1

2 **Large DNA virus promoted the endosymbiotic evolution to make a photosynthetic**

3 **eukaryote**

4

5 **Authors and Affiliations**

6 Mitsuhiro Matsuo¹, Atsushi Katahata¹, Makoto Tachikawa¹, Yohei Minakuchi², Hideki

7 Noguchi^{3,4}, Atsushi Toyoda^{2,4}, Asao Fujiyama⁴, Yutaka Suzuki⁵, Takayuki Hata¹,

8 Soichirou Satoh¹, Takuro Nakayama^{6,7}, Ryoma Kamikawa⁸, Mami Nomura^{8,9}, Yuji

9 Inagaki⁶, Ken-ichiro Ishida⁹, Junichi Obokata¹

10

11 ¹Graduate School of Life and Environmental Science, Kyoto Prefectural University,

12 Kyoto, 606-8522, Japan.

13 ²Department of Genetics and Evolutionary Biology, National Institute of Genetics,

14 Mishima, Shizuoka, 411-8540, Japan.

15 ³Center for Genome Informatics, Joint Support-Center for Data Science Research,

16 Research Organization of Information and Systems, Mishima, Shizuoka 411-8540,

17 Japan.

18 ⁴Advanced Genomics Center, National Institute of Genetics, Mishima, Shizuoka
19 411-8540, Japan.

20 ⁵Graduate school of Frontier Science, University of Tokyo, Kashiwa, Chiba, 272-8562,
21 Japan.

22 ⁶Center for Computational Sciences, University of Tsukuba, Tsukuba, Ibaraki,
23 305-8577, Japan.

24 ⁷Graduate School of Life Sciences, Tohoku University, Sendai 980-8578,
25 Japan

26 ⁸Graduate Schools Human and Environmental Studies, Kyoto University, Kyoto,
27 606-8501, Japan.

28 ⁹Graduate School of Life and Environmental Sciences, University of Tsukuba,
29 Tsukuba, Ibaraki, 305-8572, Japan.

30

31

32

33

34 Abstract

35 Chloroplasts in photosynthetic eukaryotes originated from a cyanobacterial
 36 endosymbiosis far more than 1 billion years ago¹⁻³. Due to this ancientness, it remains
 37 unclear how this evolutionary process proceeded. To unveil this mystery, we analysed
 38 the whole genome sequence of a photosynthetic rhizarian amoeba⁴, *Paulinella*
 39 *micropora*^{5,6}, which has a chloroplast-like organelle that originated from another
 40 cyanobacterial endosymbiosis⁷⁻¹⁰ about 0.1 billion years ago¹¹. Here we show that the
 41 predacious amoeba that engulfed cyanobacteria evolved into a photosynthetic organism
 42 very quickly in the evolutionary time scale, probably aided by the drastic genome
 43 reorganization activated by large DNA virus. In the endosymbiotic evolution of
 44 eukaryotic cells, gene transfer from the endosymbiont genome to the host nucleus is
 45 essential for the evolving host cell to control the endosymbiont-derived organelle¹². In *P.*
 46 *micropora*, we found that the gene transfer from the free-living and endosymbiotic
 47 bacteria to the amoeba nucleus was rapidly activated but both simultaneously ceased
 48 within the initiation period of the endosymbiotic evolution, suggesting that the genome
 49 reorganization drastically proceeded and completed. During this period, large DNA

virus appeared to have infected the amoeba, followed by the rapid amplification and diversification of virus-related genes. These findings led us to re-examine the conventional endosymbiotic evolutionary scenario that exclusively deals with the host and the symbiont, and to extend it by incorporating a third critical player, large DNA virus, which activates the drastic gene transfer and genome reorganization between them. This *Paulinella* version of the evolutionary hypothesis deserves further testing of its generality in evolutionary systems and could shed light on the unknown roles of large DNA viruses¹³ in the evolution of terrestrial life.

Main manuscript

Our laboratory culture of *P. micropora* MYN1^{5,6} (Fig. 1a) is not axenic and contains bacteria. From this culture, we prepared the chromatins of *P. micropora* by micromesh-aided cell isolation and chromatin immunoprecipitation using a canonical histone antibody. Shotgun sequencing of this chromatin DNA gave us a high-quality draft genome assembly of 967 Mb (Fig. 1, Extended Data Fig. 1, Supplementary Table 1). K-mer analysis estimated the genome size of *P. micropora* MYN1 to be 1.35 Gb

Fig. 1. An overview of the *P. micropora* MYN1 draft genome. a, A SEM image of *P. micropora* MYN1. **b,** The statistics of the draft genome. **c,** The genome composition of *P. micropora* MYN1 analysed by RepeatMasker²⁷. **d,** Simple repeats are extraordinarily rich in *P. micropora* MYN1 compared with other organisms.

(Extended Data Fig. 1); hence, our draft assembly covered 72% of the whole genome. The genome is largely composed of repeated sequences, with 19.2% unique sequences (Fig. 1c). Simple repeat sequences are extraordinarily rich, amounting to 19.6% (Fig. 1c, 1d). As much as simple repeats and transposons, 20.5% of the genome is occupied by unclassified repeat sequences that contain notable amounts of DNA virus-like fragments (Fig. 1c, Supplementary Table 2, 3).

A total of 36,763 protein gene models were predicted; on average, they were 10.4 kb long and contained eight introns, implying large and complex structures (Fig. 1b, Supplementary Table 4). Their gene ontology (GO) term analysis showed that DNA-related metabolism which associated with DNA virus is significantly over-represented compared with that of other rhizarian organisms (Extended Data Fig. 1e, Supplementary Table 5).

From the above gene set of *P. micropora*, we attempted to characterize the genes that have been pivotal for the endosymbiotic evolution. We extracted the genes derived from cyanobacteria as well as those derived from the rest of the bacteria; we refer to the former as endosymbiotic gene transfer (EGT) candidates and the latter as

horizontal gene transfer (HGT) candidates in this study. We obtained 177 EGT and 248 HGT candidates (Fig. 2, Supplementary Table 6). Half of the EGT candidates are genes for high light inducible proteins (HLIPs)¹², which are involved in the protection against excess light energy. Phylogenetic analysis of these HLIPs showed that they are polyphyletic, suggesting that HLIPs should have been acquired by multiple independent gene transfers from cyanobacteria (Extended Data Fig. 2). Thus, the gain of a light protection system should have been crucial for the predacious amoeba to evolve into a photosynthetic organism.

HGT candidates contain genes of diverse functions, including ribosome biogenesis, DNA synthesis and amino acid metabolism. These genes appear to be involved in (1) endosymbiont biogenesis and (2) changes of the cellular nutrient state from heterotrophy to photo-autotrophy. To further examine the genes essential to the evolution of a photosynthetic organism, we compared orthologs among *P. micropora*, primary photosynthetic eukaryotes and predaceous eukaryotes (Extended Data Fig. 3a, 3b); 12 orthologous groups are conserved in the former two but not in the latter, including the genes for light acclimation, organelle gene expression and changes of the

Fig. 2. *P. micropora* nuclear genes acquired by EGT/HGT. **a, b,** A functional classification of the *P. micropora* nuclear genes derived from cyanobacteria (EGT candidates) (**a**), and those from other bacteria (HGT candidates) (**b**). **c, d,** The amino acid sequence identity of EGT candidates against *P. micropora* MYN1 plastid genes (**c**) and that of HGT candidates against bacterial genes of the NCBI nr database (**d**). **e, f,** An estimation of the gene transfer age for EGT candidates (**e**) and HGT candidates (**f**). The endosymbiosis initiation period is green-highlighted. The ages of gene transfer in (**e**) and (**f**) were calculated based on the divergent time points (45.7–64.7 MYA) of two *Paulinella* species; thus, a gene transfer age younger than 60 MYA (striped phase) could not be estimated.

cellular nutrient state. Some of them were obtained horizontally from eukaryotes (Extended Data Figs. 3b–d). Therefore, *P. micropora* utilized the genes of diverse origins for endosymbiotic evolution.

The biggest challenge of this study is to elucidate the temporal sequence of the events that occurred at the birth of photosynthetic eukaryotes. To solve this puzzle, we first estimated how and when EGT occurred, based upon the sequence similarity between the EGT candidates and organelle-encoded genes. The results were surprising. We could not find any case with more than 80% amino acid sequence identity conserved between them (Fig. 2c), suggesting that plastidial EGT did not occur in a recent time period (Fig. 2e). We further searched for nuclear-localized plastid DNAs and nuclear-localized mitochondria DNAs^{15,16} in the genome and found the latter but not the former. Therefore, it is likely that plastidial EGT rapidly activated and then ceased early in the endosymbiotic evolution in *P. micropora*, while this cool down was not found for mitochondrial EGT (Extended Data Fig. 4).

