



HAL
open science

Giant virus biology and diversity in the era of genome-resolved metagenomics

Frederik Schulz, Chantal Abergel, Tanja Woyke

► **To cite this version:**

Frederik Schulz, Chantal Abergel, Tanja Woyke. Giant virus biology and diversity in the era of genome-resolved metagenomics. *Nature Reviews Microbiology*, 2022, 20 (12), pp.721-736. 10.1038/s41579-022-00754-5 . hal-03934861

HAL Id: hal-03934861

<https://amu.hal.science/hal-03934861v1>

Submitted on 11 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Public Domain

1 Giant viruses in an era of genome-resolved metagenomics

3 Authors

4 Frederik Schulz*1, Chantal Abergel2, Tanja Woyke*1

6 *corresponding authors: fschulz@lbl.gov, twoyke@lbl.gov

8 Author affiliations

9 1 DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California,
10 USA

11 2 Aix Marseille University, CNRS-AMU, Marseille, France

13 Abstract

14 The discovery of giant viruses, with capsids as large as some bacteria, megabase-range
15 genomes, and a wide variety of traits typically found only in cellular organisms, was one of the
16 most spectacular breakthroughs in biology. Until recently, most of our knowledge about these
17 giant viruses came from ~100 species-level isolates for which genome sequences were
18 available. However, these isolates were primarily derived from laboratory-based co-cultivation
19 with few cultured protists and algae and thus did not reflect the true diversity of giant viruses in
20 the environment. While virus co-cultures enabled important detailed insights into giant virus
21 biology, many questions regarding the origin, evolution and ecological importance of giant
22 viruses remain unanswered. With recent advances in sequencing technologies and
23 bioinformatics, our understanding of giant viruses has drastically expanded, as genome-
24 resolved metagenomics and environmental surveys of viral hallmark genes are beginning to
25 shed light on the taxonomic breadth of this group of viruses, their biogeography and ecology,
26 and their evolutionary trajectories. In this review, we summarize our current understanding of
27 giant virus diversity and biology based on viral isolates, as laboratory cultivation has facilitated
28 extensive insights into viral morphological features and infection strategies. We then expand on
29 how more recent cultivation-independent approaches led to an unparalleled increase of the
30 coding potential and diversity of the Nucleocytoviricota. We further elaborate on how
31 metagenomics has revolutionized our perspective of giant viruses by revealing their distribution
32 across our planet's biomes where they impact the biology and ecology of a wide range of
33 eukaryotic hosts and ultimately affect global nutrient cycles.

35 Introduction

36 Large and giant viruses are part of a group of double stranded DNA viruses, the
37 Nucleocytoplasmic Large DNA Viruses (NCLDV)^{1,2}, which constitutes the viral phylum
38 Nucleocytoviricota³. Viruses of this phylum infect a wide range of eukaryotic hosts, from the
39 tiniest known unicellular choanoflagellates to multicellular animals⁴. NCLDV typically replicate in
40 so-called viral factories built in the host cytoplasm or use the host nucleus to replicate and
41 sometimes assemble their progeny^{5,6}. Hallmark features of these viruses are large genomes,
42 greater than 100 kb and up to 2.5 Mb, and virions that can reach more than 2 microns in
43 length⁷. The term giant virus was coined in the early 2000s, when the first virus with a genome
44 in the megabase range was discovered; initial light microscopy observations led to the

45 assumption that its particles corresponded to a gram positive bacterial pathogen of amoebae^{8,9}.
46 More detailed ultrastructural analysis revealed the typical icosahedral shape of viruses and
47 genome sequencing yielded a 1.2 Mb viral genome¹⁰. This virus was named “Mimivirus”, short
48 for microbe mimicking virus, and represented an unexpected novelty in the viral world, not only
49 due to its exceptional particle and genome sizes but also its coding potential that included
50 several genes with possible roles in protein biosynthesis¹⁰. Since this discovery of giant viruses
51 their coding potential has been full of surprises and the presence of hallmark genes of cellular
52 life led to the hypothesis that these viruses might represent an enigmatic fourth domain of life^{10–}
53 ¹². Equally intriguing, much smaller viruses, so-called “virophages”, were found to infect some
54 NCLDV that have exclusively cytoplasmic infectious cycles. Virophages parasitize and
55 sometimes eradicate their viral hosts¹³. Also discovered was a third partner coined transpoviron,
56 which corresponds to a 7 kb dsDNA episome, able to propagate using both the giant virus and
57 the virophage particles as vehicles^{14,15}.

58

59 For well over a decade, giant viruses had chiefly been discovered through cultivation-based
60 approaches until very recently when virology followed the footsteps of microbial genomics by
61 applying cultivation-independent metagenomics to investigate the biology of these viruses at an
62 unparalleled pace. In the following, we build on a wealth of experimental data that has revealed
63 many insights into giant virus biology, in particular their virion structure and unique infection
64 strategies. We expand this framework by integrating the latest sequence-based studies that
65 expanded NCLDV diversity, biogeography, coding potential and putative host range. Further,
66 there is now compelling evidence that the presence of a variety of cellular hallmark genes in
67 giant virus genomes enable the virus to reprogram host metabolism, and that direct integration
68 of giant virus genetic material into host genomes may impact the biology and evolution of the
69 eukaryotic cell.

70

71 **Giant virus discovery through isolation**

72 The earliest discovered NCLDVs were the Poxviruses, the causative agents of smallpox, which
73 were the first viral particles seen under a microscope more than 130 years ago¹⁶. The genome
74 of Vaccinia virus was sequenced in the early 1990s¹⁷ and shortly thereafter additional genomes
75 of Poxviruses were sequenced (Figure 1) with sizes ranging from 120 to 360 kb¹⁸.

76 Subsequently, other viruses that infect animals, including Ascoviruses, Iridoviruses and
77 Asfarviruses were found and their genomes sequenced^{19–21}. Genomes of viruses in these
78 groups were comparably small with up to 220 kb, and even smaller in the recently discovered
79 shrimp-associated Mininucleoviridae with only 70–80 kb²². In addition to animal-infecting
80 NCLDV, a wide range of NCLDV were detected in diverse eukaryotic algae, including
81 chlorophytes, haptophytes, pelagophytes, brown algae and dinoflagellates in the early 2000s²³.
82 These algae-associated NCLDV were affiliated with the Phycodnaviruses²³ and
83 Mesomimivirinae²⁴, and although most of their genomes are around 200–500 kb^{23,25}, the
84 genomes of Tetraselmis Virus and *Prymnesium kappa* virus RF01 are 668 kb²⁶ and 1.4 Mb²⁷,
85 respectively.

86

87 After the discovery of Mimivirus in 2003⁹ other NCLDV with larger virions and genomes above
88 500 kb have been found to infect heterotrophic protists²⁸, mainly members of the Amoebozoa.

89 For more than a decade *Acanthamoeba* strains had chiefly been used as bait for the co-
90 cultivation of new viruses, leading to the frequent isolation of closely related giant viruses able to
91 infect this unicellular host²⁹. *Acanthamoeba* has proven as a particularly suitable host for many
92 Megamimivirinae and Marseilleviridae²⁹. Consequently, viruses from these taxonomic groups
93 are currently among the most thoroughly cultivated NCLDV with more than 30 genome
94 sequences readily available in public databases, including the novel Megamimivirinae lineages
95 Tupanvirus⁷ and Cotonvirus³⁰. The co-cultivation approach has been widely successful and also
96 led to the recovery of isolates from divergent NCLDV clades, facilitating the organization and
97 naming of Pithoviruses, Pandoraviruses, Molliviruses and, most recently, Medusavirus³¹. More
98 recently the use of alternative hosts, such as *Vermamoeba*, has led to the co-cultivation of
99 several new Faustovirus isolates³² and some Orpheoviruses³³, Pacmanvirus³⁴ and
100 Kaumoebaviruses³⁵, all distant relatives of Pitho- and Asfarviruses. A newly developed high-
101 throughput co-cultivation-based approach using high content screening has proven a powerful
102 tool for giant virus discovery and isolation³⁶. Yet, co-cultivation is limited by host specificity of
103 giant viruses⁴; some NCLDV lineages are able to infect only specific hosts, such as certain
104 species of *Acanthamoeba*³⁷, while others may be more versatile, exhibiting a broader host
105 range⁷. Considering the immense diversity of eukaryotes³⁸, and in particular of microeukaryotes,
106 it is likely that giant viruses that have been recovered through isolation reflect only a minute
107 fraction of NCLDV lineages extant in the wild.

108

109 **Virion structures and infection strategies in the Nucleocytoviricota**

110 One of the best studied and most well understood infectious cycles is that of the amoeba-
111 infecting Mimivirus⁸. Similar to other known members of the Mimiviridae it replicates in its host's
112 cytoplasm³⁹⁻⁴¹ (Figure 2A). The ~700 nm icosahedral virion of Mimivirus (Figure 2B) contains
113 RNA polymerase and transcript maturation machinery together with the genome. Mimivirus
114 enters its host by triggering phagocytosis upon adhering to the host cell membrane with its
115 glycosylated fibrils. Bacterial-type sugars are uniquely synthesized by the virally encoded
116 glycosylation machinery and are the building blocks of the recently determined complex 70 kDa
117 and 25 kDa polysaccharide structures that decorate the Mimivirus fibrils⁴². The structure of the
118 Mimivirus capsid has also been determined, but due to its large size, the triangulation (T)-
119 number remains unknown⁴³. Detailed Atomic Force Microscopy provided additional insights into
120 virion composition⁴⁴, further underlining the complexity of the capsid and an enclosed
121 compartment, the so-called nucleoid, that contains hundreds of proteins in addition to the
122 structural ones. It has been proposed that the non-structural proteins are required to initiate the
123 viral infectious cycle, protect the virion from oxidative stress and perform early transcription^{5,40}.
124 Once in the host vacuole, the Mimivirus virion features a specific structure, the stargate, at one
125 vertex of the icosahedron. Upon opening, the internal membrane is pulled outside the capsid to
126 fuse with the host vacuole's membrane⁴⁵. Next, the nucleoid is transferred into the host
127 cytoplasm^{46,47} where early transcription begins using the virally encoded transcription
128 machinery, which, at first, remains confined in the nucleoid⁴⁸. The accumulation of nucleic acids
129 due to active transcription and replication leads to an increase in size of the virion factory and
130 newly synthesized virions start budding at its periphery, recycling host cell membranes derived
131 from the endoplasmic reticulum^{46,48} or golgi³⁰. The last step of virion maturation, after genome

132 loading into the nucleoid, is the addition of the fibril layer to the capsids⁴⁹, and then hundreds of
133 newly synthesized virions are released after cell lysis.

134
135 Several viruses related to Mimivirus have similar infectious cycles but smaller virions. Among
136 them is *Cafeteria roenbergensis* virus, which presents an icosahedral capsid of 300 nm in
137 diameter (Figure 2B) with a lipidic membrane underneath the capsid shell. Its mode of infection
138 is not fully understood but, similar to Mimivirus, a nucleoid structure has been observed in the
139 cytoplasm as well as extracellular empty capsids, supporting an external opening of the capsids
140 followed by the fusion of the internal membrane with that of the cell, thus allowing the transfer of
141 the nucleoid into the host cytoplasm. Virions contain about 150 proteins, which either make up
142 the icosahedral capsid or are necessary to initiate the infectious cycle⁴⁶. Neo-synthesized
143 virions assemble during the late stage of infection and are released through cell lysis. The
144 structure of the complex capsid, determined by cryo-EM, corresponds to a T-number of 499 and
145 provided a new model for capsid assembly⁵⁰. Another member of the Mimiviridae, with a similar
146 icosahedral capsid of 300 nm in diameter, is *Bodo saltans* virus. Its capsid appears to be made
147 by two proteinaceous layers surrounded by 40nm long fibrils. A possible stargate-like structure
148 is present at one vertex of the capsid and there are two membranes, one lining the external
149 protein shell and one internal to the nucleoid compartment containing the genome. The
150 infectious cycle is similar to the one of Mimivirus except that the host's nuclear genome appears
151 to be degraded. The viral factory develops at the posterior pole of the cell to fill two thirds of the
152 cell space, pushing aside the nucleus and organelles. Lipid vesicles are recruited for virion
153 assembly that takes place at one side of the viral factory and mature virions detach after
154 genome loading to migrate to the posterior pole of the cell. Virions are released by budding in
155 vesicles from the host membrane and ultimately after cell lysis⁴¹.

