



Royal Netherlands Institute for Sea Research

This is a pre-copyedited, author-produced version of an article accepted for publication, following peer review.

Roitman, S.; Rozenberg, A.; Lavy, T.; Brussaard, C.P.D.; Kleifeld, O.; Bèjà, O. (2023). Isolation and infection cycle of a polinton-like virus virophage in an abundant marine alga. *Nature Microbiology* 8(2): 332-346. DOI: 10.1038/s41564-022-01305-7

Published version: <https://dx.doi.org/10.1038/s41564-022-01305-7>

NIOZ Repository: <http://imis.nioz.nl/imis.php?module=ref&refid=361150>

[Article begins on next page]

The NIOZ Repository gives free access to the digital collection of the work of the Royal Netherlands Institute for Sea Research. This archive is managed according to the principles of the [Open Access Movement](#), and the [Open Archive Initiative](#). Each publication should be cited to its original source - please use the reference as presented.

9

10 **Isolation and infection cycle of a polinton-like virus virophage in an**
11 **abundant marine alga**

12

13 Sheila Roitman^{1*}, Andrey Rozenberg¹, Tali Lavy¹, Corina P. D. Brussaard^{2,3}, Oded Kleifeld¹, Oded
14 Béjà^{1*}

15 ¹Faculty of Biology, Technion - Israel Institute of Technology, Haifa 3200003, Israel.

16 ²Department of Marine Microbiology and Biogeochemistry, NIOZ Royal Netherlands Institute for
17 Sea Research, 1797 SZ t'Horntje, The Netherlands.

18 ³Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, Amsterdam,
19 The Netherlands.

20

21 *e-mail: sheilaroitman@gmail.com, beja@technion.ac.il

22

23

24

25 **Abstract**

26 Virophages are small, dsDNA viruses that can only replicate in a host by co-infecting with
27 another virus. Marine alga are commonly associated with virophage-like elements like
28 Polinton-like viruses (PLVs), which are thought to be ancestors of dsDNA viruses but
29 remain uncharacterized. Here we isolated a PLV that co-infects the alga *Phaeocystis*
30 *globosa* with the *Phaeocystis globosa* virus-14T (PgV-14T). We name this PLV “Gezel-
31 14T” and show that it is phylogenetically distinct from the *Lavidaviridae* family where all
32 known virophages belong. Gezel-14T co-infection decreases the fitness of its viral host
33 by reducing burst sizes of PgV-14T, yet not enough to spare the cellular host population.
34 Genomic screens show Gezel-14T-like PLVs integrated into *Phaeocystis* genomes,
35 suggesting these widespread viruses are capable of integration with cellular host
36 genomes. This system presents an opportunity to better understand the evolution of
37 eukaryotic dsDNA viruses as well as the complex dynamics and implications of viral
38 parasitism.

39

40

41 Main

42 Eukaryotic genomes are a hub where viruses and selfish genetic elements (SGEs) convene.
43 Viruses and SGEs are both capable of jumping in and out of the host genome, yet SGEs usually
44 lack the structural proteins that could grant them independence from the host¹. A remarkable
45 group of SGEs are the Polintons (or Mavericks), first thought to be self-replicating large
46 transposons, and today considered viral-like mobile elements²⁻⁴. Polintons are 15-20 kbp long,
47 typically flanked by terminal inverted repeats (TIRs), and encode for a protein-primed DNA
48 polymerase (pPolB), a retrovirus-like integrase (RVE-INT), and viral capsid genes. Although there
49 are no examples of these viral-like entities becoming Polintoviruses, it has been proposed that
50 Polinton-like elements are the ancestors of most dsDNA viruses⁴⁻⁷. Recently, new groups of
51 Polinton-like viruses (PLVs) that resemble Polintons, yet lack the distinctive polinton genes
52 (pPolB, RVE-INT), have been described by genomics and metagenomics in diverse Eukaryotic
53 organisms, including algae. Some PLVs are integrated into algal genomes, while others appear
54 to be independent^{8,9}, suggesting a dual lifestyle as integrated and free-living viruses, bridging the
55 gap between Polintons and viruses. However their replication strategies remain unknown to this
56 day. . To date, particles for only one PLV have been isolated: TsV-N1, a dsDNA nuclear-
57 replicating virus infecting *Tetraselmis striata*¹⁰. PLVs encode a core set of three genes: a major
58 and a minor capsid proteins (MCP, mCP) and a packaging-ATPase, along with a variable set of
59 genes conserved among Polintons and virophages⁸.

60 Virophages are small viral parasites that depend on a virus from the nucleocytoplasmic large-
61 DNA viruses (NCLDV) for reproduction. Virophages are dsDNA viruses, have a 17-30-kbp
62 genome with a low %GC content (27-50%) and icosahedral capsids of 35-70 nm in diameter^{11,12}.
63 Most virophages were found by genomics and metagenomics, and remain uncultured¹³⁻²¹, yet a
64 few were isolated conjointly with their viral host from the *Mimiviridae* family²²⁻²⁶. All virophages
65 have been so far classified within the *Lavidaviridae* family that currently contains two genera:
66 *Sputnikvirus* and *Mavirus*¹², and the recently isolated Chlorella virus virophage SW01 also falls
67 within its scope²⁶. Interestingly, maveruses are particularly similar to Polintons in gene content^{6,23}
68 and integrated Mavirus-like elements were detected in the eukaryotic host, indicating that
69 maveruses lead a dual life-style¹⁶.

70 ***Phaeocystis globosa*** is a ubiquitous haptophyte, capable of creating enormous toxic
71 blooms that can be terminated by viral infections²⁸. In 2013, the genome of a giant virus infecting
72 *Phaeocystis globosa*, PgV-16T was sequenced which led to an unexpected discovery of a co-

73 occurring small viral genome in the assembly²⁷. The small viral genome was termed *Phaeocystis*
74 *globosa* virus virophage (PgVV)²⁷, and was later classified as a PLV⁸. Its genome was found to
75 be ~20 kbp, flanked by TIRs and low %GC content (36%)²⁷. No small viral particles were observed
76 along PgV-16T, and it was suggested that PgVV was packed as a linear plasmid within the PgV
77 capsid or integrated in the viral genome²⁷. In this work, we report the isolation and characterization
78 of a closely related virus, Gezel-14T, a PLV with a virophage life-style. We characterize the
79 dynamics of PgV-14T and Gezel-14T infection of *Phaeocystis globosa*, and analyze Polinton-like
80 genomes to create a framework to classify PLVs within the virosphere.

81

82 **RESULTS**

83 **PgV-14T and Gezel-14T genome sequencing**

84 Within the framework of analyzing infection dynamics of *P. globosa* we sequenced the genome
85 of PgV-14T. The assembly also contained a separate scaffold corresponding to a genome closely
86 related to the *Phaeocystis globosa* virus virophage (PgVV)²⁷. To avoid confusion between
87 terminology, we refer to this isolate as Gezel-14T (“gezel” meaning “companion” in Dutch, an
88 allusion to “sputnik”), while the PLV genome found associated with PgV-16T is referred to as
89 Gezel-16T. Further details on the differences between the viral isolates, as well as the gene
90 repertoire of Gezel-14T can be found in the Supplementary Information, Extended Data Fig. 1
91 and Supplementary File 2.

92 Contrary to the previous report for Gezel-16T (PgVV)⁽²⁷⁾, we did not find any indication that Gezel-
93 14T is integrated in the PgV genome: out of 539,034 trimmed Illumina reads that could be mapped
94 to the two reference genomes, no individual reads or read pairs were found to map to PgV-14T
95 and Gezel-14T simultaneously.

96

97 **Gezel-14T is a *bona-fide* virus**

98 To assess whether the Gezel-14T genome is packed within PgV-14T particles or in particles of
99 its own, we filtered a fraction of a mixed lysate through a 0.2 µm filter (twice). Half of the filtrate
100 was boiled and then all fractions (Lysate, Boiled and Filtered) were treated with DNase to
101 eliminate non-encapsidated DNA. Gezel-14T marker genes could be amplified by PCR from both
102 Lysate and Filtered fractions, while PgV-14T was only found in the Lysate (Fig. 1a,b). DNA

103 staining of PgV-14T-only, Gezel-14T-only and mixed lysates showed two distinct size populations
104 matching the expected composition of the lysates (Fig. 1d,e). In line with this, under transmission
105 electron microscopy (TEM), negatively stained lysates showed two distinct icosahedral viral
106 particles (Fig. 1c). The larger particles measured 160-215 nm in diameter (mean 188 nm, s.d. 16
107 nm, n = 24), while the smaller particles measured 50-80 nm (mean 66 nm, s.d. 7 nm, n = 77)
108 (Supplementary File 2). Taken together, along with the proteomic identification of most Gezel-
109 14T proteins in purified viral lysates (see below), we conclude that Gezel-14T is a virus with a
110 ~19.5 kbp genome packed in icosahedral capsids of 66 nm.

111

112 **Infection dynamics**

113 To get insight into the interactions between PgV-14T and Gezel-14T, we prepared a PgV-14T-
114 only lysate by using a dilution-to-extinction approach; and a Gezel-14T-only lysate by filtering a
115 mixed lysate through 0.1 μm filters followed by concentration and purification. *P. globosa* cultures
116 were infected with either a PgV-14T-only, a Gezel-14T-only lysate or their mix. Cultures infected
117 with PgV-14T (solely or in conjunction with Gezel-14T) were completely lysed after 3 or 4 days
118 (respectively), while the culture infected with Gezel-14T-only followed a similar growth pattern to
119 the control (Extended Data Fig. 2, Supplementary File 1). PgV-14T increased in abundance in
120 both infection experiments, while Gezel-14T increased only in the mixed infection. These results
121 suggest that Gezel-14T is unable to independently complete an infection cycle in *P. globosa*, and
122 depends on PgV-14T for its reproduction, the first experimental evidence that it leads a virophage
123 lifestyle.

124 The same experimental layout was followed to evaluate the effect of Gezel-14T on the
125 course of PgV-14T infection. The latent period for both viruses was 8-9 hrs, regardless of Gezel-
126 14T presence (Fig. 2a, Supplementary File 1), shorter than previously reported²⁸. We considered
127 the latent period finished when free viral particles increased by 30%. We compared the virulence
128 of PgV-14T with and without Gezel-14T, in both cases only ~10% of infections ended in lysis (Fig.
129 2b, Supplementary File 1), similarly to reports for other giant viruses²⁹⁻³¹. In mixed lysates,
130 successful co-infections are rare: less than 20% of successful PgV-14T infections are co-
131 infections with Gezel-14T (Supplementary File 1). A number of reasons could lead to this result.
132 First, *P. globosa* defense mechanisms might affect both viruses, while PgV might have additional
133 anti-virophage defense systems. Second, it is likely that in contrast to Sputnik which enters the
134 host-cell entangled in its host virus surface fibers^{32,33}, Gezel-14T has to recognize, attach and

135 enter the PgV-infected host independently. PgV particles lack long fibers (Fig. 1c) and we found
136 no homologs of the fiber-associated proteins of mimiviruses^{32,34}. However, we found phage fiber-
137 like proteins encoded in the PgV-14T and Gezel-14T genomes (see below). The timing of
138 entrance of the virophage might be critical for a successful infection, for it depends on PgV-14T
139 for transcription, as we see synchronous expression of MCPs and DNA polymerases of both
140 viruses (Fig. 2f). This is further supported by the presence of the early promoter motif of
141 mimiviruses in both viruses (Fig. 3c).

142 After inoculating the cultures with mixed and PgV-14T-only lysates, we saw immediate
143 adsorption, and the number of free viruses remained low until 6-9 hrs post-infection. *P. globosa*'s
144 DNA was usually degraded during infection, probably due to early cell lysis, however some
145 replicates showed little host DNA degradation during the first hours (Fig. 2e, Supplementary File
146 1). PgV-14T and Gezel-14T genome copies multiplied at 4 hours post-infection and increased
147 until burst (Fig. 2e). We selected two genes, representative of classical early (DNA polymerase)
148 and late (MCP) genes, and measured their expression levels during the infection (Fig. 2f). PgV-
149 14T and Gezel-14T DNAPol transcripts could be traced already at 2 hours post-infection, while
150 MCP transcripts appeared after 4 hours. These results suggest that the virophage synchronizes
151 its infection with that of its host virus and indicates that the timing of entry is critical for Gezel-14T.
152 Moreover, the ratio between PgV-14T and Gezel-14T transcripts varied between experiments,
153 even though the initial ratio was constant, which partially explains the observed variability and the
154 various Gezel-14T burst sizes (Supplementary File 1).

155 The average burst-size for PgV-14T-only was 136 viruses cell⁻¹ (range 101–181; s.d.
156 28.29), lower than previously reported²⁸. When co-infected with Gezel-14T, we saw a small but
157 significant decrease in the estimated average burst size of PgV-14T: 136 (s.d. 28.29) vs 108 (s.d.
158 33.38) viruses cell⁻¹ (Student's t-test, p-value 0.04) (Fig. 2c, Supplementary File 1). Thus, similarly
159 to other virophages^{22,23,25}, Gezel-14T inhibits PgV-14T reproduction. Since co-infections are rare,
160 and given the high variability of this experimental setup, the actual burst-size for PgV-14T from a
161 single co-infection is expected to be lower. Nevertheless, the effect might be negligible at the
162 population level since lysates derived from a single co-infection showed roughly the same PgV-
163 14T progeny as lysates without Gezel-14T (Supplementary File 1). We estimated the burst-size
164 for Gezel-14T based on the rate of successful co-infections for each experiment and found a large
165 variation, 9–321 virophages cell⁻¹ (Supplementary File 1). These values could partially explain the
166 high variability in this system in our setting (Fig. 2e,f, Supplementary File 1). *P. globosa* cultures
167 collapsed after 10 hours (for high PgV-14T/cells ratio) or 20 hours (low ratios), regardless of

168 Gezel-14T presence, or the virus/viophage ratio (Fig. 2d, Supplementary File 1). Following
169 previous findings on the *Cafeteria*-CroV-Mavirus system, where the viophage prevented CroV
170 from lysing the entire host population when CroV is added at low quantities³⁵, we reproduced
171 similar conditions by infecting *P. globosa* cultures with a Gezel-14T lysate where PgV-14T was
172 below the detection level of a standard PCR reaction. During the first four days, there was no
173 change in the viral numbers, or between infected and control cultures. At day five, we saw a slight
174 increase in the DNA copy number of both viruses; and two weeks after inoculation, PgV-14T
175 genome copy number quintuplicated, while Gezel-14T triplicated, and the infected cultures were
176 completely lysed as opposed to the stationary-phase controls (Supplementary File 1). Overall, we
177 see that although Gezel-14T replication is adverse to PgV-14T at the single-cell level, the effect
178 is not significant at the population level in our laboratory setup. These infection dynamics might
179 explain their coexistence in the environment, as proposed for other host-virus-viophage
180 systems³⁶. Changes in the local ratios between the algal-host, the giant virus and the PLV-
181 viophage might result in short-term “winners”, maintaining the equilibrium.

182

183 **Proteomics analyses**

184 We performed proteomics on samples from 4, 6 and 8 hrs post-infection with a mixed lysate,
185 uninfected *P. globosa* cultures and purified viral particles. The identification of proteins from both
186 viruses improved over the course of the infection, with most viral proteins peaking at 8 hrs.
187 Peptides for 15/18 proteins predicted in Gezel-14T were detected, six of them in all replicates of
188 purified viral particles (Fig. 3a). Three proteins were detected in at least two samples, yet below
189 the intensity threshold (PGVV01, PGVV01b, PGVV09). Curiously, despite the lack of detectable
190 signal in the proteomics data, transcripts from all of the remaining ORFs were amplified at 4 hrs
191 post-infection (Extended Data Fig. 3). The detection of capsid proteins in the early stages of
192 infection, despite MCP not being transcribed, can be explained by the high number of viruses
193 adhering/entering the cells. MCP, mCP, Ltf, the putative lipase (ABH) and the proteins of unknown
194 function PGVV05 and PGVV08 were consistently found in the viral particles, thus likely being
195 components of the Gezel-14T virion. The finding of Ltf in the particles is consistent with its
196 predicted similarity to bacteriophage tail-like fiber proteins (Supplementary Information, Extended
197 Data Fig. 4), and further suggests that it may be involved in mediation of the recognition or
198 attachment to the host cell, while the ABH might mediate the viophage entry to the host cell, as
199 shown for other small viruses³⁷. The Tlr6f protein (widespread among PLVs, lavidaviruses, giant
200 viruses and bacteriophages) was identified by MS/MS yet its MS intensity was below the intensity

201 threshold in purified viral particles (Supplementary File 3). The packaging-ATPase and PGVV16
202 were detected only in some replicates of the viral particles, where identification of low abundance
203 peptides is easier, yet they do not seem to be an integral part of the virions, similarly to PGVV13
204 and Yrec, since they are not consistently found in all viral particles replicates (Fig. 3a).

205 PgV-14T proteomic analyses can be found in Extended Data Fig. 5, Supplementary Information
206 and in Supplementary File 3.

207

208 **Gezel-14T follows the transcription pattern of PgV-14T**

209 To get a better understanding of the procession of PgV-14T infection, we divided its genes into
210 early, middle and late, based on the proteomics results (Supplementary File 3). We assume
211 peptides detected in 4 hrs post-infection samples were transcribed as early genes, while peptides
212 detected only after 8 hrs belong to late genes. Analysis of sequences upstream of these early
213 genes yielded a conserved motif with the sequence AAAATTGA at its core (Fig. 3c). AAAATTGA
214 was first reported as an early promoter motif in Mimivirus³⁸ and related motifs appear to be
215 common in other mesomimiviruses (Extended Data Fig. 6). Querying the Gezel-14T genome with
216 the early promoter motif brought 10 highly significant matches of which five fell upstream of *Yrec*,
217 *TVpol*, *pgvv05*, *seg2* and *pgvv16* (Fig. 3d). Interestingly, these results are in marked contrast to
218 the Mimivirus-Sputnik and CroV-Mavirus systems where the virophages extensively use the host-
219 virus late promoter^{16,23,39}. The presence of the early mimiviral promoter upstream of some of the
220 ORFs and of AT-rich stretches in intergenic regions indicates that Gezel-14T genes are
221 expressed as monocistronic units, despite all of the ORFs having the same orientation in the
222 genome (Fig. 4). Accordingly, we could not detect transcripts spanning adjacent gene-pairs in our
223 samples (Supplementary File 3), while individual genes could be amplified. All the above
224 observations suggest that Gezel-14T depends on PgV-14T machinery for transcription.

