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

Lingjie Meng[a], Hisashi Endo[a], Romain Blanc-Mathieu[a,b], Rodrigo Hernández-Velázquez[a,c], 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. 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. By taking advantage of the rapidly increasing metagenomic data, *in silico* host prediction approaches are expected to fill this gap between known virus–host relationships and the true but largely unknown amount of NCLDVs. In this study, we built co-occurrence networks between NCLDVs and eukaryotes using the *Tara* Oceans metagenome and metabarcoding datasets to predict virus–host interactions. Using the positive likelihood ratio to assess the performance of host prediction for NCLDVs, we demonstrated that co-occurrence approaches can increase the odds of predicting true positive relationships four-fold compared with random host predictions in the high-weight region (weight > 0.4). To refine the host predictions from high-dimensional co-occurrence networks, we employed a recently proposed phylogeny-based method, Taxon Interaction Mapper, and showed that Taxon Interaction Mapper further improved the prediction performance eight-fold using weight cut-off filtration (> 0.4). Finally, we inferred virophage and NCLDV networks that further corroborated that co-occurrence approaches are effective for predicting NCLDV hosts in marine environments.

Importance

NCLDVs infect a wide range of eukaryotes. Their life circle is less dependent on hosts than other viruses. However, our understanding of NCLDV–host systems is highly limited because few of these viruses have been isolated. Co-occurrence inference is a possible way to predict virus–host interactions. For the first time, we quantitatively assessed the effectiveness of co-occurrence analyses for NCLDV host prediction. We also improved the performance of these co-occurrence analyses with a phylogeny-guided filtering method and produced a list of candidate hosts for three NCLDV families.

Introduction

Nucleocytoplasmic large DNA viruses (NCLDVs) represent a group of double-stranded DNA viruses that belong to the viral phylum *Nucleocytoviricota*, which was previously referred to as Megavirales (Virus Taxonomy: 2019 Release) (1, 2). NCLDVs usually possess diverse gene repertoires (74 to more than 2000 proteins), large genomes (45 kb to 2.5 Mb), and outsized virions (80 nm to 1.5 µm) (3–5). NCLDVs usually 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 the host cells than other viruses (7, 8). To determine what allows these giant viruses to be more independent than other viruses, the first 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, and Dinophyceae (11, 12). Several other non-photosynthetic eukaryotic lineages, such as Bicoecea and Choanoflagellatea, were also reported as experimentally identified NCLDV hosts in marine environments (13, 14). Small to large marine organisms, including invertebrates and vertebrates, are infected by viruses that belong to the NCLDV family *Iridoviridae* (15, 16). These studies together indicate ubiquitous infectious relationships between NCLDVs and a wide range of marine eukaryotes. However, our understanding of NCLDV–host systems is highly limited because few viruses have been isolated.

69 The number of viruses and hosts isolated in the laboratory is effectively a drop in the
70 ocean compared with the amount that truly exists. NCLDVs have been found to be highly
71 diverse and abundant based on omics data (17, 18). In only a few liters of coastal seawater,
72 more than 5,000 *Mimiviridae* species were detected; by comparison, only 20 reference
73 *Mimiviridae* with known hosts have been well investigated (19). Global marine metagenomic
74 data have revealed that the richness and phylogenetic diversity of NCLDVs are even higher
75 than those of an entire prokaryotic domain (20). From biogeographical evidence, it is clear
76 that these viruses are prevalent in the oceanic environment but have a heterogeneous
77 community structure across sizes, depths, and biomes (21). Marine metatranscriptomic data
78 have also shown that NCLDVs are active everywhere in sunlit oceans and may infect hosts
79 from small piconanoplankton (0.8–5 μm) to large mesoplankton (180–2000 μm) (22).
80 Apart from quantity, previous studies also demonstrated that NCLDVs have the
81 potential to infect a greater diversity of hosts than known to date. Gene transfer between
82 NCLDVs and eukaryotes has indicated that these viruses may be able to infect most major
83 eukaryotic lineages (23, 24). NCLDVs might have started coevolving with eukaryotes even
84 before the last eukaryotic common ancestor (LECA) (25). A recent study supported this
85 hypothesis by showing that some NCLDVs encode viractins (actin-related genes in viruses),
86 which could have been acquired from proto-eukaryotes and possibly reintroduced the pre-
87 LECA eukaryotic lineage (26). Together, the findings of high abundance and a long history
88 of coevolution showcase that what is known about NCLDV hosts is just the tip of the iceberg.
89 The real host range of these viruses might be concordant with their known diversity.
90 Therefore, more effort is needed to identify hosts to elucidate the poorly known virus–host
91 relationships and the largely unknown NCLDV world.
92 Substantial effort has been made to reveal the interactions between NCLDVs and
93 their putative hosts. Apart from the classic co-culture method, some other alternative
94 methods, including high-throughput cell sorting, are also used (10, 14). Because of its high
95 efficiency, metagenomics has been increasingly used to investigate the host range of
96 NCLDVs. Genome analyses, such as to identify horizontal gene transfer (HGT), have also
97 performed well for NCLDV host prediction (23, 24).
98 Abundance-based co-occurrence analyses have been used for host prediction and are
99 supposed to be effective because viruses can only thrive in an environment where their hosts
100 exist. In addition to virus–host relationships, this method was applied to show the association
101 between NCLDVs and their "parasites" (virophages) (27). However, co-occurrence-based
102 prediction is controversial for viral host prediction, because the abundance dynamics of

103 viruses (e.g., *Emiliania huxleyi* and *Heterosigma akashiwo* viruses) and their hosts are
104 sometimes not concordant (28, 29). Usually, validation with known virus–host relationships
105 or corroboration with genomic evidence (e.g., HGT) is used to assess network-based
106 predictions (17, 30). However, the effectiveness of co-occurrence network methods has never
107 been quantitatively tested for NCLDV host prediction. The current lack of quantitative
108 assessment hinders the widespread use of this approach. Therefore, methods are needed that
109 test the accuracy of NCLDV host prediction with co-occurrence networks. Effort is also
110 needed to improve the performance of co-occurrence-based prediction.
111 The *Tara* Oceans expedition is a global-scale survey on marine ecosystems that
112 expands our knowledge of microbial diversity, organismal interactions, and ecological
113 drivers of community structure (31). The present study used *Tara* Oceans metagenomic and
114 metabarcoding datasets to predict the virus–host relationships between NCLDVs and
115 eukaryotes by constructing co-occurrence networks. To quantitatively assess the performance
116 of network-based host prediction, we employed the positive likelihood ratio (LR+) for
117 prediction with reference data for known NCLDV–host relationships. The performance of a
118 recently proposed phylogeny-based enrichment analysis approach, Taxon Interaction Mapper
119 (TIM) (32), was also assessed. We further examined NCLDV–virophage networks to
120 investigate the effectiveness of co-occurrence approaches for predicting NCLDV hosts.
121
122
123 # Results
124 **NCLDV–eukaryote co-occurrence networks**
125 From five datasets that corresponded to five size fractions, we generated five co-
126 occurrence networks. Altogether, these networks were composed of 47978 *polB*–V9
127 associations (edges). Out of these associations, 47296 had positive weights, and 682 had
128 negative weights (Fig. 2A). For the nodes, 20148 V9 and 5234 *polB* OTUs were involved in
129 the networks. The associations that involved the family *Mimiviridae* were numerically
130 dominant (n = 36830) among the different NCLDV families. The second most numerous
131 family was *Phycodnaviridae*, with 5521 edges with eukaryotes. No other family had more
132 than 2000 associations with eukaryotes. *Marseilleviridae*, the least numerous family in the
133 networks, had 132 edges with eukaryotes. Taxonomic annotation of eukaryotic OTUs
134 indicated that Alveolata, Opisthokonta, Rhizaria, and Stramenopiles were the major four
135 eukaryotic groups that were connected to NCLDVs (with 21167, 9179, 6521, and 5327
136 edges, respectively). Three of these eukaryotic groups belong to the SAR supergroup (i.e.,

137 Stramenopiles, Alveolata, and Rhizaria), which represented 68.81% of the total associations.
138 Regarding the pairs between viral families and eukaryotic lineages, *Mimiviridae* and
139 Alveolata showed the largest number of edges (n = 16548). Besides NCLDV–eukaryote
140 associations, we detected 57495 *polB–polB* associations and 234448 V9–V9 associations
141 (Fig. S2B). We also included environmental parameters in the network inference and
142 identified 25 pairs of associations between environmental parameters and *polB* OTUs (Table
143 S3).
144 The number of NCLDV–eukaryote associations generally decreased with increasing
145 size fractionation (Fig. S2A). The largest number of *polB*–V9 associations was found in the
146 0.8–5-µm fraction (n = 10647). However, the 0.8–inf-µm size fraction network was the
147 largest (n = 10477) for edges with positive weights. With the annotation of major lineages,
148 the eukaryotic community compositions in the networks varied over different size fractions
149 (Fig. 2B). In the smallest size fraction (0.8–5 µm) and the large range size fraction (0.8–inf
150 µm), Marine Alveolate Group II was the eukaryotic lineage with the largest number of
151 associations with NCLDVs (21.39% and 19.98%, respectively). Dinophyceae was the second
152 largest group connected to NCLDVs in these two size fractions and showed the largest
153 number of connections with NCLDVs in the 5–20-µm size fraction network (22.22% of total
154 interactions). The viral associations with Metazoa and Collodaria increased with increasing
155 size fractions. In the largest 180–2000-µm size fraction network, Metazoa contributed
156 39.31% of the total *polB*–V9 edges.
157 We calculated the degree of nodes (number of connected edges) for each NCLDV
158 *polB* OTU (Fig. 3A, B). Largely consistent with the total number of edges, the average
159 degree of positive associations per *polB* in all size fractions was higher than negative edges
160 and decreased along with increasing size fractions (2.69, 2.40, 2.25, and 2.10 from 0.8–5 µm
161 to 180–2000 µm, and 2.76 for 0.8–inf). Most of the *polB* nodes had more than one positive
162 association (Fig. 3A). Together with the taxonomic annotation of nodes, *polB*–V9
163 associations in the networks generated with the *Tara* Oceans data revealed their high
164 dimensionality and complexity.
165
166 **Network validation**
167 We quantitatively assessed the performance of predicting *polB*–V9 associations using
168 the positive likelihood ratio (LR+) (Fig. 1). By defining groups of metagenomic PolBs as
169 described in the Materials and Methods, 932 OTUs were recruited in the validation, and these
170 sequences contributed 6191 *polB*–V9 associations in the FlashWeave networks (Fig. S3). To