156
157 Virion morphology and infection mechanisms differ vastly between viruses of the
158 Nucleocytoviricota. In particular, viruses that infect algae have been found to form small virions.
159 Among the smallest members of the Nucleocytoviricota are Prasinoviruses with virion diameters
160 of around 120 nm. This feature is important for infecting and replicating within *Ostreococcus*,
161 which is one of the smallest free-living eukaryotes with only 0.8 μm cells⁵¹. Following viral
162 infection, genome replication begins almost immediately and within hours new virions assemble
163 in the cytoplasm; in less than 24 hours post-infection, cellular lysis of the host occurs. The host
164 cell nucleus, mitochondria and chloroplast remain intact throughout this period. Slightly larger
165 are Chloroviruses with 190 nm in diameter, an icosahedral shape (T-number 169) and an
166 internal lipid membrane. A prototype for this group of viruses is *Paramecium bursaria* chlorella
167 virus⁵²(Figure 2B). It has a spike-like structure at one vertex and a few external fibers that
168 extend from some of the capsomers⁵³. The outer capsid layer covers a single lipid bi-layered
169 membrane, which is essential for infection. Its major capsid protein (Vp54) is a glycoprotein, and
170 three Vp54s form a trimeric capsomere, which has pseudo-six-fold symmetry⁵⁴. Some of the
171 largest viruses that infect algae are affiliated with the Mimiviridae. All of them have icosahedral
172 capsids with sizes ranging from 150 nm in *Aureococcus* virus (Figure 2B) to 370 nm in the
173 recently described *Prymnesium kappia* virus²⁷. These viruses also build a viral factory in the host
174 cytoplasm but it is still unknown if the transcription machinery is loaded in the capsids allowing
175 an entirely cytoplasmic infectious cycle.

176
177 Outside of the Mimiviridae there are smaller amoeba-infecting viruses such as the
178 Melbournevirus, which has an icosahedral virion of ~250 nm diameter (Figure 2B). The Cryo-
179 EM structure of the capsid was obtained for two members of the family at various resolutions
180 revealing a T-number of 309 and a complex capsid structure⁵⁵⁻⁵⁷ with many minor capsid
181 proteins. Melbournevirus and other members of the Marseilleviruses are taken up by
182 phagocytosis and then lose their icosahedral appearance to become spherical after the
183 disappearance of the vacuole membrane. In contrast to Mimiviruses their RNA polymerase is
184 not loaded in the virion. Instead, the cell nucleus starts changing appearance and becomes
185 leaky through a still unknown mechanism triggered by the viral infection⁵⁸. The nuclear proteins
186 are recruited to the early viral factory, including the host RNA polymerase that will perform early
187 transcription. Shortly after, the nucleus integrity is restored and the virally encoded RNA
188 polymerase performs intermediate and late transcription⁵⁹ and icosahedral particles assemble
189 inside the viral factory (Figure 2A). Mature particles can gather in large vesicles⁶⁰ and cell lysis
190 leads to the release of both individual virions and filled vacuoles.

191
192 Larger amoeba-infecting viruses are Pandoraviruses, with amphora-shaped virions that can be
193 up to 1 µm in length and 500 nm in diameter (Figure 2C). There is at least one lipidic membrane
194 lining a thick tegument made of three layers including one made of cellulose⁶¹. The particles are
195 taken up through phagocytosis followed by the opening of the virion ostiole-like structure to
196 allow the fusion of the internal membrane with the phagosome and, consequently, delivery of
197 the genome and necessary proteins into the cytoplasm. In contrast to Mimiviruses the
198 Pandoraviruses build their viral factory in the host cell nucleus (Figure 2A). Although they
199 encode RNA polymerase, there is none present in the capsids and thus they rely on the host
200 cell for early transcription of viral genes. New virions start to assemble from the apex and the
201 neo-synthesized virions are released either by cell lysis or by exocytosis through membrane
202 fusion with the plasma membrane when they are in vacuoles^{62,63}.

203
204 Some of the largest virions found in the Nucleocytoviricota are those of the Pitho- and
205 Cedratviruses (Figure 2C), which have very large amphora shaped capsids that can be up to 2
206 µm long and 600 nm wide. These capsids are closed by corks, one cork for Pithovirus^{64,65}
207 (Figure 2C) and two for Cedratvirus⁶⁶, that are made by proteins organized in a honeycomb
208 array. Despite a virion morphology that closely resembles that of Pandoraviruses, the external
209 tegument is different, appears to be made of parallel strips and does not present cellulose. The
210 particle appears to be coated with short sparse fibrils^{33,64}. The infectious cycle proceeds as for
211 other amoeba-infecting viruses, by phagocytosis followed by capsid opening and membrane
212 fusion with the phagosome⁵. For the Pithoviruses and Cedratviruses, the RNA polymerase
213 loaded in the virion starts early transcription in the cytoplasm and the host nucleus remains
214 intact during the entire infectious cycle. The virions are assembled from the cork and will lead to
215 rectangular uncoated virions, the tegument being built by patches from reservoirs in the
216 cytoplasm, and the morphology changes to amphora-shaped. The neo-synthesized virions can
217 exit by exocytosis and upon cell lysis^{64,66}.

218

219 A different replication strategy has been described for Medusaviruses³¹; after uptake into the
220 host cytoplasm, its DNA is replicated in the host nucleus while virions assemble in the
221 cytoplasm. Medusaviruses have icosahedral virions that are 260 nm in diameter, covered by
222 spherical-headed spikes extending from each capsomer, and a lipidic membrane surrounds the
223 capsid interior. A low-resolution structure was determined by Cryo-EM, which returned a T-
224 number of 277⁶⁷. The mechanism of entry and egress of the Medusavirus virion from its host is
225 yet to be determined.

226

227 As these examples illustrate, there is no shared blueprint for giant virus structure and infection
228 mechanism; these characteristics vary between giant virus lineages and are likely shaped by
229 the host organisms. The host range of the currently experimentally characterized giant viruses is
230 limited to a few amoeba and algae lineages representing only a tiny fraction of the eukaryotic
231 diversity. Thus, we expect that many more unusual virions and infection strategies will be
232 revealed in the near future when new viruses will be captured together with their native hosts.

233

234 **Cultivation-independent genomic approaches to studying giant viruses**

235 Many important discoveries on giant virus biology and diversity have been made through giant
236 virus isolation and cultivation. However, such approaches are constrained by the need to satisfy
237 optimal growth requirements in a laboratory setting and are often restricted to lytic viruses.
238 Cultivation-independent methods have been proven to be an indispensable tool to discover the
239 genetic make-up of giant viruses from environmental samples.

240

241 In the earlier days of metagenomics, single-marker gene based surveys (Box 1) revealed that
242 several marine NCLDV affiliated with *Phycodnaviridae* and *Mimiviridae* were present in a wide
243 range of marine metagenomes collected during the Tara Oceans and the Sargasso Sea
244 expeditions^{68,69} and that these viruses were more abundant in the photic layer than
245 eukaryotes⁶⁹. Interestingly, in a follow-up study, data from these surveys gave rise to the
246 hypothesis that giant viruses are more diverse in the oceans than any cellular organism⁷⁰.
247 Subsequently, a large-scale analysis of the NCLDV major capsid protein (MCP), in which more
248 than 50k of these proteins were found across Earth's biomes, highlighted the global dispersal of
249 giant viruses, including in terrestrial ecosystems⁷¹.

250

251 Other approaches that facilitated the discovery of novel NCLDV are single cell genomics and
252 mini-metagenomics (Box 1). First, sorting single cells from marine samples enabled detection of
253 NCLDV particles affiliated with viruses that had previously been found in the algae
254 *Ostreococcus* sp. and *Phaeocystis globosa*⁷². This led to the sequencing of several so-called
255 gvSAGs, of which the largest was a 813 kb genome affiliated with Mimiviridae that encoded a
256 metacaspase, which potentially facilitates autocatalytic cell death of the host cell⁷³. Single-cell
257 methods were also used to identify and genome sequence 5 giant viruses associated with
258 marine choanoflagellates^{74,75}; comparative genomics together with all other NCLDV genomes
259 revealed that viruses that infect hosts with similar trophic modes, including host habitat and
260 lifestyles, express distinct genetic features^{75,76}. Further, mini-metagenomics (Box 1) on a single
261 forest soil sample led to the discovery of 15 diverse giant virus MAGs, including several
262 members of the Klosneuvirinae, revealing that giant viruses are commonly found in soil⁷⁷.

263
264 The most successful approach for obtaining NCLDV genomes from environmental sequence
265 data is genome-resolved metagenomics (Box 1). Since the early 2000s this approach has
266 become common practice for recovering genomes of bacteria and archaea from complex
267 environmental samples⁷⁸, yet it took nearly another decade before the first giant virus
268 metagenome assembled genomes (GVMAGs) appeared in public databases (Figure 1). Yau
269 and colleagues reconstructed the first GVMAGs as a byproduct of their work on virophages in
270 metagenomes from the Organic Lake in Antarctica⁷⁹. Several years later four additional
271 potentially algae-associated GVMAGs were retrieved from environmental sequence data from
272 Yellowstone lake and found to be related to phycodna- and mimiviruses, which shared some
273 genes with virophages that co-occurred in the same sample⁸⁰. Cultivation-independent
274 approaches for discovery of giant virus genome-centric sequence information gained traction
275 when members of a Mimiviridae-affiliated subfamily, the proposed Klosneuvirinae, were
276 recovered from metagenomic data⁸¹. The fact that these were found in freshwater and sewage
277 metagenomes from samples originating from four different continents suggested this novel
278 group of giant viruses is cosmopolitan⁸¹. More than 20 giant virus MAGs from deep sea
279 metagenomes were subsequently discovered, including 15 affiliated with Pithoviruses, indicating
280 a surprisingly high prevalence of pithovirus-like viruses in the ocean⁸² and additional likely
281 algae-associated freshwater giant viruses from Dishui Lake^{83,84}. The unique strength of
282 cultivation-independent approaches to viral genomics became even more evident when over
283 2000 GVMAGs were extracted from metagenome datasets generated from thousands of
284 samples collected from diverse biomes⁷¹ plus an additional 500 GVMAGs collected shortly after
285 from marine systems⁸⁵. Viruses discovered through cultivation-independent approaches
286 enabled a comprehensive update of the taxonomic framework of the Nucleocytoviricota⁸⁶
287 (Figure 3). The addition of the GVMAGs to the Nucleocytoviricota species tree led the
288 phylogenetic diversity to increase by more than 10-fold and expanded their pangenome to more
289 than 900,000 proteins⁷¹. This translated to an extensively expanded repertoire of functional
290 genes, providing not only many novel insights into how giant viruses may interact with their
291 hosts and the environment, but also generating novel hypotheses about their evolutionary
292 roles^{71,85,87,88}.