225

226 ***Phaeocystis* genomes contain PLVs related to Gezel-14T**

227 Given the close ties of Gezel-14T to integrated PLVs⁸, we reasoned that related PLVs might be
228 residing in the genome of *P. globosa*. Therefore we created a partial assembly of the *P. globosa*
229 genome⁴⁰ and developed a bioinformatic pipeline based on co-occurrence of viral marker genes
230 of PLV, laidavirus or NCLDV origin. Genomic fragments harboring viral marker genes were
231 further scrutinized for the presence of inverted repeats and drops in %GC content. Five scaffolds

232 from *P. globosa* contained at least three PLV marker genes (Fig. 4). Based on phylogeny of the
233 MCP we subdivided these fragments into four groups: Phaglo-R (shaded red) that clustered with
234 Gezel-14T, and the more distantly related Phaglo-Y (yellow), Phaglo-B1 and Phaglo-B2 (blue),
235 and Phaglo-P (purple) (Fig. 5a). The blue and yellow groups were also present in the draft-
236 genome of *P. antarctica*, while the yellow group was additionally found in *P. rex*, suggestive of
237 their distinct host range or time passed since their integration. These scaffolds contain viral
238 fragments representing PLVs of varying degrees of completeness, with short flanking regions
239 seemingly corresponding to the host genome. Similarly to the integrated maveruses in *C.*
240 *bukhardae*¹⁶, *P. globosa* PLVs have a lower %GC content than the host genome and the flanking
241 regions on the scaffolds (37.5%-54.5% vs. 64.5%) (Fig. 4). An additional scaffold, Phaglo-G with
242 a high %GC content (61.1%) appeared to contain a small virus flanked by short tandem repeats
243 with a NCLDV-type MCP gene. MCP phylogeny and gene composition of Phaglo-G revealed it to
244 be related to *Pleurochrysis* sp. “endemic viruses” (Extended Data Fig. 7). Given their size and
245 affinity to NCLDVs we provisionally refer to these viruses as “NCLDV-like dwarf viruses” (NDDVs)
246 (Supplementary Information).

247 Neither transcripts nor proteins encoded by *P. globosa* PLVs could be detected in our
248 samples (Supplementary Files 4,5). It is possible that these PLVs respond to infection by host
249 viruses other than PgV-14T. Indeed, we found transcripts for MCPs similar to Gezel-like
250 integrated PLVs (Phaglo-R and Phaglo-B) in marine metatranscriptomes (Fig. 5a, Supplementary
251 File 4). These sequencing and experimental results suggest that Gezel strains infect both PgV
252 virocells and *P. globosa* cells in a strain-specific manner (Extended Data Fig. 8a.).

253

254 **Phylogenetic analysis of the Gezel-like group**

255 Phylogenetic analysis of MCP proteins from algal and protist genome assemblies assigned to the
256 Gezel/PgVV-like group^{8,41} revealed that most of the haptophyte-associated PLVs form a single
257 well-supported clade together with MCPs from aquatic metagenomes that we refer to as the
258 “Gezel-core clade” (Fig. 5a,b, Supplementary Fig. 2). A minority of haptophyte-associated MCPs,
259 including Phaglo-P, form a separate but related clade. Interestingly, MCP proteins from green
260 algae are paraphyletic with respect to MCPs associated with haptophytes, cryptophytes and a
261 stramenopile which indicates that the Gezel-like PLVs originated as green algal viruses (Fig. 5a).
262 Gezel-like integrated PLVs could be also found in *Isochrysis galbana*, while the genome of

263 *Chrysochromulina parva*, the host of the PLVs Curly, Larry and Moe, did not possess any
264 integrated viruses (Extended Data Fig. 8b) .

265

266 **Core and common genes among Gezel-like viruses**

267 Based on profile-profile matches, we created clusters of orthologous genes across Gezel-like
268 PLVs. Five proteins comprise the set of core genes present in nearly all Gezel-like PLVs: MCP,
269 mCP, A32, PGVV05 and Tlr6F (Fig. 5b). The first three are also part of the core gene repertoire
270 in TVS-like PLVs and lavidaviruses (Fig. 5, Extended Data Fig. 9)^{8,41}. The PGVV05 cluster
271 includes the *G. theta* protein C⁸ and appears to be specific to the lineage of Gezel-like PLVs. The
272 function of PGVV05 is unknown, but we hypothesize that it constitutes a component of the virion
273 (Fig. 3). The Tlr6F-like proteins are widespread beyond the clade of Gezel-like PLVs⁸, and in
274 particular are present in PgV and other mesomimiviruses. Two further genes of unknown function
275 appear to be restricted to the Gezel-core clade: PGVV09 found in all members, and YSL1_23 that
276 is absent from Gezel itself.

277 Two other interesting gene families are ABH and Ltf. Sequence diversity and high
278 variability of ABH active-site positions (GHSQGG in Gezel, SYSDGG in Phaglo-R) in these
279 proteins is indicative of multiple independent acquisition events. Functions of alpha/beta-
280 hydrolases are difficult to predict⁴², although based on the entry mechanism of other small viruses
281 into their host cell³⁷, and the ABH being consistently found in Gezel-14T capsids (Fig. 3), it is
282 plausible that these proteins are lipases. *Ltf* is the longest ORF in Gezel genomes and encodes
283 a protein containing repeats homologous to gp36 of the Enterobacteria phage-T4, a component
284 of the long tail fibers (see Supplementary Information). This protein was also consistently found
285 in Gezel-14T viral particles, suggesting that it might be involved in mediating attachment to the
286 host cell.

287 Tyrosine recombinase (Yrec) genes represent the most widespread family involved in
288 genetic information processing among Gezel-like PLVs, and it is shared with Polintons and
289 lavidaviruses, likely mediating their integration in host genomes⁸. GIY-YIG superfamily
290 endonucleases have a more sporadic distribution. Surprisingly, Gezel-14T encodes two such
291 endonucleases: Seg1 and Seg2, with Seg2 unusually located between the major and minor
292 capsid proteins (Extended Data Fig. 10). Seg2 is likely an intronless site-specific homing
293 endonuclease, a selfish genetic element capable of integrating itself next to the recognition site

294 (see Supplementary Information). This gene might serve as a defense system of Gezel against
295 other similar virophages⁴³, or a hitchhiking SGE.

296

297

298

299 **Classification of Gezel-14T**

300 Since the defining features of a *bona-fide* virus are the possession of a coat-protein encoding
301 gene and the ability to form virions^[47,48], we can affirm that Gezel-14T is a genuine virus dependent
302 on a giant virus for reproduction. This discovery further blurs the distinction between virophages
303 and PLVs: virophage is thus a life-style and not a natural group. Viruses resembling Polintons
304 include a diverse set of lineages, at least some of which include virophages, such as the
305 *Lavidaviridae* and the Gezel-like viruses, while others, like the TVS group are independent
306 viruses. Most if not all of these lineages include viruses capable of integrating in the cellular host
307 genomes. Despite the similarities, accommodation of all of these small dsDNA viruses in the
308 *Lavidaviridae* is not possible given the vast evolutionary distances separating the different
309 lineages. Currently, no family boundaries are established among PLVs and the group as a whole
310 would require a taxonomic revision. As a first step, we propose the establishment of the new
311 genus *Gezelvirus* (genus *incertae sedis*) with the new species *Gezelvirus phaeocystis* to
312 accommodate viral isolates Gezel-14T and Gezel-16T (see Supplementary Information for
313 details).

314

315 **DISCUSSION**

316 Virophages, dsDNA viruses that parasitize an active virocell created by a NCLDV, have been
317 known for more than ten years, yet only a few were successfully cultured. How these viral
318 parasites shape and affect their host's evolution remains to be studied for each particular system.
319 Yet there are general features that seem to characterize the virophage lifestyle, like the ability to
320 integrate to the cellular host genome, and the fitness cost to the viral host. In this work we isolated
321 and characterized the Polinton-like virus Gezel-14T, a virophage co-infecting *Phaeocystis*
322 *globosa* with PgV-14T. We show that Gezel-14T is a *bona fide* virus of linear dsDNA coated by a
323 proteinaceous shell composed by major and minor capsid proteins, a putative lipase and proteins

324 of unknown function. Based on the various related PLVs found in *P. globosa* it might also be
325 capable of transiently or permanently integrating in its algal host genome (see Extended Data Fig.
326 8a). The life cycles of other Gezel-like PLVs, many of which are also associated with haptophytes
327 (Extended Data Fig. 8b), remain to be characterized. The existence of a Polinton-like virus with a
328 virophage lifestyle opens up new questions regarding the infection strategies of PLVs and their
329 potential role in parasitizing giant virus infections. As virophages from different families continue
330 to be isolated and characterized, we get to glimpse into the fascinating evolution of parasitism.

331

332 **METHODS**

333 **Cultures of *Phaeocystis globosa* and viruses.** Non-axenic *Phaeocystis globosa* strain Pg-G
334 (A), and the PgV-14T lysate from the NIOZ Culture Collection were used for this study. *P. globosa*
335 was grown in Mix-TX medium (1:1 mix of f/2⁴⁵ and ESAW⁴⁶, enriched with Tris-HCl and
336 Na₂SeO₃⁴⁶), at 15°C and 90 μmol photon m⁻² s⁻¹ in a light/dark cycle of 16:8 hours. Experiments
337 were conducted in exponentially growing cells. Large culture volumes (> 5L) were grown with
338 gentle stirring on a magnetic stirrer. PgV-14T and mixed PgV-14T/Gezel-14T lysates were
339 obtained by inoculation of *P. globosa* cultures in late-exponential phase. After full lysis the lysates
340 were gently filtered through a 0.45 μm filter (either 33 mm Millex SLHV033RS Millipore, or 75 mm
341 Nalgene rapid flow filters – Thermo Fisher Scientific, depending on the lysate volume).

342

343 **Sequencing and assembly of PgV-14T and Gezel-14T.** One ml of lysed *P. globosa* culture was
344 filtered through a 0.45 μm filter and used to extract DNA using the Promega Wizard columns
345 protocol⁴⁷. Nextera libraries were sequenced using an Illumina MiSeq sequencer at the Technion
346 Genome Center, Israel. The raw data was de-replicated with ParDRe v. 2.1.5⁴⁸ and trimmed with
347 trim_galore v. 0.6.6⁴⁹. The genome assembly was performed with spades v. 3.14.1⁵⁰. Additional
348 Sanger reads were generated to close assembly gaps in the PgV-14T genome (see
349 Supplementary File 1 for primers list). PCR was performed with Ex-Taq enzyme (TaKaRa) in a
350 total volume of 30 μl containing 1 μl viral DNA, Ex-Taq buffer (×1), 0.8 mM primers, 0.8 mM dNTPs
351 and 0.75 U polymerase. PCR conditions were as follows: 95°C – 5 min, 30 cycles of 95°C – 30
352 sec, 60°C – 30 sec, 72°C – 5 min, and a final elongation of 72°C – 5 min. PCR products were
353 cleaned from gel using NucleoSpin Gel and PCR cleanup (MN) and cloned in TOPO-TA plasmids
354 (Invitrogen) according to manufacturer's specifications. Sanger sequencing was performed by
355 MacroGen Europe.

356 The terminal inverted repeats of the Gezel-14T were represented as separate fragments
357 in the spades assembly and thus the following strategy was utilized. The Gezel-14T scaffold was
358 trimmed to include only the non-repeated region and extended with ContigExtender using the raw
359 data⁵¹. The scaffold was trimmed to include a minimal region that would contain the fragments.
360 Given the high sequence similarity between the viral isolates, ORFs could be directly transferred
361 from PgV-16T and PgVV-16T (now to be re-named Gezel-16T) to PgV-14T and Gezel-14T,
362 respectively.

363

364 **Electron Microscopy.** For transmission electron microscopy (TEM) 20 L of exponentially growing
365 (late-stage) *P. globosa* were infected with a mix of PgV-14T and Gezel-14T viruses. Upon full
366 lysis the lysate was filtered through 0.45 µm filters (Nalgene rapid flow filters – Thermo Fisher
367 Scientific) to remove cell debris. The filtrate was concentrated using a 100 kDa TFF column
368 (Repligen N06-E100-05-N) and viruses were pelleted by ultracentrifugation (141,000 × g, 2 hrs,
369 4°C). The viral pellet was resuspended in Mix-TX medium, loaded into an Optiprep (Sigma) 25-
370 40% stepped gradient, and centrifuged at 160,000 × g (SW 41-Ti rotor), for 15 hrs, 4°C. Bands
371 were pulled using a syringe to a Millipore Amicon ultra 100,000 K (Mercury), and centrifuged
372 several times at 5,000 × g to change the medium back to Mix-TX. Ten µl samples were loaded
373 into grids and stained with 10 µl 1% uranyl acetate for 1 min, followed by air-dry desiccation (3
374 hrs). Transmission electron microscopy was performed in a Talos L120C transmission electron
375 microscope at an accelerating voltage of 120 kVe at Rappaport Faculty of Medicine, Technion.
376 Particle sizes were calculated for both viruses using ImageJ v.1.53q⁵².

377

378 **Identification of Gezel-14T particles-PCR, SYBR staining.** To verify the existence of Gezel-
379 14T virions, 0.45 µm filtered viral lysates (6 ml) were separated into three fractions: One fraction
380 remained untouched (L – Lysate), the two other fractions were filtered twice through 0.2 µm filters
381 (Millex Syringe-driven filter units, SLGV033RS, Millipore) (F – filtered), one of them was then
382 boiled for 10 minutes (B – Boiled). DNase (Ambion Turbo DNase cat. AM1907) was added to all
383 three fractions in a 50 µl reaction, as follows: buffer (x1), 2U DNase, 44 ul sample, and incubated
384 30 min at 37°C. After 30 min, additional 1U of DNase was added to each sample and further
385 incubated for 30 min at 37°C. DNase inactivation was performed according to manufacturer
386 instructions. PCR was then performed on all three fractions for PgV-14T and Gezel-14T marker
387 genes with primers 7,8 (MCP) and 9,10 (TVpol) respectively (Supplementary File 1). PCR was

388 performed using the Bio-Ready Mix 2X colored (Bio-Lab) with the following parameters: 95°C – 5
389 min, 30 cycles of 95°C – 30 sec, 60°C – 30 sec, 72°C – 30 sec, and a final elongation of 72°C –
390 5 min.

391 Since the Gezel-14T genome was found to be more abundant than PgV-14T (also
392 reported in²⁷), SYBR-stained lysates were prepared for visualization. Viral lysates (PgV-14T only,
393 Gezel-14T only and Mix) were filtered through 0.45 µm filters, and stained with SYBR Green-I as
394 described elsewhere⁵³, and manually counted in an Elyra 7 eLS microscope at the Technion LS&E
395 with the Plan-Apochromat 63x/1.4 Oil DIC M27 objective, a Scientific Scmos camera and the 1.4-
396 420782 lens. Images were taken with a 488 nm excitation wavelength and 515 nm emission
397 wavelength for 100 ms, using a FITC 525/50 filter and rendered using the SIM² algorithm. Particle
398 analysis for three to four field views was performed with 3D Objects Counter v.2.0.1 for ImageJ
399 v.1.53q⁵⁴ with default settings. To make the analysis as unified as possible, a single value for
400 thresholding was used for most of the PgV-14T and Mix fields. Due to the presence of a
401 significantly brighter background and of spots that appeared to be damaged viral particles in the
402 Gezel-14T only images, for Gezel-14T analysis a separate higher threshold was chosen.

403

404 ***P. globosa* and PgV-14T counts.** *P. globosa* cells were counted using flow cytometry on the
405 basis of their scattering (SSC) and autofluorescence using a 488 nm air-cooled blue laser (530/30
406 BP filter and 505 LP filter), with a BD-LSRII flow cytometer. PgV-14T particles were counted in a
407 BD-LSRII flow cytometer and in the Cytex Aurora Flow Cytometer after fixation and staining by
408 SYBR Green-I, based on FSC and the 530/30 BP, 505 LP laser, as described elsewhere⁵⁵. Flow
409 cytometry was used to count cell and PgV-14T abundance when necessary, however, Gezel-14T
410 could not be detected using this approach. Therefore, in experiments where we compare Gezel-
411 14T to *P. globosa* and/or PgV-14T abundances, we quantified their DNA copies using qPCR (as
412 described below). This enabled us to have a uniform (yet inflated) estimation for each entity that
413 allows numerical comparisons. In all cases where qPCR was used to estimate PgV-14T DNA
414 copy numbers, the samples were also counted by flow cytometry to confirm that we observe the
415 same pattern in the experiment using both approaches. Additionally, *P. globosa* cultures were
416 monitored using chlorophyll A auto-fluorescence as a proxy for bulk biomass and livelihood of the
417 cells (excitation/emission: 440/680 nm) in a Synergy 2 microplate reader (Bio Tek) in experiments
418 where exact cell number was not required, as the flow cytometers were not always available for
419 use.

420 **Isolation of a pure PgV-14T lysate.** A mixed lysate containing PgV-14T and Gezel-14T was
421 diluted by 5×10^{-5} to ensure one viral particle per 10 μ l and used to infect 380 aliquots of *P. globosa*
422 cultures in 96-well plates (200 μ l) and incubated for 10 days. Lysed cultures were checked for
423 Gezel-14T presence by PCR.

424

425 **PgV-14T and Gezel-14T growth curve experiments.** Experimental results for this and the
426 experiments described below can be found in Supplementary File 1. To measure the latent period
427 of PgV-14T and Gezel-14T we used a PgV-14T/*P. globosa* ratio of 0.1 (counted by FACS), and
428 a PgV-14T/Gezel-14T ratio of 1 (calculated by qPCR). Experiments were performed in 25 ml of
429 *P. globosa* cultures. At every sampling point (0, 2, 4, 6, 8, 9 and 10 hrs post infection) 2 ml culture
430 was filtered through a 0.45 μ m filter (Millex, SLHV033RS, Millipore) and the filtrate was kept at
431 4°C until analysis (for a maximum of 2 weeks). Samples were treated with DNase (as described
432 above), DNA was extracted using the Promega Wizard columns as described elsewhere⁴⁷, and
433 quantified by qPCR (as described below). PgV-14T fixed particles were also counted by Cytex
434 Aurora Flow Cytometer as described elsewhere⁵⁵. We considered the latent phase finished when
435 the free viruses reached 1.3 times the free viruses at 0 hrs. n = 5. (Supplementary File 1,
436 “Infection’s latent period”).

437

438 **PgV-14T and Gezel-14T viral progeny calculation.** To assess population-level fitness costs of
439 Gezel-14T infection on PgV-14T, we used an estimate of PgV-14T progeny obtained as the yield
440 of viral particles from a single infection of a *P. globosa* cell in a 200 μ l culture volume. 96-well
441 plates with 200 μ l of exponentially growing *P. globosa* cultures were infected with a diluted PgV-
442 14T only or a mixed lysate, such that every well will be inoculated with a maximum of one PgV-
443 14T particle. After lysis PgV-14T particles were counted by flow cytometry while Gezel-14T
444 presence was confirmed by PCR. Three biological replicates were performed with 14 lysed wells
445 analyzed in total. (Supplementary File 1, “Viral progeny”).

446

447 **PgV-14T and Gezel-14T burst size calculation.** Burst size calculations were performed in
448 exponentially growing *P. globosa* cultures (5 ml) infected with a virus/cell ratio of 10-50. Lysates
449 and cultures were counted before the experiment and after full lysis, *P. globosa* and PgV-14T
450 were counted by flow cytometry as described above, while PgV-14T and Gezel-14T DNA copy

451 numbers were counted by qPCR as described below. Burst size of PgV-14T was calculated by
452 subtracting the original number of viruses from the final count (as counted by FACS) and dividing
453 the difference by the number of cells in the original culture (Supplementary File 1, “Burst size”). A
454 proxy for Gezel-14T burst size was calculated by subtracting the original Gezel-14T number from
455 the final Gezel-14T number (calculated by qPCR), and then dividing the difference by the
456 calculated number of successful PgV-14T infections from the same experiments (as calculated in
457 Supplementary File 1, “Virulence” and “Infection dynamics calc”). n = 5.