171 obtain an overall performance, we assessed the pooled associations (by removing
172 redundancy) from the five co-occurrence networks. LR+ was separately calculated for edges
173 with positive and negative weights. As shown in Fig. 4A, the LR+ of host prediction for
174 positive associations was higher than 1 (LR+ = 1 indicates no change in the likelihood of the
175 condition). The LR+ was generally higher when the cut-off for the FlashWeave weight was
176 high than when the cut-off was lower, which indicated that condition positive cases are
177 enriched in the edges with high weights. This result demonstrated that the co-occurrence-
178 based host prediction of NCLDVs outperformed random prediction (i.e., random inference of
179 virus–host pairs). In the high-weight regions: 1) weight > 0.6, the LR+ of associations was
180 higher than 10; 2) weight > 0.4, the LR+ was roughly higher than 4. Nonetheless, the false
181 discovery rate (FDR) was high (Fig. S4A), which indicated that the predictions contained
182 numerous virus–host edges that were not considered condition positive. FDR was 96.34%
183 when weight ≥ 0.4. An assessment of the host prediction for negative weight associations was
184 also carried out. There were no known NCLDV–host pairs found in the negative networks
185 (Fig. S4B). The analysis of the remaining part of our study was thus conducted for positive
186 associations.
187 Comparing the performance between different size fractions indicated that the
188 networks of small size fractions (including the 0.8–inf-μm size fraction) performed better in
189 predicting the NCLDV–host relationships (Fig. 4B). The 0.8–inf-μm size fraction had the
190 highest average LR+ out of the five size fractions (LR+ = 4.97). The LR+ of small size
191 fractions was generally higher than that of large size fractions, but there were exceptions
192 between 180–2000 and 20–180 μm. The average LR+ for the associations in the 20–180-μm
193 size fraction network was even lower than 1, which indicated that most of the predicted
194 NCLDV–eukaryote pairs in this size fraction are not yet known. Different from the average
195 results, the high-weight region (> 0.5) showed that the associations of the large size fractions
196 (5–20 and 180–2000 μm) performed better than those of small size fractions (0.8–5 μm) in
197 terms of both LR+ and FDR (Fig. S5 A, B).
198 Comparison of three filtration strategies did not show a consistent pattern in
199 prediction performance. When we compared the pooled associations from all size fractions,
200 the networks from the Q1 filtration strategy performed best (Fig. S6A). However, if we only
201 used the samples in the 0.8–inf-μm size fraction, which was shown to perform best for
202 NCLDV host prediction, the most stringent Q3 strategy performed best (Fig. S6B). Finally,
203 we compared the performance of networks inferred by three different methods: FlashWeave-

204 S (sensitive model), FlashWeave-HE (heterogeneous model), and FastSpar. FlashWeave-S
205 performed best (Fig. S7). Of note, the LR+ of these three methods were all larger than 1.
206
207 **Assessment of host prediction improvement**
208 After the phylogeny-guided TIM filtering of the host prediction, we identified 24
209 eukaryotic taxonomic groups specifically associated with NCLDV (Fig. S8). To compare the
210 performance of the TIM results with the above raw FlashWeave results, we transferred the
211 three primary eukaryotic taxonomic ranks into their associated major lineages (Table S4), and
212 the connections were plotted on a network (Fig. 5A). This network showed that three out of
213 nine NCLDV families (*Mimiviridae, Phycodnaviridae*, and *Iridoviridae*) had enriched
214 connections with eukaryotic lineages. Among the network edges, known virus–host pairs
215 were found, such as Haptophyta–*Mimiviridae*, Mamiellophyceae–*Phycodnaviridae*, and
216 Metazoa–*Iridoviridae*. The associations in the TIM-filtered results showed a sharp
217 improvement in performance from the original result with and without an edge weight cut-
218 off. The average LR+ of TIM-enriched associations was 42.22, which was higher than the
219 raw FlashWeave associations without a weight cut-off (3.43) and with a weight cut-off of 0.4
220 (5.20) (Fig. 5B). The FDR dropped to 0.74 from 0.97 (no cut-off) and 0.95 (weight cut-off of
221 0.4) (Fig. 5C).
222 It was apparent from the network that *Mimiviridae* had a diverse putative host range
223 (12 lineages) of algae, protozoans, and metazoans. Metazoa had the most enriched nodes
224 connected to *Mimiviridae*; additionally, MAST-3, Cryptophyta, Foraminifera, and Ciliophora
225 had strong relationships with *Mimiviridae*. For *Phycodnaviridae*, there were six eukaryotic
226 lineages retained after TIM filtration. Among these, Bacillariophyta, "other filosan (part of
227 filosan Cercozoa)", and Mamiellophyceae had comparatively strong connections. Moreover,
228 Rhodophyta, Ciliophora, and Dictyochophyceae had links to both *Mimiviridae* and
229 *Phycodnaviridae*. There was also a connection between *Iridoviridae* and Metazoa.
230
231 **Associations between virophages and NCLDVs**
232 Using 6818 NCLDV *polB* OTUs and 195 virophage major capsid proteins (MCPs),
233 we identified 535 FlashWeave associations (196 and 339 for pico- and femto-size fractions,
234 respectively) (Fig. 6A). Most of the associations had positive weights (n = 490), whereas
235 some had negative weights (n = 45). The average number of associations per virophage MCP
236 was different in two size fractions: 3.2 in femto- and 5.6 in pico-size fractions. The network
237 revealed that *Mimiviridae* had the largest number of virophage associations in both size

238 fractions. We also detected 84 positive associations between virophages and
239 *Phycodnaviridae*.
240 The phylogenetic tree defined three main virophage clades, and they were all found to
241 have many connections to NCLDVs. To investigate significant relationships, Fisher's exact
242 test was performed between virophage clades and NCLDV families. Families other than
243 *Phycodnaviridae* and *Mimiviridae* did not show significant associations. Therefore, we made
244 a group, "Other NCLDVs," to include all families except *Phycodnaviridae* and *Mimiviridae*.
245 First, we only used FlashWeave results with a weight > 0.4, as previous results showed that a
246 FlashWeave weight of 0.4 is a suitable cut-off that produced at least moderate performance
247 (Fig. 3A). From the femto-size fraction network, we found two significantly enriched
248 connections (Fig. 6B): one was between virophage group C and *Mimiviridae* ($p = 0.0022$)
249 and the other was between group A and "Other NCLDVs" ($p = 0.0439$). Another
250 significantly enriched relationship between virophage group B and *Phycodnaviridae* ($p =$
251 0.0410) was found in pico-size fractions when we used all associations without edge weight
252 cut-off.
253 Finally, we examined HGTs of virophage MCPs in NCLDV genomes. We found two
254 HGTs of virophage MCPs; both showed links between clade A and *Iridoviridae* (Table S3).
255 This result was consistent with Fisher's exact test, which revealed a connection between
256 virophage clade A and "Other NCLDVs", which includes *Iridoviridae*.
257

258 # Discussion

259 NCLDVs can infect a wide range of eukaryotes, from unicellular to multicellular
260 organisms (33). However, we are still far from a comprehensive understanding of their hosts
261 because few have been isolated. Therefore, to better understand the ecological functions and
262 evolutionary significance of NCLDVs, reasonable host prediction is needed. To make these
263 predictions, we constructed global co-occurrence networks based on the marine metagenome
264 and metabarcoding datasets from 85 stations of the *Tara* Oceans expedition, which cover all
265 major oceanic provinces across an extensive latitudinal gradient from pole to pole. The edges
266 (associations) between *polB* and V9 nodes (OTUs) in the networks were generated by
267 FlashWeave. The networks were huge (Fig. 2A, S2A), suggesting that numerous NCLDVs
268 interact with eukaryotes in the ocean. This was expected given the high abundance and
269 diversity of NCLDVs in marine environments (17, 20) and the identification of HGT between
270 these viruses and diverse eukaryotic lineages (23). The networks were dominated by the
271 *Mimiviridae* nodes. This result is consistent with previous reports that *Mimiviridae* is the