We confirmed this hypothesis from the different angle. Phylogenetic tree analysis of the EGT candidates showed that most of them already lost their counterparts

in the plastid genome, except for four genes; hence, we reckon that these four genes were transferred from the plastid to the nucleus relatively recently. The divergences of the four genes between their nuclear and plastid counterparts were estimated to have occurred 319.8 to 98.6 million years ago (MYA) (Extended Data Fig. 5). Considering that the photosynthetic *Paulinella* species have diverged from the heterotrophic species 141.4 to 93.6 MYA¹¹, even the latest EGT at 98.6 MYA had occurred within the initiation period of the endosymbiotic evolution. Taken together, the results of this study strongly suggest that EGT rapidly activated and ceased within the initial period of the endosymbiotic evolution (Fig. 2e), and a similar time course was also found for HGT (Fig. 2d, 2f).

What does this rapid and simultaneous cool-down of EGT and HGT (Figs. 2c–2f) mean? The most simple and likely explanation is that the predaceous *Paulinella* shrank and lost phagocytic activity at this time to become a photosynthetic organism, accompanied by the shut-down of phagocytosis-aided EGT/HGT. In reality, HGT from prey cyanobacteria occurred in the predaceous *Paulinella* species¹⁷. If our assumption is correct, the predaceous *Paulinella* should have changed its cellular, genomic and

metabolic systems very quickly in terms of the evolutionary time scale. How was this drastic change possible? To examine this, we re-focused this study on DNA virus-like fragments frequently found in the *P. micropora* genome.

Fig. 3a shows a genomic scaffold containing putative virus fragments that are characterized by having from a dozen to a hundred copies, high GC content, ORFs similar to eukaryotic virus genes, and many intron-less genes of heterogeneous origins with unknown functions (Extended Data Fig. 6, Supplementary Table 3). In addition, they are often intermingled with simple repeats and mobile genetic elements, i.e., Maverick/Polinton-type giant transposons^{18,19} and retrotransposons. Most notably, the maximum fragment size reaches 300 kb (Extended Data Fig. 6). These structural features of the DNA virus-like fragments resemble those of nucleocytoplasmic large DNA viruses (NCLDV)²⁰ whose genome size ranges from 100 kbp to 2.5 Mbp and who have many genes of heterogeneous origins with unknown functions. However, a phylogenetic analysis based on DNA polymerases shows that those genes, encoded by the putative viral fragments, form a monophyletic clade distant from the genes of eukaryotes, prokaryotes and known NCLDVs (Fig. 3b). Therefore, we assume that they

Fig. 3. Putative DNA virus and mobile elements in *P. micropora* MYN1. **a**, A schematic view of DNA virus-like fragments and mobile elements in the *P. micropora* draft genome (Scaffold 1104). The genomic regions were coloured according to the sequence characteristics; putative dsDNA virus (pink), Polinton (light blue), retrotransposon (brass yellow) and simple repeat-rich region (grey). The copy number of the interspersed repeat elements was analysed by BLASTN against the simple-repeat-masked *P. micropora* draft genome. **b**, ML phylogenetic tree of DNA polymerases of viruses, eukaryotes and prokaryotes. **c**, Divergent time analysis of the virus-type GPCR in *Paulinella*'s lineage. **d**, **e**, Divergent time analysis of DNA polymerase genes of metazoa-type (**d**) and rhizarian-type (**e**) Polintons. Asterisks: the branch point of *P. micropora* and *P. chromatophora* set at 45.7–64.7 MYA. Green bands: initiation periods of endosymbiosis with cyanobacteria (93.6–141.4 MYA). *P. micropora*; *Paulinella micropora* MYN1, *P. chromatophora*; *Paulinella chromatophora* CCAC0185, *P. brassicae*; *Plasmodiophora brassicae*, *R. filosa*; *Reticulomyxa filosa*, *S. purpuratus*; *Strongylocentrotus purpuratus*.

are from a novel large DNA virus but share several properties with known NCLDV.

Our next question was when the putative virus infected the *Paulinella* lineage.

Although ancient infection hallmarks were already smeared, we found a suggestive case

in a *Paulinella*-specific gene family (Fig. 3c, Extended Data Fig. 7). The G-protein

coupled receptor (GPCR) genes rapidly expanded and diversified within a short

evolutionary period around the endosymbiosis initiation point. Noteworthy, two genes

of this family were found only in the putative viral fragment regions (yellow squares in

Fig. 3c and Extended Data Fig. 7). This suggests that these two genes diverged from the

rest of the family around the endosymbiosis initiation point and have been inherited

from the virus genome. This indicates that the putative virus has infected the *Paulinella*

lineage around the endosymbiosis initiation point or earlier.

To further prove this, we investigated the Maverick/Polinton-type transposons

derived from a virophage²¹, which parasitizes giant viruses (extremely large NCLDVs)

with its propagation depending on the host virus^{21,22}. Virophages are also integrated into

the nuclear genome and could function as an anti-DNA virus system to protect

eukaryotic cells from the DNA virus^{23,24}. Therefore, we hypothesized that the

emergence of Mavericks/Polintons and their amplifications have occurred concomitant with the DNA virus infection. In the *P. micropora* genome, two distinct Mavericks/Polintons, metazoan- and rhizaria-type, were detected in abundance (Extended Data Fig. 8, Supplementary Table 7). The divergent time analysis showed that both started amplification in the endosymbiosis initiation period (Fig. 3d, 3e). These results of the Maverick/Polinton-type transposons support the hypothesis that the large DNA virus infected the *Paulinella* lineage around the endosymbiosis initiation period.

Recent studies of NCLDV and giant DNA virus have drastically changed our conventional view of viruses¹³, especially their huge potential to incorporate diverse genetic materials of heterogenous origins^{25,26} and to mediate their shuffling. Considering these properties of large DNA virus and the results of this study, we could reconstruct the initial evolutionary process of the photosynthetic *Paulinella* species as shown in Fig. 4. In this hypothetical model, large DNA virus contributed to the endosymbiotic evolution as a critical player, in addition to the original players of the host and symbiont. We should note that, in general, the detection of ancient infection

Fig. 4. Initial process of the endosymbiotic evolution of photosynthetic *Paulinella* species modeled from the *P. micropora* genomic data. 1) Predacious ancestors digested prey bacteria via phagocytosis with continuing low levels of HGT. 2) The infection of large DNA virus triggered the massive HGT/EGT to promote the rapid endosymbiotic evolution. 3) Acquiring photosynthetic competency shrunk the phagocytic activity to shut down the source of HGT/EGT. 4) The photo-autotrophic *Paulinella* sp. contains one photosynthetic organelle per cell, hence, release of the organelle DNA hardly occurred without losing photosynthetic activity. In this final stage, virus-mediated gene transfer continued at trace level.

hallmarks of large DNA virus seems difficult because (1) it could be easily lost due to its harmful and undesirable effects on host proliferation, (2) of poor information of the virus sequences and (3) of repeated sequences that are apt to be omitted in the assembly process of the genomic sequencing projects. This *Paulinella* version of the endosymbiotic evolution hypothesis deserves further examination to test its generality in many evolutionary systems.

Methods

Data availability

The sequences of the *P. micropora* draft genome, plastid (chromatophore) genome, mitochondria genome and the raw reads data set were deposited to the DDBJ (Accession No. are shown in Supplemental Table 8) and DDBJ reads archives (DRA Accession No. DRA003059, DRA003106, DRA008524).

***P. micropora* culture and cell isolation.** The *P. micropora* MYN1 strain (NIES Collection, Tsukuba, JAPAN, NIES-4060) was cultured according to Nomura et al.

194 (2014)⁵ and harvested by low-speed centrifugation ($500 \times g$, 2 min) at 4 °C. The
195 harvested cells were resuspended in the culture medium and filtrated through a 20 μ m
196 mesh nylon filter (HD-20, Nippon Rikagaku Kikai Co., Ltd., Tokyo, Japan) to remove
197 dead cell aggregates with high bacterial contamination. Recovered healthy cells were
198 repeatedly washed with culture medium and subjected to RNA extraction. For
199 extraction of chromatin DNA, the cells were subsequently washed three times with 10
200 mM Tris-HCL (pH 8.0), six times with 10 mM Tris-HCL (pH 8.0) plus 10 mM EDTA,
201 and recovered by a 5 μ m mesh nylon filter (PP-5n, Kyoshin Rikoh Inc., Tokyo, Japan)
202 to give clean cells largely free of bacterial contamination.