458

459 **Virulence of PgV-14T.** To calculate virulence (proportion of infections ending in lysis),
460 exponentially growing *P. globosa* cultures were infected with PgV-14T only and a mix of PgV-14T
461 and Gezel-14T at a virus/cell ratio of 30-50. Since adsorption of viruses was very fast
462 (Supplementary File 1), and at 6 hrs we already see lysis of the cells, we chose 3 hrs post-infection
463 as the time point for analysis as it gives enough time for viral adsorption, yet short enough to have
464 intact cells by the end of the sorting. Three hours after infection the cells were pelleted at 5,000 ×
465 g for 3 minutes and washed with fresh media three times to remove all free viral particles. Infected
466 *P. globosa* cells were sorted into 96-well plates of fresh exponentially growing *P. globosa* cells
467 using the FACS Aria III sorter as described elsewhere⁵⁶ and incubated until full lysis. Gezel-14T
468 presence in lysates was confirmed by PCR. 1149 wells were surveyed for PgV-14T-only, 1824
469 for PgV-14T + Gezel-14T. An uninfected culture of *P. globosa* was subjected to the same
470 treatment as a control for the livelihood of the cells. n = 3 infected cultures. (Supplementary File
471 1, “Virulence”).

472

473 ***P. globosa* cell survival.** Cell survival experiments were conducted on exponentially growing *P.*
474 *globosa* cultures infected with either PgV-14T only or a mix of PgV-14T and Gezel-14T at a
475 virus/host ratio of 3-10 and incubated for two days. Cell survival was measured by chlorophyll A
476 auto-fluorescence as a proxy for bulk biomass and livelihood of the cells (excitation/emission:
477 440/680 nm) in a Synergy 2 microplate reader (Bio Tek). n = 5. (Supplementary File 1, “Cells
478 survival”)

479

480 **Gezel-14T solo infection of *P. globosa*.** Exponentially growing *P. globosa* cultures were split
481 each into four 200 µl cultures: Uninfected control, infected with PgV-14T only, infected with Gezel-

482 14T only and infected with a mix of both. A Gezel-14T pure lysate was obtained by 0.2 µm filtering
483 of a mixed lysate (Millex, SLGV033RS, Merck Millipore), TFF concentration (Repligen N06-E100-
484 05-N, 100 kDa) optiprep (Sigma) gradient separation (as described above) and filtration through
485 a 0.1 µm filter (Millex, SLVV033RS, Merck Millipore). PCR of the resulting lysate showed PgV-
486 14T below the detection level for 30 cycles. The mixed lysate was obtained by combining the
487 PgV-14T-only and Gezel-14T-only lysates. Cultures were infected with lysates in a 20% v/v ratio
488 (ensuring at least 3 viruses per cell) and incubated until the control culture declined (4 days after
489 infection). Growth of *P. globosa* was monitored by measuring chlorophyll A OD
490 (excitation/emission: 440/680 nm) in a Synergy 2 microplate reader (Bio Tek). Samples from the
491 infected cultures were diluted in TE and kept at -20°C until analysis (maximum of 2 weeks).
492 Samples were further diluted in DDW (final dilution 1:100) and analyzed by real-time qPCR as
493 described below. The same setup was used for a Gezel-14T /PgV-14T co-infection at high
494 virophage/virus ratio (proxy of 20 virophages per giant virus, calculated using qPCR copies, yet
495 PgV-14T was below detection for a standard PCR reaction of 30 cycles). Cultures were incubated
496 for two weeks. n = 3 for each experiment. (Supplementary File 1, “Gezel-14T-only infection”).

497

498 **Course of PgV-14T and Gezel-14T infection.** Infection experiments were performed to obtain
499 intracellular DNA, RNA and proteins during the course of an infection cycle. Exponentially growing
500 *P. globosa* cells were infected with a PgV-14T/Gezel-14T mixed lysate, at virus/host ratio of 3-10.
501 For intracellular DNA 1.5 ml culture was pelleted by centrifugation at 10,000 × g, 10 °C for 7 min,
502 and resuspended in 2 ml media, three times. Washed pellets were flash-frozen and kept at -80°C
503 until DNA extraction (maximum of 2 months). DNA was extracted using the GenElute – Plant
504 Genomic DNA Miniprep Kit (Sigma), samples were cleaned by DNase (as described above), and
505 analyzed by qPCR as described below. For RNA extraction 1.5 ml culture was pelleted by
506 centrifugation (7 min, 10,000 × g, 10°C), flash-frozen and kept at -80°C (maximum of 2 months).
507 RNA extraction was performed with the Monarch Total RNA Miniprep Kit (NEB). DNase treatment
508 was performed as described above and cDNA was synthesized using LunaScript RT Supermix
509 Kit (NEB). RT and non-RT vials were checked by PCR with primers 7,8 for PgV (Supplementary
510 File 1) . cDNA was diluted by ×100 and used as template for qPCR. 50 ml of culture was pelleted
511 for proteomic analyses by 20 min centrifugation at 5,000 × g, followed by another 5 min at 10,000
512 × g. Samples were flash-frozen and kept at -80°C (for up to 12 months). The *P. globosa* - PgV-
513 14T - Gezel-14T system showed high variability between biological replicates, especially
514 regarding relative RNA amounts. In experiments where averaging replicates resulted in a graph

515 that does not represent the diverging trends of the experiments, we present one experiment per
516 trend separately (Fig. 2). n = 5.

517

518 **Real-Time qPCR.** qPCR reactions were performed on extracted DNA & RNA, and on diluted
519 lysates according to the description for each experiment (see above). The PerfeCTa SYBR Green
520 Fast Mix (QuantaBio) was used in a volume of 20 μ l: 5 μ l template, MasterMix \times 1 and 0.25 mM
521 primers. Reactions were carried out on a LightCycler 480 Real-Time PCR system (Roche) as
522 follows: 95°C – 10 min, 45 cycles of 95°C – 10 sec, 60°C – 30 sec (annealing and elongation). At
523 the end of each cycle the plate fluorescence was read and the point at which the fluorescence of
524 a well raised above the background was calculated using the LightCycler 480 software (release
525 1.5.0) using the absolute quantification/second-derivative maximum analysis package. Specificity
526 of each amplified qPCR product was verified by melting curve analysis on the LightCycler 480
527 instrument. DNA and RNA concentration of each sample for every single experiment were
528 measured using Qubit (dsDNA and RNA high sensitivity kits - Thermo Fisher Scientific) and
529 equalized to enable reliable and uniform quantification. In addition, standard curves were
530 generated for each qPCR run from known copy numbers of a linearized pGEM-T Easy plasmid
531 (Promega) containing the amplified PCR product to calculate the number of absolute copies in
532 the samples.

533

534 **Proteomics sample preparation.** Proteome analysis was performed to free purified viral
535 particles, infected and control cells. For viral particles three independent *P. globosa* cultures (20L,
536 20L and 4L) were grown until late exponential phase and infected with three independent PgV-
537 14T and Gezel-14T mixed lysates. An extra lysate of 10L was filtered through 0.45 μ m and 0.2
538 μ m (x2) to obtain a Gezel-14T pure sample. Samples were prepared following the protocol
539 described above for electron microscopy lysates preparation. For infection experiments, three
540 time points were used for proteomic analysis, representing the early (4 hrs), middle (6 hrs) and
541 late (8 hrs) stages of infection. For a preliminary experiment one replicate at 6 hrs post-infection
542 was fractionated in a gel and seven fractions were run and analyzed separately, while a viral
543 lysate (4L) was cut into three fractions from the gel. All other samples were analyzed whole, since
544 the fractionation led to little improvement in resolution. Overall, three purified mixed viral lysates,
545 one Gezel-14T only lysate, three replicates for infected cells at each time point and one replicate
546 for each time point of control cells were run in the mass spectrometer.

547 Frozen cells were resuspended with Tris-HCl pH 7.4 with a final concentration of 50 mM
548 and 5% SDS and incubated at RT for 30-60 minutes. Cells were disrupted by beat beating for 2
549 minutes using 0.4-0.6 mm glass beads (Sartorius). Beads were pelleted by centrifugation (21,000
550 × g, 5 min, RT). Sonication of the samples was performed with a UP200St ultrasonic processor
551 connected to vialtweeter (Hielscher, Germany) using a cycle of 80% with amplitude of 100% for
552 10 min. The samples were centrifuged (21,000 × g, 10 min, RT) and 90% of the sample volume
553 was recovered. For viral particles, 1 ml of purified viral particles were mixed with Tris-HCl pH 7.4
554 to a final concentration of 50 mM and 5% SDS, and incubated 60 min at RT. Protein SDS PAGE-
555 sample buffer with 10 mM β-mercaptoethanol was added to 10 ug protein of viral lysates, control
556 and infection samples, and incubated overnight at RT. Samples were run in a GeneScript gel (4-
557 20%). Purified viral lysates were prepared and analyzed following in-gel digestion as described
558 before⁵⁷. Infected and control samples were digested in-solution using S-TRAP™ (Protifi, USA)⁵⁸.
559 Seventy five ug of each lysate were digested with sequencing-grade trypsin (Promega) 1:100
560 ratio for 12 hours at 37°C. Resulting peptides were desalted using C18 StageTips⁵⁹ or TopTip™
561 (PolyLC, USA) and eluted with 50 μL of 50% acetonitrile, 0.1% FA, dried to completeness and
562 resuspended in 2% acetonitrile, 0.1% FA.

563

564 **LC-MS.** Desalted peptides of the different samples were subjected to LC-MS/MS analysis using
565 Q-Exactive-Plus or Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) coupled to nano
566 HPLC. The peptides were resolved by reverse-phase chromatography on 0.075 × 180 mm fused
567 silica capillaries (J&W) packed with Reprosil reversed-phase material (*Dr. Maisch; GmbH,*
568 *Germany*). The peptides of in-gel digested samples (purified viruses) were eluted with a linear
569 60 min gradient of 5–28%, followed by a 15 min gradient of 28–95%, and a 10 min wash at 95%
570 acetonitrile with 0.1% formic acid in water (at flow rates of 0.15 μl/min). The peptides of in-solution
571 digestion (infected and control samples) were eluted with a linear 120 min gradient of 6–30%,
572 followed by a 15 min gradient of 30–95%, and a 15 min wash at 95% acetonitrile with 0.1% formic
573 acid in water (at flow rates of 0.15 μl/min). Mass spectrometry analysis by Q Exactive Plus mass
574 spectrometer (Thermo Fisher Scientific) was in positive mode using a range of m/z 300–1800,
575 MS1 resolution 70,000 with AGC target: 3E6; maximum IT: 20 ms. These were followed by high
576 energy collisional dissociation (HCD) of the 10 most dominant ions selected from the first MS
577 scan. MS2 scans were done at 17,500 resolution, AGC target 1E5, maximum IT: 100msec,
578 isolation window: 1.4 m/z; and HCD Collision Energy: 25%. Dynamic exclusion was set to
579 20 seconds and the “exclude isotopes” option was activated. Mass spectrometry analysis by Q

580 Exactive HF mass spectrometer (Thermo Fisher Scientific) was in positive mode using a range of
581 m/z 300–1800, MS1 resolution 120,000 with AGC target: 3E6; maximum IT: 20 ms. These were
582 followed by high energy collisional dissociation (HCD) of the 20 most dominant ions selected from
583 the first MS scan. MS2 scans were done at 15,000 resolution, AGC target 1E5, maximum IT:
584 60msec, isolation window: 1.3 m/z; and HCD Collision Energy: 27%. Dynamic exclusion was set
585 to 20 seconds and the “exclude isotopes” option was activated.

586

587 **Proteomics Data analysis.** Infected cultures and purified viral particles samples MS/MS spectra
588 were analyzed with MSFragger v. 3.5⁶⁰, via FragPipe v. 18.0 (<https://fragpipe.nesvilab.org/>) while
589 using IonQuant (v. 1.8) and Philosopher (v. 4.3). The searches were conducted using Fragpipe
590 LFQ-MBR configuration. Precursor mass tolerance was set to 20 ppm, fragment mass tolerance
591 was set to 20 ppm, cleavage type set to “Enzymatic”, the enzyme was defined as strict trypsin
592 and 2 missed cleavages were allowed. Cysteine carbamidomethylation was set as fixed
593 modification and methionine oxidation and protein N-terminal acetylation were set as variable
594 modifications. Peptide length was set to be between 7 to 50 amino acids and using default settings
595 for label-free quantification. The searches were conducted against a database composed of viral
596 proteins (Gezel, PgV), proteins encoded in the *P. globosa* integrated PLVs and *P. globosa*
597 proteins. The set of proteins representing *P. globosa* was created by combining amino acid
598 sequences encoded in the transcriptomes of three closely related strains: RCC851 (NCBI TSA
599 HBRH000000000), RCC678 (HBRF000000000) and RCC739 (HBRB000000000). ORFs were
600 predicted with TransDecoder v. 5.5.0 (<https://github.com/TransDecoder>) and the protein
601 sequences were clustered at 100% identity level with cdhit v. 4.8.1⁶¹. Orthogroups were identified
602 with ProteinOrtho v. 6.0.25⁶² using DIAMOND v.2.0.6.144⁶³, identity threshold of 90% and
603 coverage threshold of 25%. Representative proteins were selected from orthogroups appearing
604 in at least two of the three strains.

605 For relative analysis of the infection course-proteomics we calculated and combined Max-
606 LFQ intensity for all peptides found for a single protein at each time point. The time point with the
607 highest intensity was arbitrarily set as 100% and the other two time points were normalized
608 accordingly. For viral particles we used a yes/no approach, so that the 100% represents proteins
609 found in the 5 samples run in the mass spectrometer, 80% for proteins found in 4 samples, etc.
610 We analyzed only proteins with a Max-LFQ value (a protein with a Max-LFQ intensity value was
611 considered a true finding, above the threshold). However, we also present results for proteins

612 whose peptides were identified by MS/MS, but their Max-LFQ intensity value could not be
613 determined. These proteins are presented as a small dot for future reference (Fig. 3a).

614

615 **Bioinformatic analyses.** Unless specified, all programs were run with default parameters. A
616 scheme of the bioinformatic workflow used to analyze Gezel-like PLVs and other viruses can be
617 found in Supplementary Fig. 1 (Supplementary Information).

618

619 **Modeling of Gezel capsid proteins.** Gezel mCP (penton) and MCP proteins were folded with
620 alphafold2 via ColabFold v.1.3.0^{64,65}.

621

622 **Genome assembly of *Phaeocystis* species.** In order to extract genes of viral origin and
623 eventually complete viral segments from the genomes of *Phaeocystis* species we assembled the
624 raw data available for genomes of *P. globosa* Pg-G (Bioproject PRJNA265550), *P. antarctica*
625 CCMP1374 (PRJNA34537)⁴⁰ and *P. rex* CCMP2000 (PRJNA534927), publicly available via NCBI
626 SRA. Mate-pair libraries of *P. globosa* and *P. antarctica* were processed by detecting and
627 replacing the junction linker with cutadapt v. 4.1 and custom scripts and categorizing reads with
628 nxtrim v. 0.4.3⁶⁶. All reads were trimmed trim_galore v. 0.6.7
629 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The trimmed data were
630 assembled with megahit v. 1.29⁶⁷ using default settings. The genomic data for *P. rex* were
631 assembled with spades v. 3.14.1. To extract complete viruses integrated in the *P. globosa*
632 genome, the megahit assembly was scaffolded with SoapDenovo v. 2.40⁶⁸ with K-mer size of 127
633 and remaining scaffold overlaps were joined by a round of assembly with mira v. 5rc2 (clustering,
634 accurate)⁶⁹. Long viral segments of at least 6000 bp were extracted by searching for MCP genes
635 with the HMM profile for PLVs and virophages⁹ and the NCLDV-specific profile VOG01840 from
636 VOGDB (<http://vogdb.org/>). The extracted scaffolds were extended with Contig Extender v. 0.1⁵¹
637 and then polished and gap-filled with pilon v. 1.24⁷⁰.

638

639 **Identification of MCP genes in assemblies.** Gezel-type major capsid protein (MCP) genes were
640 searched for in a collection of eukaryotic genomes and custom set of transcriptomes by extracting
641 ORFs and searching using hmmsearch from HMMER v. 3.3.2⁷¹ with a HMM profile based on an

642 alignment of MCPs from Gezel- and *Phaeocystis* endognized PLVs with an E-value threshold of
643 1e-8 and sequences at least 200 amino acid residues were extracted for downstream analyses.
644 The same analysis was implemented for NCLDV-type MCPs: initial search was performed with a
645 HMM profile built from MCP sequences from mesomimiviruses and pre-extracted MCP
646 sequences from endemic viral elements. The collection covered 186 species (see Supplementary
647 File 4) with the four largest groups represented by green plants (96 species), Stramenopiles (46),
648 Alveolata (11) and Haptophyta (10).

649

650 **Phylogenetic analysis.** The collected MCP protein sequences were combined with MCPs from
651 previously reported PgVV-group PLVs and TVS-group PLVs⁸ for rooting and aligned with
652 hmalign from HMMER using the virophage and PLV MCP profile from⁹. For the NCLDV MCPs
653 the reference set included a representative set of MCP genes from *Mimiviridae* and
654 *Phycodnaviridae* with *Iridoviridae*, *Ascoviridae*, *Marseilliviridae* and *Asfarviridae* and the
655 alignment was performed with a HMM profile created from a mafft v. 7.475 alignment⁷² of
656 reference NCLDV MCPs using hmalign from HMMER v. 3.3.2. Alignments were trimmed to
657 include aligned positions and resulting sequences longer than 300 residues after trimming were
658 selected for phylogenetic analysis using iqtree v.2.1.2 with 1000 ultrafast bootstrap iterations^{73,74}.
659 Shorter sequences were placed on the resulting trees by evaluating the trees with ng-raxml
660 v.1.0.1⁷⁵ and performing phylogenetic placement with epa-ng v.0.3.8⁷⁶.

661 A different set of phylogenetic reconstructions was performed for a panel of genes
662 frequently appearing among Gezel-like viruses: MCP, mCP, A32 ATPase, Tyr recombinase and
663 PGVV05 (Supplementary Information, Supplementary Fig. 2). This analysis was restricted to
664 complete viral genomes and a selection of PLVs integrated in algal genomes. Homologous genes
665 were aligned with mafft (--localpair --maxiterate 1000), the alignments were trimmed with trimal v.
666 1.4.5 (-gt 0.9)⁷⁷. Phylogenetic analysis was done using iqtree as described above.

667

668 **Identification of *Phaeocystis* PLVs in metatranscriptome data.** Gezel-14T MCP protein
669 sequence was used as a query for BLAST against a collection of JGI freshwater and marine
670 metatranscriptomes (updated by August 2021). Only hits to publicly available unrestricted
671 databases were used.

672

673 **Gene homology and functional annotation.** Homology between genes encoded by Gezel-14T
674 and related integrated and free PLVs was established by using profile-profile matches. All
675 predicted protein sequences from Gezel-like PLVs, unrelated reference PLVs and lavidaviruses,
676 integrated NCLDV-like dwarf viruses (NDDVs) and mesomimiviruses were merged together and
677 clustered with mmseqs v. 13.45111⁷⁸ at a minimum of 30% identity and 80% coverage and the
678 clustered sequences were aligned with result2msa. Each sequence was searched against the
679 resulting database with hhblits from HH-utils v. 3.3.0⁷⁹ (three iterations, E-value threshold of 1e-
680 5) and secondary structure was predicted with addss.pl. The a3m database obtained this way
681 was searched against itself with hhsearch from HH-utils. The results of the hhsearch matches
682 were filtered to include hits with probabilities of at least 90 and a coverage of at least 60% of the
683 query and the template and the resulting match pairs were clustered with MCL v. 14.137⁸⁰. The
684 same a3m database was used to query a profile database based on Pfam v. 34.0 available from
685 the HH-suite webserver with hhsearch. Manually curated rules based on Pfam profile matches as
686 well previously published annotations for individual genes were created to assign functions to the
687 resulting clusters.