293 294 **Exploring the host ranges of giant viruses**

295 Genome-resolved metagenomics facilitated the discovery of thousands of viral genomes, of
296 which many represented lineages divergent from viruses recovered by isolation or co-
297 cultivation^{71,85} (Figure 3). However, NCLDV discovered through metagenomics typically lack
298 direct connections to host organisms⁸⁹. Approaches that have been used to overcome this
299 limitation are the detection of viruses and potential eukaryotic hosts co-occurring in the same
300 sample. Further, as horizontal swapping of genetic material between viruses and their hosts is a
301 common phenomenon, the analysis of viral genes that may have been acquired through recent
302 horizontal gene transfer (HGT) can potentially identify host organisms. In the early days of giant
303 virus metagenomics, read-mapping based co-occurrence analysis (Box 1) revealed that the
304 presence of giant virus sequences in some marine samples was positively correlated with those
305 of Oomycetes⁶⁹, which is a eukaryote that has not yet been found associated with NCLDV. More
306 recently, co-occurrences of sequences of NCLDV and microalgae, including haptophytes,

307 chlorophytes, dinophytes, pelagophytes and raphidophytes, were detected across a wide range
308 of marine samples collected during the Tara Oceans expedition⁹⁰. The strongest sequence co-
309 occurrences were found between a virus affiliated with the *Mimiviridae* and Chrysophytes as its
310 potential host⁹⁰. Subsequent detection of putative HGT events between giant virus
311 metagenome-assembled genomes and Chrysophyte genomes and transcriptomes provided
312 further support for this host-virus connection⁹⁰. A systematic analysis of HGT candidates
313 present in more than 2000 NCLDV genomes, of which most were MAGs from diverse global
314 sampling sites, revealed thousands of genes likely derived from host genomes through recent
315 HGT⁷¹. Based on these results it was possible to propose connections between NCLDV and
316 members of all major eukaryotic phyla⁷¹. Though most of these predicted hosts have not yet
317 been found to be infected by giant viruses, more than 20 distinct connections could be
318 confirmed through existing isolates⁷¹ (Figure 4A).

319
320 While sequence-based computational host predictions provide a means to expand the range of
321 putative NCLDV hosts, the approaches have some potential challenges and biases. For
322 example, co-occurrence analysis is dependent on sufficient host genome coverage for detection
323 in metagenome data and HGT analysis likewise requires the availability of the host genomic
324 sequences. Further, it is difficult to detect ancient HGT from previous hosts, and the integration
325 of NCLDV genes into host genomes whose sequences were deposited in public databases can
326 potentially yield misleading results. The quality of databases used for the analyses is another
327 important factor and the presence of NCLDV genomes misannotated as bacteria, archaea, or
328 eukarya is a major weakness when employing automated tools for HGT detection^{71,91}. Despite
329 some of these limitations, expanding the putative host range of metagenome-derived NCLDV
330 provides a basis for targeted sampling of putative hosts, for studying virus-host co-evolution,
331 and for identifying viral-encoded functions for targeted modulation of host metabolism.
332 Sequence-based inferences of viruses and their hosts may then be extrapolated to assess the
333 impact of such interactions on the environment to enable predictive ecosystem modeling.

334
335 Not only is horizontal transfer of single genes a common phenomenon but some giant viruses
336 can even integrate their entire genomes into the host's chromosome (Figure 4). This so-called
337 endogenization is a term initially coined for a mechanism discovered in retroviruses^{92,93}. Arrays
338 of NCLDV genes have occasionally been found in genomes of eukaryotes, in particular in algae,
339 plants⁹⁴⁻⁹⁶ and amoebae⁹⁷⁻⁹⁹. A recent survey of published eukaryotic genomes and
340 transcriptomes revealed the presence of giant virus genes in 66 different eukaryotes including
341 several *Acanthamoeba* species, flagellates, ciliates, stramenopiles, oomycetes, fungi,
342 arthropods and diverse unicellular and multicellular algae¹⁰⁰ (Figure 4). Yet, for many of these
343 eukaryotes giant virus infections have not been observed. The integration of NCLDV genes
344 often appears to be highly host specific, with viral genes detected in one eukaryotic species
345 being unrelated to viral genes found in its sister species¹⁰⁰. Among the integrated genes were
346 NCLDV hallmark genes that were, in some instances, dispersed over the host chromosome
347 whereas in other cases genes were co-localized in islands composed of more than 100
348 genes¹⁰⁰. The integration of complete viral genomes has been described for Phaeoviruses; for
349 example, *Ectocarpus siliculosus* virus integrated into its brown algal host more than 20 years
350 ago⁹⁶ likely through use of integrases¹⁰¹. Phaeoviruses are lysogenic viruses leading to

351 persistent infections^{102,103}, which is in stark contrast to many other known NCLDV lineages that
352 were successfully isolated based on the fact that they lyse their amoeba host⁵. In a recent study,
353 the analysis of existing algal genomes and transcriptome data revealed other examples of
354 whole giant virus genomes integrated into eukaryotic host genomes¹⁰⁴. Some regions encoded
355 more than 1,500 viral genes, up to 10% of the genes of the green algal host¹⁰⁴. Several of the
356 detected viral genes were annotated as enzymes with roles in carbohydrate metabolism,
357 chromatin remodeling, signal transduction, energy production and translation¹⁰⁴.

358
359 It remains unknown if integrated giant viruses are dormant with no or minimal benefit to the
360 host, or if the host cell benefits from some of the viral genes that may provide or finetune
361 metabolic capabilities. Another unanswered question is whether there are mechanisms
362 encoded in the integrated viral genome that may reinstate an active infection after transcribing
363 and translating some of the integrated viral genes. This would then be followed by the release of
364 the giant virus genetic material during host replication and effective dispersal to new hosts. If
365 there is no reactivation of viral infection, it is conceivable that some giant virus genes would
366 decay over time leading to pseudogenization. Genomics, metagenomics and
367 metatranscriptomics enabled the discovery of NCLDV integration and revealed potential model
368 systems, such as the green algae *Tetraabaena socialis* and *Chlamydomonas eustigma*¹⁰⁴ for
369 experimental work. Future investigation of integration of giant virus genes is expected to provide
370 some answers for how endogenization has shaped and continues to shape the evolution and
371 ecology of eukaryotic organisms.

372 373 **Reprogramming of the host and its impact on host populations**

374 Reprogramming of the host can occur when the virus transforms its host into a so-called
375 virocell¹⁰⁵ in which particular viral auxiliary metabolic genes augment host metabolism to further
376 promote viral replication^{106,107}. Analogous to bacteriophages, giant viruses seem to contribute
377 genes to their hosts to augment and/or modulate metabolic capabilities of the host cell (Figure
378 5). One of the first described examples was a host-derived nitrogen transporter in *Ostreococcus*
379 *taurus* virus that had been found expressed during the infection of its green algal host¹⁰⁸.
380 Experimental characterization provided evidence that this transporter may enhance the uptake
381 of nitrogen by the host cell¹⁰⁸. Other studies revealed the presence of fermentation genes in the
382 Tetraselmis virus genome with possible implications for host metabolism in nutrient-limited
383 marine systems²⁶. A survey of giant virus isolates and metagenome-assembled genomes
384 revealed the widespread presence of genes for cytochrome P450 monooxygenases, potentially
385 enabling or modulating complex metabolic processes such as synthesis of sterols and other
386 fatty acids⁸⁸. Distant homologs of actins and also myosin have been found in NCLDV
387 genomes^{87,109} indicating a viral impact on cell structure, motility and intracellular transport
388 processes, but further functional validation is needed. Furthermore, a giant virus affiliated with
389 Mesomimivirinae that infects heterotrophic Choanoflagellates was found to encode type-1
390 rhodopsins together with the pathway for synthesis of the required pigment, β -carotene⁷⁴.
391 Experimental characterization of the NCLDV rhodopsin showed that the putative rhodopsin
392 likely functions as a proton pump - generating energy from light⁷⁴. A phylogenetically distinct
393 NCLDV rhodopsin was found in the Organic Lake virus metagenome-assembled genome and
394 experimental characterization of this protein revealed that it may function as a light-gated

395 pentameric ion channel, potentially impacting ion homeostasis and phototaxis of the host cell¹¹⁰.
396 Further, through global metagenomics it was predicted that genes encoding various substrate
397 transport processes, energy generation, i.e. photosynthesis, carbon fixation and also glycolysis
398 are commonly found in giant virus metagenome-assembled genomes affiliated with diverse
399 lineages of the Nucleocytoviricota^{71,85} (Figure 5). More detailed phylogenetic analysis revealed
400 that some auxiliary metabolic genes encoding transporters for iron, phosphate, magnesium and
401 ammonium originated in eukaryotic hosts and were likely recently acquired by giant viruses
402 through horizontal gene transfer^{71,74,85}. However, other genes encoding several rhodopsins,
403 succinate hydrogenase, aconitase, and glyceraldehyde 3-phosphate dehydrogenase showed a
404 pattern that suggested a viral origin or a common evolutionary origin in one of the ancestral
405 hosts^{71,74,85}. Taken together, the widespread presence of metabolic genes in diverse NCLDV
406 lineages implies that augmenting host metabolic capacities is likely a strategy more commonly
407 used by members of the NCLDV than initially assumed. However, the current lack of
408 experimental evidence on the functions and activities of most of these genes and pathways, and
409 their effects on the host cell, demand further experimental investigation.

410
411 Metabolic reprogramming has direct consequences on host population structure and dynamics.
412 One striking example is the cosmopolitan marine coccolithophore *Emiliana huxleyi* which forms
413 massive blooms that play important roles in global carbon and sulfur cycles¹¹¹. *E. huxleyi*
414 populations are subject to persistent but ultimately lytic infections by the *E. huxleyi*
415 virus²³(Figure 4B). Once lysis is induced this leads to the termination of the algal bloom coupled
416 to the deposition of massive amounts of calcite and nutrients into the ocean which increases the
417 marine pool of dissolved organic matter¹¹²⁻¹¹⁴. Importantly, viral infections do not only lead to
418 host lysis but also enhance viral replication by rewiring host physiology, in particular the
419 turnover of sugars and synthesis of fatty acids and lipids¹¹⁵⁻¹¹⁷. Comparably little is known about
420 the impact of other large and giant viruses on host populations but it is conceivable that similar
421 principles are omnipresent and important in shaping our planet's biomes and biogeochemical
422 cycles.

423 424 **Towards independence from the host**

425 Among the most intriguing features found in giant virus genomes are signatures of cellular life,
426 such as genes involved in protein biosynthesis. This phenomenon was first described upon
427 sequencing the Mimivirus genome⁸. Subsequent analyses revealed the phylogenetic placement
428 of virus-encoded cellular genes between bacteria and eukaryotes suggesting an ancient
429 origin¹⁰. Other cellular hallmark genes with similarly deep branching patterns were found in
430 other giant virus genomes and led to the hypotheses that giant viruses may either represent a
431 fourth domain of life¹² or are remnants of a highly degraded eukaryotic cell derived by reductive
432 evolution¹¹. However, the use of more complex phylogenetic models revealed that many of
433 these genes had most likely been acquired from different eukaryotic hosts^{118,119}. Some of these
434 genes might represent ancient transfers from undiscovered eukaryotic hosts. This finding
435 provided evidence for the hypothesis that giant viruses may have evolved from smaller
436 viruses¹¹⁹. Yet, other studies have reported alternative topologies for some housekeeping and
437 other metabolic genes of cellular organisms, including rhodopsins^{71,74,85} and cytochrome P450⁸⁸.
438 It cannot be ruled out that these genes were transferred from ancestral giant viruses to past

439 eukaryotic hosts, or even a proto-eukaryote, highlighting a potentially integral role of giant
440 viruses in the evolution of the eukaryotic cell^{120,121}. Further, it is possible that some genes that
441 may function as part of the eukaryotic core metabolism were introduced upon integration of
442 giant virus genetic material into the genome of an ancient eukaryotic cell - further shaping
443 eukaryotic evolution^{120,122}. The presence of genes for aminoacyl tRNA synthetases (aaRS) and
444 eukaryotic translation factors has been recorded multiple times in newly recovered giant virus
445 genomes. Indeed, a nearly complete set of 20 aaRS has been reported in Klosneuvirus from
446 metagenomic data⁸¹. Shortly after two Tupanviruses were isolated with genomes that contained
447 a full set of aaRS⁷ and subsequently the first isolates affiliated with the Klosneuvirinae were
448 described, which also contained a complete set of aaRS¹²³. Especially in Klosneuviruses, the
449 presence of aaRS with lineage-specific evolutionary histories provided additional support that
450 these genes derived from different eukaryotic hosts⁸¹. The presence of genes for extended sets
451 of aaRS is currently limited to members of the Mimiviridae and information on the role of giant
452 virus aaRS in host interaction is limited, however, some have been experimentally studied and
453 were indeed functional¹²⁴. Nevertheless, there is some experimental evidence for potential roles
454 of these genes in making giant viruses less dependent on host machinery, e.g., during
455 shutdown of host translation in response to viral infection or other adverse conditions¹²⁵. On the
456 other hand, a suspected role in enhancing viral translation by providing additional copies of
457 aaRS to support host translation has not yet been confirmed. Another path towards host
458 independence is implied by encoding translation system components, in addition to recently
459 discovered giant virus genes predicted to be involved in energy generation^{26,85}. A recent study
460 reported on an active membrane potential in *Pandoravirus massiliensis* virions together with the
461 expression of several remote homologs of TCA genes¹²⁶. Despite encoding expanded genetic
462 complements for functions that were recently thought to be exclusively present in cellular
463 organisms, there is currently no evidence that giant viruses may perform protein translation
464 without host-derived ribosomes or host-independent energy generation.