203

204 **Chromatin and genomic DNA extraction.** Genomic DNA used for paired-end (300 b,
205 500 b) and mate-pair (3 kb, 5 kb) libraries for HiSeq sequencing were purified by
206 chromatin immunoprecipitation (ChIP) as follows. Ten milligrams of *P. micropora* cells
207 were homogenized in 500 μ l homogenizing buffer (20 mM Tris-HCl (pH 7.6), 10 mM
208 NaCl, 10 mM KCl, 2.5 mM EDTA, 250 mM sucrose, 0.1 mM spermine, 0.5 mM
209 spermidine and 1 mM DTT) using a 30 μ m clearance glass homogenizer (RD440911,

210 Teraoka Co., Ltd., Osaka, Japan). After centrifugation ($1000 \times g$, 10 min, 4 °C), the
211 pellets were resuspended in 300 μ l ChIP buffer (50 mM Tris-HCl (pH 8.0), 500 mM
212 NaCl, 10 mM EDTA, 0.1% SDS, 0.5% Na-deoxycholate, 1% Triton X-100, 1 mM DTT
213 and 10% glycerol) with 20 μ l Dynabeads protein G (Thermo Fisher Scientific, MA,
214 U.S.A.) charged with 1 μ g anti-histone H3 antibody (Ab1791) (Abcam plc, Cambridge,
215 UK) and incubated at 4 °C for 20 min. Dynabeads were then washed twice with ChIP
216 buffer, twice with glycerol-free ChIP buffer and finally suspended in DNA extraction
217 buffer (10 mM Tris-HCl (pH 8.0), 1 mM EDTA and 1% SDS). After RNase A (10
218 μ g/ml) and proteinase K (200 μ g/ml) treatment, DNAs were purified by using Plant
219 DNeasy Mini Kit (Qiagen, Hilden, Germany). For the construction of the long mate-pair
220 libraries (12 kb, 15 kb, 18 kb and 20 kb), the total *P. micropora* genome was extracted
221 without ChIP purification.

222

223 **Genome sequencing and assembly.** Sequencing libraries were prepared using a
224 TruSeq DNA PCR-Free Library Preparation Kit and a Nextera Mate Pair
225 Library Prep Kit (Illumina, San Diego, CA). Two paired-end libraries with

226 300 and 500bp inserts and six mate pair libraries (3kb, 5kb, 12kb, 15kb,
227 18kb, and 20kb) were constructed and sequenced on the Illumina HiSeq
228 2500 sequencers with 151 cycles per run. The nuclear draft genome was
229 assembled by SOAPdenovo v2.04-r240²⁸ with a k-mer size of 121 after removing the
230 sequence reads of the plastid (chromatophore) and mitochondria genomes, and those of
231 two contaminating bacteria genomes. After the genome assembly, we checked for the
232 contamination of the organelle genome and the bacteria genomes again, and we
233 removed the contaminants from the draft genome. K-mer frequency analysis was
234 performed by Jellyfish²⁹. Genome scaffolds longer than 1 kb were analysed in this
235 study.

236

237 **RNA-seq and Iso-Seq analysis.** RNAs were extracted from *P. micropora* cells at 0, 4,
238 8, 12, 16 and 20 hr of 14L/10D photoperiod by Trizol® reagent (Thermo Fisher
239 Scientific), and further purified using Plant RNeasy Mini Kit (Qiagen) with RNase-free
240 DNase I treatment (Qiagen). Samples of the above time points were equally mixed and
241 subjected to RNA-seq analysis using Agilent Strand Specific RNA Library Preparation

242 Kit (Agilent Technologies, CA, U.S.A) and Illumina HiSeq 2500 (Illumina). The paired
243 end reads of RNA-seq were *de novo* assembled by Trinity³⁰ with the default setting or
244 by CLC Genomics Workbench 7.0.3 (Qiagen) using a K-mer value of 54. The RNA-seq
245 reads were mapped on the genome with Tophat³¹ and assembled with Cufflinks³² or
246 Trinity³⁰. For isoform-sequencing (Iso-Seq) of full-length transcripts, cDNAs were
247 prepared from polyA+RNA using SMARTer® PCR cDNA Synthesis Kit (Takara Bio
248 Inc., Shiga, Japan). The cDNA samples were size-fractionated with the BluePippin™
249 system (Sage Science, MA, USA) and 700–1500 bp, 1500–3000 bp and 3000–6000 bp
250 fractions were analysed with a PacBio®RSII sequencer (Pacific Biosciences, CA,
251 U.S.A.). Iso-Seq-contigs were constructed using the RS_IsoSeq protocol in SMRT
252 Analysis (v2.3.0) with the parameter of estimated cDNA size. The *de novo* assembled
253 RNA-seq-contigs and Iso-Seq-contigs were mapped to the genome by BLAT³³, and
254 each contig was annotated when at least 80% of its sequence was mapped. The mapped
255 transcript data were used to make longer transcript models with PASA2 v. 2.0.2³⁴.
256
257 **Annotation of repeat sequences.** Repeat sequences of *P. micropora* were identified

using the RepeatModeller package (v. open-1-0-8, <http://www.repeatmasker.org>) and masked by RepeatMasker v. 4.0²⁷, using the identified-model repeat sequences and the repeat sequences of Repbase (ver. 20150807)³⁵ (<https://www.girinst.org/repbase/>). The model repeat sequences were annotated based on the sequence homology search against the NCBI nr database by BLASTX³⁶. Representative Polintons and retrotransposons in Supplementary Table 3 were identified by manual inspection of the *P. micropora* draft genome aided by a sequence similarity search using BLAST software³⁶.

Gene annotation. Protein genes of *P. micropora* were annotated by a combination of transcriptome-based gene modelling, *ab initio* gene prediction and protein homology-based gene prediction. In the transcriptome-based gene modelling, the transcripts were masked first by RepeatMasker because many spurious repetitive sequences, which were not removed by genome-repeat-masking, were detected. We discarded the transcript models when more than 80% of the region was masked by RepeatMasker. Exceptions were made when ORFs (> 50 aa) were predicted from the unmasked region. The ORFs and coding sequences of transcripts were predicted with

the Transdecoder Utility of Trinity³⁷. *Ab initio* gene prediction was performed by Augustus³⁸, whose training was performed using Iso-Seq data. Since *P. micropora* protein genes often contain simple repeat sequences in the exon regions, we avoided masking them for the *ab initio* gene prediction. The protein homology-based gene prediction was performed by Exonerate³⁹ after masking both the simple repeat and the interspersed repeat sequences, because the homology search in the presence of simple repeat sequences generated an extraordinary number of meaningless candidates (data not shown). In the Exonerate analysis, we used the protein sequences that passed the prescreening by BLASTX search (*P. micropora* genome vs. a local protein database composed of Uniprot and 4 rhizarian organisms, *B. natans*, *P. brassicae*, *P. chromatophore* and *R. filosa*). All gene models described above were combined, and the best one for each gene locus was chosen according to the bit score of the BLAST search, the presence/absence of transcript and ORF length. The quality of the genome assembly and the annotation was assessed by BUSCO v. 1⁴⁰.

Detection of DNA virus-like fragments. Genomic segments with discernible

boundaries and distinct from the rest of the genome by the following four criteria were denoted as DNA virus-like fragments or putative DNA virus fragments: 1) repeat sequences detected by RepeatModeller but distinct from retrotransposons and the Maverick/Polinton-type transposons, 2) higher GC content, 3) large heterogeneous intron-less gene clusters and 4) ORFs similar to DNA virus proteins are encoded; DNA virus genes are found within the top 100 by BLASTP search against the NCBI nr database. The detailed procedure is as follows. Genomic scaffolds containing repeat sequences of the unknown class²⁷ were subjected to ORFfinder⁴¹, and ORFs detected were annotated by BLASTP³⁶ search against the NCBI nr database. The GC% distribution was analysed with CLC genomic workbench 7.0.3 (Qiagen). The virus copy number was analysed using the BLASTN³⁶ program searching the simple repeat-masked draft genome for virus coding sequences (Supplementary Table 3) or 100 b fragments generated by slicing the virus-containing scaffolds (Fig. 3, Extended Data Fig. 6). BLASTN-redundant hits were manually removed.

GO-term-, metabolism-pathway-, orthogroup- and protein-domain analysis.

GO-terms were acquired by InterproScan 5⁴², and the enrichment analysis was performed by web-based GStat⁴³ (<http://gostat.wehi.edu.au/>) using 27,653 GO-terms of 12,007 *P. micropora* genes and 88,634 GO-terms of 39,773 genes of 4 rhizarian organisms (*B. natans*, *P. brassicae*, *R. filosa* and *P. micropora*). Metabolism-pathways were analysed using KAAS of KEGG⁴⁴ (<https://www.genome.jp/kegg/kaas/>). Orthogroup-analysis of Extended Data Fig. 3 was performed by Orthofinder⁴⁵ using a local protein database (Supplementary Table 10). Protein domain information in Extended Data Fig. 7 was acquired by NCBI conserved domain (CD) search (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>)⁴⁶.

nuclear-localized plastid DNA (nupDNA) and mitochondria DNA (numDNA) analysis. Using the *P. micropora* plastid genome (DDBJ Accession No. LC490351) and mitochondria genome sequence (DDBJ Accession No. LC490352) for queries, nupDNA and numDNA were searched for by BLASTN against the *P. micropora* draft genome and the raw sequence reads of Illumina HiSeq2500. In the raw read-based nupDNA and numDNA analysis, chimeric segments of the organelle-like and non-organelle

sequences, which represent the junction region of nuclear localized organelle DNA, were surveyed. The detected reads were assembled by CodonCode Aligner (CodonCode Corporation, MA, U.S.A.). The chimera artefacts due to sequencing adaptors, or contaminated bacterial and mitochondria genomes, were identified by BLAST analysis using the NCBI database (nt, nr) and the mitochondria genome sequence (LC490352).