688 Bipartite network of genomes and shared gene clusters was created based on the MCL
689 clusters. vcontact2 clustering and genome network were obtained with vcontact2 v.0.9.19⁸¹.

690

691 **Analysis of PGVV14 sequence.** Protein sequence coded by *pgvv14* (*Ltf*) was analyzed by
692 searching it against PDB_mmCIF70_21_Mar, Pfam-A_v35 and UniProt-SwissProt-
693 viral70_3_Nov_2021 databases with hhsearch via the HHpred Server⁸² and by searching for
694 repeats with RADAR via the EBI tools server⁸³. PGCG_00042 and other PLV-encoded proteins
695 matching the Pfam profile of T4 tail-fiber protein gp36 in local hhsearch (see above) with a
696 probability as low as 80 were classified as proteins containing gp36-like domains.

697

698 **Promoter motifs in PgV.** PgV-14T genes were classified as early, middle or late genes based
699 on their proteomic profile. Genes whose peptides were detected at 4 hours were considered
700 “early”, while genes with detected peptides only in 8 hours post-infection samples were labeled
701 “late”. The rest, genes whose detected peptide abundance did not change between the samples
702 at 6 and 8 hours post-infection samples were considered “middle”. Many genes had no peptides
703 detected and were classified as “none”. Although this division is not expected to mirror the exact

704 RNA expression pattern of the genes (for example the MCP protein is detected at 4 hrs, yet it is
705 RNA is only starting to be transcribed), we expected the majority of genes classified as “early” to
706 be in this category, and enable better resolution for the motifs analysis. 150 bp upstream from the
707 starting codon of each gene were extracted and every category (early, middle and late) was
708 analyzed separately using meme from the MEME suite v. 5.3.0 for a promoter motif. An additional
709 analysis for motif identification was carried for all PgV-genes, yielding the same early motif (93
710 sites, E-value: 1.7e-61).

711 Promoter motifs were searched across mesomimiviruses with meme from the MEME suite
712 v. 5.3.0. Up to 10 promoter motifs with a length of 6-16 nt were searched for in the 150 nt upstream
713 of each ORF. Sequences similar to the AAAATTGA-containing putative early promoter motif of
714 PgV were searched in the whole genome of Gezel-14T using fimo from the MEME package using
715 the meme output as query. The matches were filtered to include hits with q values < 0.05 and
716 required the matched sequence to include the highly conserved TG dinucleotide. For middle and
717 late genes we could not detect a significant promoter motif.

718

719 **Data availability.** The sequencing data are available from NCBI SRA SRR20333090
720 (Bioproject PRJNA835735). PgV-14T and Gezel-14T (as PgVV-14T) genome assemblies were
721 deposited in NCBI Genbank under accession numbers OP080611 and OP080612. Annotated
722 fragments of complete PLVs and NDDV from *P. globosa* and other algae are provided as
723 Supplementary File 6. Source data are provided with this paper. The mass spectrometry
724 proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE
725 partner repository with the dataset identifier PXD036892. Additional material is supplied in the
726 Figshare repository <https://doi.org/10.6084/m9.figshare.21294852>.

727

728 **Code availability.** Code used for bioinformatic analyses is available from
729 <https://github.com/BejaLab/Gezelvirus> and [https://github.com/BejaLab/phaeocystis-viral-](https://github.com/BejaLab/phaeocystis-viral-elements)
730 [elements](https://github.com/BejaLab/phaeocystis-viral-elements).

731

732

733 **Acknowledgments.** We thank Anna Noordeloos for providing advice on how to culture *P.*
734 *globosa* and PgV-14T, Lihi Shaulov for expert technical assistance with TEM sample preparations
735 and imaging, Irena Pekarsky and Nitzan Dahan for their help with light microscopy, Ilana Navon
736 and the Smoler Proteomics Center for their help with the mass spectrometry analyses, the ICTV
737 Virophage study group for nomenclature discussions, and Shirley Larom for technical assistance.
738 This work was funded by a European Commission ERC Advanced Grant 321647 (O.B.), Israel
739 Science Foundation grants 143/18 (O.B.), 1623/17 and 2167/17 (T.L. and O.K.), and the Ariane
740 de Rothschild Women Doctoral Program (S.R.). O.B. holds the Louis and Lyra Richmond Chair
741 in Life Sciences.

742

743

744

745 **Author contributions.** S.R. conceived the project, designed the experiments and
746 performed the experimental work. A.R. performed bioinformatic analyses. S.R., T.L. and O.K.
747 performed proteomics. C.P.D.B supplied the algal and viral strains. O.B. supervised the project.
748 S.R. drafted the paper, which was critically revised and approved by all authors.

749

750 **Conflicts of interest.** The authors declare that they have no conflicts of interest.

751

752

753

754 **FIGURE CAPTIONS**

755 Fig. 1. **Gezel-14T is a bona-fide virus.** a. PCR assay with a PgV-14T marker (MCP1, pgv157)
756 and b. a Gezel-14T marker (TVpol, pgvv04). M, Molecular Marker; NTC, No template control; +,
757 Positive control (DNA); +F, 0.2 μ m filtered DNA; L, lysate; B, boiled filtered lysate; F, 0.2 μ m
758 filtered lysate. L, B and F were subjected to DNase treatment before the PCR. Two independent
759 lysates were analyzed with similar results, only one is shown. C. Transmission electron
760 microscopy images of a negatively stained PgV-14T/Gezel-14T mixed lysate. Green and yellow

761 arrows denote particles of PgV-14T and Gezel-14T expected size, respectively. A single lysate
762 from each virus were combined for the TEM analysis. d. SYBR green stained PgV-14T only,
763 mixed PgV-14T and Gezel-14T, and Gezel-14T only lysates under Elyra 7 eLS microscope.
764 Green and yellow arrows denote particles of PgV-14T and Gezel-14T size, respectively. Two
765 biological replicates of each lysate were stained. e. Quantification of dots in SYBR green stained
766 samples by apparent volume. Y-axis denotes the number of points counted.

767

768 **Fig. 2. Infection dynamics of *P. globosa*, PgV-14T and Gezel-14T.** a. Latent period of PgV-
769 14T (Pink, in a mixed lysate with Gezel-14T ; Green, PgV-14 only lysate) and Gezel-14T (yellow).
770 We considered the latent period finished when the number of free virions reached 1.3 than t_0 . n
771 = 4 biologically independent cultures and lysates. Data are presented as mean values +/- SD. T-
772 test (two sided) showed no statistical significance (ns). b. Virulence of PgV-14T in mixed (pink)
773 and solo (green) infections. Virulence was calculated as how many individual infections end in
774 lysis. n = 1149 (PgV-14 only), n = 1824 (PgV-14T / Gezel-14T mix), derived from 2 biologically
775 independent cell cultures and 5 independent lysates as described in Supplementary File 1
776 "virulence". Data are presented as mean values +/- SD. T-test (two sided) showed no statistical
777 significance (ns). c. Burst size of PgV-14T in a mixed (pink) and solo (green) infections. n = 5
778 biologically independent lysates in 2 biologically independent cultures as described in
779 Supplementary File 1 "burst size". Data are presented as mean values +/- SD. * T-test (two-sided)
780 = 0.04. d. Cell survival (measured by chlorophyll A autofluorescence) of *P. globosa* in control and
781 infected cultures. n = 5 biologically independent cell cultures and lysates. Data are presented as
782 mean values +/- SD. e. Intracellular DNA copies (absolute copy numbers) of *P. globosa*, PgV-14T
783 and Gezel-14T during infection. Three representative replicates are shown (each with a different
784 Y-axis scale) A single outlier is marked with a circle in III. f. RNA copies (normalized to RNA
785 concentration) of MCP and DNA polymerase of PgV-14T and Gezel-14T during infection. Three
786 representative biological replicates are shown (same as in panel e.). Results for all six replicates
787 can be found in Supplementary File 1 ("Infection's latent period").

788

789 **Fig. 3. Gezel-14T proteomic features.** a. Proteins found by mass spectrometry at 4, 6 and 8
790 hours post-infection (circles) and in purified viral particles (rhomboids). Relative quantification as
791 described in the methods section. Proteins for whom peptides were found, yet below the intensity
792 threshold are marked with a point (0%). Tlr-6f, conserved uncharacterized protein; Seg1, GIY-

793 YIG family nuclease; Yrec, OLV11-like tyrosine recombinase; TVpol, hybrid transposon-viral
794 polymerase, superfamily 3 helicase; ABH, alpha-beta hydrolase (putative lipase); A32, ATPase;
795 mCP, minor capsid protein; Seg2, homing endonuclease; MCP, major capsid protein; Ltf, L-
796 shaped tail fiber-like protein; putative proteins of unknown function are denoted by serial number.
797 b. Structural models of Gezel major (MCP) and penton (mCP) capsid proteins, featuring a double
798 and a single jelly-roll respectively. Secondary structure elements are colored green for beta-
799 strands, and orange for alpha-helices. c. Early genes promoter motif of PgV-14T and motif
800 location relative to the start codon of Gezel genes (marked as 0). Numbering denotes bp. d.
801 Schematic map of Gezel-14T genome, PgV early-genes promoter motif is marked with red arrows
802 oriented according to their strand. ORFs color coding according to 3a, colored proteins were found
803 in our proteomic data, white-colored ORFs were not significantly detected.

804

805 **Fig. 4. PLVs associated with *Phaeocystis globosa*.** Schematic representation of the Gezel-
806 14T PLV genome and *P. globosa* genomic contigs including viral-like elements. Gaps in the
807 assemblies are marked with a dotted line. Homologous proteins are marked in color by their
808 cluster family (Supplementary Table 1), according to the legend. Repeats are marked in pale-
809 yellow with dotted borders. Blue lines denote %GC content, the gray line marks 50% GC. Color-
810 coding of the contigs names match the shading on Fig. 5a. The bottom bar designates bp. DNA
811 sequences for the PLVs retrieved in this work can be found in Supplementary File 4. S1H;
812 superfamily 1 Helicase; RT, reverse transcriptase; FkbM, methyltransferase; RuvC, RuvC
813 nuclease; Yqaj, Yqaj-like recombinase; VLTF3, late transcription factor; DUF2738, unknown
814 protein with DUF2738/5871 domain. The other designations are the same as in Fig. 3.

815

816 **Fig. 5. Phylogeny and gene content of Gezel-like PLVs.** a. Phylogenetic analysis of MCPs from
817 the Gezel-group PLVs. MCP sequences were extracted from previously published PgVV-group
818 PLVs^{8,9,19,21} and newly discovered viral elements found in eukaryotic genome assemblies. The
819 tree is rooted by MCP sequences from the TVS group. Associated host groups are indicated when
820 known. Four clades of PLVs discovered in *Phaeocystis* genome assemblies are highlighted with
821 color (see Fig. 4). Parsimonious predictions of three gene acquisition events characteristic to the
822 Gezel-core clade are indicated. Cultured viruses are indicated with a green hexagon. Numbers of
823 sequences for collapsed clades are shown in parentheses. A complementary phylogenetic tree
824 for the TVS group can be found elsewhere⁴⁴. b. Core genes of Gezel-like group PLVs. Protein

825 sequences were clustered based on profile-profile matches and the clusters were further grouped
826 into protein families based on shared Pfam matches. Each color hue represents one cluster, such
827 that genomes may contain multiple clusters from the same family. Descriptions of the gene
828 families are provided in Supplementary Table 1. Only genes appearing in at least three subgroups
829 of Gezel-like PLVs are shown, and are arranged by the number of genomes they are observed
830 in. The genomes of TvV-S1, Iavdaviruses Sputnik and Mavirus, NDDVs and mesomimiviruses
831 are provided for reference. Asterisks mark partial sequences. AaV– Aureococcus
832 anophagefferens virus; CeV-01B– Chrysochromulina ericina virus CeV-01B; CpV-BQ2–
833 Chrysochromulina parva virus BQ2; PgV-16T– Phaeocystis globosa virus 16T; PoV-01B–
834 Pyramimonas orientalis virus 01B; TetV– Tetraselmis virus 1; PsEV1a, PsEV1b, PsEV2–
835 Pleurochrysis sp. endemic viruses 1a, 1b and 2; TvV-S1– Tetraselmis viridis virus S1. Species
836 prefixes for integrated PLVs are as follows: Botryococcus– Botryococcus braunii; Chromochloris–
837 Chromochloris zofingiensis; Dialut– Diacronema lutheri; Guillardia– Guillardia theta;
838 Monoraphidium– Monoraphidium neglectum; Isogal– Isochrysis galbana; Phaant– Phaeocystis
839 antarctica; Phaglo– P. globosa; Pharex– P. rex. Metagenome prefixes are as follows: ACE– Ace
840 Lake; Chesapeake– Chesapeake Bay; Delaware– Delaware Bay, Etoliko– Etoliko Lagoon; Han–
841 Han River, Montjoie– Lake Montjoie; RED– Red Sea; SAF– South Africa; Soyang– Lake Soyang;
842 YSL– Yellowstone Lakes. For the complete list of viral genomes and all cluster and family
843 assignments see Supplementary Files 5 and 6, respectively.

844

845

846 Bibliography

- 847 1. Koonin, E. V. & Dolja, V. V. Virus World as an Evolutionary Network of Viruses and Capsidless Selfish
848 Elements. *Microbiol. Mol. Biol. Rev. MMBR* **78**, 278–303 (2014).
- 849 2. Pritham, E. J., Putliwala, T. & Feschotte, C. Mavericks, a novel class of giant transposable elements
850 widespread in eukaryotes and related to DNA viruses. *Gene* **390**, 3–17 (2007).
- 851 3. Kapitonov, V. V. & Jurka, J. Self-synthesizing DNA transposons in eukaryotes. *Proc. Natl. Acad. Sci.*
852 **103**, 4540–4545 (2006).
- 853 4. Krupovic, M. & Koonin, E. V. Polintons: a hotbed of eukaryotic virus, transposon and plasmid

- 854 evolution. *Nat. Rev. Microbiol.* **13**, 105–115 (2015).
- 855 5. Koonin, E. V., Krupovic, M. & Yutin, N. Evolution of double-stranded DNA viruses of eukaryotes:
856 from bacteriophages to transposons to giant viruses. *Ann. N. Y. Acad. Sci.* **1341**, 10–24 (2015).
- 857 6. Yutin, N., Raoult, D. & Koonin, E. V. Virophages, polintons, and transpovirons: a complex
858 evolutionary network of diverse selfish genetic elements with different reproduction strategies.
859 *Viol. J.* **10**, 158 (2013).
- 860 7. Krupovic, M., Bamford, D. H. & Koonin, E. V. Conservation of major and minor jelly-roll capsid
861 proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. *Biol. Direct* **9**, 6
862 (2014).
- 863 8. Yutin, N., Shevchenko, S., Kapitonov, V., Krupovic, M. & Koonin, E. V. A novel group of diverse
864 Polinton-like viruses discovered by metagenome analysis. *BMC Biol.* **13**, 95 (2015).
- 865 9. Bellas, C. M. & Sommaruga, R. Polinton-like viruses are abundant in aquatic ecosystems.
866 *Microbiome* **9**, 13 (2021).
- 867 10. Pagarete, A., Grébert, T., Stepanova, O., Sandaa, R.-A. & Bratbak, G. Tsv-N1: A Novel DNA Algal Virus
868 that Infects *Tetraselmis striata*. *Viruses* **7**, 3937–3953 (2015).
- 869 11. Bekliz, M., Colson, P. & La Scola, B. The Expanding Family of Virophages. *Viruses* **8**, 317 (2016).
- 870 12. Fischer, M. G. The Virophage Family *Lavidaviridae*. *Curr. Issues Mol. Biol.* 1–24 (2021)
871 doi:10.21775/cimb.040.001.
- 872 13. Desnues, C. *et al.* Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc.*
873 *Natl. Acad. Sci.* **109**, 18078–18083 (2012).
- 874 14. Campos, R. K. *et al.* Samba virus: a novel mimivirus from a giant rain forest, the Brazilian Amazon.
875 *Viol. J.* **11**, 95 (2014).
- 876 15. Gaia, M. *et al.* Broad spectrum of mimiviridae virophage allows its isolation using a mimivirus
877 reporter. *PLoS One* **8**, e61912 (2013).

- 878 16. Hackl, T., Duponchel, S., Barenhoff, K., Weinmann, A. & Fischer, M. G. Virophages and
879 retrotransposons colonize the genomes of a heterotrophic flagellate. *Elife* **10**, e72674 (2021).
- 880 17. Yau, S. *et al.* Virophage control of antarctic algal host-virus dynamics. *Proc. Natl. Acad. Sci.* **108**,
881 6163–8 (2011).
- 882 18. Gong, C. *et al.* Novel Virophages Discovered in a Freshwater Lake in China. *Front. Microbiol.* **7**, 5
883 (2016).
- 884 19. Zhou, J. *et al.* Three Novel Virophage Genomes Discovered from Yellowstone Lake Metagenomes. *J.*
885 *Virolog.* **89**, 1278–1285 (2014).
- 886 20. Yutin, N., Kapitonov, V. V. & Koonin, E. V. A new family of hybrid virophages from an animal gut
887 metagenome. *Biol. Direct* **10**, 19 (2015).
- 888 21. Stough, J. M. A. *et al.* Genome and Environmental Activity of a *Chrysochromulina parva* Virus and Its
889 Virophages. *Front. Microbiol.* **10**, 703 (2019).
- 890 22. La Scola, B. *et al.* The virophage as a unique parasite of the giant Mimivirus. *Nature* **455**, 100–4
891 (2008).
- 892 23. Fischer, M. G. & Suttle, C. A. A Virophage at the Origin of Large DNA Transposons. *Science* **332**, 231–
893 234 (2011).
- 894 24. Gaia, M. *et al.* Zamilon, a Novel Virophage with Mimiviridae Host Specificity. *PLoS One* **9**, e94923
895 (2014).
- 896 25. Mougari, S. *et al.* Guarani Virophage, a New Sputnik-Like Isolate From a Brazilian Lake. *Front.*
897 *Microbiol.* **10**, (2019).
- 898 26. Sheng, Y., Wu, Z., Xu, S. & Wang, Y. Isolation and Identification of a Large Green Alga Virus (*Chlorella*
899 *Virus* XW01) of Mimiviridae and Its Virophage (*Chlorella Virus* Virophage SW01) by Using Unicellular
900 Green Algal Cultures. *J. Virol.* **96**, e02114-21 (2022).
- 901 27. Santini, S. *et al.* Genome of *Phaeocystis globosa* virus PgV-16T highlights the common ancestry of

- 902 the largest known DNA viruses infecting eukaryotes. *Proc. Natl. Acad. Sci.* **110**, 10800–10805 (2013).
- 903 28. Baudoux, A. C. & Brussaard, C. P. D. Characterization of different viruses infecting the marine
904 harmful algal bloom species *Phaeocystis globosa*. *Virology* **341**, 80–90 (2005).
- 905 29. Tarutani, K., Nagasaki, K. & Yamaguchi, M. Virus adsorption process determines virus susceptibility
906 in *Heterosigma akashiwo* (Raphidophyceae). *Aquat. Microb. Ecol.* **42**, 209–213 (2006).
- 907 30. Gann, E. R., Gainer, P. J., Reynolds, T. B. & Wilhelm, S. W. Influence of light on the infection of
908 *Aureococcus anophagefferens* CCMP 1984 by a “giant virus”. *PLoS ONE* **15**, (2020).
- 909 31. Van Etten, J. L., Burbank, D. E., Xia, Y. & Meints, R. H. Growth cycle of a virus, PBCV-1, that infects
910 *Chlorella*-like algae. *Virology* **126**, 117–125 (1983).
- 911 32. Boyer, M. *et al.* Mimivirus shows dramatic genome reduction after intraamoebal culture. *Proc. Natl.*
912 *Acad. Sci.* **108**, 10296–10301 (2011).
- 913 33. Desnues, C. & Raoult, D. Inside the Lifestyle of the Virophage. *Intervirology* **53**, 293–303 (2010).
- 914 34. Sobhy, H., Scola, B. L., Pagnier, I., Raoult, D. & Colson, P. Identification of giant Mimivirus protein
915 functions using RNA interference. *Front. Microbiol.* **6**, (2015).
- 916 35. Fischer, M. G. & Hackl, T. Host genome integration and giant virus-induced reactivation of the
917 virophage mavirus. *Nature* **540**, 288–291 (2016).
- 918 36. Wodarz, D. Evolutionary dynamics of giant viruses and their virophages. *Ecol. Evol.* **3**, 2103–2115
919 (2013).
- 920 37. Farr, G. A., Zhang, L. & Tattersall, P. Parvoviral virions deploy a capsid-tethered lipolytic enzyme to
921 breach the endosomal membrane during cell entry. *Proc. Natl. Acad. Sci.* **102**, 17148–17153 (2005).
- 922 38. Suhre, K., Audic, S. & Claverie, J.-M. Mimivirus gene promoters exhibit an unprecedented
923 conservation among all eukaryotes. *Proc. Natl. Acad. Sci.* **102**, 14689–14693 (2005).
- 924 39. Legendre, M. *et al.* mRNA deep sequencing reveals 75 new genes and a complex transcriptional
925 landscape in Mimivirus. *Genome Res.* **20**, 664–674 (2010).