272 most abundant and has the widest array of transcribed genes out of NCLDV families in
273 marine environments (21, 22). *Mimiviridae* was known to infect amoebae, algae, and
274 stramenopiles (3). In our study, these three eukaryotic groups were all found to have
275 numerous associations with *Mimiviridae*. *Phycodnaviridae* has been known to infect many
276 species of aquatic organisms, such as *Emiliania huxleyi* (Haptophyta), *Ectocarpus siliculosus*
277 (Phaeophyceae), *Chlorella heliozoae* (Trebouxiophyceae), and *Ostreococcus tauri*
278 (Mamiellophyceae) (34–36). Correspondingly, plenty of associations of *Phycodnaviridae*
279 were found in the co-occurrence networks. For the eukaryotic nodes, all high taxonomic rank
280 groups, including the SAR supergroup (i.e., Stramenopiles, Alveolata, Rhizaria),
281 Opisthokonta, Archaeplastida, Amoebozoa, Excavata, and other eukaryotes, have
282 associations with NCLDVs. Among these groups, the SAR supergroup contributed the most
283 (~68%) *polB*–V9 associations. However, this is still lower than in other microbial co-
284 occurrence analyses; for example, in aquatic microbial interactions, they dominated ~92% of
285 the total associations (37). A substantially proportion (~32%) of NCLDV–eukaryote
286 interactions were from non-SAR groups, which covered the known NCLDV host range, such
287 as Archaeplastida and Haptophyta. Our result provides a list of candidate eukaryotes for
288 NCLDV host prediction.
289 However, it is difficult to directly predict NCLDV hosts from constructed networks
290 because of the high degree of associations per *polB* OTU (Fig. 3A). One node connected to
291 multiple edges was expected in the co-occurrence analysis. In previous NCLDV host
292 prediction studies, additional processing was performed to filter the high dimensional
293 associations to predict the meaningful interactions, such as weight cut-off or a combination of
294 different co-occurrence network inference methods (17, 27). Moreover, no previous study
295 quantitively assessed the performance of co-occurrence networks when predicting NCLDV–
296 host relationships. Qualitatively identifying known pairs and detecting HGT (without
297 validation) have been commonly used to assess prediction reliability (17, 30). Therefore, we
298 aimed to 1) quantitatively assess the performance of co-occurrence-based host prediction for
299 NCLDVs and 2) improve the prediction results using filtering methods.
300 In a previous study of bacteriophage host prediction, ROC curves were used as an
301 assessment metric to compare different prediction methods (38). However, the number of
302 known virus–host pairs of NCLDVs is not sufficient to generate a dataset for ROC
303 assessment. Therefore, in this study, we carried out an alternative method, the LR+, to assess
304 the performance. LR+ is calculated with two relative values, sensitivity and specificity (Fig.
305 1). The LR+ of co-occurrence-based host predictions for positive associations was higher

306 than 1 and increased along with increasing cut-off values for the edge weights (Fig. 4A),
307 which demonstrated that positive prediction results were more likely to be true positives than
308 those based on random prediction. In high-weight regions (> 0.6 and > 0.4), LR+ values were
309 larger than 10 and 4, respectively. These LR+ values indicate that FlashWeave can increase
310 the probability of predicting true positives (39). However, both the true positive rate (<
311 18.9%) and false positive rate (< 6.37%) were very low. These low rates were from
312 FlashWeave with a cut-off of alpha < 0.01 (as described in the Materials and Methods),
313 which excluded a large proportion of the *polB*–V9 pairs from the results. The FDR of co-
314 occurrence was even higher than 90% (Fig. S4A). Such a high FDR in co-occurrence
315 networks demonstrates that condition positive connections (i.e., known interactions) are
316 embedded in an immense pool of condition negative connections. However, these negative
317 signals can correspond to unidentified (i.e., currently unknown true interactions), indirect, or
318 false relationships.
319 We also found that true positive predictions only existed in positive weight
320 associations, whereas negative weight associations did not contribute to NCLDV–host
321 detection (Fig. S4B). This result indicates that the abundance dynamics of NCLDVs and their
322 potential hosts were positively correlated with each other in the analyzed samples, which
323 were collected at a global scale; this might be because NCLDVs detected in the dataset were
324 active viruses that replicate locally in their hosts. Similar results were obtained in other co-
325 occurrence-based host prediction studies (30, 40). However, several experimental studies
326 showed that the abundance dynamics of NCLDVs and hosts showed a delay in time (28, 29).
327 It is possible that the global-scale samples did not have sufficiently high resolution to detect
328 negative correlations (or correlations with a time delay) due to lack of time-resolution (41).
329 Therefore, further studies, especially those that focus on a high temporal resolution, are
330 needed to better understand the detailed dynamics of virus–host associations and the capacity
331 of co-occurrence-based methods for host prediction.
332 The networks of different size fractions showed different performance patterns in
333 predicting NCLDV–host relationships (Fig. 4B). Generally, the associations from small-sized
334 networks performed better than the associations from large-sized fractions. To date, most of
335 the known NCLDV hosts are small, such as the genera *Prymnesium* (10 μm), *Aureococcus* (2
336 μm), and *Ostreococcus* (1 μm). Because of this, our assessment method might be biased
337 toward small size fractions. However, it is also possible that NCLDV infections are more
338 prevalent in smaller size fractions. Notably, the 0.8–inf-μm size fraction network, which

339 covered all four individual size fractions, performed best. This might be because NCLDVs
340 can infect not only small hosts but also hosts from a broad size range.
341 According to a previous study, trimming of low-abundance OTUs was recommended
342 to improve the prediction of true interactions (42). In our study, however, we did not achieve
343 such performance improvement by treating input abundance data with a rigorous filtration
344 (upper quartile) (Fig. S6). This result might be because the true positive and false positive
345 rates defined in this study were too low; therefore, the validation may not be sufficiently
346 sensitive to reflect the change between different abundance trimming strategies. However, it
347 is also possible that low-abundance NCLDV OTUs are indeed network participants, as was
348 demonstrated in a study showing that rare cyanobacterial species might play fundamental
349 roles in blooming (43). Our result also revealed that FlashWeave-S was better than
350 FlashWeave-HE and FastSpar at predicting NCLDV–host interactions (Fig, S7). The
351 difference between FlashWeave-HE and FlashWeave-S is that HE mode can remove
352 structural zeros during network inference. Structural zero is a typical property of
353 heterogeneous datasets, like *Tara* Oceans datasets, and may lead to false-positive edges (44).
354 Conversely, our results demonstrated that retaining structural zeros helped improve NCLDV–
355 eukaryote prediction, which indicates that the "presence–absence" pattern is more
356 informative than the "more–less" pattern when identifying NCLDV–host relationships. Our
357 result is consistent with a previous "K-r-strategist" hypothesis: NCLDVs, like mimiviruses,
358 are K-strategists that decay slowly and can form stable associations with their hosts (45). A
359 recent report supported these non-"boom and bust" dynamics of prasinoviruses and their
360 hosts with an experiment-based mathematical model (46). Overall, all these results show that
361 co-occurrence is an acceptable method for predicting NCLDV–host interactions in the marine
362 metagenome, and the LR+ is an effective quantitative metric for assessing the performance of
363 co-occurrence analysis.
364 Although the results generated by FlashWeave already increased the probability of
365 prediction, the condition positive interactions were still embedded in many noise edges, as
366 shown by the high FDR (Fig. S4A, S5B). We used a phylogeny-guided and network-based
367 association enrichment tool, TIM, to reduce the high dimension of associations and improve
368 NCLDV host prediction (32). The results showed that NCLDVs had enriched connections
369 with 14 major eukaryotic lineages, which included 24 taxonomic groups in three different
370 ranks (order, class, and phylum) (Table S4) (Fig. 5A, S7). Using the LR+ as a metric,
371 NCLDV host prediction improved 12-fold with TIM filtration (Fig. 5B). FDR dropped below
372 23% after TIM treatment (Fig. 5C). In TIM-enriched connections, some are known NCLDV–

host pairs, such as *Phycodnaviridae* and Mamiellophyceae, *Mimiviridae* and Haptophyta, and *Iridoviridae* and Metazoa. Some other studies revealed that *Mimiviridae* can exclusively infect diverse putative hosts (23, 47). Our results support this assumption that *Mimiviridae* has connections with 12 eukaryotic lineages out of 14 total lineages. Among the lineages, *Mimiviridae* had the most numerous links to Metazoa. Some mimiviruses (namaoviruses) are known to infect freshwater sturgeon, *Acipenser fulvescens* (48). Metazoans are presumed to be susceptible to mimiviruses, because the choanoflagellates, a group of eukaryotes that is phylogenetically close to metazoans, were recently proven to be the host of a species of *Mimiviridae* (14). Moreover, the TIM result revealed that *Phycodnaviridae* is closely connected to Bacillariophyta, which consists of three NCBI taxonomic groups: Thalassiophysales, Cymbellales, and Bacillariophyceae. Thalassiophysales was shown to have many HGT candidates with a large range of NCLDVs, and Bacillariophyceae also has a significant HGT candidate with phaeoviruses (23). Although Dictyochophyceae itself has not been proven to be a phycodnavirus host, its sister group Pelagophyceae was experimentally identified as an AaV host (49). Additionally, it is interesting to note the connection between Metazoa (Calanoida) and *Iridoviridae*. Calanoida is an order of arthropods commonly found as zooplankton; most of the sizes are 500–2000 µm. The viruses of the family *Iridoviridae* infect many Arthropod species, including insects and crustaceans (16).

