiol Rev 40:258–272.
835 40. Salkić NN. 2008. Objective assessment of diagnostic tests validity : a short review for
836 clinicians and other mortals . Part II. Acta Med Acad 39–42.
837 41. Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard
838 CPD, Dutilh BE, Thompson FL. 2017. Marine viruses discovered via metagenomics
839 shed light on viral strategies throughout the oceans. Nat Commun 8:1–12.
840 42. Martin-Platero AM, Cleary B, Kauffman K, Preheim SP, McGillicuddy DJ, Alm EJ,
841 Polz MF. 2018. High resolution time series reveals cohesive but short-lived
842 communities in coastal plankton. Nat Commun 9:266.
843 43. Huete-Ortega M, Cermeño P, Calvo-Díaz A, Marañón E. 2012. Isometric size-scaling
844 of metabolic rate and the size abundance distribution of phytoplankton. Proc R Soc B
845 Biol Sci 279:1815–1823.
846 44. Berry D, Widder S. 2014. Deciphering microbial interactions and detecting keystone
847 species with co-occurrence networks. Front Microbiol 5:1–14.
848 45. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. 2015. Sparse
849 and Compositionally Robust Inference of Microbial Ecological Networks. PLoS
850 Comput Biol 11:1–25.
851 46. Xue Y, Chen H, Yang JR, Liu M, Huang B, Yang J. 2018. Distinct patterns and

852 processes of abundant and rare eukaryotic plankton communities following a reservoir
853 cyanobacterial bloom. ISME J 12:2263–2277.
854 47. Tackmann J, Matias Rodrigues JF, von Mering C. 2019. Rapid Inference of Direct
855 Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial
856 Sequencing Data. Cell Syst 9:286-296.e8.
857 48. Suttle CA. 2007. Marine viruses - Major players in the global ecosystem. Nat Rev
858 Microbiol 5:801–812.
859 49. Blanc-Mathieu R, Dahle H, Hofgaard A, David B, Kalinowski J, Ogata H, Sandaa R-
860 A. 2020. The genome of a persistent giant algal virus encodes an unprecedented
861 number of genes involved in energy metabolism. bioRxiv
862 doi.org/10.1101/2020.07.30.228163.
863 50. Yau S, Krasovec M, Felipe Benites L, Rombauts S, Groussin M, Vancaester E, Aury
864 JM, Derelle E, Desdevises Y, Escande ML, Grimsley N, Guy J, Moreau H, Sanchez-
865 Brosseau S, van de Peer Y, Vandepoele K, Gourbiere S, Piganeau G. 2020. Virus-host
866 coexistence in phytoplankton through the genomic lens. Sci Adv 6:eaay2587.
867 51. Claverie JM, Grzela R, Lartigue A, Bernadac A, Nitsche S, Vacelet J, Ogata H,
868 Abergel C. 2009. Mimivirus and Mimiviridae: Giant viruses with an increasing
869 number of potential hosts, including corals and sponges. J Invertebr Pathol 101:172–
870 180.
871 52. Clouthier SC, Vanwalleghem E, Copeland S, Klassen C, Hobbs G, Nielsen O,
872 Anderson ED. 2013. A new species of nucleo-cytoplasmic large DNA virus (NCLDV)
873 associated with mortalities in Manitoba lake sturgeon Acipenser fulvescens. Dis Aquat
874 Organ 102:195–209.
875 53. Gastrich MD, Leigh-Bell JA, Gobler CJ, Anderson OR, Wilhelm SW, Bryan M. 2004.
876 Viruses as potential regulators of regional brown tide blooms caused by the alga,
877 Aureococcus anophagefferens. Estuaries 27:112–119.
878 54. Duponchel S, Fischer MG. 2019. Viva lavidaviruses! five features of virophages that
879 parasitize giant DNA viruses. PLoS Pathog 15:1–7.
880 55. Xu S, Zhou L, Liang X, Zhou Y, Chen H, Yan S, Wang Y. 2020. Novel Cell-Virus-
881 Virophage Tripartite Infection Systems Discovered in the Freshwater Lake Dishui
882 Lake in Shanghai, China. J Virol 94:e00149-20.
883 56. Sun S, La Scola B, Bowman VD, Ryan CM, Whitelegge JP, Raoult D, Rossmann MG.
884 2010. Structural Studies of the Sputnik Virophage. J Virol 84:894–897.
885 57. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A,