Phylogenetic analysis and the divergent time analysis. Multiple sequence alignment analyses were performed by MUSCLE⁴⁷ and MAFFT⁴⁸. The phylogenetic trees were constructed using MEGA packages (version 6⁴⁹, 7⁵⁰ and -CC⁵¹) and IQ-tree⁵². The divergent time analysis was performed using the RelTime method⁵³ implemented in MEGA 6. Parameters used for phylogenetic analyses are shown in Supplementary Table 6 and 9, respectively.

Analysis of EGT/HGT candidates. *P. micropora* nuclear genes derived from cyanobacteria are referred to as EGT, and those derived from the rest of the bacteria are defined as HGT. To screen the EGT/HGT candidates, *P. micropora* MYN1 genes were

338 used as queries for the BLASTP search against the NCBI nr database with e-value $\leq$
339 $1e^{-10}$, and top-hitting genes of the *Paulinella* plastid genes or prokaryotic protein
340 sequences were selected. These genes were subjected to multiple alignment analyses by
341 MUSCLE using the protein sequences of a local protein dataset in Supplementary Table
342 10 and of the best BLASTP hit sequences. The obtained data sets were subjected to a
343 neighbour joining (NJ) phylogenetic analysis (MEGA6, 7, CC) to choose *P. micropora*
344 nuclear genes that form sister groups with prokaryotes or photosynthetic eukaryotes.
345 After fine taxon re-samplings from the NCBI nr database, the NJ analysis selection was
346 conducted again. The selected genes were finally subjected to maximum likelihood
347 (ML) analysis using MEGA packages (MEGA6, 7, CC). *P. micropora* nuclear genes
348 that satisfied at least one of the following three criteria were used as EGT and HGT
349 candidates. 1) BLAST analysis of the gene gave no hint of eukaryote genes in the NCBI
350 nr database. 2) EGT and HGT are supported by ML phylogenetic analysis with a high
351 bootstrap value. We adopted a bootstrap value of 95 as threshold when phylogenetically
352 available protein alignment sequence positions were long enough (≥ 100 aa). We
353 lowered the threshold to 70 when the available sites were less than 100 aa. 3) The gene

is included in the clade of photosynthetic organisms. To confirm the validity of the above mentioned EGT candidate selection, we also screened the EGT candidates by another independent procedure. The EGT candidates obtained by these two independent analyses were almost overlapping and, therefore, used for the analysis (Supplementary Table 6). The alternative analysis procedure is as follows. We performed a phylogenetic analysis using the software of multiple alignment (MAFFT) and ML phylogenetic analysis (IQ-tree), and alignment trimming tools (trimAI⁵⁴, BMGE⁵⁵). We selected genes hitting alpha-type cyanobacteria^{56,57} within the top 100 by BLAST search against a custom database, which consists of the NCBI nr database supplemented with the protein sequences of 14 phylogenetically informative protists (Supplemental Table 10).

Estimation of EGT/HGT timing. To estimate the timing of EGT/HGT, we used *P. micropora* genes whose counter genes of *P. chromatophora* CCAC0185 were reported as EGT/HGT candidates^{58,59}. In addition, we restricted the analysis to *Paulinella* ortholog's pairs that form a monophyletic sister group with a high bootstrap value ($\geq$ 70). The nearest protein sequences of HGT/EGT candidates were surveyed from NCBI

nr database by BLASTP analysis. Within the top 2000 sequences of BLASTP hits, the phylogenetically nearest gene sequences were estimated using NJ and ML phylogenetic analysis. To estimate gene transfer timing, the branching time point when *P. micropora* MYN1 separated from the nearest non-rhizarian organisms in the ML phylogenetic tree was calculated using the RelTime method⁵⁵. We used an estimated value of the divergence of *P. micropora* and *P. chromatophora* (45.7–64.7 MYA) based on the 18S rRNA phylogenetic tree corrected by fossil information^{11,60}.

Analysis of Mavericks/Polinton transposons. *P. micropora*'s Mavericks/Polintons transposons were detected from the draft genome by tBLASTN using the sequence of the DNA polymerase (DNA-pol) domain. Thousands of DNA polymerase ORFs, predicted from the genome sequences by Transdecoder³⁷, were subjected to NJ phylogenetic analyses. We grouped highly similar copies. Representative sequences that have long ORFs and less ambiguous amino acid residues were selected from each group and subjected to ML phylogenetic analysis.

386 **Divergent time analysis of the mobile elements.** In the divergent time analysis of
387 Marvericks/Polintons DNA polymerases, representative DNA-pol sequences of *P.*
388 *micropora* MYN1 Polintons having the traits of recent amplification (nucleotide
389 sequence identity between the copies $\geq 90\%$) were used. For the DNA-pol sequences of
390 *P. chromatophora* CCAC0185 Polintons, the genome sequence-reads (SRR3217293.sra,
391 SRR3217303.sra) were searched by BLASTX (e-value $< 1e^{-20}$) using the sequence of *P.*
392 *micropora* MYN1 Polintons, and then, the hit-reads were assembled into contigs with
393 CLC Genomic Workbench 7.0 (Qiagen). For the analysis of virus-type GPCR genes, in
394 addition to *P. micropora* MYN1 genes and *P. micropora* putative virus genes, the
395 translated ORFs of *P. chromatophora* CCAC0185 transcripts were analysed. The GPCR
396 gene family was detected by Orthofinder⁴⁵ and BLASTP³⁶ search. In this analysis, the
397 genes encoding seven intact trans-membrane domain sequences, of which all seven
398 trans-membrane helices can be identified by CD search⁴⁶, were used. Furthermore,
399 several genes predicted *ab initio* by Augustus without any supporting experimental data
400 were removed from the analysis, because their gene models appeared to be artificial
401 from applying eukaryotic splicing rules to virus-like fragments. The divergent time

points of the mobile elements and GPCR genes were calculated by setting the nearest branching point of *P. micropora* MYN1 and *P. chromatophora* CCAC0185 at 45.7–64.7 MYA.

Statistical analysis

In GO-term enrichment analysis, Fisher's exact test (two tailed test) was performed and the p-values corrected with false discovery rate (Benjamini) were calculated. In the phylogenetic analysis, bootstrap test with ≥ 100 replicates was performed.

References

1

Yoon, H.S., Hackett, J.D., Ciniglia, C., Pinto, G. & Bhattacharya, D. (2004) A molecular timeline for the origin of photosynthetic eukaryotes. *Mol. Biol. Evol.* **21**, 809-18, <https://doi.org/10.1093/molbev/msh075> (2004).

2

- 418 Sánchez-Baracaldo, P., Raven, J. A., Pisani, D. & Knoll, A. H. Early photosynthetic
419 eukaryotes inhabited low-salinity habitats. *Proc. Natl. Acad. Sci. USA* **114**,
420 E7737-E7745. <https://doi.org/10.1073/pnas.1620089114> (2017).
- 421
- 422 3
- 423 McFadden, G. I. Origin and evolution of plastids and photosynthesis in eukaryotes.
424 *CSH Perspect Biol.* **6**, a016105, <https://doi.org/10.1101/cshperspect.a016105> (2014).
- 425
- 426 4
- 427 Burki, F. & Keeling, P. J. Rhizaria. *Curr. Biol.* **24**, R103-R107,
428 <https://doi.org/10.1016/j.cub.2013.12.025> (2014).
- 429
- 430 5
- 431 Nomura, M., Nakayama, T. & Ishida, K. Detailed process of shell construction in the
432 photosynthetic testate amoeba *Paulinella chromatophora* (euglyphid, Rhizaria). *J.*
433 *Eukaryot. Microbiol.* **61**, 317-321. <https://doi.org/10.1111/jeu.12102> (2014).

434

435 6

436 Matsuo, M. *et al.* Characterization of spliced leader trans-splicing in a photosynthetic437 rhizarian amoeba, *Paulinella micropora*, and its possible role in functional gene transfer.438 *PLoS One* **13**, e0200961, <https://doi.org/10.1371/journal.pone.0200961> (2018).

439

440 7

441 Nowack, E. C., Melkonian, M. & Glöckner, G. Chromatophore genome sequence of

442 *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. *Curr. Biol.* **18**,443 410-418, <https://doi.org/10.1016/j.cub.2008.02.051> (2008).

444

445 8

446 Yoon, H. S. *et al.* A single origin of the photosynthetic organelle in different *Paulinella*447 lineages. *BMC Evol. Biol.* **9**, 98, <https://doi.org/10.1186/1471-2148-9-98> (2009).

448

449 9

450 Reyes-Prieto, A. *et al.* Differential gene retention in plastids of common recent origin.
451 *Mol. Biol. Evol.* **27**, 1530-1537, <https://doi.org/10.1093/molbev/msq032> (2010).
452
453 10
454 Nakayama, T. & Ishida, K. Another acquisition of a primary photosynthetic organelle is
455 underway in *Paulinella chromatophora*. *Curr. Biol.* **19**, R284-R285,
456 <https://doi.org/10.1016/j.cub.2009.02.043> (2009).
457
458 11
459 Delaye, L., Valadez-Cano, C. & Pérez-Zamorano, B. (2016) How Really Ancient Is
460 *Paulinella Chromatophora*? *PLOS Curr. Tree. Life*.
461 <https://doi.org/10.1371/currents.tol.e68a099364bb1a1e129a17b4e06b0c6b> (2016).
462
463 12
464 Martin, W. *et al.* Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast
465 genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the

466 nucleus. *Proc. Natl. Acad. Sci. USA.* **99**, 12246-12251,

467 <https://doi.org/10.1073/pnas.182432999> (2002).