465 466 **Conclusions**

467
468 Nearly 20 years of giant virus isolation has yielded viral isolates representing highly diverse
469 lineages. Complementary detailed research on the biology of these viruses has revealed many
470 important details of virion structures and infection strategies. It has become obvious that there
471 are stark differences in virion size and structure and, while there are some similarities in how
472 these viruses enter and exit the host cell, most giant viruses employ contrasting strategies for
473 replicating within and exploiting their host cells. Sequencing of viral isolates has led to the
474 largest and smallest known genomes of viruses of the Nucleocytoviricota.

475 Cultivation-independent approaches have accelerated the discovery of genome sequences of
476 new giant viruses and other large viruses in the NCLDV, providing novel insights into their
477 phylogenetic diversity and functional potential. These approaches also revealed that these
478 viruses can be found anywhere on Earth, are affiliated with diverse eukaryotes and are likely
479 manipulating host physiology through metabolic reprogramming - ultimately altering structure
480 and function of host communities in the environment. At the same time, estimates based on
481 NCLDV hallmark genes in metagenomic datasets indicated that only a small fraction of giant
482 virus genomes have been discovered so far⁷¹, and that the diversity of giant viruses may be far

483 greater than that of bacteria, at least in the oceans⁷⁰. A controlled metagenomic binning
484 experiment where giant viruses were spiked into an environmental sample showed that genome
485 fragments of many giant viruses that are present in a given sample likely remain below the
486 detection limit, highlighting the need for ultra-deep metagenome sequencing¹²⁷ or targeted
487 isolation efforts¹²⁸. Furthermore, there is a strong bias towards detecting giant viruses that are
488 similar to those already known, as tools used to identify viruses from metagenomes rely heavily
489 on features observed in sequenced NCLDV genomes, such as large sets of conserved
490 genes^{71,82,85,129,130}. However, giant virus genomes exhibit extensive plasticity, such that viruses
491 within the same clade quickly diverge and share very few genes²⁸. A recent stunning example of
492 NCLDV diversity is Yaravirus, which was isolated with its native amoeba host¹³¹, yet no close
493 relative of Yaravirus was detectable in public metagenomic datasets. Its placement within
494 NCLDV was challenging due to greater than 90 percent of its genes lacking similarity to those in
495 public databases and the paucity of most viral hallmark genes¹³¹. Such excessive gene novelty
496 of viruses in the NCLDV, observed through both cultivation and cultivation-independent
497 methods, further underlines that many giant viruses are likely to be hiding in plain sight, awaiting
498 discovery.

499

500 **Author contributions**

501 FS analyzed and visualized the data. Figures were prepared by FS and CA (Figure 2). FS, CA
502 and TW wrote the manuscript.

503

504 **Acknowledgements**

505 This work was conducted by the US Department of Energy Joint Genome Institute, a DOE
506 Office of Science User Facility, under Contract No. DE-AC02–05CH11231 and made use of
507 resources of the National Energy Research Scientific Computing Center, also supported by the
508 DOE Office of Science under Contract No. DE-AC02–05CH11231. We thank Dr. Xiao (River)
509 Chuan from Department of Chemistry and Biochemistry, University of Texas El Paso for
510 providing cryo electron micrographs for AaV, Mimivirus and CroV. Further we acknowledge R.
511 Watanabe and K. Murata, ExCELLS, NINS, Japan who provided cryo electron micrographs for
512 Medusavirus, and Dr. R.N. Burton-Smith and Prof. K. Murata, ExCELLS, NINS, Japan for cryo
513 electron micrographs of Melbournevirus.

514

515 **Competing interests**

516 The authors declare no competing interests.

517

518

519 **References**

- 520 1. Fischer, M. G. Giant viruses come of age. *Curr. Opin. Microbiol.* **31**, 50–57 (2016).
- 521 2. Iyer, L. M., Balaji, S., Koonin, E. V. & Aravind, L. Evolutionary genomics of nucleo-
522 cytoplasmic large DNA viruses. *Virus Res.* **117**, 156–184 (2006).
- 523 3. Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World.

- 524 *Microbiol. Mol. Biol. Rev.* **84**, (2020).
- 525 4. Sun, T.-W. *et al.* Host Range and Coding Potential of Eukaryotic Giant Viruses. *Viruses* **12**,
526 (2020).
- 527 5. Abergel, C., Legendre, M. & Claverie, J.-M. The rapidly expanding universe of giant
528 viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol. Rev.* **39**, 779–
529 796 (2015).
- 530 6. Iyer, L. M., Aravind, L. & Koonin, E. V. Common origin of four diverse families of large
531 eukaryotic DNA viruses. *J. Virol.* **75**, 11720–11734 (2001).
- 532 7. Abrahão, J. *et al.* Tailed giant Tupanvirus possesses the most complete translational
533 apparatus of the known virosphere. *Nat. Commun.* **9**, 749 (2018).
- 534 8. Colson, P., La Scola, B., Levasseur, A., Caetano-Anollés, G. & Raoult, D. Mimivirus:
535 leading the way in the discovery of giant viruses of amoebae. *Nat. Rev. Microbiol.* **15**, 243–
536 254 (2017).
- 537 9. La Scola, B. *et al.* A giant virus in amoebae. *Science* **299**, 2033 (2003).
- 538 10. Raoult, D. *et al.* The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344–
539 1350 (2004).
- 540 11. Legendre, M., Arslan, D., Abergel, C. & Claverie, J.-M. Genomics of Megavirus and the
541 elusive fourth domain of Life. *Commun. Integr. Biol.* **5**, 102–106 (2012).
- 542 12. Colson, P., de Lamballerie, X., Fournous, G. & Raoult, D. Reclassification of giant viruses
543 composing a fourth domain of life in the new order Megavirales. *Intervirology* **55**, 321–332
544 (2012).
- 545 13. La Scola, B. *et al.* The virophage as a unique parasite of the giant mimivirus. *Nature* **455**,
546 100–104 (2008).
- 547 14. Jeudy, S. *et al.* Exploration of the propagation of transpovirons within Mimiviridae reveals a
548 unique example of commensalism in the viral world. *ISME J.* **14**, 727–739 (2020).
- 549 15. Desnues, C. *et al.* Provirophages and transpovirons as the diverse mobilome of giant

- 550 viruses. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 18078–18083 (2012).
- 551 16. Fenner, F. Adventures with poxviruses of vertebrates. *FEMS Microbiol. Rev.* **24**, 123–133
552 (2000).
- 553 17. Goebel, S. J. *et al.* The complete DNA sequence of vaccinia virus. *Virology* **179**, 247–66,
554 517–63 (1990).
- 555 18. Oliveira, G. P., Rodrigues, R. A. L., Lima, M. T., Drumond, B. P. & Abrahão, J. S. Poxvirus
556 Host Range Genes and Virus–Host Spectrum: A Critical Review. *Viruses* **9**, 331 (2017).
- 557 19. İnce, İ. A., Özcan, O., Ilter-Akulke, A. Z., Scully, E. D. & Özgen, A. Invertebrate Iridoviruses:
558 A Glance over the Last Decade. *Viruses* **10**, (2018).
- 559 20. Piégu, B., Asgari, S., Bideshi, D., Federici, B. A. & Bigot, Y. Evolutionary relationships of
560 iridoviruses and divergence of ascoviruses from invertebrate iridoviruses in the superfamily
561 Megavirales. *Mol. Phylogenet. Evol.* **84**, 44–52 (2015).
- 562 21. Dixon, L. K., Chapman, D. A. G., Netherton, C. L. & Upton, C. African swine fever virus
563 replication and genomics. *Virus Res.* **173**, 3–14 (2013).
- 564 22. Subramaniam, K. *et al.* A New Family of DNA Viruses Causing Disease in Crustaceans
565 from Diverse Aquatic Biomes. *MBio* **11**, (2020).
- 566 23. Wilson, W. H., Van Etten, J. L. & Allen, M. J. The Phycodnaviridae: the story of how tiny
567 giants rule the world. *Curr. Top. Microbiol. Immunol.* **328**, 1–42 (2009).
- 568 24. Gallot-Lavallée, L., Blanc, G. & Claverie, J.-M. Comparative Genomics of
569 Chrysochromulina Ericina Virus and Other Microalga-Infecting Large DNA Viruses
570 Highlights Their Intricate Evolutionary Relationship with the Established Mimiviridae Family.
571 *J. Virol.* **91**, (2017).
- 572 25. Claverie, J.-M. & Abergel, C. Mimiviridae: An Expanding Family of Highly Diverse Large
573 dsDNA Viruses Infecting a Wide Phylogenetic Range of Aquatic Eukaryotes. *Viruses* **10**,
574 (2018).
- 575 26. Schvarcz, C. R. & Steward, G. F. A giant virus infecting green algae encodes key

- 576 fermentation genes. *Virology* **518**, 423–433 (2018).
- 577 27. Blanc-Mathieu, R. *et al.* A persistent giant algal virus, with a unique morphology, encodes
578 an unprecedented number of genes involved in energy metabolism. *J. Virol.* (2021)
579 doi:10.1128/JVI.02446-20.
- 580 28. Koonin, E. V. & Yutin, N. Evolution of the Large Nucleocytoplasmic DNA Viruses of
581 Eukaryotes and Convergent Origins of Viral Gigantism. *Adv. Virus Res.* **103**, 167–202
582 (2019).
- 583 29. Pagnier, I. *et al.* A decade of improvements in Mimiviridae and Marseilleviridae isolation
584 from amoeba. *Intervirology* **56**, 354–363 (2013).
- 585 30. Takahashi, H., Fukaya, S., Song, C., Murata, K. & Takemura, M. Morphological and
586 Taxonomic Properties of the Newly Isolated Cottonvirus japonicus , a New Lineage of the
587 Subfamily Megavirinae. *Journal of Virology* vol. 95 (2021).
- 588 31. Yoshikawa, G. *et al.* Medusavirus, a Novel Large DNA Virus Discovered from Hot Spring
589 Water. *J. Virol.* **93**, (2019).
- 590 32. Reteno, D. G. *et al.* Faustovirus, an asfarvirus-related new lineage of giant viruses infecting
591 amoebae. *J. Virol.* **89**, 6585–6594 (2015).
- 592 33. Andreani, J. *et al.* Orpheovirus IHUMI-LCC2: A New Virus among the Giant Viruses. *Front.*
593 *Microbiol.* **8**, 2643 (2017).
- 594 34. Andreani, J. *et al.* Pacmanvirus, a New Giant Icosahedral Virus at the Crossroads between
595 Asfarviridae and Faustoviruses. *J. Virol.* **91**, (2017).
- 596 35. Bajrai, L. H. *et al.* Kaumoebavirus, a new virus that clusters with faustoviruses and
597 Asfarviridae. *Viruses* **8**: 278. (2016).
- 598 36. Francis, R., Ominami, Y., Bou Khalil, J. Y. & La Scola, B. High-throughput isolation of giant
599 viruses using high-content screening. *Commun Biol* **2**, 216 (2019).
- 600 37. Dornas, F. P. *et al.* Isolation of new Brazilian giant viruses from environmental samples
601 using a panel of protozoa. *Front. Microbiol.* **6**, 1086 (2015).