- 926 40. Smith, D. R., Arrigo, K. R., Alderkamp, A.-C. & Allen, A. E. Massive difference in synonymous
927 substitution rates among mitochondrial, plastid, and nuclear genes of *Phaeocystis* algae. *Mol.*
928 *Phylogenet. Evol.* **71**, 36–40 (2014).
- 929 41. Krupovic, M., Kuhn, J. H. & Fischer, M. G. A classification system for virophages and satellite viruses.
930 *Arch. Virol.* **161**, 233–247 (2016).
- 931 42. Suplatov, D. A., Besenmatter, W., Svedas, V. K. & Svendsen, A. Bioinformatic analysis of alpha/beta-
932 hydrolase fold enzymes reveals subfamily-specific positions responsible for discrimination of
933 amidase and lipase activities. *Protein Eng. Des. Sel.* **25**, 689–697 (2012).
- 934 43. Burt, A. & Koufopanou, V. Homing endonuclease genes: the rise and fall and rise again of a selfish
935 element. *Curr. Opin. Genet. Dev.* **14**, 609–615 (2004).
- 936 44. Chase, E., Desnues, C. & Blanc, G. *Integrated viral elements unveil the dual lifestyle of Tetraselmis*
937 *spp. polinton-like viruses*. Preprint at doi:10.1101/2022.05.02.489867 (2022).
- 938 45. Guillard, R. R. L. *Culture of Marine Invertebrate Animals: Proceedings — 1st Conference on Culture of*
939 *Marine Invertebrate Animals Greenport*. Springer US (1975).
- 940 46. Cottrell, M. & Suttle, C. Wide-spread occurrence and clonal variation in viruses which cause lysis of a
941 cosmopolitan, eukaryotic marine phytoplankter *Micromonas pusilla*. *Mar. Ecol. Prog. Ser.* **78**, 1–9
942 (1991).
- 943 47. Sullivan, M. B. DNA Extraction of Cesium Chloride-Purified Viruses using Wizard Prep Columns.
944 Protocol at <https://dx.doi.org/10.17504/protocols.io.c26yhd> (2016).
- 945 48. González-Domínguez, J. & Schmidt, B. ParDRe: faster parallel duplicated reads removal tool for
946 sequencing studies. *Bioinformatics* **32**, 1562–1564 (2016).
- 947 49. Krueger, F., James, F., Ewels, P., Afyounian, E. & Schuster-Boeckler, B. FelixKrueger/TrimGalore:
948 v0.6.7 - DOI via Zenodo. doi:10.5281/zenodo.5127899 (2021).
- 949 50. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell

- 950 Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
- 951 51. Deng, Z. & Delwart, E. ContigExtender: a new approach to improving de novo sequence assembly
952 for viral metagenomics data. *BMC Bioinformatics* **22**, 119 (2021).
- 953 52. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis.
954 *Nat. Methods* **9**, 671–675 (2012).
- 955 53. Patel, A. *et al.* Virus and prokaryote enumeration from planktonic aquatic environments by
956 epifluorescence microscopy with SYBR Green I. *Nat. Protoc.* **2**, 269–276 (2007).
- 957 54. Bolte, S. & Cordelières, F. P. A guided tour into subcellular colocalization analysis in light
958 microscopy. *J. Microsc.* **224**, 213–232 (2006).
- 959 55. Brussaard, C. P. D. Optimization of Procedures for Counting Viruses by Flow Cytometry. *Appl.*
960 *Environ. Microbiol.* **70**, 1506–1513 (2004).
- 961 56. Kirzner, S., Barak, E. & Lindell, D. Variability in progeny production and virulence of cyanophages
962 determined at the single-cell level. *Environ. Microbiol. Rep.* **8**, 605–613 (2016).
- 963 57. Ziv, I. *et al.* A Perturbed Ubiquitin Landscape Distinguishes Between Ubiquitin in Trafficking and in
964 Proteolysis. *Mol. Cell. Proteomics MCP* **10**, M111.009753 (2011).
- 965 58. HaileMariam, M. *et al.* S-Trap, an Ultrafast Sample-Preparation Approach for Shotgun Proteomics. *J.*
966 *Proteome Res.* **17**, 2917–2924 (2018).
- 967 59. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-
968 fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906
969 (2007).
- 970 60. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger:
971 ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat.*
972 *Methods* **14**, 513–520 (2017).
- 973 61. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or

974 nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

975 62. Lechner, M. *et al.* Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC*
976 *Bioinformatics* **12**, 124 (2011).

977 63. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using
978 DIAMOND. *Nat. Methods* **18**, 366–368 (2021).

979 64. Mirdita, M. *et al.* ColabFold - Making protein folding accessible to all. Preprint at
980 <https://doi.org/10.1101/2021.08.15.456425> (2022).

981 65. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589
982 (2021).

983 66. O’Connell, J. *et al.* NxTrim: optimized trimming of Illumina mate pair reads. *Bioinformatics* **31**, 2035–
984 2037 (2015).

985 67. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for
986 large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–
987 1676 (2015).

988 68. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo
989 assembler. *GigaScience* **1**, 2047-217X-1–18 (2012).

990 69. Chevreux, B., Wetter, T. & Suhai, S. Genome Sequence Assembly Using Trace Signals and Additional
991 Sequence Information. in *Proceedings of the German Conference on Bioinformatics, GCB 1999*,
992 *October 4-6, 1999, Hannover, Germany* 45–56 (1999).

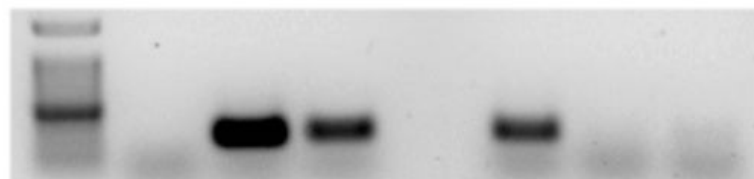
993 70. Walker, B. J. *et al.* Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and
994 Genome Assembly Improvement. *PLoS One* **9**, e112963 (2014).

995 71. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

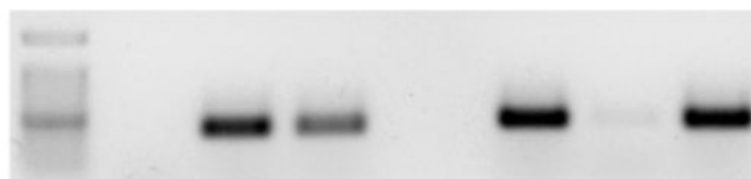
996 72. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence
997 alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–66 (2002).

- 998 73. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the
999 Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
- 1000 74. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the
1001 Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
- 1002 75. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAXML-NG: a fast, scalable and user-
1003 friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
- 1004 76. Barbera, P. *et al.* EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Syst. Biol.*
1005 **68**, 365–369 (2019).
- 1006 77. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment
1007 trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
- 1008 78. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis
1009 of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
- 1010 79. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation.
1011 *BMC Bioinformatics* **20**, 473 (2019).
- 1012 80. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of
1013 protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
- 1014 81. Bolduc, B. *et al.* vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect
1015 *Archaea* and *Bacteria*. *PeerJ* **5**, e3243 (2017).
- 1016 82. Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred
1017 Server at its Core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
- 1018 83. Heger, A. & Holm, L. Rapid automatic detection and alignment of repeats in protein sequences.
1019 *Proteins Struct. Funct. Bioinforma.* **41**, 224–237 (2000).
- 1020 84. Egge, E. S., Eikrem, W. & Edvardsen, B. Deep-branching Novel Lineages and High Diversity of
1021 Haptophytes in the Skagerrak (Norway) Uncovered by 454 Pyrosequencing. *J. Eukaryot. Microbiol.*

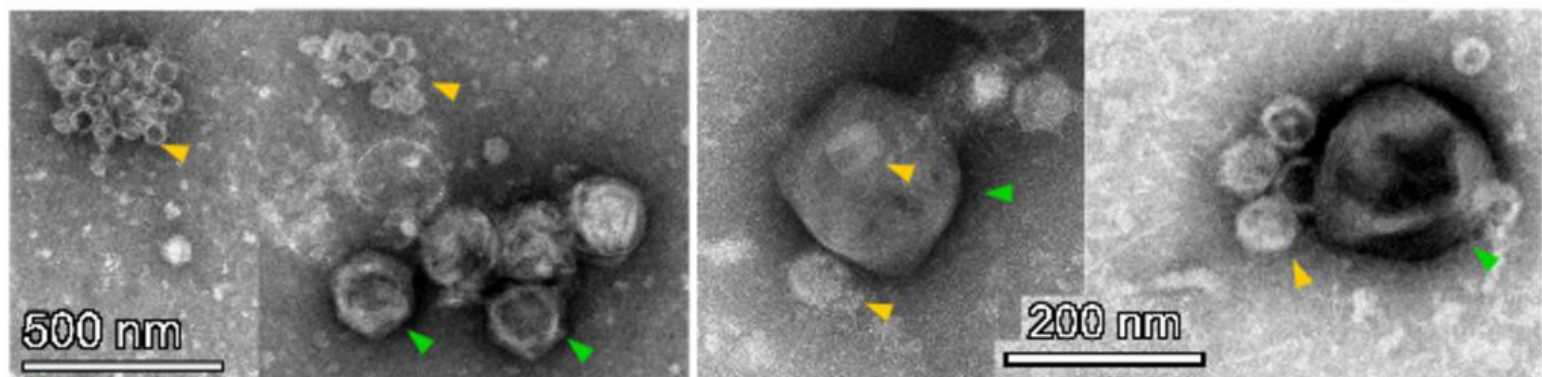
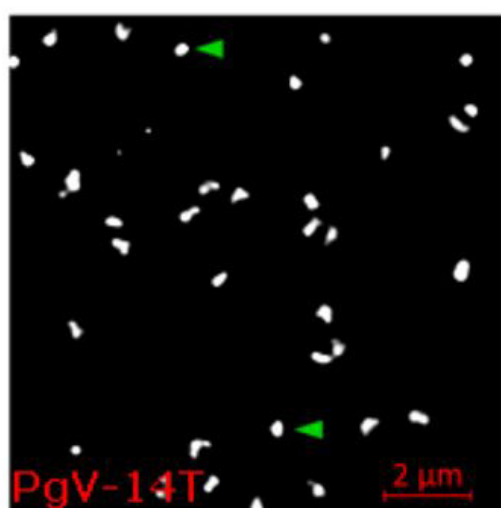
- 1022 **62**, 121–140 (2015).
- 1023 85. Hovde, B. T. *et al.* Chrysochromulina: Genomic assessment and taxonomic diagnosis of the type
1024 species for an oleaginous algal clade. *Algal Res.* **37**, 307–319 (2019).
- 1025 86. Andersen, R. A., Bailey, J. C., Decelle, J. & Probert, I. *Phaeocystis rex* sp. nov. (Phaeocystales,
1026 Prymnesiophyceae): a new solitary species that produces a multilayered scale cell covering. *Eur. J.*
1027 *Phycol.* **50**, 207–222 (2015).
- 1028 87. Stepanova, O. A. Black Sea algal viruses. *Russ. J. Mar. Biol.* **42**, 123–127 (2016).
- 1029 88. Alarcón-Schumacher, T., Guajardo-Leiva, S., Antón, J. & Díez, B. Elucidating Viral Communities
1030 During a Phytoplankton Bloom on the West Antarctic Peninsula. *Front. Microbiol.* **10**, 1014 (2019).
- 1031

a -14T

M NTC + +F L B F

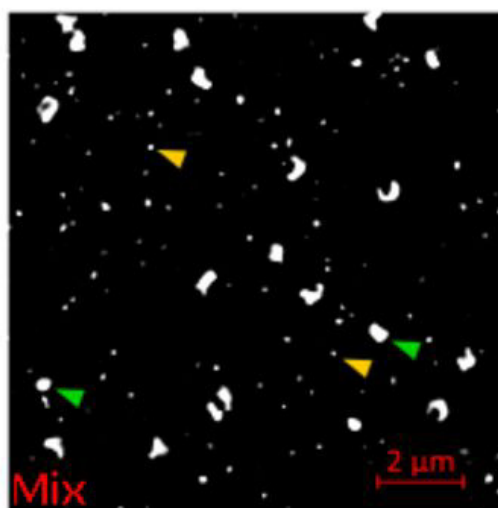
b**Gezel-14T**

M NTC + +F L B F

c**d**

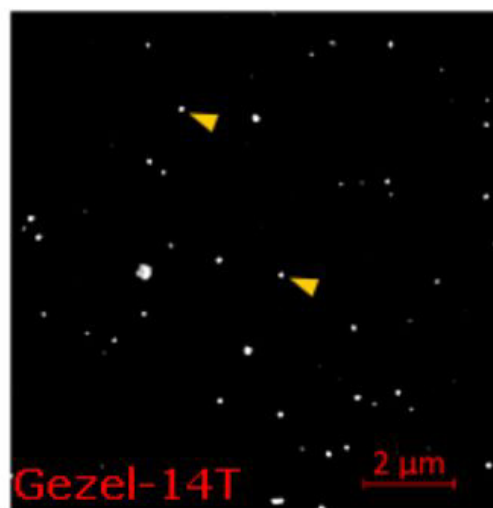
PgV-14T

2 μm



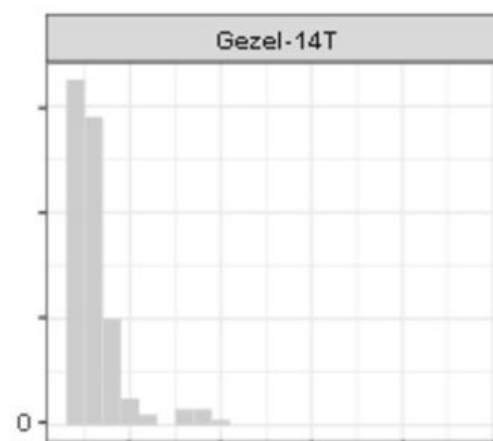
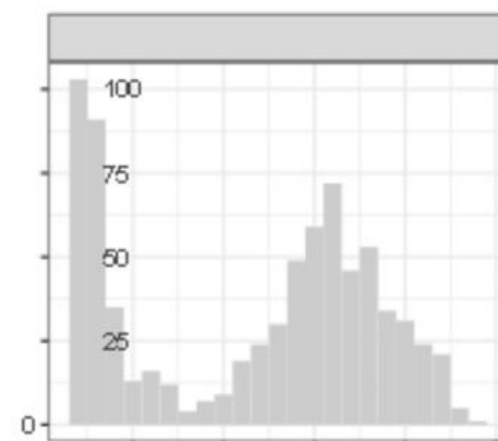
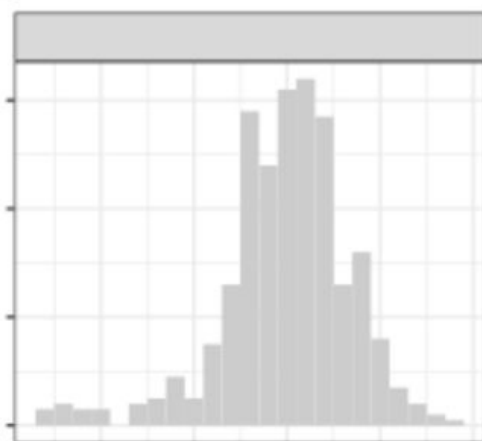
Mix