Furthermore, we constructed the connections between virophages and NCLDVs. To date, all isolated virophages are only known to infect *Mimiviridae* (50). As expected, *Mimiviridae* was the family with dominant connections to virophages (Fig. 6A). Recently, *in silico* evidence demonstrated that virophages can infect *Phycodnaviridae*, which indicated that the virophage host range might be more extensive than we know (51). Consistently, a relatively large number of virophage OTUs were found to be associated with *Phycodnaviridae* in our study. The enrichment analysis also revealed significant connections between three virophage clades and NCLDV families (Fig. 6B). To support the enrichment analysis, we conducted an HGT analysis because gene transfers have previously been found between Sputnik virophages and giant viruses (52). Our HGT analysis indicated a previously unknown infectious relationship between virophage clade A and *Iridoviridae*. Overall, the results of virophage–NCLDV associations support our previous statement that co-occurrence is an appropriate tool for investigating NCLDV interactions in marine metagenomic data.

Materials and Methods

Metagenomic and metabarcoding data

407 The microbial metagenomic and eukaryotic metabarcoding data used in this study
408 were previously generated from plankton samples collected by the *Tara* Oceans expedition
409 from 2008 to 2013 (53, 54). Because our research requires paired metagenomic and
410 metabarcoding datasets, we used data derived from the euphotic zone samples, namely those
411 from the surface (SRF) and Deep Chlorophyll Maximum (DCM) layers (55). Type B DNA
412 polymerase (*polB*) was used as the marker gene for NCLDVs. A total of 6818 NCLDV *polB*
413 OTUs were extracted from the metagenomic datasets (i.e., the second version of the Ocean
414 Microbial Reference Gene Catalog, OM-RGC.v2) using the pplacer phylogenetic placement
415 method (56, 57). These *polB* sequences were classified into seven NCLDV families
416 (*Mimiviridae*, *Phycodnaviridae*, *Marseilleviridae*, *Ascoviridae*, *Iridoviridae*, *Asfarviridae*,
417 and *Poxviridae*) and two other giant virus groups ("Medusaviridae" and "Pithoviridae"). We
418 employed the metabarcoding data for eukaryotes, which targets the 18S ribosomal RNA gene
419 hypervariable V9 region (V9) (58). Taxonomic annotation of the eukaryotic metabarcoding
420 data was previously performed by the *Tara* Oceans consortium with an extensive V9_PR2
421 reference database, which was derived from the original Protist Ribosomal Reference (PR2)
422 database.
423
424 **Data processing**
425 A relative abundance matrix for the NCLDV *polB* OTUs was extracted from OM-
426 RGC.v2 for the samples derived from the pico-size fractions (0.22–1.6 or 0.22–3.0 μm). We
427 converted the relative abundances of *polB* OTUs back to absolute read counts based on gene
428 length and read length (assumed to be 100 nt). This process was required because small
429 decimal numbers cannot be used by FlashWeave and because relative abundance data suffer
430 from apparent correlations, which reduce the specificity of co-occurrence networks in
431 revealing microbial interactions (42). To build comprehensive interaction networks involving
432 eukaryotes of different sizes, we extracted the V9 read count matrices from the
433 metabarcoding dataset for the following five size fractions: 0.8–5 μm; 5–20 μm and 3–20 μm
434 (hereafter referred to as "5–20 μm" for simplicity); 20–180 μm; 180–2000 μm; and > 0.8 μm
435 (hereafter referred to as 0.8–inf μm). To create the input files for network inference, the *polB*
436 matrix was combined with each of the V9 matrices (corresponding to different size fractions),
437 and only the samples represented by both *polB* and V9 files were placed in new files. In total,
438 samples from 84 *Tara* Oceans stations (a total of 560 samples for two depths and five size
439 fractions) widely distributed across oceans were used in this study (Fig. S1A). Depending on
440 the individual size fractions, 84–127 samples were retained and included in the co-occurrence

441 analysis (Fig. S1B). Read counts in the newly generated matrices were normalized using
442 centered log-ratio (*clr*) transformation after adding a pseudo count of one to all matrix
443 elements because zero cannot be transformed in *clr.* Following *clr* normalization, we filtered
444 out low-abundance OTUs with a lower quartile (Q1) filtering approach. Specifically, OTUs
445 were retained in the matrices when their *clr*-normalized abundance was higher than Q1
446 (among the non-zero counts in the original count matrix prior to the addition of a pseudo
447 count of one) in at least five samples. Normalization and filtering were separately applied to
448 *polB* and V9. The numbers of OTUs in the final matrices are provided in Fig. S1C.
449
450 **Co-occurrence-based network inference**
451 Network inference was performed using FlashWeave [v0.15.0 (44)]. FlashWeave is a
452 fast and compositionally robust tool for ecological network inference based on the local-to-
453 global learning framework. Meta-variables (such as environmental parameters) can be
454 included in the FlashWeave network to remove potential indirect associations. We used
455 temperature, salinity, nitrate, phosphate, and silicate concentrations as meta-variables in our
456 network inferences to determine their correlations with *polB* OTUs. FlashWeave has a
457 heterogeneous mode (FlashWeave-HE), which helps overcome sample heterogeneity.
458 However, FlashWeave-HE was not considered appropriate for the *Tara* Oceans data because
459 it was shown to predict an insufficient number of known planktonic interactions (44).
460 Therefore, we mainly used FlashWeave-S with default settings except for the FlashWeave
461 normalization step and comparison between FlashWeave-S and FlashWeave-HE. A threshold
462 to determine the statistical significance was set to alpha < 0.01. All detected pairwise
463 associations were assigned a value called "weight" that ranged between −1 and +1. Edges
464 with weights > 0 or < 0 were referred to as positive and negative associations, respectively.
465 To compare the performance of FlashWeave-S to other co-occurrence methods, we used
466 FlashWeave-HE and FastSpar (59). The FlashWeave-HE settings were the same as
467 FlashWeave-S but with a command "heterogeneous." In FastSpar, we used 50 iterations, 20
468 excluded iterations, and a threshold of 0.1 to generate associations. To reduce the high
469 dimensionality of the datasets, upper quartile (Q3) filtered matrices were used for
470 comparison.
471
472 **Network validation**
473 We validated the virus–host associations in inferred networks based on a confusion
474 matrix defined by the known NCLDV–host information (Fig. 1). Briefly, we manually

475 compiled 69 known virus–host relationships for NCLDVs (Table S1). In the validation
476 process, eukaryotic taxonomic groups were annotated at the level of the "Major lineages" in
477 the extensive PR2 database (updated after publication) (58). The "Major lineages" were used
478 in the present study because 1) the deficiency of known virus–host relationships limited the
479 use of lower eukaryotic taxonomy ranks, such as genus, for assessment, and 2) these lineages
480 adequately represented marine eukaryotes by covering the full spectrum of cataloged
481 eukaryotic V9 diversity at a comparable phylogenetic depth (58). Then, we performed
482 BLASTp [2.10.1 (60)] searches from the *Tara* Oceans PolB sequences against the NCLDV
483 reference database to define groups of metagenomic PolBs with a threshold of 65% sequence
484 identity by retaining only the best hit for each environmental PolB sequence. This threshold
485 was determined because, by using reference PolB sequences and RefSeq protein sequence
486 databases, we found that 60–70% of sequence identity could distinguish whether the
487 NCLDVs infected hosts of the same major lineages; this was mainly tested for
488 *Phycodnaviridae* because of the lack of host information for closely related viruses in other
489 NCLDV families (Table S2). Then, 65% was chosen because it could provide a better LR+
490 (as described below) than 60% and 70%.
491 The likelihood ratio was used in for assessment. This approach is commonly used in
492 diagnostic testing to assess if a test (host prediction in this study) usefully changes the
493 probability of the existence of condition positive (true positive). In this study, the LR+ was
494 used because host prediction is a test to discover condition positive states (61). LR+ is
495 calculated by dividing the true-positive rate (sensitivity) by the false-positive rate (1 –
496 specificity). If LR+ is close to 1, the performance of the prediction is comparable to a random
497 prediction. If LR+ >> 1, a positive prediction result is more likely to be a true positive than
498 that based on random prediction. From the set of detected associations between a given *polB*
499 OTU and V9 OTUs that belong to a given major eukaryotic lineage, we only kept the best
500 positive and negative associations (i.e., the edges with the highest absolute weights) to
501 simplify the prediction scheme. As an auxiliary assessment, the FDR was also calculated by
502 dividing the number of false positives by the number of positive predictions (Fig. 1).
503
504 **Assessment of phylogeny-guided filtering for host prediction**
505 We assessed the performance of TIM (https://github.com/RomainBlancMathieu/TIM)
506 (32) in improving host predictions by the FlashWeave co-occurrence approach. TIM is a
507 recently proposed tool for assigning the most likely associations based on the assumption that
508 evolutionarily related viruses tend to infect evolutionarily related hosts. TIM requires a