- 602 38. Burki, F., Roger, A. J., Brown, M. W. & Simpson, A. G. B. The New Tree of Eukaryotes.
603 *Trends Ecol. Evol.* **35**, 43–55 (2020).
- 604 39. Renesto, P. *et al.* Mimivirus giant particles incorporate a large fraction of anonymous and
605 unique gene products. *J. Virol.* **80**, 11678–11685 (2006).
- 606 40. Fischer, M. G., Allen, M. J., Wilson, W. H. & Suttle, C. A. Giant virus with a remarkable
607 complement of genes infects marine zooplankton. *Proc. Natl. Acad. Sci. U. S. A.* **107**,
608 19508–19513 (2010).
- 609 41. Deeg, C. M., Chow, C.-E. T. & Suttle, C. A. The kinetoplastid-infecting Bodo saltans virus
610 (BsV), a window into the most abundant giant viruses in the sea. *Elife* **7**, (2018).
- 611 42. Notaro, A. *et al.* Expanding the Occurrence of Polysaccharides to the Viral World: The
612 Case of Mimivirus. *Angew. Chem. Int. Ed Engl.* **60**, 19897–19904 (2021).
- 613 43. Klose, T. *et al.* The three-dimensional structure of Mimivirus. *Intervirology* **53**, 268–273
614 (2010).
- 615 44. Kuznetsov, Y. G. *et al.* Atomic force microscopy investigation of the giant mimivirus.
616 *Virology* **404**, 127–137 (2010).
- 617 45. Zauberman, N. *et al.* Distinct DNA exit and packaging portals in the virus Acanthamoeba
618 polyphaga mimivirus. *PLoS Biol.* **6**, e114 (2008).
- 619 46. Fischer, M. G., Kelly, I., Foster, L. J. & Suttle, C. A. The virion of Cafeteria roenbergensis
620 virus (CroV) contains a complex suite of proteins for transcription and DNA repair. *Virology*
621 **466-467**, 82–94 (2014).
- 622 47. Arslan, D., Legendre, M., Seltzer, V., Abergel, C. & Claverie, J.-M. Distant Mimivirus
623 relative with a larger genome highlights the fundamental features of Megaviridae. *Proc.*
624 *Natl. Acad. Sci. U. S. A.* **108**, 17486–17491 (2011).
- 625 48. Mutsafi, Y., Zauberman, N., Sabanay, I. & Minsky, A. Vaccinia-like cytoplasmic replication
626 of the giant Mimivirus. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5978–5982 (2010).
- 627 49. Kuznetsov, Y. G., Klose, T., Rossmann, M. & McPherson, A. Morphogenesis of Mimivirus

- 628 and Its Viral Factories: an Atomic Force Microscopy Study of Infected Cells. *Journal of*
629 *Virology* vol. 88 3055–3055 (2014).
- 630 50. Fischer, M. G. & Hackl, T. Host genome integration and giant virus-induced reactivation of
631 the virophage mavirus. *Nature* **540**, 288–291 (2016).
- 632 51. Bellec, L., Grimsley, N., Moreau, H. & Desdevises, Y. Phylogenetic analysis of new
633 Prasinoviruses (Phycodnaviridae) that infect the green unicellular algae *Ostreococcus*,
634 *Bathycoccus* and *Micromonas*. *Environ. Microbiol. Rep.* **1**, 114–123 (2009).
- 635 52. Van Etten, J. L., Agarkova, I. V. & Dunigan, D. D. Chloroviruses. *Viruses* **12**, (2019).
- 636 53. Cherrier, M. V. *et al.* An icosahedral algal virus has a complex unique vertex decorated by a
637 spike. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 11085–11089 (2009).
- 638 54. De Castro, C. *et al.* Structure of the chlorovirus PBCV-1 major capsid glycoprotein
639 determined by combining crystallographic and carbohydrate molecular modeling
640 approaches. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E44–E52 (2018).
- 641 55. Burton-Smith, R. N. *et al.* The 4.4 Å structure of the giant Melbournevirus virion belonging
642 to the Marseilleviridae family. *bioRxiv* (2021) doi:10.1101/2021.07.14.452405.
- 643 56. Chihara, A. *et al.* A novel capsid protein network allows the characteristic inner membrane
644 structure of Marseilleviridae giant viruses. *bioRxiv* 2021.02.03.428533 (2021)
645 doi:10.1101/2021.02.03.428533.
- 646 57. Fang, Q. *et al.* Near-atomic structure of a giant virus. *Nat. Commun.* **10**, 388 (2019).
- 647 58. Rodrigues, R. A. L. *et al.* Analysis of a Marseillevirus Transcriptome Reveals Temporal
648 Gene Expression Profile and Host Transcriptional Shift. *Frontiers in Microbiology* vol. 11
649 (2020).
- 650 59. Oliveira, G. P. *et al.* The investigation of promoter sequences of marseilleviruses highlights
651 a remarkable abundance of the AAATATTT motif in intergenic regions. *J. Virol.* **91**, (2017).
- 652 60. Okamoto, K. *et al.* Cryo-EM structure of a Marseilleviridae virus particle reveals a large
653 internal microassembly. *Virology* **516**, 239–245 (2018).

- 654 61. Legendre, M. *et al.* In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus
655 infecting Acanthamoeba. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5327–35 (2015).
- 656 62. Scheid, P., Balczun, C. & Schaub, G. A. Some secrets are revealed: parasitic keratitis
657 amoebae as vectors of the scarcely described pandoraviruses to humans. *Parasitol. Res.*
658 **113**, 3759–3764 (2014).
- 659 63. Akashi, M. & Takemura, M. Co-isolation and characterization of two pandoraviruses and a
660 Mimivirus from a riverbank in Japan. *Viruses* **11**, 1123 (2019).
- 661 64. Levasseur, A. *et al.* Comparison of a Modern and Fossil Pithovirus Reveals Its Genetic
662 Conservation and Evolution. *Genome Biol. Evol.* **8**, 2333–2339 (2016).
- 663 65. Andreani, J. *et al.* Cedratvirus, a Double-Cork Structured Giant Virus, is a Distant Relative
664 of Pithoviruses. *Viruses* **8**, (2016).
- 665 66. Bertelli, C. *et al.* Cedratvirus lausannensis - digging into Pithoviridae diversity. *Environ.*
666 *Microbiol.* **19**, 4022–4034 (2017).
- 667 67. Yoshida, K. *et al.* Draft Genome Sequence of Medusavirus Stheno, Isolated from the
668 Tatakai River of Uji, Japan. *Microbiol Resour Announc* **10**, (2021).
- 669 68. Ghedin, E. & Claverie, J.-M. Mimivirus relatives in the Sargasso sea. *Virology* **2**, 62 (2005).
- 670 69. Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans
671 microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013).
- 672 70. Mihara, T. *et al.* Taxon Richness of ‘Megaviridae’ Exceeds those of Bacteria and Archaea
673 in the Ocean. *Microbes Environ.* **33**, 162–171 (2018).
- 674 71. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics.
675 *Nature* **578**, 432–436 (2020).
- 676 72. Martínez Martínez, J., Swan, B. K. & Wilson, W. H. Marine viruses, a genetic reservoir
677 revealed by targeted viromics. *ISME J.* **8**, 1079–1088 (2014).
- 678 73. Wilson, W. H. *et al.* Genomic exploration of individual giant ocean viruses. *ISME J.* **11**,
679 1736–1745 (2017).

- 680 74. Needham, D. M. *et al.* A distinct lineage of giant viruses brings a rhodopsin photosystem to
681 unicellular marine predators. *Proc. Natl. Acad. Sci. U. S. A.* 201907517 (2019).
- 682 75. Needham, D. M. *et al.* Targeted metagenomic recovery of four divergent viruses reveals
683 shared and distinctive characteristics of giant viruses of marine eukaryotes. *Philos. Trans.*
684 *R. Soc. Lond. B Biol. Sci.* **374**, 20190086 (2019).
- 685 76. Sun, T.-W. & Ku, C. Unraveling gene content variation across eukaryotic giant viruses
686 based on network analyses and host associations. *Virus Evol* **7**, veab081 (2021).
- 687 77. Schulz, F. *et al.* Hidden diversity of soil giant viruses. *Nat. Commun.* **9**, 4881 (2018).
- 688 78. Allen, E. E. & Banfield, J. F. Community genomics in microbial ecology and evolution. *Nat.*
689 *Rev. Microbiol.* **3**, 489–498 (2005).
- 690 79. Yau, S. *et al.* Virophage control of antarctic algal host–virus dynamics. *Proc. Natl. Acad.*
691 *Sci. U. S. A.* **108**, 6163–6168 (2011).
- 692 80. Zhang, W. *et al.* Four novel algal virus genomes discovered from Yellowstone Lake
693 metagenomes. *Sci. Rep.* **5**, 15131 (2015).
- 694 81. Schulz, F. *et al.* Giant viruses with an expanded complement of translation system
695 components. *Science* **356**, 82–85 (2017).
- 696 82. Bäckström, D. *et al.* Virus Genomes from Deep Sea Sediments Expand the Ocean
697 Megavirome and Support Independent Origins of Viral Gigantism. *MBio* **10**, (2019).
- 698 83. Chen, H. *et al.* The genome of a prasinoviruses-related freshwater virus reveals unusual
699 diversity of phycodnaviruses. *BMC Genomics* **19**, 49 (2018).
- 700 84. Xu, S. *et al.* Novel Cell-Virus-Virophage Tripartite Infection Systems Discovered in the
701 Freshwater Lake Dishui Lake in Shanghai, China. *J. Virol.* **94**, (2020).
- 702 85. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic
703 genome evolution and complex virocell metabolism of globally-distributed giant viruses.
704 *Nat. Commun.* **11**, 1710 (2020).
- 705 86. Aylward, F. O., Moniruzzaman, M., Ha, A. D. & Koonin, E. V. A phylogenomic framework

706 for charting the diversity and evolution of giant viruses. *PLoS Biol.* **19**, e3001430 (2021).

707 87. Da Cunha, V., Gaia, M., Ogata, H., Jaillon, O. & Delmont, T. O. Giant viruses encode novel
708 types of actins possibly related to the origin of eukaryotic actin: the viractins. *bioRxiv*
709 (2020).

710 88. Lamb, D. C. *et al.* On the occurrence of cytochrome P450 in viruses. *Proc. Natl. Acad. Sci.*
711 *U. S. A.* **116**, 12343–12352 (2019).

712 89. Mihara, T. *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, 66 (2016).

713 90. Endo, H. *et al.* Biogeography of marine giant viruses reveals their interplay with eukaryotes
714 and ecological functions. *Nat Ecol Evol* **4**, 1639–1649 (2020).

715 91. Andreani, J., Verneau, J., Raoult, D., Levasseur, A. & La Scola, B. Deciphering viral
716 presences: two novel partial giant viruses detected in marine metagenome and in a mine
717 drainage metagenome. *Virolog. J.* **15**, 66 (2018).