468

469 13

470 Koonin, E. V. & Yutin, N. Evolution of the Large Nucleocytoplasmic DNA Viruses of

471 Eukaryotes and Convergent Origins of Viral Gigantism. *Adv. Virus Res.* **103**, 167-202,

472 <https://doi.org/10.1016/bs.aivir.2018.09.002> (2019).

473

474 14

475 Komenda, J. & Sobotka, R. Cyanobacterial high-light-inducible proteins--Protectors of

476 chlorophyll-protein synthesis and assembly. *Biochim. Biophys. Acta.* **1857**, 288-295,

477 <https://doi.org/10.1016/j.bbabbio.2015.08.011> (2016).

478

479 15

480 Matsuo, M., Ito, Y., Yamauchi, R. & Obokata, J. The rice nuclear genome continuously

481 integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear

482 DNA flux. *Plant Cell* **17**, 665-675, <https://doi.org/10.1105/tpc.104.027706> (2005).

483

484 16

485 Smith, D. R., Crosby, K. & Lee, R. W. Correlation between nuclear plastid DNA
486 abundance and plastid number supports the limited transfer window hypothesis.

487 *Genome Biol. Evol.* **3**, 365-71, <https://doi.org/10.1093/gbe/evr001> (2011).

488

489 17

490 Bhattacharya, D. *et al.* Single cell genome analysis supports a link between phagotrophy
491 and primary plastid endosymbiosis. *Sci Rep.* **2**, 356, <https://doi.org/10.1038/srep00356>
492 (2012).

493

494 18

495 Feschotte, C. & Pritham, E. J. Non-mammalian c-integrases are encoded by giant
496 transposable elements. *Trends Genet.* **21**, 551-552,
497 <https://doi.org/10.1016/j.tig.2005.07.007> (2005).

498

499 19

500 Kapitonov, V. V. & Jurka, J. Self-synthesizing DNA transposons in eukaryotes. *Proc.*501 *Natl. Acad. Sci. USA* **103**, 4540-4545, <https://doi.org/10.1073/pnas.0600833103> (2006).

502

503 20

504 Koonin, E. V. & Yutin, N. Multiple evolutionary origins of giant viruses. *F1000Res.* **7**,505 F1000 Faculty Rev-1840, <https://doi.org/10.12688/f1000research.16248.1> (2018).

506

507 21

508 Fischer, M. G. & Suttle, C. A. A virophage at the origin of large DNA transposons.

509 *Science* **332**, 231-234, <https://doi.org/10.1126/science.1199412> (2011).

510

511 22

512 La Scola, B. *et al.* The virophage as a unique parasite of the giant mimivirus. *Nature*513 **455**, 100-104. <https://doi.org/10.1038/nature07218> (2008).

514

515 23

516 Fischer, M. G. & Hackl, T. Host genome integration and giant virus-induced
517 reactivation of the virophage mavirus. *Nature* **540**, 288-291.
518 <https://doi.org/10.1038/nature20593> (2016).

519

520 24

521 Blanc, G., Gallot-Lavallée, L. & Maumus, F. Provirophages in the Bigelowiella genome
522 bear testimony to past encounters with giant viruses. *Proc. Natl. Acad. Sci. USA* **112**,
523 E5318-5326, <https://doi.org/10.1073/pnas.1506469112> (2015).

524

525 25

526 Boyer, M. *et al.* Giant Marseillevirus highlights the role of amoebae as a melting pot in
527 emergence of chimeric microorganisms. *Proc. Natl. Acad. Sci. USA* **106**, 21848-21853,
528 <https://doi.org/10.1073/pnas.0911354106> (2009).

529

530 26

531 Piskurek, O. & Okada, N. Poxviruses as possible vectors for horizontal transfer of
532 retroposons from reptiles to mammals. *Proc. Natl. Acad. Sci. USA* **104**, 12046-12051
533 <https://doi.org/10.1073/pnas.0700531104> (2007).

534

535 27

536 Smit, AFA, Hubley, R. & Green, P. *RepeatMasker Open-4.0*. 2013-2015,
537 <http://www.repeatmasker.org>.

538

539 **Method-only references**

540 28

541 Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de
542 novo assembler. *GigaScience* **1**, 18, <https://doi.org/10.1186/2047-217X-1-18> (2012).

543

544 29

545 Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of

546 occurrences of k-mers. *Bioinformatics* **27**, 764–770,
547 <https://doi.org/10.1093/bioinformatics/btr011> (2011).
548
549 30
550 Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a
551 reference genome. *Nat. Biotechnol.* **29**, 644–652, <https://doi.org/10.1038/nbt.1883>
552 (2011).
553
554 31
555 Trapnell C., Pachter L. & Salzberg, S. L. TopHat: discovering splice junctions with
556 RNA-Seq. *Bioinformatics* **25**, 1105–1111, <https://doi.org/10.1093/bioinformatics/btp120>
557 (2009).
558
559 32
560 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals
561 unannotated transcripts and isoform switching during cell differentiation. *Nat.*

562 *Biotechnol.* **28**, 511–515, <https://doi.org/10.1038/nbt.1621> (2010).

563

564 33

565 Kent W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656-664,

566 <https://doi.org/10.1101/gr.229202> (2002).

567

568 34

569 Haas, B.J. *et al.* Improving the Arabidopsis genome annotation using maximal

570 transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654-5666,

571 <https://doi.org/10.1093/nar/gkg770> (2003).

572

573 35

574 Bao, W., Kojima, K. K., & Kohany, O. Repbase Update, a database of repetitive

575 elements in eukaryotic genomes. *Mob. DNA* **6**, 11,

576 <https://doi.org/10.1186/s13100-015-0041-9> (2015).

577

578 36

579 Camacho C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421,
580 <https://doi.org/10.1186/1471-2105-10-421> (2009).

581

582 37

583 Haas, B.J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the
584 Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494-1512,
585 <https://doi.org/10.1038/nprot.2013.084> (2013).

586

587 38

588 Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically
589 mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–
590 644, <https://doi.org/10.1093/bioinformatics/btn013> (2008).

591

592 39

593 Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence

594 comparison. *BMC Bioinformatics* **6**, 31, <https://doi.org/10.1186/1471-2105-6-31> (2005).

595

596 40

597 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov E. M.

598 BUSCO: assessing genome assembly and annotation completeness with single-copy

599 orthologs. *Bioinformatics* **31**, 3210–3212, <https://doi.org/10.1093/bioinformatics/btv351>

600 (2015).

601

602 41

603 David, L. *et al.* Database resources of the National Center for Biotechnology. *Nucleic*

604 *Acids Res.* **31**, 28–33, <https://doi.org/10.1093/nar/gkg033> (2003).

605

606 42

607 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification.

608 *Bioinformatics* **30**, 1236–1240, <https://doi.org/10.1093/bioinformatics/btu031> (2014).

609

610 43

611 Beissbarth, T. & Speed, T. P. GStat: Find statistically overrepresented Gene
612 Ontologies within a group of genes. *Bioinformatics* **20**, 1464-1465,
613 <https://doi.org/10.1093/bioinformatics/bth088> (2004).

614

615 44

616 Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A., & Kanehisa, M. KAAS: an automatic
617 genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**,
618 W182-W185, <https://doi.org/10.1093/nar/gkm321> (2007).

619

620 45

621 Emms, D. M. & Kelly S. OrthoFinder: solving fundamental biases in whole genome
622 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**,
623 157, <https://doi.org/10.1186/s13059-015-0721-2> (2015).

624

625 46

626 Marchler-Bauer, A. et al. CDD/SPARCLE: functional classification of proteins via
627 subfamily domain architectures. *Nucleic Acids Res.* **45**, D200-D203,
628 <https://doi.org/10.1093/nar/gkw1129> (2017).

629

630 47

631 Edgar R. C. MUSCLE: multiple sequence alignment with high accuracy and high
632 throughput. *Nucleic Acids Res.* **32**, 1792-1797, <https://doi.org/10.1093/nar/gkh340>
633 (2004).

634

635 48

636 Katoh, K., Misawa, K., Kuma, K. & Miyata T. MAFFT: a novel method for rapid
637 multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**,
638 3059-3066, <https://doi.org/10.1093/nar/gkf436> (2002).

639

640 49

641 Tamura, K., Stecher, G., Peterson, D., Filipowski, A. & Kumar, S. MEGA6: Molecular

642 Evolutionary Genetics Analysis Version 6.0. *Mol. Biol. Evol.* **30**, 2725-2729,
643 <https://doi.org/10.1093/molbev/mst197> (2013).