509 phylogenetic tree of viruses and a set of connections between viruses and eukaryotes, and
510 then tests whether a node of the virus tree is enriched with a specific predicted host group
511 compared with the rest of the tree using Fisher's exact test and Benjamini–Hochberg
512 adjustment (32).
513 We pooled network associations using FlashWeave analysis for five size fractions. To
514 build a concise and credible viral phylogenetic tree, we removed all of the PolB sequences
515 that were absent in the FlashWeave network associations, and the remaining sequences were
516 filtered by the amino acid sequence length (≥ 500 aa). Protein alignment was conducted using
517 MAFFT-*linsi* [version 7.471 (62)], and 18 sequences were manually removed because they
518 were not well aligned with other PolB sequences. A total of 501 PolB sequences were used to
519 make a maximum likelihood phylogenetic tree with FastTree [version 2.1.11 (63)]. Then, the
520 PolB–V9 associations were mapped on the tree to calculate the significance of the enrichment
521 of specific associations using TIM. TIM provides a list of nodes in the viral tree and
522 associated NCBI taxonomies (order, class, and phylum) of eukaryotes that show a significant
523 enrichment in the leaves under the nodes. The TIM result was visualized with iTOL [version
524 5 (64)]. The TIM result was converted to a network, in which nodes correspond to the major
525 eukaryotic lineages. The network edge weight was defined by the number of tree nodes in
526 each viral family subtree enriched with a specific major eukaryotic lineage. The network was
527 visualized with Cytoscape [version 3.7.1] using prefuse force directed layout (65). To assess
528 the effectiveness of TIM in improving prediction, we extracted all the associations predicted
529 by TIM and compared their performance with the associations without TIM filtration.
530
531 **Virophage–NCLDV associations**
532 We inferred the networks between NCLDVs and virophages. *mcp* was used as the
533 marker gene of virophages. First, 47 reference MCP amino acid sequences were collected
534 from public databases and used to build an HMM profile. The HMM profile was used to
535 search against the amino acid sequences of OM-RGC v2 using HMMER hmmsearch [version
536 3.3.1] with the threshold of E-value < 1E−90 (66). This threshold was determined by
537 searching reference sequences against the GenomeNet nr-aa database. The search detected
538 195 *Tara* Oceans virophage MCP sequences in the OM-RGC database. Together with 47
539 reference MCPs, a phylogenetic tree of MCP amino acid sequences was built using MAFFT
540 and FastTree.
541 We extracted the abundance profiles for the 195 MCP sequences from the pico-
542 (0.22–1.6 or 0.22–3.0 μm) and femto-size (< 0.22 μm) fractions. We used samples from the

SRF and DCM depths. PolB and MCP abundance profiles were merged into two matrices corresponding to the two virophage size fractions. Then, network inference was conducted using the FlashWeave default settings after Q1 filtration. In the MCP phylogenetic tree, three virophage clades contributed most of the NCLDV connections. Thus, an NCLDV enrichment analysis for the three clades was carried out using Fisher's exact test, and the *p*-value was adjusted by the Benjamin–Hochberg method. This approach was the same as TIM, but we did not use the TIM software because TIM only works for eukaryotic nodes with NCBI taxonomy annotations.

We used another approach, HGT, to predict the virophage-NCLDV interactions. First, we generated an NCLDV genome database, which includes 56 reference NCLDV genomes corresponding to our *polB* dataset and 2074 metagenome-assembled genomes from a previous study (23). A total of 827548 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.

References

577 1. Colson P, De Lamballerie X, Yutin N, Asgari S, Bigot Y, Bideshi DK, Cheng XW,
578 Federici BA, Van Etten JL, Koonin E V., La Scola B, Raoult D. 2013. "Megavirales",
579 a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Arch Virol
580 158:2517–2521.
581 2. Koonin E V., Dolja V V., Krupovic M, Varsani A, Wolf YI, Yutin N, Zerbini FM,
582 Kuhn JH. 2020. Global Organization and Proposed Megataxonomy of the Virus
583 World. Microbiol Mol Biol Rev 84:e00061-19.
584 3. Koonin E V., Yutin N. 2019. Evolution of the Large Nucleocytoplasmic DNA Viruses
585 of Eukaryotes and Convergent Origins of Viral Gigantism. Advances in Virus
586 Research 103:167-202.
587 4. Legendre M, Alempic JM, Philippe N, Lartigue A, Jeudy S, Poirot O, Ta NT, Nin S,
588 Couté Y, Abergel C, Claverie JM. 2019. Pandoravirus celtis illustrates the
589 microevolution processes at work in the giant Pandoraviridae genomes. Front
590 Microbiol 10:1–11.
591 5. Boratto PVM, Oliveira GP, Machado TB, Andrade ACSP, Baudoin J-P, Klose T,
592 Schulz F, Azza S, Decloquement P, Chabrière E, Colson P, Levasseur A, La Scola B,
593 Abrahão JS. 2020. Yaravirus: A novel 80-nm virus infecting Acanthamoeba
594 castellanii. Proc Natl Acad Sci 117:16579–16586.
595 6. Yoshikawa G, Blanc-Mathieu R, Song C, Kayama Y, Mochizuki T, Murata K, Ogata
596 H, Takemura M. 2019. Medusavirus, a Novel Large DNA Virus Discovered from Hot
597 Spring Water. J Virol 93:1–3.
598 7. Forterre P. 2011. Manipulation of cellular syntheses and the nature of viruses: The
599 virocell concept. Comptes Rendus Chim 14:392–399.
600 8. Mougari S, Sahmi-Bounsiar D, Levasseur A, Colson P, Scola B La. 2019. Virophages
601 of giant viruses: An update at eleven. Viruses 11:1–28.
602 9. Koonin E V., Yutin N. 2010. Origin and evolution of eukaryotic large nucleo-
603 cytoplasmic DNA viruses. Intervirology 53:284–292.
604 10. La Scola B, Audic S, Robert C, Jungang L, De Lamballerie X, Drancourt M, Birtles R,
605 Claverie JM, Raoult D. 2003. A giant virus in amoebae. Science 299:2033.
606 11. Nagasaki K, Yamaguchi M. 1997. Isolation of a virus infectious to the harmful bloom
607 causing microalga Heterosigma akashiwo (Raphidophyceae). Aquat Microb Ecol
608 13:135–140.
609 12. Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH,
610 Wilhelm SW. 2014. Genome of brown tide virus (AaV), the little giant of the

611 Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution.
612 Virology 466–467:60–70.
613 13. Colson P, Gimenez G, Boyer M, Fournous G, Raoult D. 2011. The giant Cafeteria
614 roenbergensis virus that infects a widespread marine phagocytic protist is a new
615 member of the fourth domain of life. PLoS One 6:13–17.
616 14. Needham DM, Yoshizawa S, Hosaka T, Poirier C, Choi CJ, Hehenberger E, Irwin
617 NAT, Wilken S, Yung CM, Bachy C, Kurihara R, Nakajima Y, Kojima K, Kimura-
618 Someya T, Leonard G, Malmstrom RR, Mende DR, Olson DK, Sudo Y, Sudek S,
619 Richards TA, DeLong EF, Keeling PJ, Santoro AE, Shirouzu M, Iwasaki W, Worden
620 AZ. 2019. A distinct lineage of giant viruses brings a rhodopsin photosystem to
621 unicellular marine predators. Proc Natl Acad Sci U S A 116:20574–20583.
622 15. Williams T. 2008. Natural Invertebrate Hosts of Iridoviruses (Iridoviridae). Neotrop
623 Entomol 37:615–632.
624 16. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, Hingamp
625 P, Goto S, Ogata H. 2016. Linking virus genomes with host taxonomy. Viruses 8:10–
626 15.
627 17. Hingamp P, Grimsley N, Acinas SG, Clerissi C, Subirana L, Poulain J, Ferrera I,
628 Sarmento H, Villar E, Lima-Mendez G, Faust K, Sunagawa S, Claverie JM, Moreau
629 H, Desdevises Y, Bork P, Raes J, De Vargas C, Karsenti E, Kandels-Lewis S, Jaillon
630 O, Not F, Pesant S, Wincker P, Ogata H. 2013. Exploring nucleo-cytoplasmic large
631 DNA viruses in Tara Oceans microbial metagenomes. ISME J 7:1678–1695.
632 18. Monier A, Claverie JM, Ogata H. 2008. Taxonomic distribution of large DNA viruses
633 in the sea. Genome Biol 9:R106.
634 19. Li Y, Hingamp P, Watai H, Endo H, Yoshida T, Ogata H. 2018. Degenerate PCR
635 primers to reveal the diversity of giant viruses in coastal waters. Viruses 10:496.
636 20. Mihara T, Koyano H, Hingamp P, Grimsley N, Goto S, Ogata H. 2018. Taxon richness
637 of "Megaviridae" exceeds those of bacteria and archaea in the ocean. Microbes
638 Environ 33:162–171.
639 21. Endo H, Blanc-Mathieu R, Li Y, Salazar G, Henry N, Labadie K, de Vargas C,
640 Sullivan MB, Bowler C, Wincker P, Karp-Boss L, Sunagawa S, Ogata H. 2020.
641 Biogeography of marine giant viruses reveals their interplay with eukaryotes and
642 ecological functions. Nat Ecol Evol https://doi.org/10.1038/s41559-020-01288-w.
643 22. Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R,
644 Lima-Mendez G, Rocha F, Tirichine L, Labadie K, Kirilovsky A, Bertrand A, Engelen