2 μm



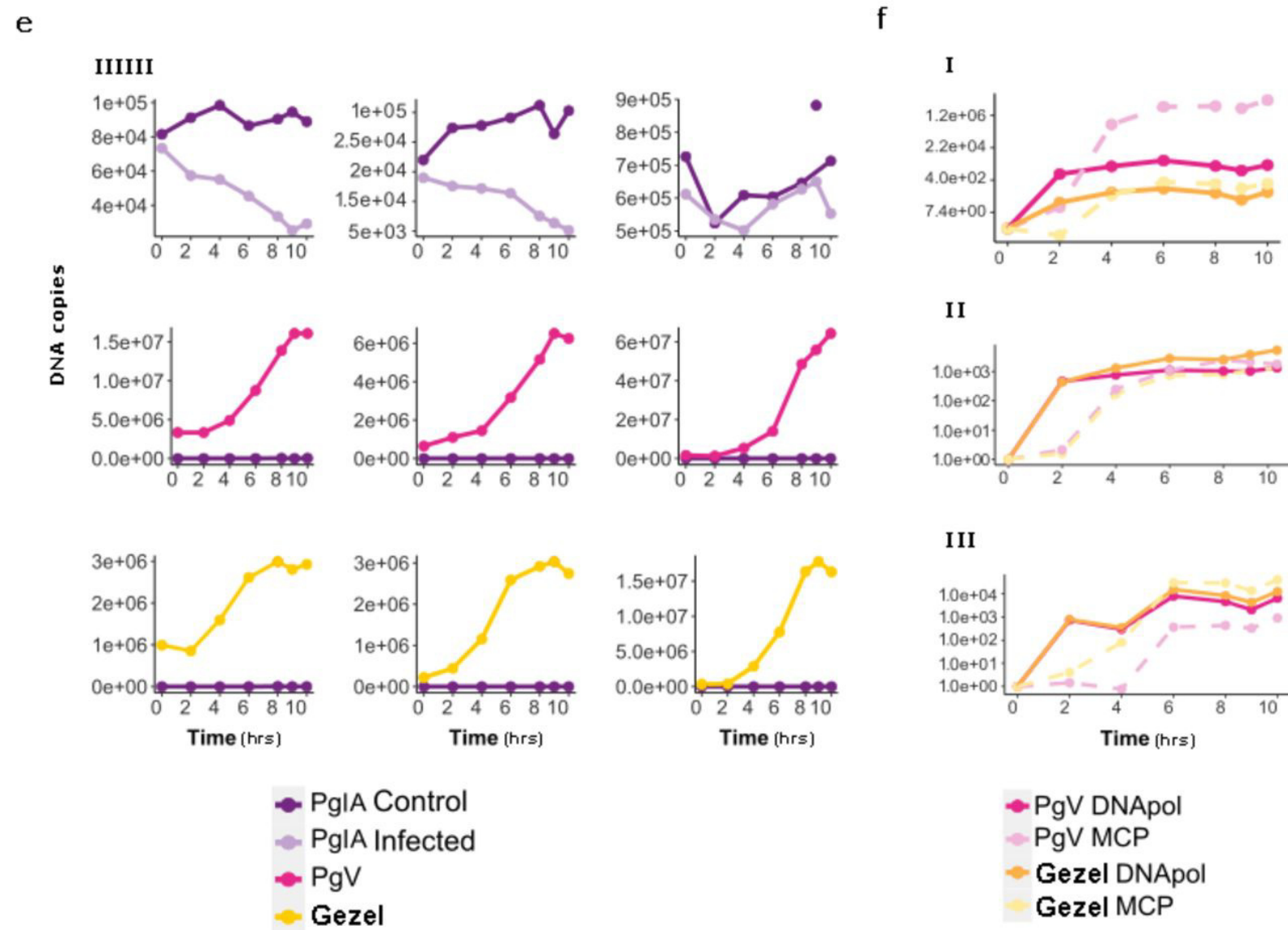
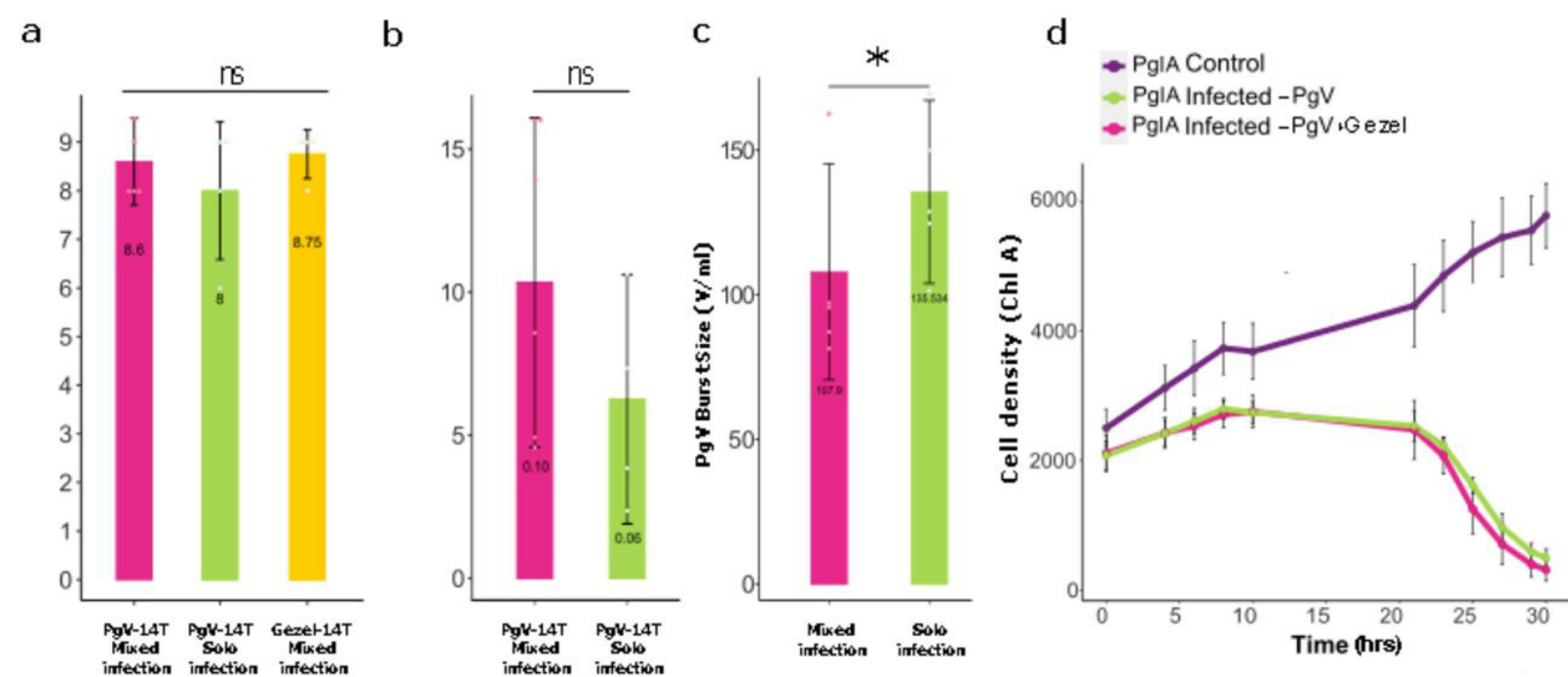
Gezel-14T

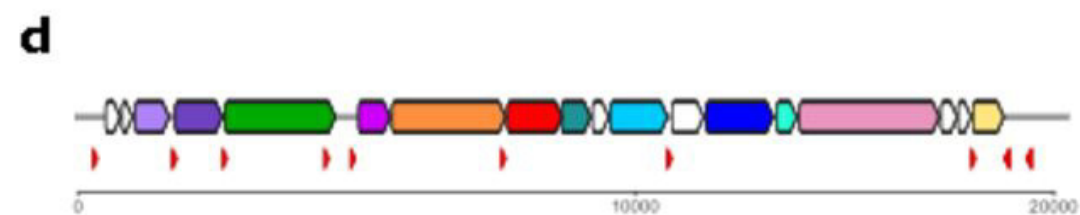
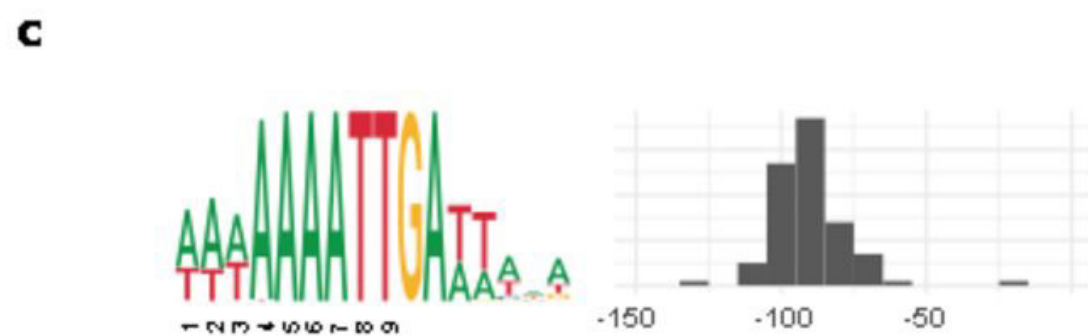
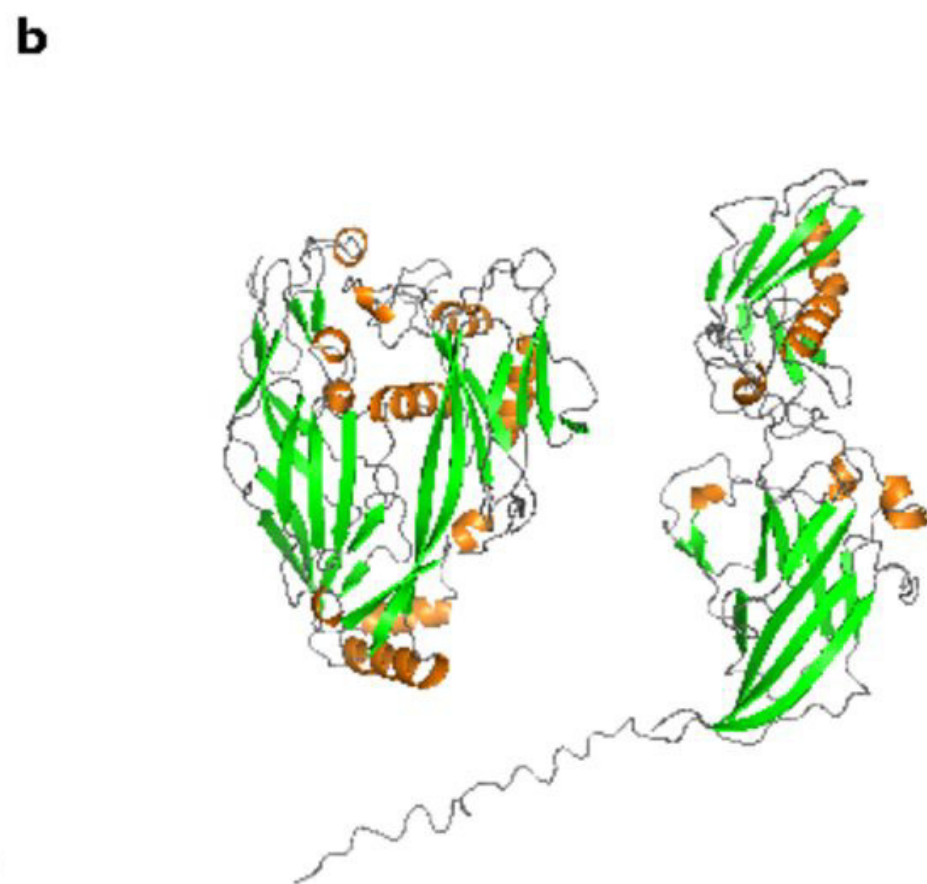
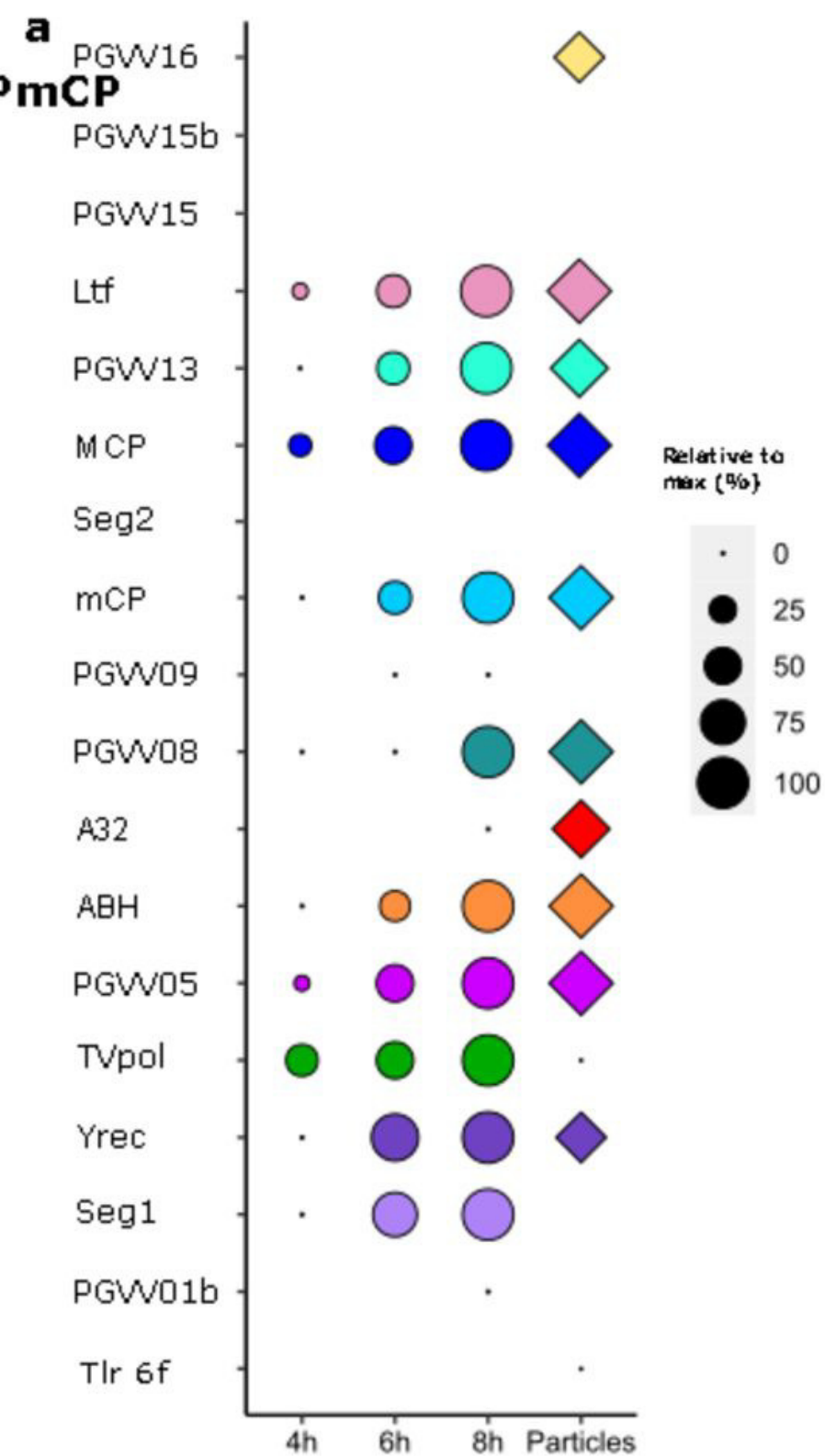
2 μm