718 92. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on
719 host biology. *Nat. Rev. Genet.* **13**, 283–296 (2012).

720 93. Chiba, S. *et al.* Widespread endogenization of genome sequences of non-retroviral RNA
721 viruses into plant genomes. *PLoS Pathog.* **7**, e1002146 (2011).

722 94. Maumus, F., Epert, A., Nogu e, F. & Blanc, G. Plant genomes enclose footprints of past
723 infections by giant virus relatives. *Nature Communications* vol. 5 (2014).

724 95. Wang, L. *et al.* Endogenous viral elements in algal genomes. *Acta Oceanologica Sinica* vol.
725 33 102–107 (2014).

726 96. Delaroque, N., Maier, I., Knippers, R. & M ller, D. G. Persistent virus integration into the
727 genome of its algal host, *Ectocarpus siliculosus* (Phaeophyceae). *J. Gen. Virol.* **80 (Pt 6)**,
728 1367–1370 (1999).

729 97. Maumus, F. & Blanc, G. Study of Gene Trafficking between *Acanthamoeba* and Giant
730 Viruses Suggests an Undiscovered Family of Amoeba-Infecting Viruses. *Genome Biol.*
731 *Evol.* **8**, 3351–3363 (2016).

- 732 98. Clarke, M. *et al.* Genome of *Acanthamoeba castellanii* highlights extensive lateral gene
733 transfer and early evolution of tyrosine kinase signaling. *Genome Biol.* **14**, R11 (2013).
- 734 99. Chelkha, N. *et al.* *Vermamoeba vermiformis* CDC-19 draft genome sequence reveals
735 considerable gene trafficking including with candidate phyla radiation and giant viruses.
736 *Scientific Reports* vol. 10 (2020).
- 737 100. Gallot-Lavallée, L. & Blanc, G. A Glimpse of Nucleo-Cytoplasmic Large DNA Virus
738 Biodiversity through the Eukaryotic Genomics Window. *Viruses* **9**, (2017).
- 739 101. Delaroque, N. & Boland, W. The genome of the brown alga *Ectocarpus siliculosus* contains
740 a series of viral DNA pieces, suggesting an ancient association with large dsDNA viruses.
741 *BMC Evol. Biol.* **8**, 110 (2008).
- 742 102. Stevens, K. *et al.* A novel evolutionary strategy revealed in the phaeoviruses. *PLoS One* **9**,
743 e86040 (2014).
- 744 103. Cock, J. M. *et al.* The *Ectocarpus* genome and the independent evolution of multicellularity
745 in brown algae. *Nature* **465**, 617–621 (2010).
- 746 104. Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A. & Aylward, F. O.
747 Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* **588**,
748 141–145 (2020).
- 749 105. Forterre, P. The virocell concept and environmental microbiology. *ISME J.* **7**, 233–236
750 (2013).
- 751 106. Howard-Varona, C. *et al.* Phage-specific metabolic reprogramming of virocells. *The ISME*
752 *Journal* vol. 14 881–895 (2020).
- 753 107. Hurwitz, B. L., Hallam, S. J. & Sullivan, M. B. Metabolic reprogramming by viruses in the
754 sunlit and dark ocean. *Genome Biol.* **14**, R123 (2013).
- 755 108. Monier, A. *et al.* Host-derived viral transporter protein for nitrogen uptake in infected marine
756 phytoplankton. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E7489–E7498 (2017).
- 757 109. Kijima, S. *et al.* Discovery of Viral Myosin Genes With Complex Evolutionary History Within

758 Plankton. *Front. Microbiol.* **12**, 683294 (2021).

759 110. Bratanov, D. *et al.* Unique structure and function of viral rhodopsins. *Nat. Commun.* **10**,
760 4939 (2019).

761 111. Paasche, E. A review of the coccolithophorid *Emiliana huxleyi* (Prymnesiophyceae), with
762 particular reference to growth, coccolith formation, and calcification-photosynthesis
763 interactions. *Phycologia* **40**, 503–529 (2001).

764 112. Kuhlisch, C. *et al.* Viral infection of algal blooms leaves a unique metabolic footprint on the
765 dissolved organic matter in the ocean. *Sci Adv* **7**, (2021).

766 113. Breitbart, M. Marine viruses: truth or dare. *Ann. Rev. Mar. Sci.* **4**, 425–448 (2012).

767 114. Wilhelm, S. W. & Suttle, C. A. Viruses and Nutrient Cycles in the Sea: Viruses play critical
768 roles in the structure and function of aquatic food webs. *Bioscience* **49**, 781–788 (1999).

769 115. Malitsky, S. *et al.* Viral infection of the marine alga *Emiliana huxleyi* triggers lipidome
770 remodeling and induces the production of highly saturated triacylglycerol. *New Phytol.* **210**,
771 88–96 (2016).

772 116. Schleyer, G. *et al.* In plaque-mass spectrometry imaging of a bloom-forming alga during
773 viral infection reveals a metabolic shift towards odd-chain fatty acid lipids. *Nature*
774 *Microbiology* **4**, 527–538 (2019).

775 117. Rosenwasser, S. *et al.* Rewiring Host Lipid Metabolism by Large Viruses Determines the
776 Fate of *Emiliana huxleyi*, a Bloom-Forming Alga in the Ocean. *Plant Cell* **26**, 2689–2707
777 (2014).

778 118. Williams, T. A., Embley, T. M. & Heinz, E. Informational gene phylogenies do not support a
779 fourth domain of life for nucleocytoplasmic large DNA viruses. *PLoS One* **6**, e21080 (2011).

780 119. Yutin, N., Wolf, Y. I. & Koonin, E. V. Origin of giant viruses from smaller DNA viruses not
781 from a fourth domain of cellular life. *Virology* **466–467**, 38–52 (2014).

782 120. Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P. & Gaia, M. Diversification of giant and
783 large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc. Natl.*

784 *Acad. Sci. U. S. A.* (2019) doi:10.1073/pnas.1912006116.

785 121. Bell, P. J. Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus?
786 *Journal of molecular evolution* vol. 53 251–256 (2001).

787 122. Cheng, S., Wong, G. K.-S. & Melkonian, M. Giant DNA viruses make big strides in
788 eukaryote evolution. *Cell host & microbe* vol. 29 152–154 (2021).

789 123. Hussein Bajrai, L. *et al.* Isolation of Yasminevirus, the first member of Klosneuvirinae
790 isolated in coculture with *Vermamoeba vermiformis*, demonstrates an extended arsenal of
791 translational apparatus components. *J. Virol.* (2019) doi:10.1128/JVI.01534-19.

792 124. Abergel, C., Rudinger-Thirion, J., Giegé, R. & Claverie, J.-M. Virus-encoded aminoacyl-
793 tRNA synthetases: structural and functional characterization of mimivirus TyrRS and
794 MetRS. *J. Virol.* **81**, 12406–12417 (2007).

795 125. Silva, L. C. F. *et al.* Modulation of the expression of mimivirus-encoded translation-related
796 genes in response to nutrient availability during *Acanthamoeba castellanii* infection. *Front.*
797 *Microbiol.* **6**, 539 (2015).

798 126. Aherfi, S. *et al.* Incomplete tricarboxylic acid cycle and proton gradient in Pandoravirus
799 massiliensis: is it still a virus? *The ISME Journal* (2021) doi:10.1038/s41396-021-01117-3.

800 127. Schulz, F. *et al.* Advantages and Limits of Metagenomic Assembly and Binning of a Giant
801 Virus. *mSystems* **5**, (2020).

802 128. Christo-Foroux, E. *et al.* Characterization of Mollivirus kamchatka, the First Modern
803 Representative of the Proposed Molliviridae Family of Giant Viruses. *J. Virol.* **94**, (2020).

804 129. Aylward, F. O. & Moniruzzaman, M. ViralRecall: A Flexible Command-Line Tool for the
805 Detection of Giant Virus Signatures in Omic Data. doi:10.1101/2020.12.15.422924.

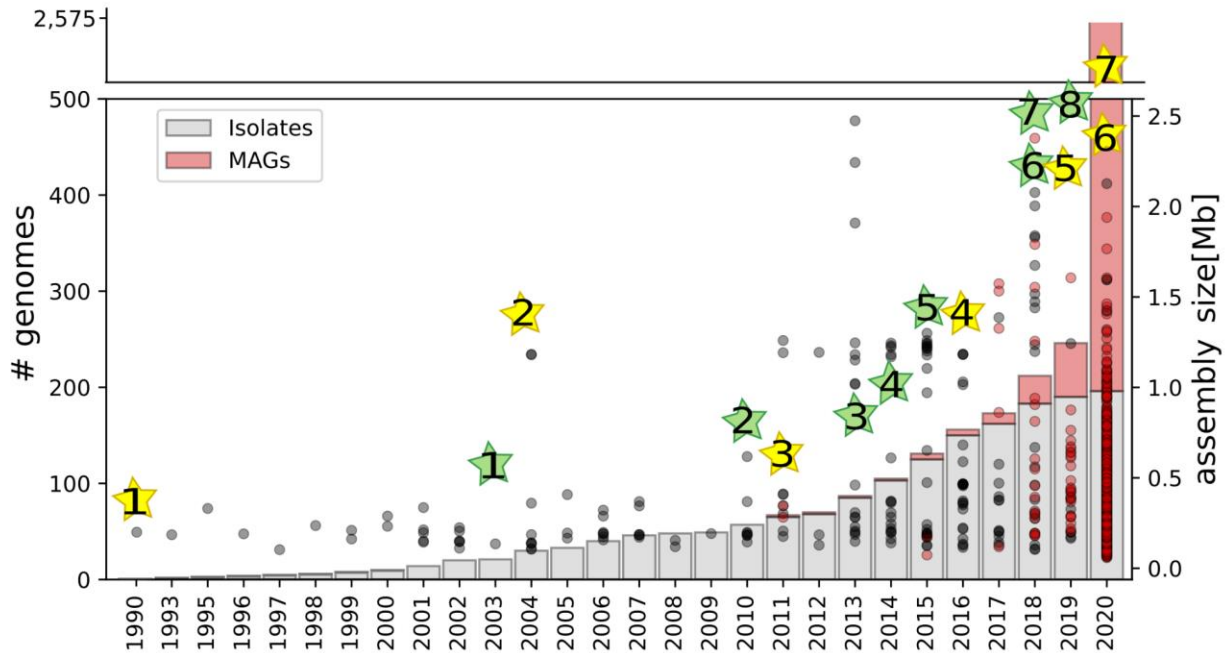
806 130. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-
807 assembled viral genomes. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-00774-7.

808 131. Boratto, P. V. M. *et al.* Yaravirus: A novel 80-nm virus infecting *Acanthamoeba castellanii*.
809 *Proc. Natl. Acad. Sci. U. S. A.* **117**, 16579–16586 (2020).