645 S, Madoui MA, Méheust R, Poulain J, Romac S, Richter DJ, Yoshikawa G, Dimier C,
646 Kandels-Lewis S, Picheral M, Searson S, Acinas SG, Boss E, Follows M, Gorsky G,
647 Grimsley N, Karp-Boss L, Krzic U, Pesant S, Reynaud EG, Sardet C, Sieracki M,
648 Speich S, Stemmann L, Velayoudon D, Weissenbach J, Jaillon O, Aury JM, Karsenti
649 E, Sullivan MB, Sunagawa S, Bork P, Not F, Hingamp P, Raes J, Guidi L, Ogata H,
650 De Vargas C, Iudicone D, Bowler C, Wincker P. 2018. A global ocean atlas of
651 eukaryotic genes. Nat Commun 9:373.
652 23. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, McMahon KD,
653 Konstantinidis KT, Eloe-Fadrosh EA, Kyrpides NC, Woyke T. 2020. Giant virus
654 diversity and host interactions through global metagenomics. Nature 578:432–436.
655 24. Gallot-Lavallée L, Blanc G. 2017. A glimpse of nucleo-cytoplasmic large DNA virus
656 biodiversity through the eukaryotic genomicswindow. Viruses 9:17.
657 25. Guglielmini J, Woo AC, Krupovic M, Forterre P, Gaia M. 2019. Diversification of
658 giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes.
659 Proc Natl Acad Sci U S A 116:19585–19592.
660 26. Cunha V Da, Gaïa M, Ogata H, Jaillon O, Delmont TO, Forterre P. 2020. Giant
661 viruses encode novel types of actins possibly related to the origin of eukaryotic actin:
662 the viractins. bioRxiv doi.org/10.1101/2020.06.16.150565.
663 27. Roux S, Chan LK, Egan R, Malmstrom RR, McMahon KD, Sullivan MB. 2017.
664 Ecogenomics of virophages and their giant virus hosts assessed through time series
665 metagenomics. Nat Commun 8:858.
666 28. Tomaru Y, Katanozaka N, Nishida K, Shirai Y, Tarutani K, Yamaguchi M, Nagasaki
667 K. 2004. Isolation and characterization of two distinct types of HcRNAV, a single-
668 stranded RNA virus infecting the bivalve-killing microalga Heterocapsa
669 circularisquama. Aquat Microb Ecol 34:207–218.
670 29. Martínez Martínez J, Schroeder DC, Larsen A, Bratbak G, Wilson WH. 2007.
671 Molecular dynamics of Emiliania huxleyi and cooccurring viruses during two separate
672 mesocosm studies. Appl Environ Microbiol 73:554–562.
673 30. Moniruzzaman M, Wurch LL, Alexander H, Dyhrman ST, Gobler CJ, Wilhelm SW.
674 2017. Virus-host relationships of marine single-celled eukaryotes resolved from
675 metatranscriptomics. Nat Commun 8:16054.
676 31. Sunagawa S, Acinas SG, Bork P, Bowler C, Eveillard D, Gorsky G, Guidi L, Iudicone
677 D, Karsenti E, Lombard F, Ogata H, Pesant S, Sullivan MB, Wincker P, de Vargas C.
678 2020. Tara Oceans: towards global ocean ecosystems biology. Nat Rev Microbiol

679 18:428–445.
680 32. Blanc-Mathieu R, Kaneko H, Endo H, Chaffron S, Hernández-Velázquez R, Nguyen
681 CH, Mamitsuka H, Henry N, De Vargas C, Sullivan MB, Suttle CA, Guidi L, Ogata H.
682 2019. Viruses of the eukaryotic plankton are predicted to increase carbon export
683 efficiency in the global sunlit ocean. bioRxiv https://doi.org/doi.org/10.1101/710228.
684 33. Fischer MG. 2016. Giant viruses come of age. Curr Opin Microbiol 31:50–57.
685 34. Delaroque N, Wolf S, Muller DG, Knippers R. 2000. The brown algal virus EsV-1
686 particle contains a putative hybrid histidine kinase. Virology 273:383–390.
687 35. Fitzgerald LA, Graves M V., Li X, Hartigan J, Pfitzner AJP, Hoffart E, Van Etten JL.
688 2007. Sequence and annotation of the 288-kb ATCV-1 virus that infects an
689 endosymbiotic chlorella strain of the heliozoon Acanthocystis turfacea. Virology
690 362:350–361.
691 36. Clerissi C, Desdevises Y, Grimsley N. 2012. Prasinoviruses of the Marine Green Alga
692 Ostreococcus tauri Are Mainly Species Specific. J Virol 86:4611–4619.
693 37. Bjorbækmo MFM, Evenstad A, Røsæg LL, Krabberød AK, Logares R. 2020. The
694 planktonic protist interactome: where do we stand after a century of research? ISME J
695 14:544–559.
696 38. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. 2016. Computational approaches
697 to predict bacteriophage-host relationships. FEMS Microbiol Rev 40:258–272.
698 39. Salkić NN. 2008. Objective assessment of diagnostic tests validity : a short review for
699 clinicians and other mortals . Part II. Acta Med Acad 39–42.
700 40. Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard
701 CPD, Dutilh BE, Thompson FL. 2017. Marine viruses discovered via metagenomics
702 shed light on viral strategies throughout the oceans. Nat Commun 8:1–12.
703 41. Martin-Platero AM, Cleary B, Kauffman K, Preheim SP, McGillicuddy DJ, Alm EJ,
704 Polz MF. 2018. High resolution time series reveals cohesive but short-lived
705 communities in coastal plankton. Nat Commun 9:266.
706 42. Berry D, Widder S. 2014. Deciphering microbial interactions and detecting keystone
707 species with co-occurrence networks. Front Microbiol 5:1–14.
708 43. Xue Y, Chen H, Yang JR, Liu M, Huang B, Yang J. 2018. Distinct patterns and
709 processes of abundant and rare eukaryotic plankton communities following a reservoir
710 cyanobacterial bloom. ISME J 12:2263–2277.
711 44. Tackmann J, Matias Rodrigues JF, von Mering C. 2019. Rapid Inference of Direct
712 Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial

713 Sequencing Data. Cell Syst 9:286-296.e8.
714 45. Suttle CA. 2007. Marine viruses - Major players in the global ecosystem. Nat Rev
715 Microbiol 5:801–812.
716 46. Yau S, Krasovec M, Felipe Benites L, Rombauts S, Groussin M, Vancaester E, Aury
717 JM, Derelle E, Desdevises Y, Escande ML, Grimsley N, Guy J, Moreau H, Sanchez-
718 Brosseau S, van de Peer Y, Vandepoele K, Gourbiere S, Piganeau G. 2020. Virus-host
719 coexistence in phytoplankton through the genomic lens. Sci Adv 6:eaay2587.
720 47. Claverie JM, Grzela R, Lartigue A, Bernadac A, Nitsche S, Vacelet J, Ogata H,
721 Abergel C. 2009. Mimivirus and Mimiviridae: Giant viruses with an increasing
722 number of potential hosts, including corals and sponges. J Invertebr Pathol 101:172–
723 180.
724 48. Clouthier SC, Vanwalleghem E, Copeland S, Klassen C, Hobbs G, Nielsen O,
725 Anderson ED. 2013. A new species of nucleo-cytoplasmic large DNA virus (NCLDV)
726 associated with mortalities in Manitoba lake sturgeon Acipenser fulvescens. Dis Aquat
727 Organ 102:195–209.
728 49. Gastrich MD, Leigh-Bell JA, Gobler CJ, Anderson OR, Wilhelm SW, Bryan M. 2004.
729 Viruses as potential regulators of regional brown tide blooms caused by the alga,
730 Aureococcus anophagefferens. Estuaries 27:112–119.
731 50. Duponchel S, Fischer MG. 2019. Viva lavidaviruses! five features of virophages that
732 parasitize giant DNA viruses. PLoS Pathog 15:1–7.
733 51. Xu S, Zhou L, Liang X, Zhou Y, Chen H, Yan S, Wang Y. 2020. Novel Cell-Virus-
734 Virophage Tripartite Infection Systems Discovered in the Freshwater Lake Dishui
735 Lake in Shanghai, China. J Virol 94:e00149-20.
736 52. Sun S, La Scola B, Bowman VD, Ryan CM, Whitelegge JP, Raoult D, Rossmann MG.
737 2010. Structural Studies of the Sputnik Virophage. J Virol 84:894–897.
738 53. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A,
739 Ardyna M, Arkhipova K, Carmichael M, Cruaud C, Dimier C, Domínguez-Huerta G,
740 Ferland J, Kandels S, Liu Y, Marec C, Pesant S, Picheral M, Pisarev S, Poulain J,
741 Tremblay JÉ, Vik D, Acinas SG, Babin M, Bork P, Boss E, Bowler C, Cochrane G, de
742 Vargas C, Follows M, Gorsky G, Grimsley N, Guidi L, Hingamp P, Iudicone D,
743 Jaillon O, Kandels-Lewis S, Karp-Boss L, Karsenti E, Not F, Ogata H, Poulton N,
744 Raes J, Sardet C, Speich S, Stemmann L, Sullivan MB, Sunagawa S, Wincker P,
745 Culley AI, Dutilh BE, Roux S. 2019. Marine DNA Viral Macro- and Microdiversity
746 from Pole to Pole. Cell 177:1109-1123.e14.