644

645 50

646 Kumar, S., Stecher, G., & Tamura, K. (2016) MEGA7: Molecular Evolutionary
647 Genetics Analysis Version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870-1874,
648 <https://doi.org/10.1093/molbev/msw054> (2016).

649

650 51

651 Kumar, S., Stecher, G., Peterson, D. & Tamura, K. MEGA-CC: Computing Core of
652 Molecular Evolutionary Genetics Analysis Program for Automated and Iterative Data
653 Analysis. *Bioinformatics* **28**, 2685-2686, <https://doi.org/10.1093/bioinformatics/bts507>
654 (2012).

655

656 52

657 Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. IQ-TREE: A fast and

658 effective stochastic algorithm for estimating maximum likelihood phylogenies. *Mol.*

659 *Biol. Evol.* **32**, 268-274, <https://doi.org/10.1093/molbev/msu300> (2015).

660

661 53

662 Tamura, K, Tao, Q. & Kumar S. Theoretical Foundation of the RelTime Method for

663 Estimating Divergence Times from Variable Evolutionary Rates. *Mol. Biol. Evol.* **35**,

664 1770–1782, <https://doi.org/10.1093/molbev/msy044> (2018).

665

666 54

667 Capella-Gutiérrez S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated

668 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973,

669 <https://doi.org/10.1093/bioinformatics/btp348> (2009).

670

671 55

672 Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a

673 new software for selection of phylogenetic informative regions from multiple sequence

alignments. *BMC Evol. Biol.* **10**, 210, <https://doi.org/10.1186/1471-2148-10-210> (2010).

56

Badger M. R., Hanson, D. & Price G. D. Evolution and diversity of CO₂ concentrating mechanisms in cyanobacteria. *Funct. Plant Biol.* **29**, 161–173, <https://doi.org/10.1071/PP01213> (2002).

57

Marin, B., Nowack, E. C., Glöckner, G. & Melkonian, M. The ancestor of the Paulinella chromatophore obtained a carboxysomal operon by horizontal gene transfer from a Nitrococcus-like gamma-proteobacterium. *BMC Evol. Biol.* **7**, 85, <https://doi.org/10.1186/1471-2148-7-85> (2007).

58

Nowack, E. C. *et al.* Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of Paulinella chromatophora. *Proc. Natl. Acad.*

690 *Sci. USA* **113**, 12214-12219, <https://doi.org/10.1073/pnas.1608016113> (2016).

691

692 59

693 Nowack, E. C. et al. Endosymbiotic gene transfer and transcriptional regulation of
694 transferred genes in *Paulinella chromatophora*. *Mol. Biol. Evol.* **28**, 407-422,
695 <https://doi.org/10.1093/molbev/msq209> (2011).

696

697 60

698 Lhee, D. *et al.* Diversity of the Photosynthetic *Paulinella* Species, with the Description
699 of *Paulinella micropora* sp. nov. and the Chromatophore Genome Sequence for strain
700 KR01. *Protist* **168**, 155–170, <https://doi.org/10.1016/j.protis.2017.01.003> (2017).

701

702

703 **Acknowledgements**

704 We thank Dr. Terabayashi T. for his kind instructions in SEM analysis.

705 Computations were partially performed on the NIG supercomputer at ROIS National

706 Institute of Genetics. This work was supported by JSPS KAKENNHI grants: 221S0002,
707 15K14554, 1 and 6K14788, and by a Grant-in-Aid for Scientific Research on
708 Innovative Areas 3308 of the Ministry of Education, Culture, Sports, Science and
709 Technology of Japan.

710

711 **Author contributions**

712 M. M. prepared the *P. micropora* MYN1 genome samples, identified the
713 putative virus sequence in the *P. micropora* MYN1 genome and wrote the draft
714 manuscript. M. M. and A. K. performed the RNA sample preparation. Y. M., H. N., A.
715 T. and A. F. performed the genome sequencing, the genome assembly and the Iso-Seq
716 analysis. Y. S. performed the RNA-seq sequencing analysis. M. N. and K. I. cultured *P.*
717 *micropora* MYN1. M. M., A. K., M. T., H. N., H. T., S. S., M. N. and K. I. annotated
718 the *P. micropora* MYN1 genome. M. M., T. N. and R. K. performed the phylogenetic
719 analysis. T. N., R. K. and Y. I. analysed the organelle (chromatophore, mitochondria)
720 genome. J. O. organized and managed the *P. micropora* MYN1 genome project and
721 finalized the manuscript. We thank the members of Comparative Genomics

722 Laboratory in National Institute of Genetics for technical and computational
723 assistance.

724

725 **Competing interests**

726 The authors declare no competing interests.

727

728 **Materials & Correspondence**

729 Junichi Obokata

730 Email: obokata@kpu.ac.jp

731 Tel/Fax: +81-75-703-5164

732 Graduate School of Life and Environmental Science,

733 Kyoto Prefectural University, Kyoto, 606-8522, Japan

734

735

736

737

738 **Figure Legends**

739 **Fig. 1. An overview of the *P. micropora* MYN1 draft genome.** **a**, A SEM image of *P.*
 740 *micropora* MYN1. **b**, The statistics of the draft genome. **c**, The genome composition of
 741 *P. micropora* MYN1 analysed by RepeatMasker²⁷. **d**, Simple repeats are extraordinarily
 742 rich in *P. micropora* MYN1 compared with other organisms.

743

744 **Fig. 2. *P. micropora* nuclear genes acquired by EGT/HGT.** **a, b**, A functional
 745 classification of the *P. micropora* nuclear genes derived from cyanobacteria (EGT
 746 candidates) (**a**), and those from other bacteria (HGT candidates) (**b**). **c, d**, The amino
 747 acid sequence identity of EGT candidates against *P. micropora* MYN1 plastid genes (**c**)
 748 and that of HGT candidates against bacterial genes of the NCBI nr database (**d**). **e, f**, An
 749 estimation of the gene transfer age for EGT candidates (**e**) and HGT candidates (**f**). The
 750 endosymbiosis initiation period is green-highlighted. The ages of gene transfer in (**e**)
 751 and (**f**) were calculated based on the divergent time points (45.7–64.7 MYA) of two
 752 *Paulinella* species; thus, a gene transfer age younger than 60 MYA (striped phase)
 753 could not be estimated.

754

755 **Fig. 3. Putative DNA virus and mobile elements in *P. micropora* MYN1. a, A**
 756 schematic view of DNA virus-like fragments and mobile elements in the *P. micropora*
 757 draft genome (Scaffold 1104). The genomic regions were coloured according to the
 758 sequence characteristics; putative dsDNA virus (pink), Polinton (light blue),
 759 retrotransposon (brass yellow) and simple repeat-rich region (grey). The copy number
 760 of the interspersed repeat elements was analysed by BLASTN against the
 761 simple-repeat-masked *P. micropora* draft genome. **b**, ML phylogenetic tree of DNA
 762 polymerases of viruses, eukaryotes and prokaryotes. **c**, Divergent time analysis of the
 763 virus-type GPCR in *Paulinella*'s lineage. **d**, **e**, Divergent time analysis of DNA
 764 polymerase genes of metazoa-type (**d**) and rhizarian-type (**e**) Polintons. Asterisks: the
 765 branch point of *P. micropora* and *P. chromatophora* set at 45.7–64.7 MYA. Green
 766 bands: initiation periods of endosymbiosis with cyanobacteria (93.6–141.4 MYA). *P.*
 767 *micropora*; *Paulinella micropora* MYN1, *P. chromatophora*; *Paulinella*
 768 *chromatophora* CCAC0185, *P. brassicae*; *Plasmodiophora brassicae*, *R. filosa*;
 769 *Reticulomyxa filosa*, *S. purpurus*; *Strongylocentrotus purpuratus*.

770

771 **Fig. 4. Initial process of the endosymbiotic evolution of photosynthetic *Paulinella***
772 **species modeled from the *P. micropora* genomic data. 1)** Predacious ancestors
773 digested prey bacteria via phagocytosis with continuing low levels of HGT. **2)** The
774 infection of large DNA virus triggered the massive HGT/EGT to promote the rapid
775 endosymbiotic evolution. **3)** Acquiring photosynthetic competency shrunk the
776 phagocytic activity to shut down the source of HGT/EGT. **4)** The photo-autotrophic
777 *Paulinella* sp. contains one photosynthetic organelle per cell, hence, release of the
778 organelle DNA hardly occurred without losing photosynthetic activity. In this final
779 stage, virus-mediated gene transfer continued at trace level.