- 810 132.Philippe, N. *et al.* Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching
811 That of Parasitic Eukaryotes. *Science* (2013) doi:10.1126/science.1239181.
- 812 133.Legendre, M. *et al.* Thirty-thousand-year-old distant relative of giant icosahedral DNA
813 viruses with a pandoravirus morphology. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 4274–4279
814 (2014).
- 815 134.Gann, E. R. *et al.* Structural and Proteomic Studies of the *Aureococcus anophagefferens*
816 Virus Demonstrate a Global Distribution of Virus-Encoded Carbohydrate Processing. *Front.*
817 *Microbiol.* **11**, 2047 (2020).
- 818 135.Xiao, C. *et al.* Structural studies of the giant mimivirus. *PLoS Biol.* **7**, e92 (2009).
- 819 136.Xiao, C. *et al.* Cryo-EM reconstruction of the Cafeteria roenbergensis virus capsid suggests
820 novel assembly pathway for giant viruses. *Sci. Rep.* **7**, 5484 (2017).
- 821 137.Kerepesi, C. & Grolmusz, V. The ‘Giant Virus Finder’ discovers an abundance of giant
822 viruses in the Antarctic dry valleys. *Arch. Virol.* **162**, 1671–1676 (2017).
- 823 138.Chatterjee, A. & Kondabagil, K. Giant viral genomic signatures in the previously reported
824 gut metagenomes of pre-school children in rural India. *Arch. Virol.* **164**, 2819–2822 (2019).
- 825 139.Pires de Souza, G. A., Rolland, C., Nafeh, B., La Scola, B. & Colson, P. Giant virus-related
826 sequences in the 5300-year-old Ötzi mummy metagenome. *Virus Genes* **57**, 222–227
827 (2021).
- 828 140.Verneau, J., Levasseur, A., Raoult, D., La Scola, B. & Colson, P. MG-Digger: An
829 Automated Pipeline to Search for Giant Virus-Related Sequences in Metagenomes. *Front.*
830 *Microbiol.* **7**, 428 (2016).
- 831 141.Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
832 assessing the quality of microbial genomes recovered from isolates, single cells, and
833 metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
- 834 142.Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a
835 metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**,

- 836 725–731 (2017).
- 837 143.Woyke, T., Doud, D. F. R. & Schulz, F. The trajectory of microbial single-cell sequencing.
838 *Nat. Methods* **14**, 1045–1054 (2017).
- 839 144.Martínez, J. M., Martinez-Hernandez, F. & Martinez-Garcia, M. Single-virus genomics and
840 beyond. *Nat. Rev. Microbiol.* **18**, 705–716 (2020).
- 841 145.Khalil, J. Y. B. *et al.* High-Throughput Isolation of Giant Viruses in Liquid Medium Using
842 Automated Flow Cytometry and Fluorescence Staining. *Front. Microbiol.* **7**, 26 (2016).
- 843 146.Yu, F. B. *et al.* Microfluidic-based mini-metagenomics enables discovery of novel microbial
844 lineages from complex environmental samples. *Elife* **6**, (2017).

845



846

847

Figure 1. Timeline of genomic and metagenomic discoveries in the Nucleocytoviricota.

848

Stacked bars indicate the number of virus isolate genomes (grey, left y-axis) and giant virus

849

metagenome-assembled genomes (GVMAGs; red, left y-axis) of members of the

850

Nucleocytoviricota that have been published and/or became available at the NCBI Genbank

851

database for each year on the x-axis. Filled circles indicate genome size (right y-axis) of virus

852

isolates (grey) and GVMAGs (red). Important genomic and metagenomic events are highlighted

853

by yellow stars: 1. First genome of a member of the NCLDV became available¹⁷; 2. Sequencing

854

of the first giant virus with a genome size above 1 Mb, *Acanthamoeba polyphaga* Mimivirus¹⁰; 3.

855

First-time recovery of GVMAGs (from Organic Lake, Antarctica)⁷⁹; 4. Viral subfamily

856

“Klosneuvirinae” proposed based on GVMAGs recovered from environmental sequence data⁸¹;

857

5. Single-cell genomics enabled discovery of Choanovirus from marine choanoflagellates⁷⁴; 6.

858

First large-scale global metagenomics study leading to the recovery of over 2000 GVMAGs

859

facilitating an 11-fold increase in phylogenetic diversity and a 10-fold expansion in functional

860

diversity⁷¹; 7. Detection of whole giant virus genomes integrated in host chromosomes¹⁰⁴.

861

Important cultivation based events are highlighted by green stars: 1. Isolation of Mimivirus in

862

amoeba co-cultivation⁹ 2. Isolation of a giant virus together with its *Cafeteria roenbergensis*

863

host⁴⁰ 3. Isolation of amphora-shaped Pandoravirus with 2.5Mb genome¹³² 4. Recovery of

864

Pithovirus, from a 30,000 year old ice core, through co-cultivation with an amoeba¹³³ 5. Isolation

865

of Faustovirus in co-cultivation with *Vermamoeba*³² 6. Isolation of *Bodo saltans* virus the first

866

isolated member of the Klosneuvirinae with its native Kinetoplastid host⁴¹ 7. Isolation of

867

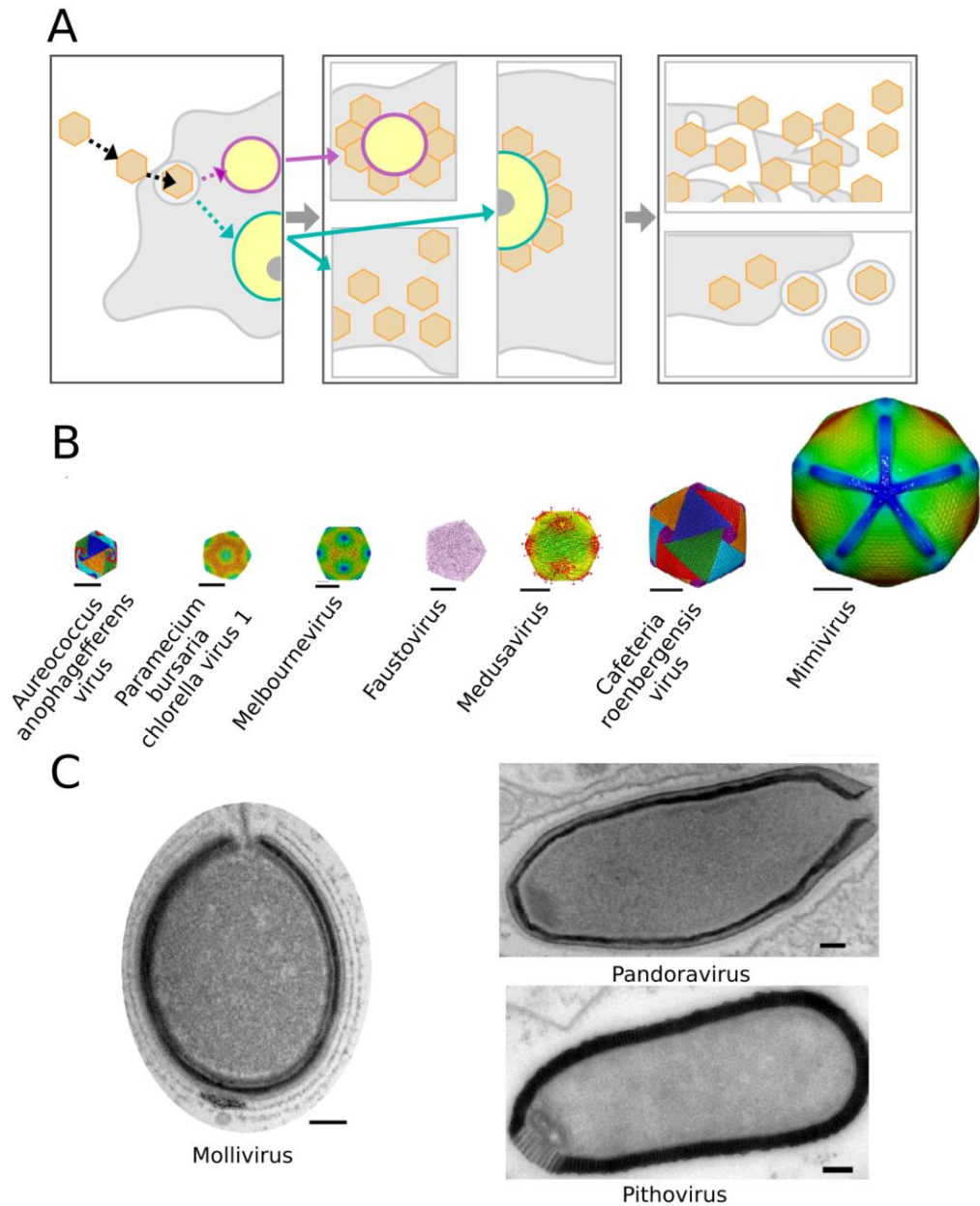
Tupanvirus in amoeba - currently the largest giant virus based on its capsid diameter and

868

length⁷ 8. Isolation of Medusavirus in amoeba represents a divergent new lineage in the

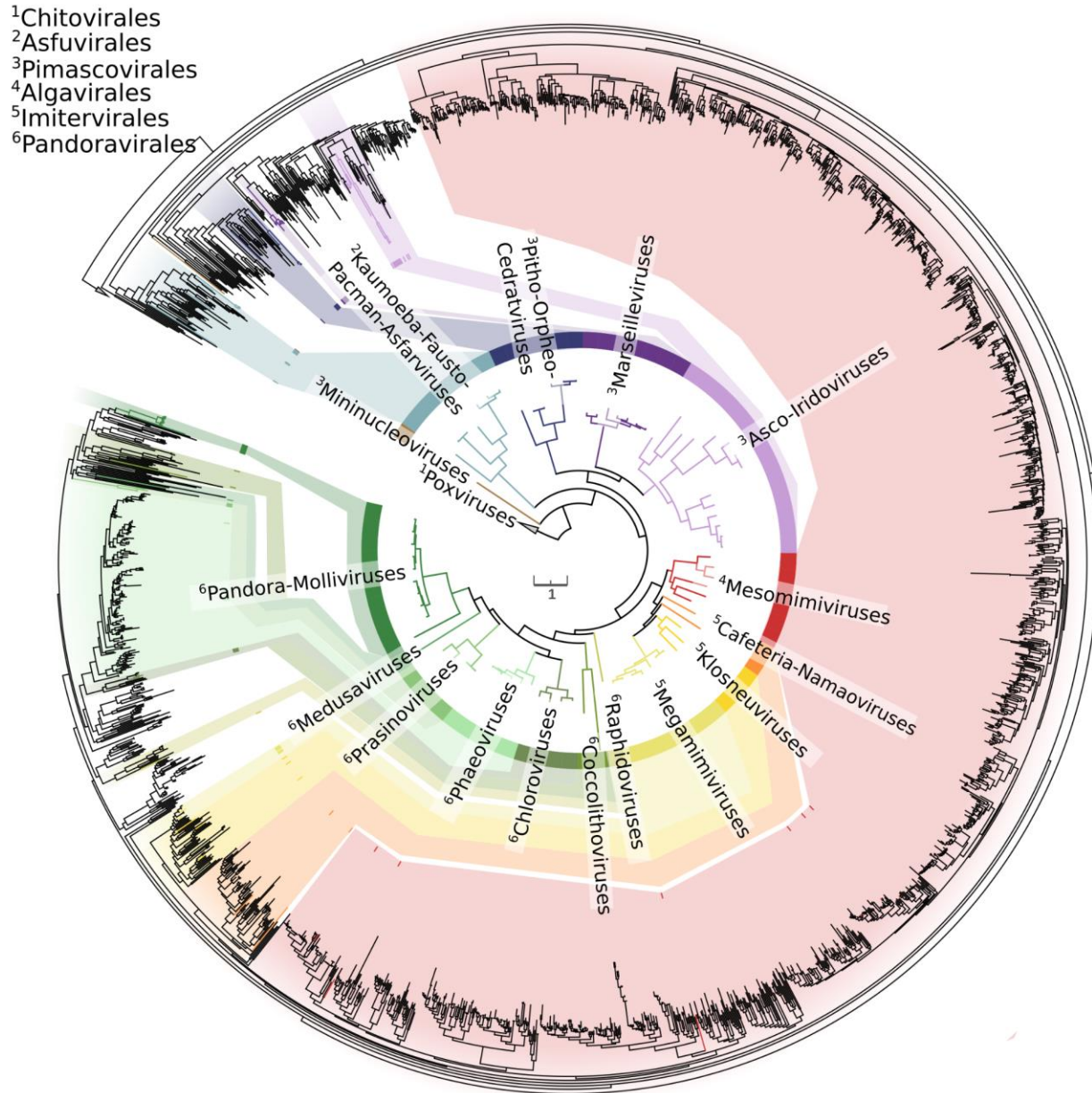
869

Nucleocytoviricota³¹.