747 54. Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, Iudicone D,
748 Karsenti E, Speich S, Trouble R, Dimier C, Searson S. 2015. Open science resources
749 for the discovery and analysis of Tara Oceans data. Sci Data 2:150023.
750 55. de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le
751 Bescot N, Probert I, Carmichael M, Poulain J, Romac S, Colin S, Aury J-M, Bittner L,
752 Chaffron S, Dunthorn M, Engelen S, Flegontova O, Guidi L, Horák A, Jaillon O,
753 Lima-Mendez G, Lukeš J, Malviya S, Morard R, Mulot M, Scalco E, Siano R, Vincent
754 F, Zingone A, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Acinas SG, Bork P,
755 Bowler C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S,
756 Raes J, Sieracki ME, Speich S, Stemmann L, Sunagawa S, Weissenbach J, Wincker P,
757 Karsenti E. 2015. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean.
758 Science 348:1261605.
759 56. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri
760 B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C,
761 D'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F,
762 Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H,
763 Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Boss E, Follows
764 M, Karp-Boss L, Krzic U, Reynaud EG, Sardet C, Sieracki M, Velayoudon D, Bowler
765 C, De Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F,
766 Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P,
767 Karsenti E, Raes J, Acinas SG, Bork P. 2015. Ocean plankton. Structure and function
768 of the global ocean microbiome. Science 348:1261359.
769 57. Matsen FA, Kodner RB, Armbrust EV. 2010. pplacer: linear time maximum-
770 likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference
771 tree. BMC Bioinformatics 11:538.
772 58. De Vargas C, Engelen S, Hingamp P, Sieracki M, Vargas C, Audic S, Henry N,
773 Decelle J, Mahé F, Logares R, Lara E, Berney C, Bescot N, Probert I, Carmichael M,
774 Poulain J, Romac S. 2015. Eukaryotic plankton diversity in the sunlit ocean. Science
775 348:1261605–1/11.
776 59. Watts SC, Ritchie SC, Inouye M, Holt KE. 2019. FastSpar: Rapid and scalable
777 correlation estimation for compositional data. Bioinformatics 35:1064–1066.
778 60. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.
779 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10:1–9.
780 61. Yang WT, Parikh JR, Thomas Stavros A, Otto P, Maislin G. 2018. Exploring the

781 negative likelihood ratio and how it can be used to minimize false-positives in breast
782 imaging. Am J Roentgenol 210:301–306.
783 62. Katoh K, Misawa K, Kuma KI, Miyata T. 2002. MAFFT: A novel method for rapid
784 multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res
785 30:3059–3066.
786 63. Price MN, Dehal PS, Arkin AP. 2009. Fasttree: Computing large minimum evolution
787 trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641–1650.
788 64. Letunic I, Bork P. 2019. Interactive Tree of Life (iTOL) v4: Recent updates and new
789 developments. Nucleic Acids Res 47:256–259.
790 65. Paul Shannon, Andrew Markiel, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N,
791 Schwikowski B, Ideker T. 2003. Cytoscape: A Software Environment for Integrated
792 Models. Genome Res 13:426.
793 66. Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755–763.
794

795 **Author contributions**

796 LM and HO designed the study. LM performed most of the bioinformatics analysis.
797 HE generated the primary NCLDV data. RBM, RHV, and HK contributed to the
798 bioinformatics analysis. All authors contributed to the writing of the manuscript.
799

800 **Materials & Correspondence**

801 Correspondence and material requests should be addressed to HO (email:
802 ogata@kuicr.kyoto-u.ac.jp).
803

804 **Competing financial interests**

805 The authors declare no competing financial interests.
806
807
808
809
810
811
812
813
814

815 **Figure 1. Overall workflow for inferring co-occurrence networks and quantitative**
816 **assessment.** This figure shows how the input data (*Tara* Oceans metagenomics and
817 metabarcoding data) were used in this study. The definition of the confusion matrix for
818 quantitative assessment is shown in the table. The LR+ and FDR equations are given at the
819 lower right corner of the plot.
820
821 **Figure 2. *polB*–V9 co-occurrence network.** We performed co-occurrence analysis at the
822 OTU level and constructed the network with pooled ***polB*** –V9 associations from five size
823 fraction networks (A). To better display co-occurrence patterns, PolB OTUs were grouped at
824 the family or family-like level, and V9 OTUs were grouped using annotation at high
825 taxonomic ranks. The size of each node indicates the number of OTUs that belong to the
826 group, and the width of each edge indicates the number of associations between two
827 connected groups. Associations with positive weight are shown in red and negative
828 associations are shown in blue. (B) Number of associations connected to NCLDVs for each
829 major eukaryotic lineage in five size fractions. The top 10 lineages were retained, and other
830 lineages were omitted and shown as "others." Size fractions are presented in μm.
831
832 **Figure 3. Density plots for the degree of NCLDV nodes in co-occurrence networks.** The
833 degree of an NCLDV node is given by the associations between this node and eukaryotes in
834 the networks. The amount of NCLDV nodes are given on the top of the density values. (A)
835 Positive degree (number of positive associations per node) for NCLDV nodes in five size
836 fraction networks. (B) Negative degree (number of negative associations per node) for
837 NCLDV nodes in five size fraction networks. Size fractions are presented in μm.
838
839 **Figure 4. Positive likelihood ratios (LR+) in the NCLDV virus–host validation.** (A)
840 General performance of co-occurrence networks is shown with the LR+ calculated with
841 associations pooled from five size fractions networks. To show the relationship between LR+
842 and FlashWeave association weight, the LR+ values are plotted along with the association
843 weight. (B) Performance of each size fraction network is shown with the violin plot by
844 ggplot2 with a bandwidth of 2. Size fractions are presented in μm.
845
846 **Figure 5. Prediction of NCLDV virus–host relationships with TIM.** (A) Undirected
847 network that shows the relationships between NCLDVs and eukaryotes after TIM filtration.
848 The size of each node indicates the number of predicted interactions of this group. The
849 weight of network edges as defined by the number of tree nodes enriched in each viral family
850 subtree to specific eukaryotic major lineages in the TIM analysis. Known virus–host
851 relationships are highlighted in red, and the pairs found to have horizontal gene transfer are
852 highlighted in yellow (1). (B) Performance of networks on NCLDV host prediction for
853 original FlashWeave results without a weight cut-off, weight cut-off > 0.4, and TIM
854 filtration, plotted by ggplot2 with a bandwidth of 2. (C) FDR of networks for NCLDV host
855 prediction with the original FlashWeave results without a weight cut-off, weight cut-off >
856 0.4, and TIM filtration.
857
858 **Figure 6. Associations between virophages and NCLDVs.** (A) Number of associations
859 with virophages is shown for seven NCLDV families and two unclassified groups,
860 "Medusaviruses" and "Pithoviruses." Associations in the femto-size fraction network are
861 shown in yellow, and in the pico-size fraction network are shown in green. The number of
862 positive associations is above the zero axis, and the number of negative associations is below
863 the zero axis. (B) Phylogenetic tree was constructed from 195 environmental virophages and
864 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.

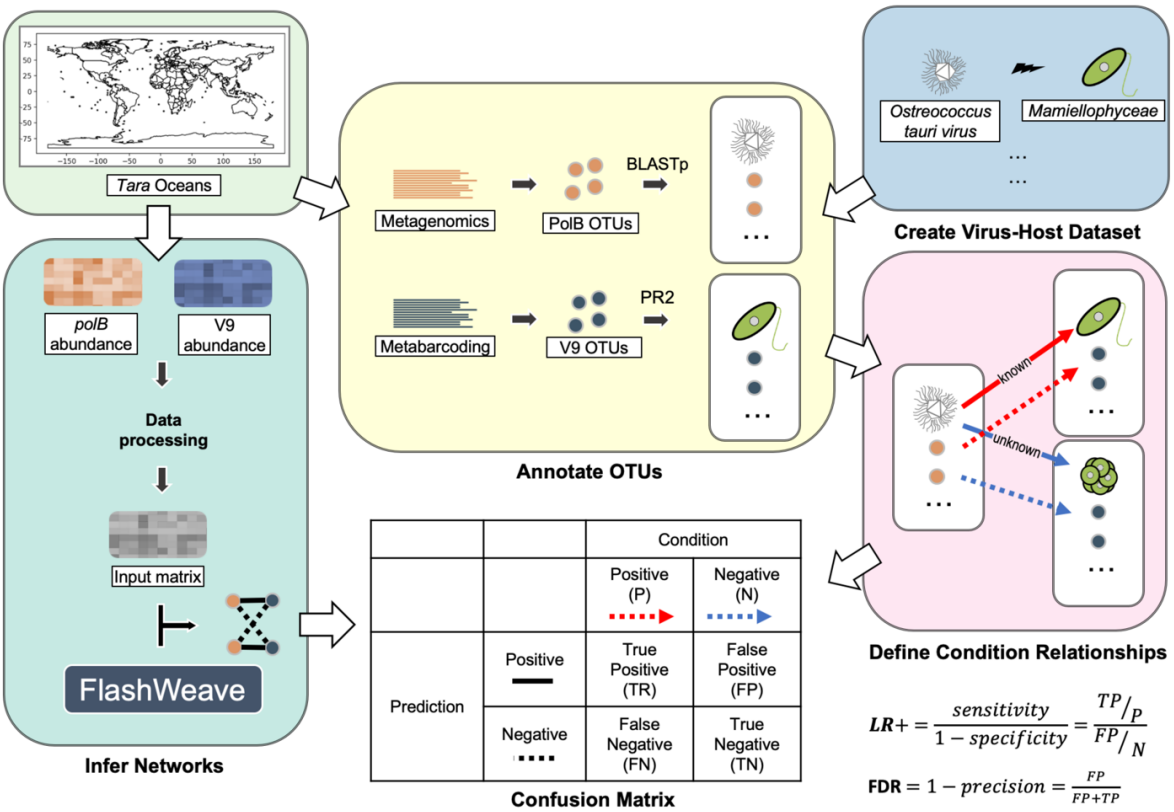
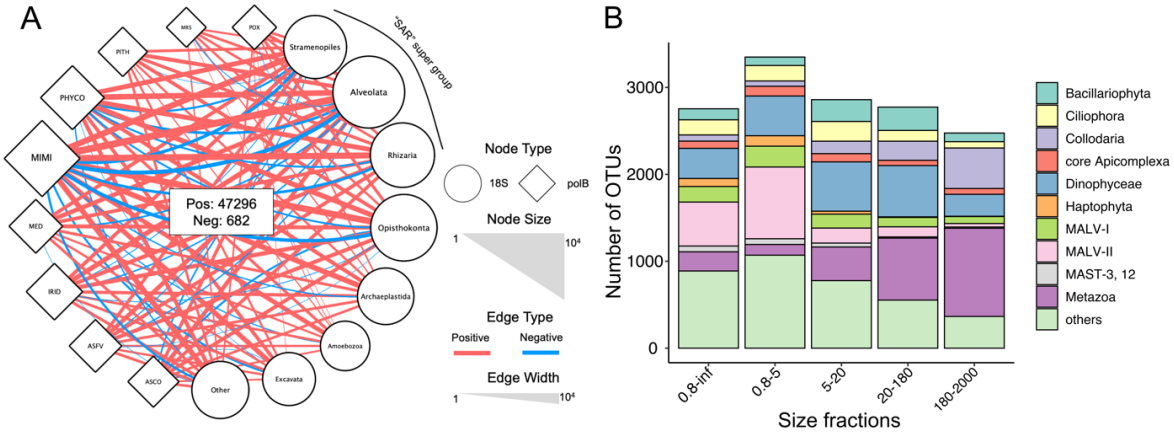