780

781

782

783

784

785

Extended Data Fig. 1. Information of the *P. micropora* MYN1 genome. **a**, Estimation of the *P. micropora* MYN1 genome by K-mer frequency analysis. Single peak at 39 of the multiplicity are detected with 31-mer, utilizing 52,592,411,100 b from Illumina 500b pair-end reads. From the peak value and the used reads length, 1.35 Gb genome size was estimated. **b**, Summary of the *P. micropora* MYN1 organelle genome. **c**, Validation of the genome assembly by mapping of the sequences of the isoform sequencing (Iso-Seq) transcripts. 13787 non-redundant Iso-Seq sequences of the intron-containing genes, either hitting protein sequences of the Swiss-Prot database by BLASTX search (e-value $\leq 1e^{-60}$) or containing long ORFs (≥ 300 amino acids), were mapped on the draft genome. **d**, Assessment of the genome assembly using 429 BUSCO v. 1 (Benchmarking Universal Single-Copy Orthologs) genes. In comparison with other eukaryote genome assemblies, the BUSCO values of a photosynthetic organism (*Arabidopsis thaliana*) and that of the well-assembled genome of a rhizarian organism (*Plasmodiophora brassicae*) are shown. **e**, The top 10 GO-terms that are significantly overrepresented in the *P. micropora* MYN1 genome compared with other rhizarian organisms (*Bigelowiella natans*, *Plasmodiophora brassicae*, *Reticulomyxa filosa*). Fisher's exact test p-values corrected with false discovery rate (Benjamini) are represented.

Extended Data Fig. 2. A ML phylogenetic tree of Hlips of *P. micropora* MYN1 and *P. chromatophora*. Hlips of cyanobacteria, photosynthetic eukaryotes and cyanophages are used as a reference for operational taxonomy units. Sixty-four *P. micropora* Hlip genes were subjected to phylogenetic analysis and grouped according to the clade. The redundant paralogs of identical sequences were removed. Asterisks: *Paulinella* ortholog-pair supported by a bootstrap value >70. Blue and red circle: nuclear encoded Hlip of *P. micropora* and *P. chromatophora*. Light-blue triangle: *P. micropora* plastid (chromatophore) Hlip. Phylogenetic branch of the *Paulinella* gene (light blue), virus (grey) and eukaryote Hlip-like gene (orange) are highlighted.

Extended Data Fig. 3. Comparison of the orthogroups of *P. micropora*, primary photosynthetic eukaryotes and predator-type eukaryotes. **a**, Venn diagram of *P. micropora* orthogroups with the predator-type eukaryote orthogroups and the conserved orthogroups in primary photosynthetic eukaryotes. Orthogroups were detected by Orthofinder from the gene sets of *P. micropora* MYN1, 14 predator eukaryotes (*Acanthamoeba castellanii*, *Dictyostelium discoideum*, *Entamoeba histolytica* HM-1, *Bodo saltans*, *Naegleria gruberi*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Monosiga brevicollis* MX1, *Oxytricha trifallax*, *Paramecium tetraurelia*, *Stylonychia lemnae*, *Tetrahymena thermophila*, *Reticulomyxa filosa*, *Thecamonas trahens* ATCC50062) and 13 photosynthetic eukaryotes (*Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Coccomyxa subellipsoidea* C-169, *Klebsormidium nitens*, *Micromonas pusilla* CCMP1545, *Oryza sativa*, *Ostreococcus tauri*, *Physcomitrella patens*, *Selaginella moellendorffii*, *Chondrus crispus*, *Cyanidioschyzon merolae*, *Galdieria sulphuraria*, *Cyanophora paradoxa*). **b**, Annotations of 12 orthogroups conserved in primary photosynthetic eukaryotes, but not in predator eukaryotes. The functional annotation and the cellular localization were based on UniprotKB information of *Arabidopsis thaliana* orthologs, and manual annotations are represented in parenthesis. The estimated origins were based on the ML phylogenetic analysis. Genes involved in light acclimation (cyan), nutrient auxotrophy (yellow) and organelle gene expression (magenta) are highlighted. **c** and **d**, ML phylogenetic trees of orthogroup genes acquired by HGT from other eukaryotes; Haptophyta (**c**, OG0001742) and Stramenopiles (**d**, OG0002023). The phylogenetic analysis of **c** and **d** were performed using a LG+G (**c**) and a LG+G+I model (**d**), respectively. Photosynthetic *Paulinella* species are boxed.

Extended Data Fig. 4. Analysis of nuclear-localized plastid DNA (nupDNA) and mitochondria DNA (numDNA) in *P. micropora*. **a**, Analysis of the junction sequences of nupDNA and numDNA using Illumina raw-reads sequences. 902,121,079 quality-trimmed HiSeq paired-end sequence reads (insert size: 300 b, 500 b) were analysed. **b**, numDNA sequence structures in the *P. micropora* draft genome. **c**, Distribution of numDNA sequences in mitochondrial genome positions. **d**, Nucleotide percent identity of numDNA compared with the mitochondria genome sequence. NumDNA directions are represented by right and left arrows which denote clockwise- and anti-clockwise directions of the mitochondria genome sequence (**c**), respectively.

PsaK; N18495 [LG+G+I]

CcmL; N15137, N17777 [rtREV+G]

Hlip; N30377 [LG+G]

Hypothetical protein; N11415 [LG+I]

Extended Data Fig. 5. ML phylogenetic trees of four EGT candidates whose counterpart orthologs are found in the plastid (chromatophore) genome of *P. micropora*. The clades of the *Paulinella* plastid and cyanobacteria are highlighted in green. Blue circle: *P. micropora* nuclear gene, red circle: *P. chromatophora* nuclear gene, blue triangle: *P. micropora* plastid gene, red triangle: *P. chromatophora* plastid gene, unmarked: cyanobacterium gene. The branch of *P. micropora* and *P. chromatophora* nuclear genes supported by high bootstraps (numbers in the parenthesis) is represented as a red line. The divergence time of *Paulinella* nuclear genes from the plastid- or other cyanobacterial-genes (yellow stars) were estimated by RelTime methods of MEGA6 using the divergent time of *P. micropora* and *P. chromatophora* (45.7–64.7 MYA) (blue cross). PsaK: Photosystem I subunit K, CcmL: CO₂ concentrating mechanism protein, Hlip: High light inducible protein. Brackets mean the substitution model used in the phylogenetic analysis. Trees with species names are available at the repository (<https://figshare.com/s/a665678c48d0af073894>).

Extended Data Fig. 6. Putative DNA virus fragments detected in the *P. micropora* draft genome. ORF structures, GC%, copy number and sequence gaps are represented as described in the legend to Fig. 3.

Extended Data Fig. 7. Characteristics of virus-type GPCRs of *P. micropora* and *P. chromatophora*. **a**, A ML phylogenetic tree of seven transmembrane domains of GPCRs. GPCRs in *P. micropora*, *P. chromatophora*, *P. brassicae* (CEO98393.1, CEO98395.1), and those detected in DNA viral fragments were analysed using a LG+G+F model. Branches with bootstrap values <50 are condensed. **b**, Existence of the transcript. **c**, GC% of the coding sequence of GPCR genes. **d**, Exon/intron structures of the genes. **e**, Protein domain structures. Protein domains were detected by CD search⁴⁸. GPCRs in putative DNA viruses are boxed with a blue line.

Extended Data Fig. 8. ML phylogenetic tree of DNA polymerases of Polintons. a, DNA polymerase sequences of *P. micropora* Polintons (No. 2766 and No. 3055), *R. filosa* Polintons (Genbank, ETO12468.1, ETO19856.1, ETO13433.1), metazoan Polintons in Repbase³⁷ and Mavirus (Genbank, YP 004300281.1) were analysed with a LG+G+I model. The metazoan and rhizarian groups are highlighted in green and yellow, respectively.

786 **Extended Data legends**

787 **Extended Data Fig. 1. Information of the *P. micropora* MYN1 genome. a,**

788 Estimation of the *P. micropora* MYN1 genome size by K-mer frequency analysis.

789 Single peak at 39 of the multiplicity are detected with 31-mer, utilizing 52,592,411,100

790 b from Illumina 500b pair-end reads. From the peak value and the used reads length,

791 1.35 Gb genome size was estimated. **b**, Summary of the *P. micropora* MYN1 organelle

792 genome. **c**, Validation of the genome assembly by mapping of the sequences of the

793 isoform sequencing (Iso-Seq) transcripts. 13787 non-redundant Iso-Seq sequences of

794 the intron-containing genes, either hitting protein sequences of the Swiss-Prot database

795 by BLASTX search (e-value $\leq 1e^{-60}$) or containing long ORFs (≥ 300 amino acids),

796 were mapped on the draft genome. **d**, Assessment of the genome assembly using 429

797 BUSCO v. 1 (Benchmarking Universal Single-Copy Orthologs) genes. In comparison

798 with other eukaryote genome assemblies, the BUSCO values of a photosynthetic

799 organism (*Arabidopsis thaliana*) and that of the well-assembled genome of a rhizarian

800 organism (*Plasmodiophora brassicae*) are shown. **e**, The top 10 GO-terms that are

801 significantly overrepresented in the *P. micropora* MYN1 genome compared with other

rhizarian organisms (*Bigelowiella natans*, *Plasmodiophora brassicae*, *Reticulomyxa filosa*). Fisher's exact test p-values corrected with false discovery rate (Benjamini) are represented.