870
871
872
873
874
875
876
877
878
879
880
881

Figure 2. Giant virus infection mechanisms and virion structures. **A** Attachment to the host cell envelope followed by uptake by endocytosis and activation of giant virus transcription in the nucleus (turquoise arrow) or the cytoplasm / viral factory (purple arrow). This is followed by giant replication and assembly of new virions in the periphery of the cytoplasmic viral factory, in the host nucleus or scattered in the virocell. Virions are released after host cell lysis or exocytosis of membrane-bound virions. **B** Structures of isolated giant viruses resolved by cryo-electron microscopy^{55,134–136}. Note the blue-colored stargate structure on Mimivirus. **C** Transmission electron micrographs of ultrathin sections of non icosahedral viruses embedded in resin. The scale bars in B and C are 100 nm.

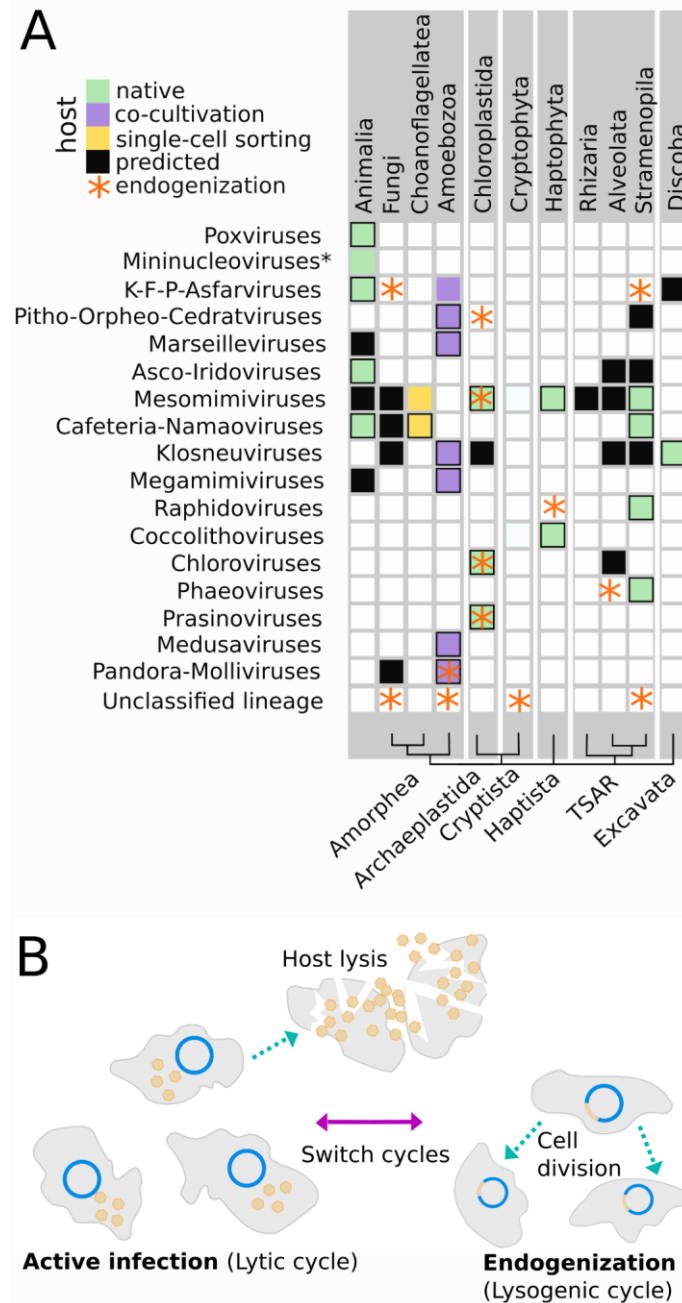


882

883

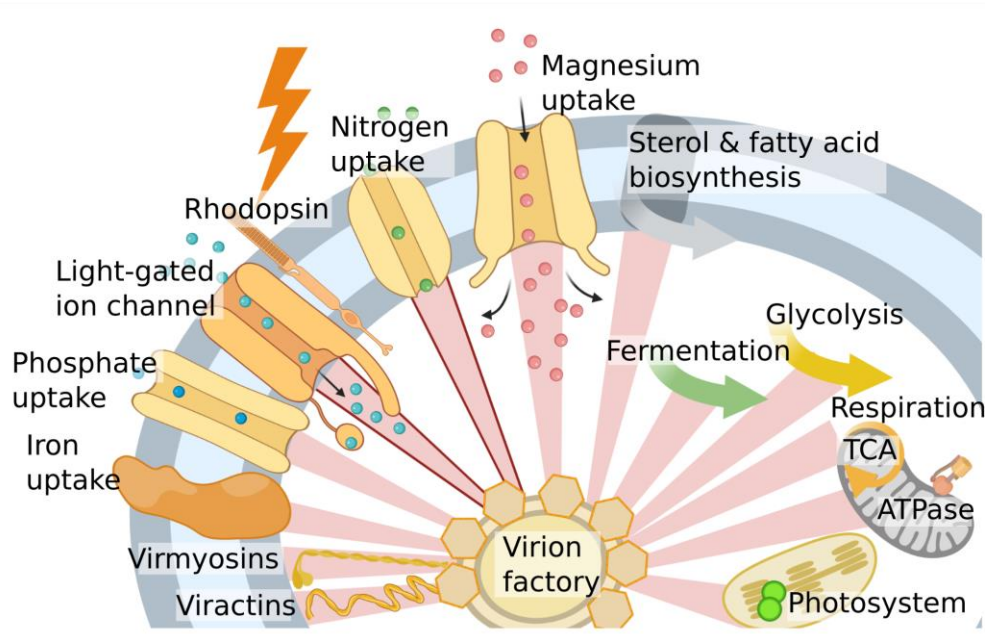
884 **Figure 3. Expansion of the Nucleocytoviricota phylogenetic diversity through**
 885 **metagenomics.** Shown are two species trees of the Nucleocytoviricota: the inner tree depicts
 886 the diversity of NCLDV based on the genomes of viral isolates and the outer inverted tree
 887 highlights the expansion of species diversity through genomes derived from cultivation-
 888 independent sequencing approaches. Branches are colored based on affiliation to taxonomic
 889 groups consisting of isolates and extrapolated to the expanded diversity of the outer tree.
 890 Superscripts map to the proposed updated taxonomic framework for Nucleocytoviricota
 891 lineages⁸⁶.

892



893
894
895
896
897
898
899
900
901
902
903

Figure 4. Experimentally verified and computationally predicted host ranges of the Nucleocytoviricota. **A** Shown are host lineages identified through isolation with the native host, co-cultivation, single-cell sorting, and *in silico* HGT-based predictions. Black outline of colored boxes indicates that experimentally verified interaction has been also predicted computationally. K-F-P-Asfarviruses: Kaumoeba-Fausto-Pacman-Asfarviruses; *genomes of this group were not included in the HGT analysis. Topology of the eukaryotic species tree adapted from³⁸. **B** Lysis of host cells and entire host populations are the typical outcomes of a giant virus infection. There is growing evidence that long term persistent infections may be prevalent in algae populations, including the integration of giant virus genomes, so-called endogenization, into the host chromosome.



904
 905 **Figure 5: Predicted metabolic reprogramming of a giant virus-derived virocell and**
 906 **consequences of giant virus infection for host populations.** Shown is a hypothetical virocell
 907 with a combination of metabolic roles that different giant viruses are predicted to have during
 908 host infection, based on the presence of auxiliary metabolic genes in giant virus genomes. Dark
 909 red lines denote metabolic roles that are supported by some functional data obtained through
 910 experiments.
 911

912 **Box 1:**

913 **Toolkits for giant virus discovery: cultivation independent genomic approaches**

914 ***Read mapping-based approaches***

915 Mapping metagenomic reads to giant virus reference genomes has been successfully applied to
916 detect giant viruses and estimate their abundances in the environment^{41,69,137–139} and several
917 tools have been published^{137,140}. Read mapping-based approaches are advantageous because
918 they are sensitive enough to detect lowly abundant giant viruses¹³⁹; however, they typically do
919 not lead to the recovery of viral genomes and thus cannot provide information on genome
920 features and coding potential. Moreover, mapping approaches are highly dependent on the
921 quality of the reference genome database, and if low mapping stringency is used, false positive
922 hits may occur. Detection of giant viruses may also be hindered if taxonomic classification of
923 metagenome-assembled genomes was performed using automated tools; this has resulted in
924 several NCLDV MAGs that have been misclassified as being of bacterial, archaeal or eukaryotic
925 origin^{71,91}. In addition, genes that have recently been integrated into NCLDV genomes after
926 being horizontally acquired from bacteria or eukaryotes may produce sequence reads that map
927 to cellular genomes, resulting in false positive hits.

928 ***Marker gene surveys***

929 Detection and phylogenetic analysis of signature genes in complex environmental datasets is a
930 commonly used approach to assess viral diversity in metagenome data. For Nucleocytoviricota
931 genes that encode major capsid protein, DNA Polymerase B, or viral packaging ATPase have
932 been used as marker genes. The approach is less error-prone than read mapping as it can be
933 coupled with phylogenetic analysis to confirm the monophyly of the respective gene homologs
934 found in known NCLDV genomes. This approach has been successfully applied in several
935 studies^{69–71,90} and, while being less sensitive than read mapping, it can detect viruses that were
936 not abundant enough in a metagenome to be successfully assembled and binned¹²⁷.

937 ***Genome-resolved metagenomics***

938 The reconstruction of metagenome-assembled genomes through metagenomic binning is an
939 established approach to recover microbial genomes. Due to their virion sizes giant viruses are
940 often present in environmental samples that have been selectively filtered to target microbes. In
941 contrast to smaller viruses such as most bacteriophages, the large genomes of most members
942 of the Nucleocytoviricota typically require metagenomic binning to increase genome
943 completeness¹²⁷. However, in most microbe-centric metagenome projects giant virus genome
944 bins were frequently ignored, as tools that estimate genome quality¹⁴¹ predict viral genomes to
945 be of low completeness based on their lack of cellular marker genes¹²⁷ which then leads to their
946 exclusion from downstream analyses¹⁴². Several recent studies employed custom workflows to
947 identify giant virus MAGs and to estimate completeness and contamination by, for example,
948 identifying copy numbers of conserved NCLDV genes^{74,75,77,81,85,130} or inferring deviations from
949 lineage specific copy numbers of low-copy orthologs⁷¹. It is important to note that genome sizes
950 of GVMAGs typically do not represent the size of the complete viral genome.

951 ***Single cell genomic approaches***

952 Flow-sorting and sequencing of single viruses can be used to detect viruses in environmental
953 samples^{143,144}, yet only a few such studies have discovered novel giant viruses^{72–75,77}. Due to
954 large virion sizes and a bright signal using DNA stains^{73,145} giant viruses are a promising target
955 for sorting. A drawback of this approach is that the subsequent whole genome amplification, if

956 performed on a single virus, may lead to low genome recovery¹⁴³. An alternative approach to
957 direct sorting of giant viruses from an environmental sample is targeted sorting of host cells^{74,75}.
958 Viruses actively replicating inside a host cell can produce hundreds to thousands of virions with
959 clonal copies of viral genomes, which would greatly improve whole genome amplification¹⁴³.
960 Further, if successful, this approach enables identification of the virus and its native host.
961 Similarly, mini metagenomics uses fluorescence activated single cell sorting or microfluidics to
962 collect tens to hundreds of cell-sized particles^{77,146}. The presence of many identical viral
963 particles, either through repeated sorting of clonal single viruses, an infected host cell, or the
964 sorting of vacuoles filled with giant viruses, would then increase genome recovery.