Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of *Paulinella chromatophora*

Eva C. M. Nowack[a,b,1], Dana C. Price[c], Debashish Bhattacharya[d], Anna Singer[b], Michael Melkonian[e], and Arthur R. Grossman[a]

[a]Department of Plant Biology, Carnegie Institution for Science, Stanford, CA 94305; [b]Department of Biology, Heinrich-Heine-Universität Düsseldorf, 40225 Dusseldorf, Germany; [c]Department of Plant Biology and Pathology, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901; [d]Department of Ecology, Evolution and Natural Resources, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901; and [e]Biozentrum, Universität zu Köln, 50674 Koln, Germany

Edited by John M. Archibald, Dalhousie University, Halifax, Canada, and accepted by Editorial Board Member W. Ford Doolittle September 6, 2016 (received for review May 19, 2016)

Plastids, the photosynthetic organelles, originated >1 billion y ago via the endosymbiosis of a cyanobacterium. The resulting proliferation of primary producers fundamentally changed global ecology. Endosymbiotic gene transfer (EGT) from the intracellular cyanobacterium to the nucleus is widely recognized as a critical factor in the evolution of photosynthetic eukaryotes. The contribution of horizontal gene transfers (HGTs) from other bacteria to plastid establishment remains more controversial. A novel perspective on this issue is provided by the amoeba *Paulinella chromatophora*, which contains photosynthetic organelles (chromatophores) that are only 60–200 million years old. Chromatophore genome reduction entailed the loss of many biosynthetic pathways including those for numerous amino acids and cofactors. How the host cell compensates for these losses remains unknown, because the presence of bacteria in all available *P. chromatophora* cultures excluded elucidation of the full metabolic capacity and occurrence of HGT in this species. Here we generated a high-quality transcriptome and draft genome assembly from the first bacteria-free *P. chromatophora* culture to deduce rules that govern organelle integration into cellular metabolism. Our analyses revealed that nuclear and chromatophore gene inventories provide highly complementary functions. At least 229 nuclear genes were acquired via HGT from various bacteria, of which only 25% putatively arose through EGT from the chromatophore genome. Many HGT-derived bacterial genes encode proteins that fill gaps in critical chromatophore pathways/processes. Our results demonstrate a dominant role for HGT in compensating for organelle genome reduction and suggest that phagotrophy may be a major driver of HGT.

endosymbiosis | genome evolution | organellogenesis | horizontal gene transfer | coevolution

Plastids are photosynthetic organelles in algae and plants that originated >1 billion y ago in the protistan ancestor of the Archaeplastida (red, glaucophyte, and green algae plus plants) via the primary endosymbiosis of a β-cyanobacterium (1, 2). Subsequently, plastids spread through eukaryote–eukaryote (i.e., secondary and tertiary) endosymbioses to other algal groups (3). The resulting proliferation of primary producers fundamentally changed our planet's history, allowing for the establishment of human populations. Plastid evolution was accompanied by a massive size reduction of the endosymbiont genome and the transfer of thousands of endosymbiont genes into the host nuclear genome, a process known as endosymbiotic gene transfer (EGT) (4). Proteins encoded by the transferred genes are synthesized in the cytoplasm and many are posttranslationally translocated into the plastid through the TIC/TOC protein import complex (5). EGT is widely recognized as a major contributor to the evolution of eukaryotes, and in particular the transformation of an endosymbiont into an organelle. More recently, it was proposed that horizontal gene transfers (HGTs) from cooccurring intracellular bacteria also supplied genes that facilitated plastid establishment (6). However, the extent and sources of HGTs and their importance to organelle evolution remain controversial topics (7, 8).

The chromatophore of the cercozoan amoeba *Paulinella chromatophora* (Rhizaria) represents the only known case of acquisition of a photosynthetic organelle other than the primary endosymbiosis that gave rise to the Archaeplastida (9). The chromatophore originated much more recently than plastids (~60–200 Ma) via the uptake of an α-cyanobacterial endosymbiont related to *Synechococcus*/*Cyanobium* spp. (9, 10). In contrast to heterotrophic *Paulinella* species that feed on bacteria, their phototrophic sister, *P. chromatophora*, lost its phagotrophic ability and relies primarily on photosynthetic carbon fixation for survival (11, 12). The chromatophore genome is reduced to 1 Mbp, approximately one-third the size of the ancestral cyanobacterial genome. Genome reduction was accompanied by the

Significance

Eukaryotic photosynthetic organelles (plastids) originated >1 billion y ago via the endosymbiosis of a β-cyanobacterium. The resulting proliferation of primary producers fundamentally changed our planet's history, allowing for the establishment of human populations. Early stages of plastid integration, however, remain poorly understood, including the role of horizontal gene transfer from nonendosymbiotic bacteria. Rules governing organellogenesis are difficult, if not impossible, to evaluate using the highly derived algal and plant systems. Insights into this issue are provided by the amoeba *Paulinella chromatophora*, which contains more recently established photosynthetic organelles of α-cyanobacterial origin. Here we show that the impact of Muller's ratchet that leads to endosymbiont genome reduction seems to drive the fixation of horizontally acquired "compensatory" bacterial genes in the host nuclear genome.