805

Extended Data Fig. 2. A ML phylogenetic tree of Hlips of *P. micropora* MYN1 and *P. chromatophora*. Hlips of cyanobacteria, photosynthetic eukaryotes and cyanophages are used as a reference for operational taxonomy units. Sixty-four *P. micropora* Hlip genes were subjected to phylogenetic analysis and grouped according to the clade. The redundant paralogs of identical sequences were removed. Asterisks: *Paulinella* ortholog-pair supported by a bootstrap value >70. Blue and red circle: nuclear encoded Hlip of *P. micropora* and *P. chromatophora*. Light-blue triangle: *P. micropora* plastid (chromatophore) Hlip. Phylogenetic branch of the *Paulinella* gene (light blue), virus (grey) and eukaryote Hlip-like gene (orange) are highlighted.

815

Extended Data Fig. 3. Comparison of the orthogroups of *P. micropora*, primary photosynthetic eukaryotes and predator-type eukaryotes. a, Venn diagram of *P.*

818 *micriopora* orthogroups with the predator-type eukaryote orthogroups and the
 819 conserved orthogroups in primary photosynthetic eukaryotes. Orthogroups were
 820 detected by Orthofinder from the gene sets of *P. micropora* MYN1, 14 predator
 821 eukaryotes (*Acanthamoeba castellanii*, *Dictyostelium discoideum*, *Entamoeba*
 822 *histolytica* HM-1, *Bodo saltans*, *Naegleria gruberi*, *Caenorhabditis elegans*,
 823 *Drosophila melanogaster*, *Monosiga brevicollis* MX1, *Oxytricha trifallax*, *Paramecium*
 824 *tetraurelia*, *Stylonychia lemnae*, *Tetrahymena thermophila*, *Reticulomyxa filosa*,
 825 *Thecamonas trahens* ATCC50062) and 13 photosynthetic eukaryotes (*Arabidopsis*
 826 *thaliana*, *Chlamydomonas reinhardtii*, *Coccomyxa subellipsoidea* C-169,
 827 *Klebsormidium nitens*, *Micromonas pusilla* CCMP1545, *Oryza sativa*, *Ostreococcus*
 828 *tauri*, *Physcomitrella patens*, *Selaginella moellendorffii*, *Chondrus crispus*,
 829 *Cyanidioschyzon merolae*, *Galdieria sulphuraria*, *Cyanophora paradoxa*). **b**,
 830 Annotations of 12 orthogroups conserved in primary photosynthetic eukaryotes, but not
 831 in predator eukaryotes. The functional annotation and the cellular localization were
 832 based on UniprotKB information of *Arabidopsis thaliana* orthologs, and manual
 833 annotations are represented in parenthesis. The estimated origins were based on the ML

phylogenetic analysis. Genes involved in light acclimation (cyan), nutrient auxotrophy (yellow) and organelle gene expression (magenta) are highlighted. **c** and **d**, ML phylogenetic trees of orthogroup genes acquired by HGT from other eukaryotes; Haptophyta (**c**, OG0001742) and Stramenopiles (**d**, OG0002023). The phylogenetic analysis of **c** and **d** were performed using a LG+G (**c**) and a LG+G+I model (**d**), respectively. Photosynthetic *Paulinella* species are boxed.

Extended Data Fig. 4. Analysis of nuclear-localized plastid DNA (nupDNA) and mitochondria DNA (numDNA) in *P. micropora*. **a**, Analysis of the junction sequences of nupDNA and numDNA using Illumina raw-reads sequences. 902,121,079 quality-trimmed HiSeq paired-end sequence reads (insert size: 300 b, 500 b) were analysed. **b**, numDNA sequence structures in the *P. micropora* draft genome. **c**, Distribution of numDNA sequences in mitochondrial genome positions. **d**, Nucleotide percent identity of numDNA compared with the mitochondria genome sequence. NumDNA directions are represented by right and left arrows which denote clockwise- and anti-clockwise directions of the mitochondria genome sequence (**c**), respectively.

850

851 **Extended Data Fig. 5. ML phylogenetic trees of four EGT candidates whose**
852 **counterpart orthologs are found in the plastid (chromatophore) genome of *P.***
853 ***micropora*.** The clades of the *Paulinella* plastid and cyanobacteria are highlighted in
854 green. Blue circle: *P. micropora* nuclear gene, red circle: *P. chromatophora* nuclear
855 gene, blue triangle: *P. micropora* plastid gene, red triangle: *P. chromatophora* plastid
856 gene, unmarked: cyanobacterium gene. The branch of *P. micropora* and *P.*
857 *chromatophora* nuclear genes supported by high bootstraps (numbers in the parenthesis)
858 is represented as a red line. The divergence time of *Paulinella* nuclear genes from the
859 plastid- or other cyanobacterial-genes (yellow stars) were estimated by RelTime
860 methods of MEGA6 using the divergent time of *P. micropora* and *P. chromatophora*
861 (45.7–64.7 MYA) (blue cross). Psak: Photosystem I subunit K, CcmL: CO₂
862 concentrating mechanism protein, Hlip: High light inducible protein. Brackets mean the
863 substitution model used in the phylogenetic analysis. Trees with species names are
864 available at the repository (<https://figshare.com/s/a665678c48d0af073894>).

865

Extended Data Fig. 6. Putative DNA virus fragments detected in the *P. micropora* draft genome. ORF structures, GC%, copy number and sequence gaps are represented as described in the legend to Fig. 3.

Extended Data Fig. 7. Characteristics of virus-type GPCRs of *P. micropora* and *P. chromatophora*. **a**, A ML phylogenetic tree of seven transmembrane domains of GPCRs. GPCRs in *P. micropora*, *P. chromatophora*, *P. brassicae* (CEO98393.1, CEO98395.1), and those detected in DNA viral fragments were analysed using a LG+G+F model. Branches with bootstrap values <50 are condensed. **b**, Existence of the transcript. **c**, GC% of the coding sequence of GPCR genes. **d**, Exon/intron structures of the genes. **e**, Protein domain structures. Protein domains were detected by CD search⁴⁸. GPCRs in putative DNA viruses are boxed with a blue line.

Extended Data Fig. 8. ML phylogenetic tree of DNA polymerases of Polintons. **a**, DNA polymerase sequences of *P. micropora* Polintons (No. 2766 and No. 3055), *R. filosa* Polintons (Genbank, ETO12468.1, ETO19856.1, ETO13433.1), metazoan

882 Polintons in Repbase³⁷ and Mavirus (Genbank, YP 004300281.1) were analysed with a
883 LG+G+I model. The metazoan and rhizarian groups are highlighted in green and yellow,
884 respectively.

885

886

887 **Supplementary information**

888 **Supplementary Table 1. Summary of the raw sequence data of the *P. micropora***
889 **genome.**

890

891 **Supplementary Table 2. Annotation and categorization of the *P. micropora***
892 **RepeatModeller sequences.** *The categorization is based on the sequence similarities
893 with the manually curated repeat sequences in Supplementary Table 3 by BLASTX
894 search (e-value < 1e⁻⁵).

895

896 **Supplementary Table 3. Annotation of the repeat elements manually identified**
897 **from the *P. micropora* draft genome.** *The copy number of the repeat elements was

estimated by BLASTN search against the *P. micropora* genome. Redundant BLASTN hits at the same genome locus were manually removed.

Supplementary Table 4. *P. micropora* MYN1 gene list.

Supplementary Table 5. GO-terms over- and under-represented in the *P. micropora* MYN1 genome compared with the genome of other rhizarian organisms (*B. natans*, *P. brassicae* and *R. filosa*). Significantly enriched GO-terms with Fisher's exact test p-value less than 0.01 are represented.

Supplementary Table 6. EGT and HGT candidates in *P. micropora* MYN1. Genes satisfying at least one of the following criteria were considered as EGT and HGT candidates. 1) No hits to eukaryote genes of the NCBI nr database by BLAST analysis* 2) EGT and HGT are supported by ML phylogenetic analysis with a high bootstrap value (≥ 95) except when the phylogenetically available protein alignment sequences were short (< 100 amino acids)** 3) *P. micropora* genes were embedded in the clade of

914 photosynthetic organisms in the phylogenetic tree. * BLASTP top 1000 hits (NCBI nr,
 915 e-value $< 1e^{-10}$) were classified. A: Archaea, B: Bacteria, E: Eukaryotes, V: Viruses, O:
 916 Others, U: Unknown. ** We lowered the threshold of the bootstrap value to 70 when
 917 the alignment sequence positions available for the phylogenetic analysis were less than
 918 100 amino acids.

919

920 **Supplementary Table 7. Detection of *P. micropora* Polintons by tBLASTn search**
 921 **using DNA polymerase domain sequences.** A tBLASTn search was performed against
 922 the *P. micropora* draft genome (e-value $< 1e^{-10}$). The draft genome sequences and the
 923 query sequences are available at the repository
 924 (<https://figshare.com/s/a665678c48d0af073894>).

925

926 **Supplementary Table 8. Accession numbers of *P. micropora* MYN1 genome**
 927 **sequences.**

928

929 **Supplementary Table 9. Parameters of the ML phylogenetic analysis by MEGA6.**

930

931 **Supplementary Table 10. Sequence data used in this study.**

932

933

934

935

936

937

938

939

940

941

942

943

944

945