886 Ardyna M, Arkhipova K, Carmichael M, Cruaud C, Dimier C, Domínguez-Huerta G,
887 Ferland J, Kandels S, Liu Y, Marec C, Pesant S, Picheral M, Pisarev S, Poulain J,
888 Tremblay JÉ, Vik D, Acinas SG, Babin M, Bork P, Boss E, Bowler C, Cochrane G, de
889 Vargas C, Follows M, Gorsky G, Grimsley N, Guidi L, Hingamp P, Iudicone D,
890 Jaillon O, Kandels-Lewis S, Karp-Boss L, Karsenti E, Not F, Ogata H, Poulton N,
891 Raes J, Sardet C, Speich S, Stemmann L, Sullivan MB, Sunagawa S, Wincker P,
892 Culley AI, Dutilh BE, Roux S. 2019. Marine DNA Viral Macro- and Microdiversity
893 from Pole to Pole. Cell 177:1109-1123.e14.
894 58. Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, Iudicone D,
895 Karsenti E, Speich S, Trouble R, Dimier C, Searson S. 2015. Open science resources
896 for the discovery and analysis of Tara Oceans data. Sci Data 2:150023.
897 59. de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le
898 Bescot N, Probert I, Carmichael M, Poulain J, Romac S, Colin S, Aury J-M, Bittner L,
899 Chaffron S, Dunthorn M, Engelen S, Flegontova O, Guidi L, Horák A, Jaillon O,
900 Lima-Mendez G, Lukeš J, Malviya S, Morard R, Mulot M, Scalco E, Siano R, Vincent
901 F, Zingone A, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Acinas SG, Bork P,
902 Bowler C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S,
903 Raes J, Sieracki ME, Speich S, Stemmann L, Sunagawa S, Weissenbach J, Wincker P,
904 Karsenti E. 2015. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean.
905 Science 348:1261605.
906 60. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri
907 B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C,
908 D'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F,
909 Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H,
910 Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Boss E, Follows
911 M, Karp-Boss L, Krzic U, Reynaud EG, Sardet C, Sieracki M, Velayoudon D, Bowler
912 C, De Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F,
913 Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P,
914 Karsenti E, Raes J, Acinas SG, Bork P. 2015. Ocean plankton. Structure and function
915 of the global ocean microbiome. Science 348:1261359.
916 61. Matsen FA, Kodner RB, Armbrust EV. 2010. pplacer: linear time maximum-
917 likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference
918 tree. BMC Bioinformatics 11:538.
919 62. De Vargas C, Engelen S, Hingamp P, Sieracki M, Vargas C, Audic S, Henry N,

Decelle J, Mahé F, Logares R, Lara E, Berney C, Bescot N, Probert I, Carmichael M, Poulain J, Romac S. 2015. Eukaryotic plankton diversity in the sunlit ocean. Science 103:167–202.

63. Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, De Vargas C, Decelle J, Del Campo J, Dolan JR, Dunthorn M, Edvardsen B, Holzmann M, Kooistra WHCF, Lara E, Le Bescot N, Logares R, Mahé F, Massana R, Montresor M, Morard R, Not F, Pawlowski J, Probert I, Sauvadet AL, Siano R, Stoeck T, Vaulot D, Zimmermann P, Christen R. 2013. The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res 41:597–604.

64. Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos PM, Stevens MHH, & Wagner H. 2008. The vegan package. Community Ecol Packag 190.

65. Watts SC, Ritchie SC, Inouye M, Holt KE. 2019. FastSpar: Rapid and scalable correlation estimation for compositional data. Bioinformatics 35:1064–1066.

66. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, Vijaykumar A, Bardelli A Pietro, Rothberg A, Hilboll A, Kloeckner A, Scopatz A, Lee A, Rokem A, Woods CN, Fulton C, Masson C, Häggström C, Fitzgerald C, Nicholson DA, Hagen DR, Pasechnik D V., Olivetti E, Martin E, Wieser E, Silva F, Lenders F, Wilhelm F, Young G, Price GA, Ingold GL, Allen GE, Lee GR, Audren H, Probst I, Dietrich JP, Silterra J, Webber JT, Slavič J, Nothman J, Buchner J, Kulick J, Schönberger JL, de Miranda Cardoso JV, Reimer J, Harrington J, Rodríguez JLC, Nunez-Iglesias J, Kuczynski J, Tritz K, Thoma M, Newville M, Kümmerer M, Bolingbroke M, Tartre M, Pak M, Smith NJ, Nowaczyk N, Shebanov N, Pavlyk O, Brodtkorb PA, Lee P, McGibbon RT, Feldbauer R, Lewis S, Tygier S, Sievert S, Vigna S, Peterson S, More S, Pudlik T, Oshima T, Pingel TJ, Robitaille TP, Spura T, Jones TR, Cera T, Leslie T, Zito T, Krauss T, Upadhyay U, Halchenko YO, Vázquez-Baeza Y. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272.

67. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10:1–9.

954 68. Yang WT, Parikh JR, Thomas Stavros A, Otto P, Maislin G. 2018. Exploring the
955 negative likelihood ratio and how it can be used to minimize false-positives in breast
956 imaging. Am J Roentgenol 210:301–306.
957 69. Faust K, Lima-Mendez G, Lerat JS, Sathirapongsasuti JF, Knight R, Huttenhower C,
958 Lenaerts T, Raes J. 2015. Cross-biome comparison of microbial association networks.
959 Front Microbiol 6:1–13.
960 70. Roux S, Adriaenssens EM, Dutilh BE, Koonin E V., Kropinski AM, Krupovic M,
961 Kuhn JH, Lavigne R, Brister JR, Varsani A, Amid C, Aziz RK, Bordenstein SR, Bork
962 P, Breitbart M, Cochrane GR, Daly RA, Desnues C, Duhaime MB, Emerson JB,
963 Enault F, Fuhrman JA, Hingamp P, Hugenholtz P, Hurwitz BL, Ivanova NN, Labonté
964 JM, Lee KB, Malmstrom RR, Martinez-Garcia M, Mizrachi IK, Ogata H, Páez-Espino
965 D, Petit MA, Putonti C, Rattei T, Reyes A, Rodriguez-Valera F, Rosario K, Schriml L,
966 Schulz F, Steward GF, Sullivan MB, Sunagawa S, Suttle CA, Temperton B, Tringe
967 SG, Thurber RV, Webster NS, Whiteson KL, Wilhelm SW, Wommack KE, Woyke T,
968 Wrighton KC, Yilmaz P, Yoshida T, Young MJ, Yutin N, Allen LZ, Kyrpides NC,
969 Eloe-Fadrosh EA. 2019. Minimum information about an uncultivated virus genome
970 (MIUVIG). Nat Biotechnol 37:29–37.
971 71. Katoh K, Misawa K, Kuma KI, Miyata T. 2002. MAFFT: A novel method for rapid
972 multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res
973 30:3059–3066.
974 72. Price MN, Dehal PS, Arkin AP. 2009. Fasttree: Computing large minimum evolution
975 trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641–1650.
976 73. Letunic I, Bork P. 2019. Interactive Tree of Life (iTOL) v4: Recent updates and new
977 developments. Nucleic Acids Res 47:256–259.
978 74. Paul Shannon, Andrew Markiel, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N,
979 Schwikowski B, Ideker T. 2003. Cytoscape: A Software Environment for Integrated
980 Models. Genome Res 13:426.
981 75. Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755–763.
982

983 # Author contributions

984 LM and HO designed the study. LM performed most of the bioinformatics analysis.
985 HE generated the primary NCLDV data. RBM, SC, RHV, and HK contributed to the
986 bioinformatics analysis. All authors contributed to the writing of the manuscript.
987

Materials & Correspondence

Correspondence and material requests should be addressed to HO (email: ogata@kuicr.kyoto-u.ac.jp).

Competing financial interests

The authors declare no competing financial interests.