e

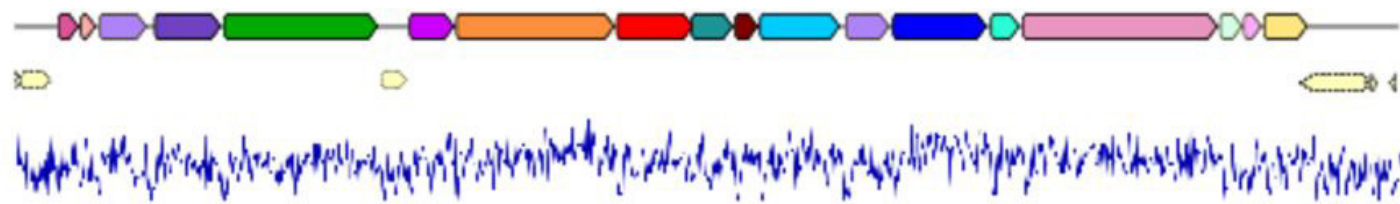
Gezel-14T

Apparent volume, μm^3

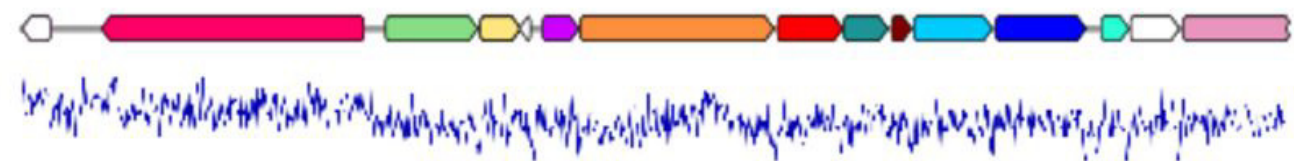




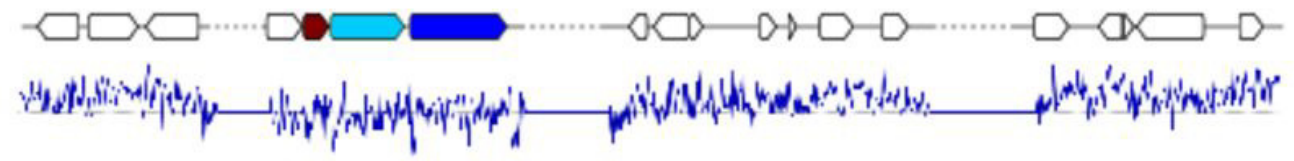
Gezel-14T



Phaglo-R



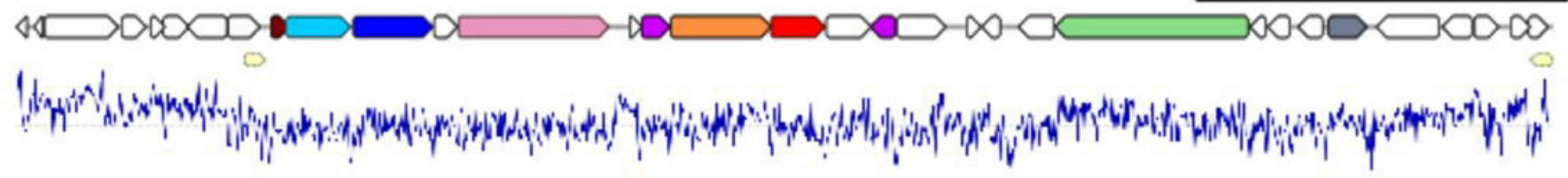
Phaglo-B1



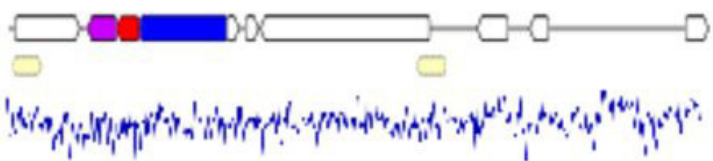
Phaglo-B2



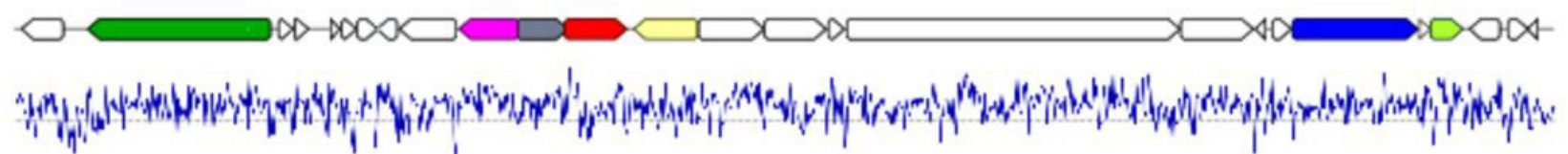
Phaglo-Y



Phaglo-P

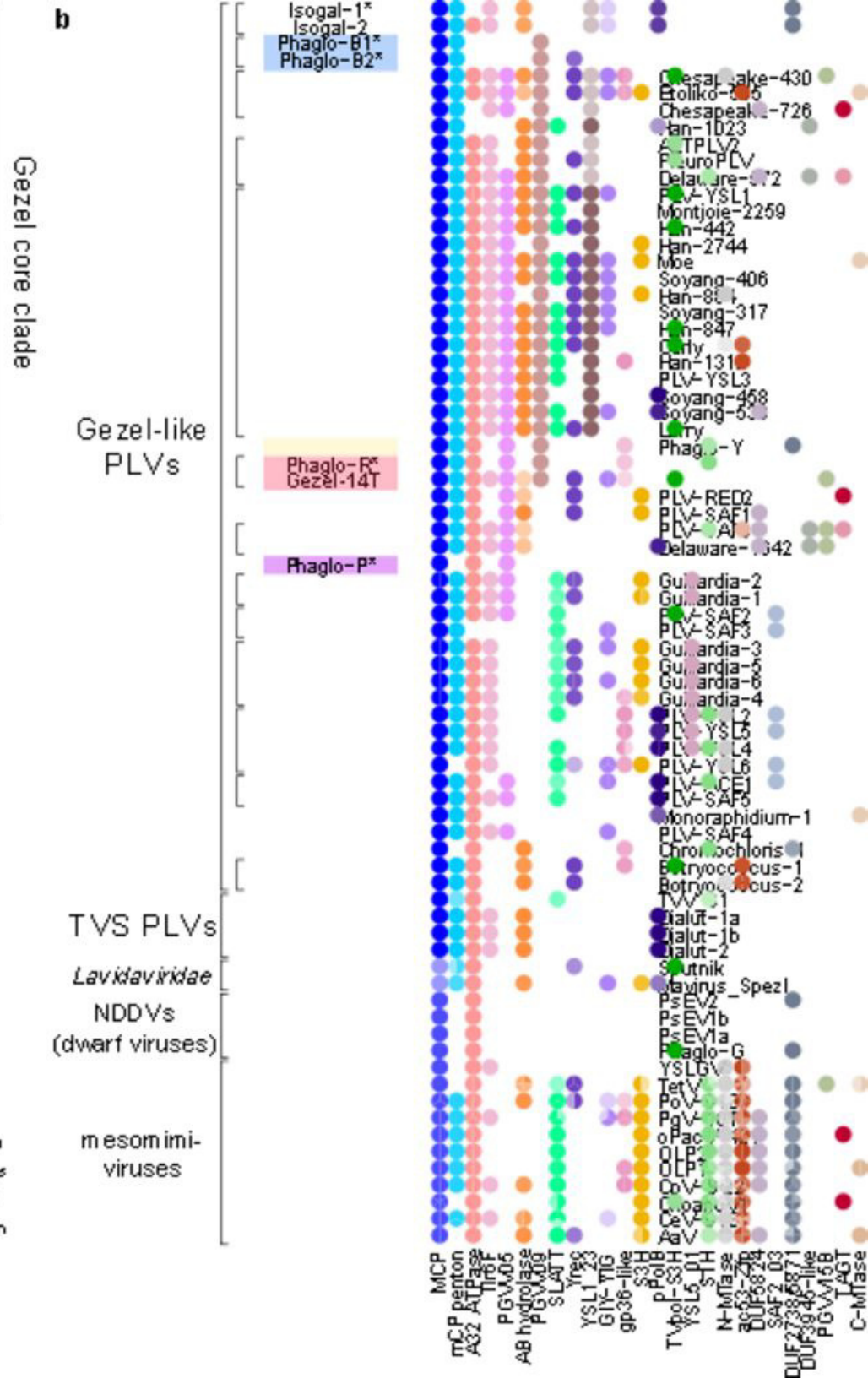
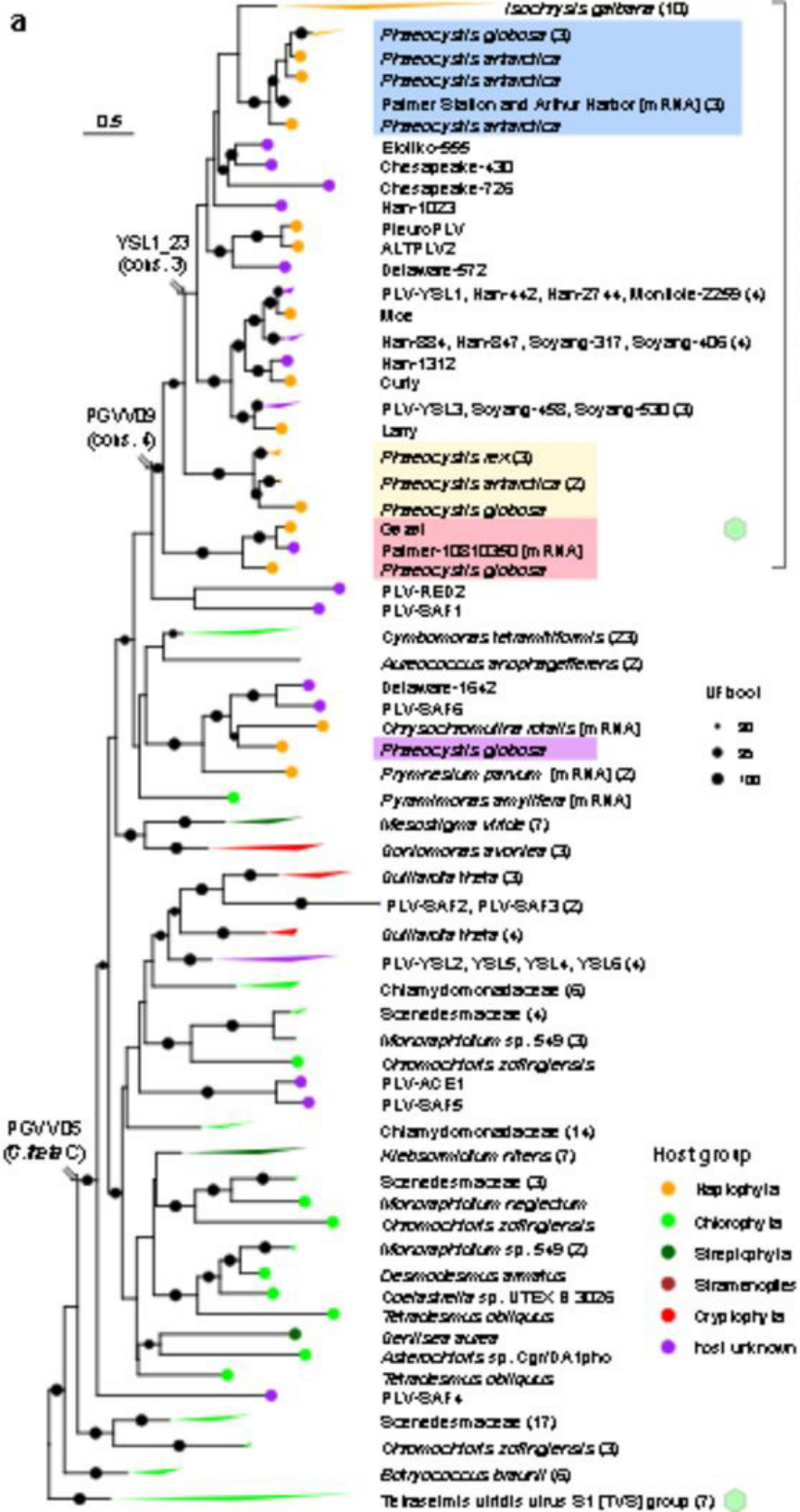


Phaglo-G



Family	
Tlr6F	gp36-like
PGV01b	PGV15
GIY-YIG	PGV15b
Yrec	PGV16
TVpd-S3H	S1H
PGV05	RT
ABH	FkbM
A32	RuvC
PGV08	Yqj
PGV09	VLTF3
mCP	DUF2738
MCP	Hyp
PGV13	Repeat

1 2,000 4,000 6,000 8,000 10,000 12,000 14,000 16,000 18,000 20,000 22,000 24,000 26,050



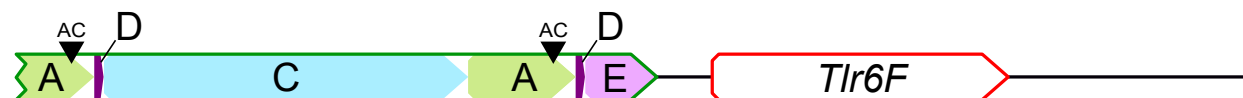
Gezel-16T 5' TIR (+ strand)



Gezel-16T 3' TIR (- strand)



Gezel-14T 5' TIR (+ strand)




Gezel-14T 3' TIR (- strand)



 TIRs flanking the genome

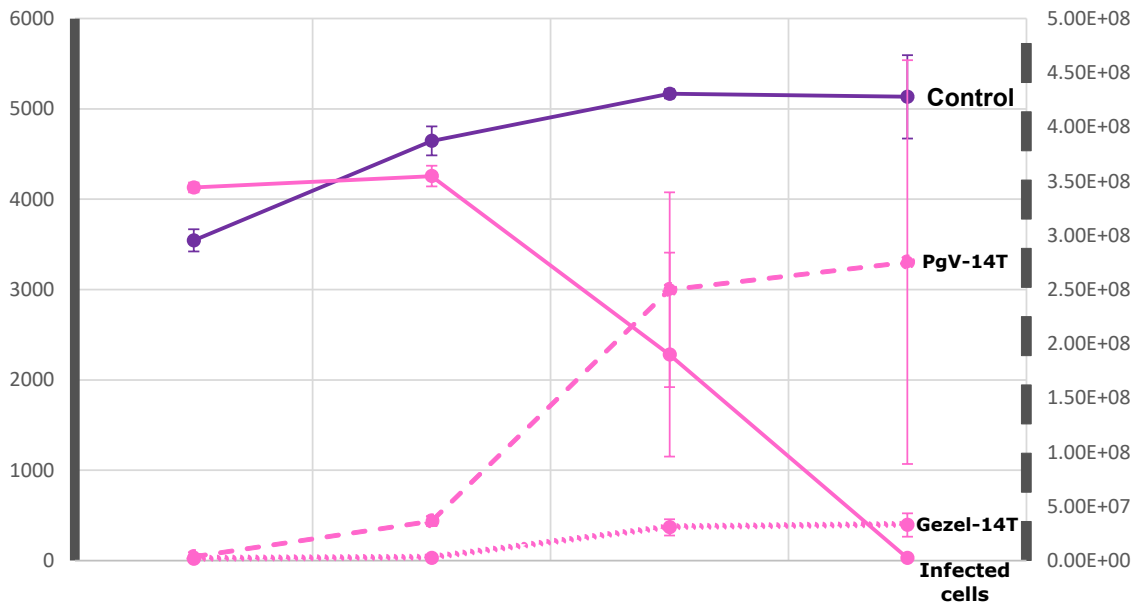
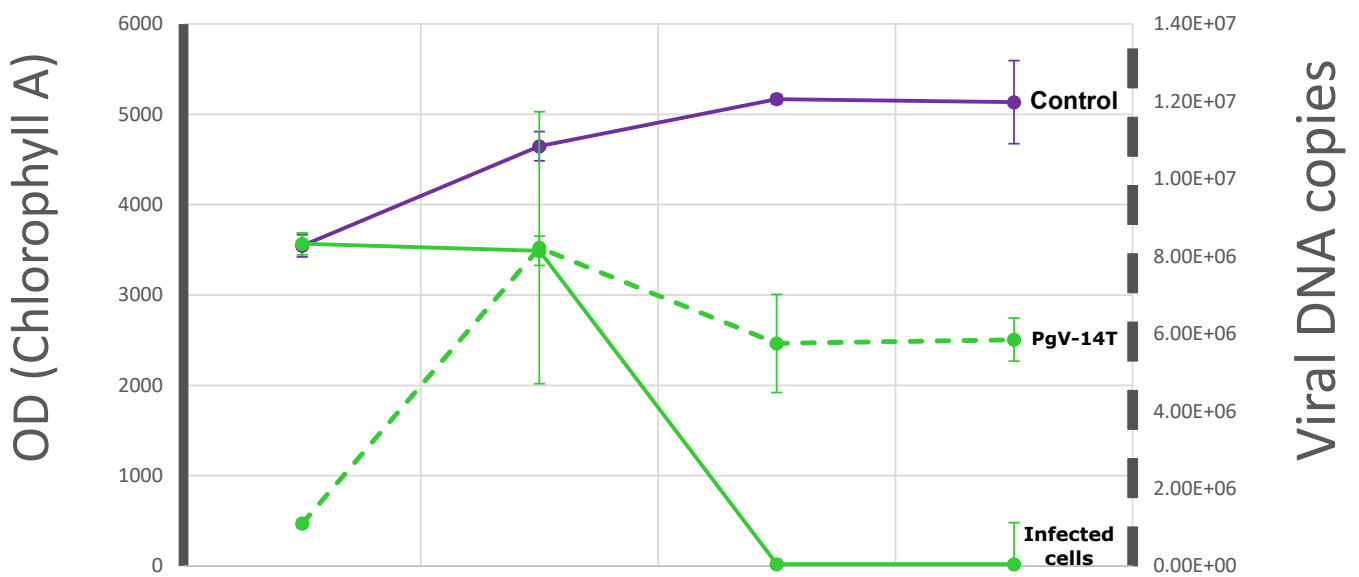
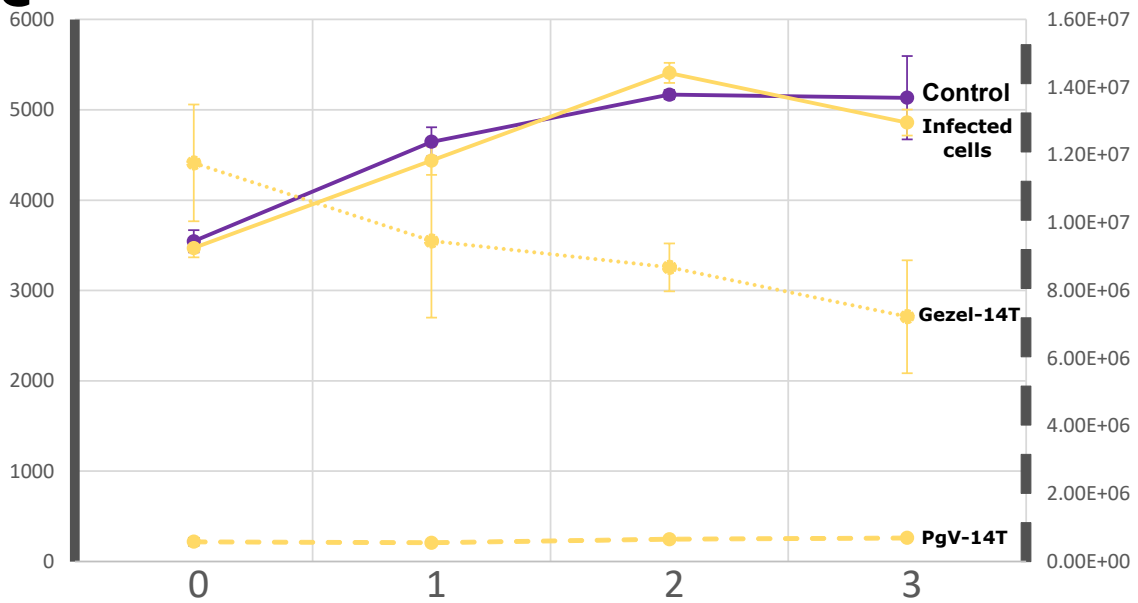
 Repeat units

 ORFs

 Internal inverted repeat

200 nt



a**b****c**

Days

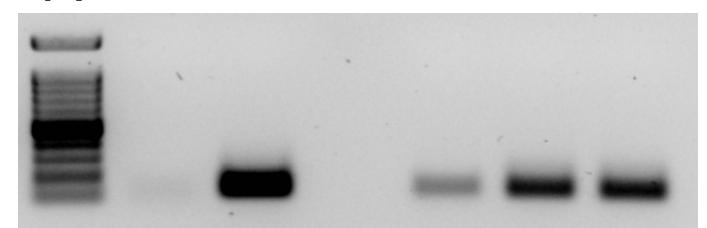
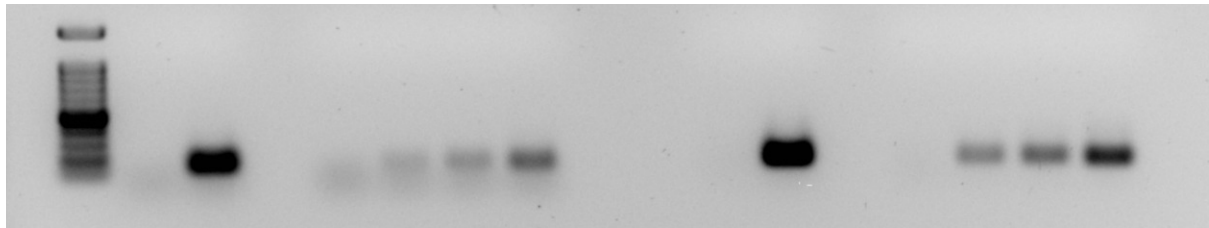
pgvv01 (Tlr 6f)

pgvv09

pgvv01b

M - + 2 4 6 8 (hs) - + 2 4 6 8 (hs)

M - + 2 4 6 (hs)



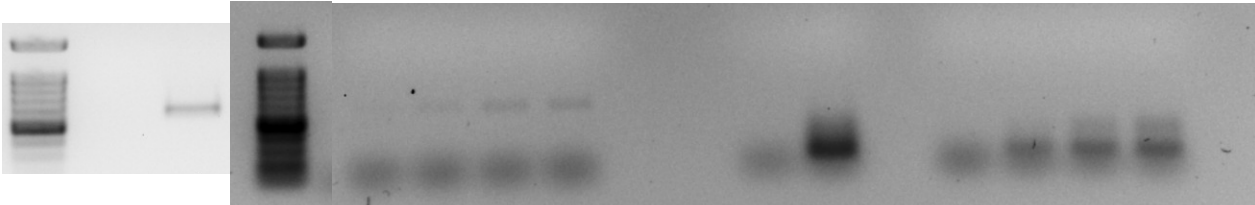
pgvv11

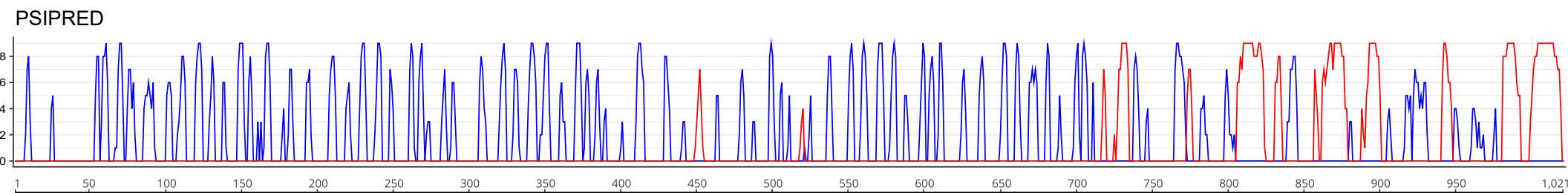
pgvv15

pgvv15b

M - + M 2 4 6 8 (hs) - + 2 4 6 8 (hs)

M - + 2 4 6 (hs)