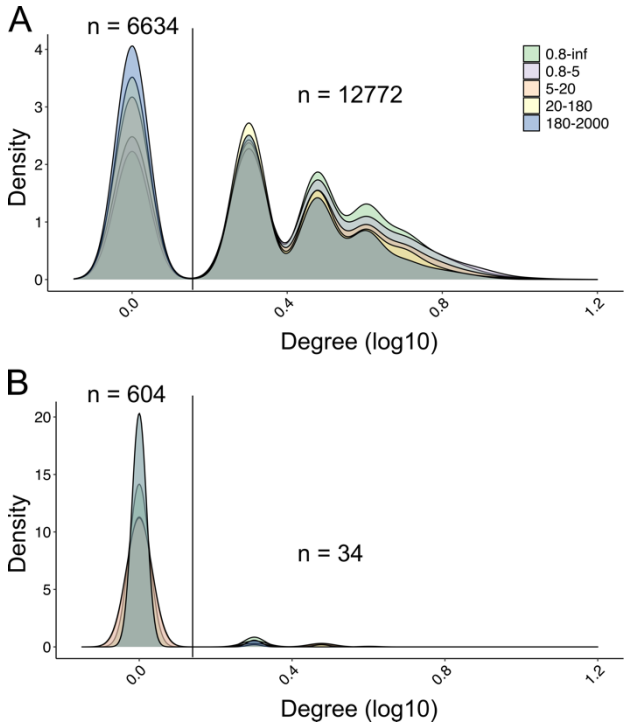
Figure 1.

873



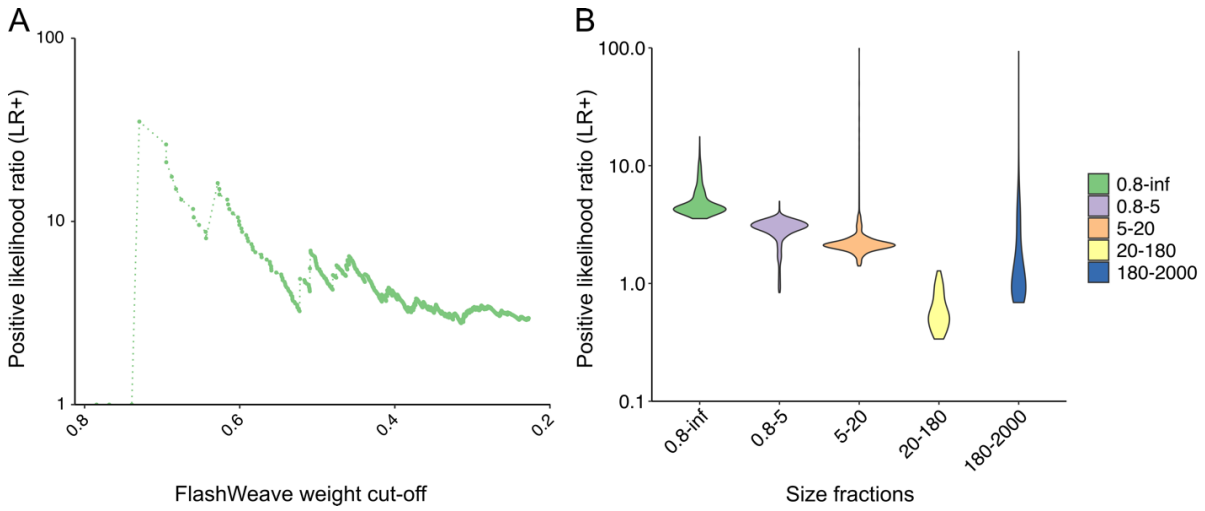
874 Figure 2.

875



876 Figure 3.

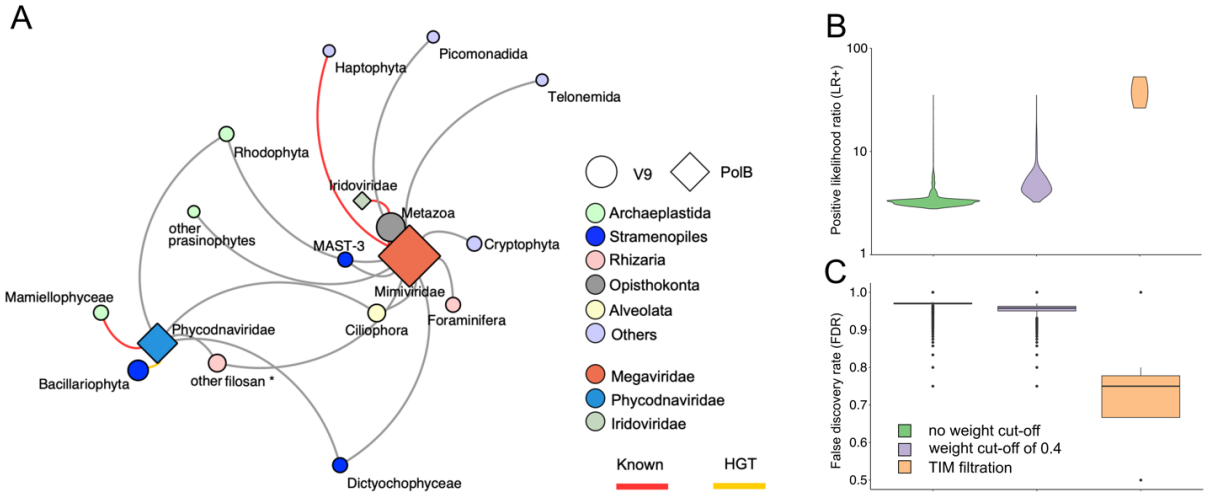
877



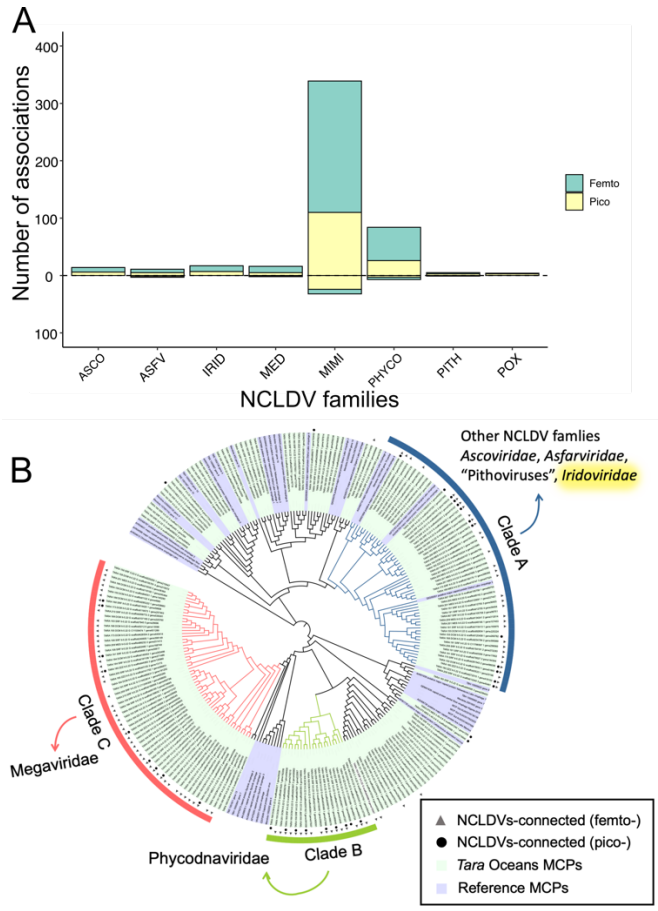
878 Figure 4.

879

880



881 Figure 5.


A
Number of associations
400
300
200
100
0
100
Femto
Pico
ASCO
ASFV
IRID
MED
MIMI
PHYCO
PITH
POX
NCLDV families
B
Other NCLDV famlies
Ascoviridae, Asfarviridae,
"Pithoviruses", Iridoviridae
Clade A
Clade C
Megaviridae
Phycodnaviridae
Clade B
NCLDVs-connected (femto-)
NCLDVs-connected (pico-)
Tara Oceans MCPs
Reference MCPs

882 Figure 6.