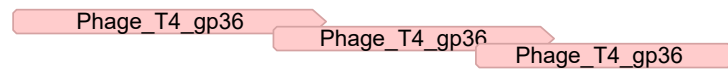
Deletion in Gezel-14T



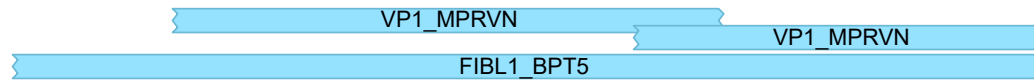
RADAR

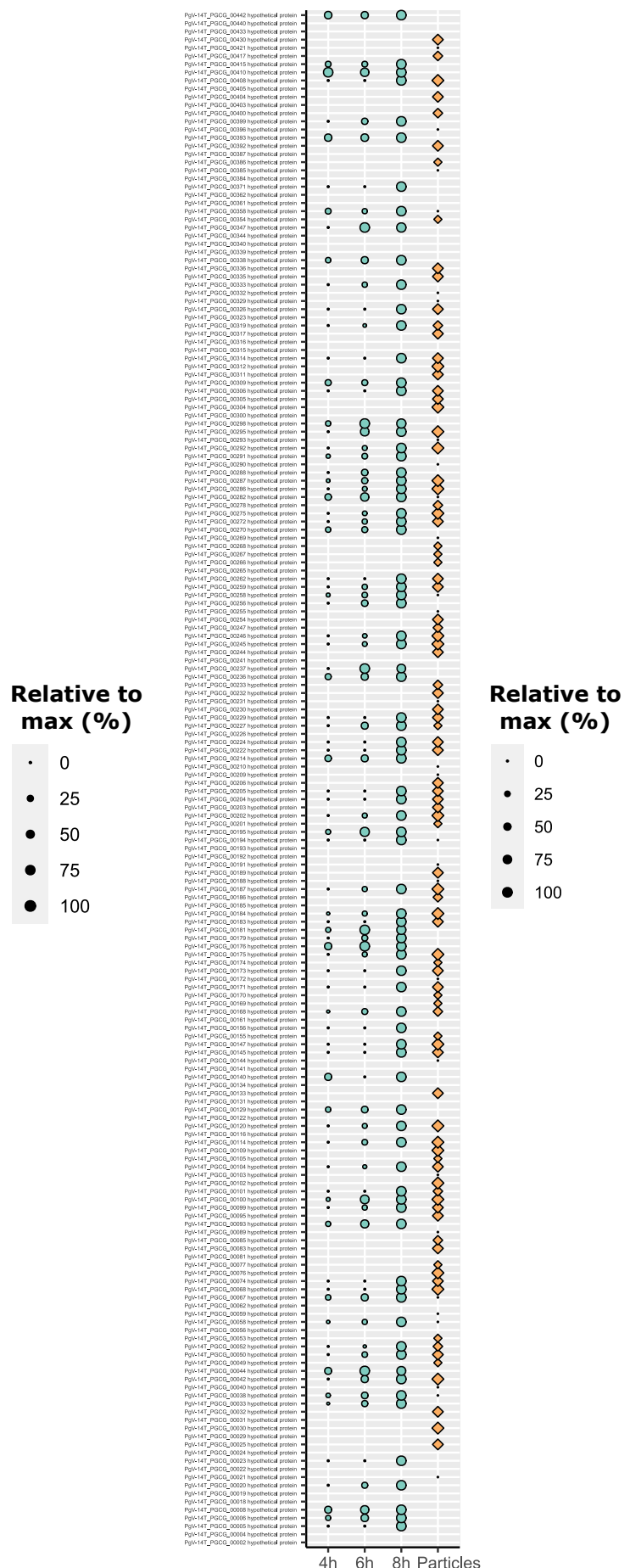
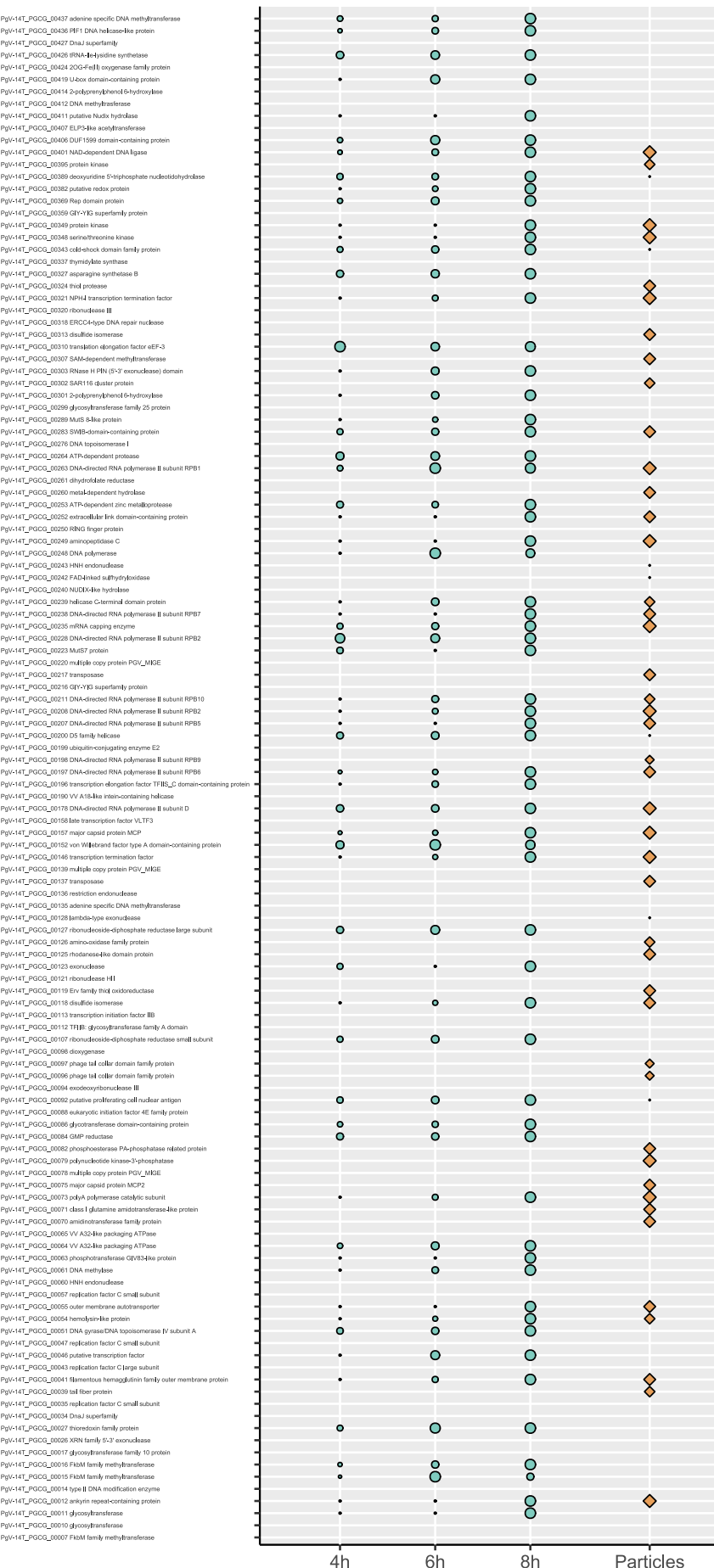


hhsearch: Pfam

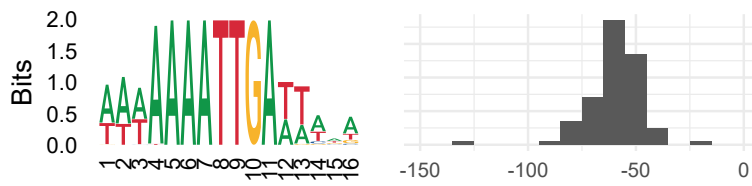


hhsearch: Uniprot

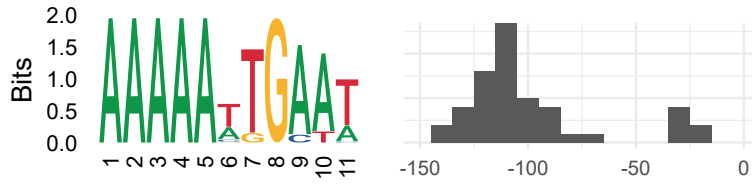




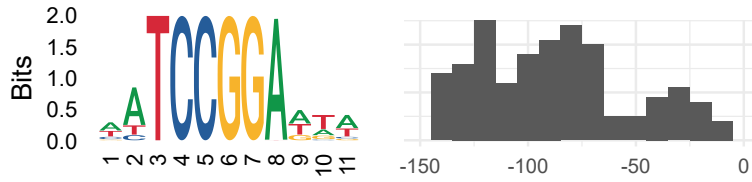
PgV-16T
WAWAAAATTGAWTWWA
Sites = 93 (21.4%)
E-value = 1.7e-61



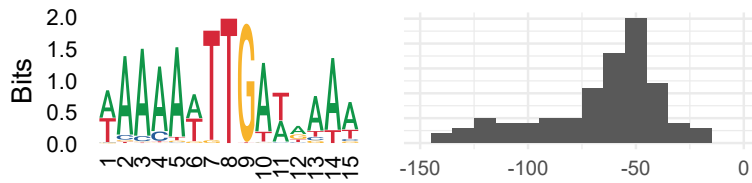
Tetraselmis virus 1
AAAAAWTGAAT
Sites = 47 (7.2%)
E-value = 3.2e-15



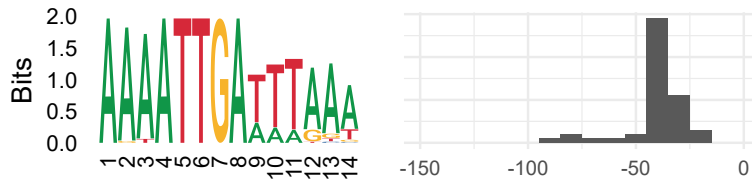
Tetraselmis virus 1
WATCCGGAWTW
Sites = 198 (30.3%)
E-value = 9.3e-237



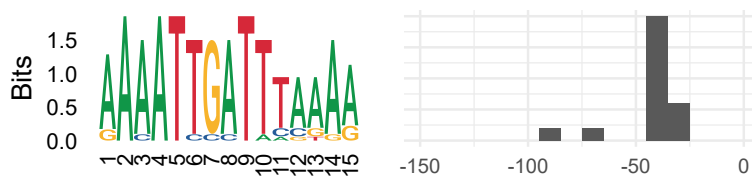
Pyramimonas orientalis virus 01B
WAAAATTGAWRAAA
Sites = 103 (20.4%)
E-value = 1.7e-21



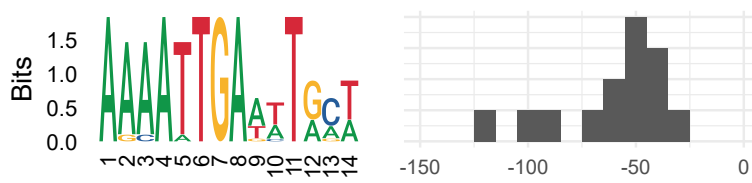
Organic Lake phycodnavirus 1
AAAATTGATTTAAA
Sites = 50 (12.5%)
E-value = 2.1e-42



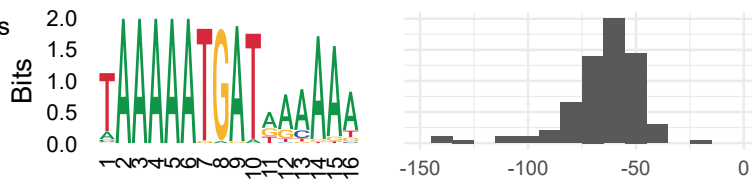
Organic Lake phycodnavirus 2
AAAATTGATTTAAA
Sites = 15 (4.6%)
E-value = 9800



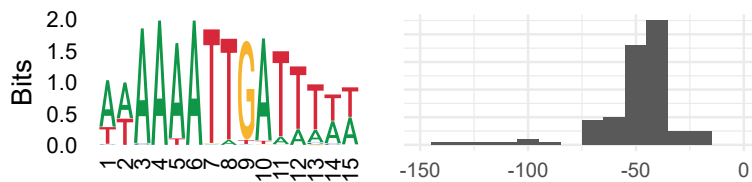
Yellowstone lake mimivirus
AAAATTGAWTRCW
Sites = 14 (13.7%)
E-value = 0.26



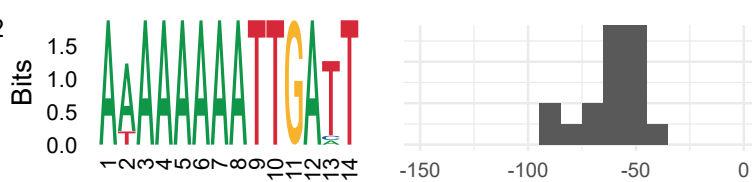
Aureococcus anophagefferens virus
TAAAATGATRAAAA
Sites = 128 (34.1%)
E-value = 1.6e-241

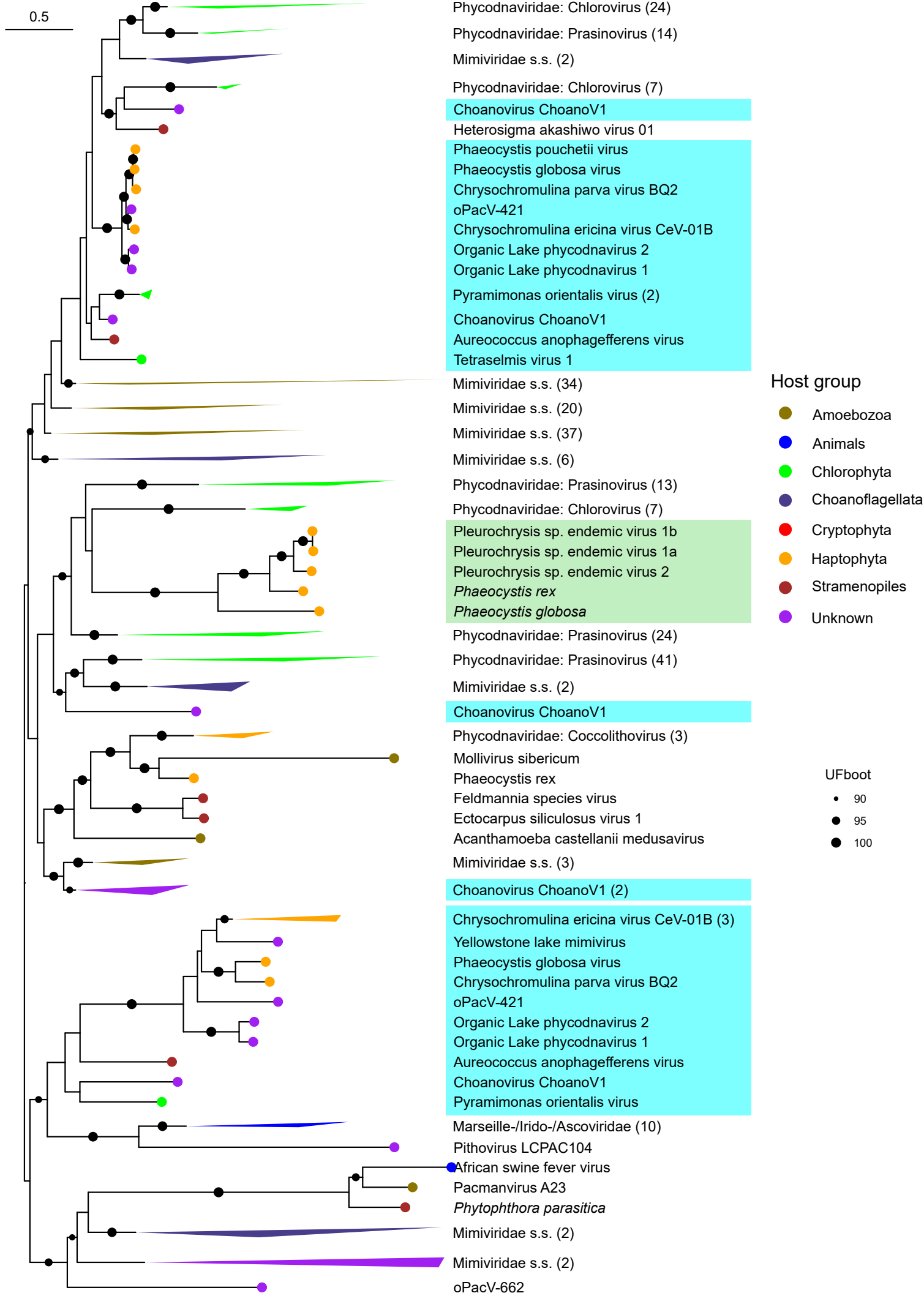


Choanovirus ChoanoV1
AWAAAATTGATTTWW
Sites = 118 (13.8%)
E-value = 7.3e-20

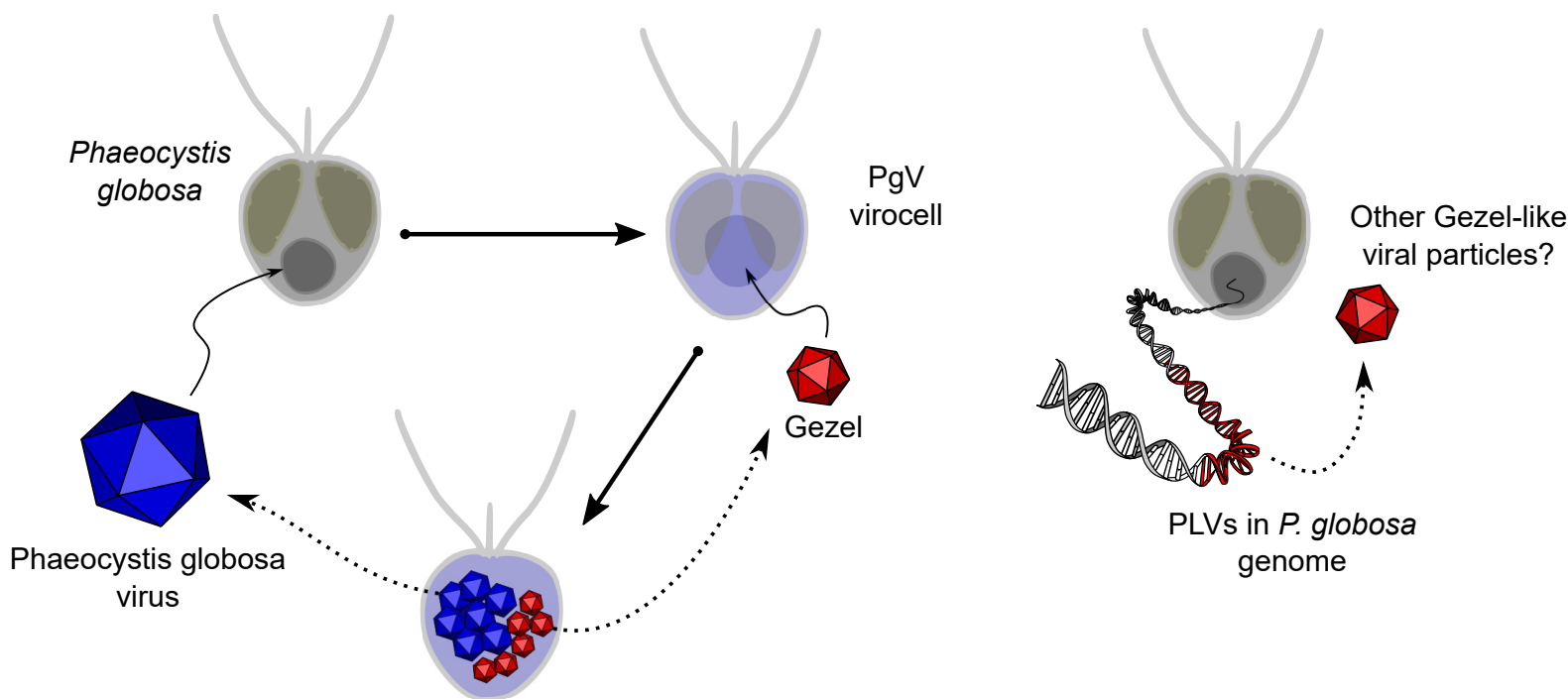


Chrysochromulina parva virus BQ2
AAAAAAAATTGATT
Sites = 18 (4.4%)
E-value = 0.00091





a



b