Author contributions: E.C.M.N., D.B., and A.R.G. designed research; E.C.M.N. and D.C.P. performed research; M.M. helped to establish the axenic *Paulinella* culture; E.C.M.N., D.C.P., D.B., A.S., and A.R.G. analyzed data; and E.C.M.N., D.C.P., D.B., and A.R.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. J.M.A. is a Guest Editor invited by the Editorial Board.

Data deposition: All sequence and assembly data generated in this project can be accessed via NCBI BioProject PRJNA311736. Sequence raw data reported in this paper have been deposited in the NCBI Sequence Read Archive [accession nos. SRX1624577 (cDNA reads) and SRX1624478, SRX1624478, and SRX1624515 (gDNA reads)].

[1]To whom correspondence should be addressed. Email: E.Nowack@uni-duesseldorf.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1608016113/-/DCSupplemental.

complete loss of many biosynthetic pathways, including those for various amino acids and cofactors. In other pathways, genes for single metabolic enzymes were lost (13). How the host compensates for the loss of metabolic functions from the chromatophore remains unknown. Previous studies identified >30 nuclear genes of α-cyanobacterial origin that were likely acquired via EGT from the chromatophore (14–16). However, most of these genes encoded functions related to photosynthesis and light adaptation and do not seem to complement gaps in chromatophore-encoded metabolic pathways. Three EGT-derived genes that encode the photosystem I (PSI) subunits PsaE, PsaK1, and PsaK2 were shown to be synthesized on cytoplasmic ribosomes and traffic (likely via the Golgi) into the chromatophore, where they assemble with chromatophore-encoded PSI subunits (17). Even though details of the protein translocation mechanism remain to be elucidated, these findings demonstrate that cytoplasmically synthesized proteins can be imported into chromatophores. Owing to the large number of bacteria associated with *P. chromatophora* in all available laboratory cultures, the full metabolic capacity of *P. chromatophora* is unknown and the occurrence of HGTs remains uncertain because of the inability to distinguish genes from contaminating bacteria from true HGT.

Results and Discussion

Transcriptome and Genome Datasets from Axenic *P. chromatophora*. To deduce the rules that govern organelle integration into cellular metabolism, we focused on exploring the extent of HGT in *P. chromatophora* and the putative functions of proteins derived from HGT. For this purpose, we established a bacteria-free (i.e., axenic) culture of *P. chromatophora*. These cells were used to generate the transcriptome and genome data discussed here. The *P. chromatophora* transcriptome dataset comprises 49.5 Mbp of assembled sequence with a contig N50 of 1.1 kbp. These contigs encode homologs of 442/458 (97%) of the core eukaryotic proteins in the Core Eukaryotic Genes Mapping Approach (CEGMA) database (16). Preliminary analyses indicate that the nuclear genome has a surprisingly large estimated size of ~9.6 Gbp (Fig. S1 and *Materials and Methods*). Thus, despite generating 147.4 Gbp of data from paired-end and mate-pair libraries (*Materials and Methods*), our initial assembly remained highly fragmented (N50 of 711 bp). All contigs >15 kbp in size were chromatophore- or mitochondrion-derived sequences. A potentially circular contig of 47.4 kbp with an average read coverage of 12,903× (0.82% of total genomic mapped reads) was identified as the complete, or nearly complete, *P. chromatophora* mitochondrial genome (Fig. S2). This contig contains 22 protein-coding genes, 27 tRNAs, and two (large + small) ribosomal RNA subunits.

Chromatophore and Host Genomes Encode Complementary Functions. Metabolic reconstruction of the amoeba gene inventory revealed the presence of genes for many metabolic pathways on the nuclear genome that were originally also present on, but then lost from, the chromatophore genome (e.g., Met, Ser, Gly, and purine biosynthesis; Fig. 1*A* and Figs. S3 and S4). In other instances, gaps in chromatophore-encoded pathways are filled by proteins encoded on the nuclear genome (e.g., Arg, His, and aromatic amino acid biosynthesis; Fig. 1*B* and Fig. S3). Interestingly, chromatophore genome reduction also involved the loss of genes essential for bacteria-specific functions that cannot be replaced by eukaryotic genes. One such

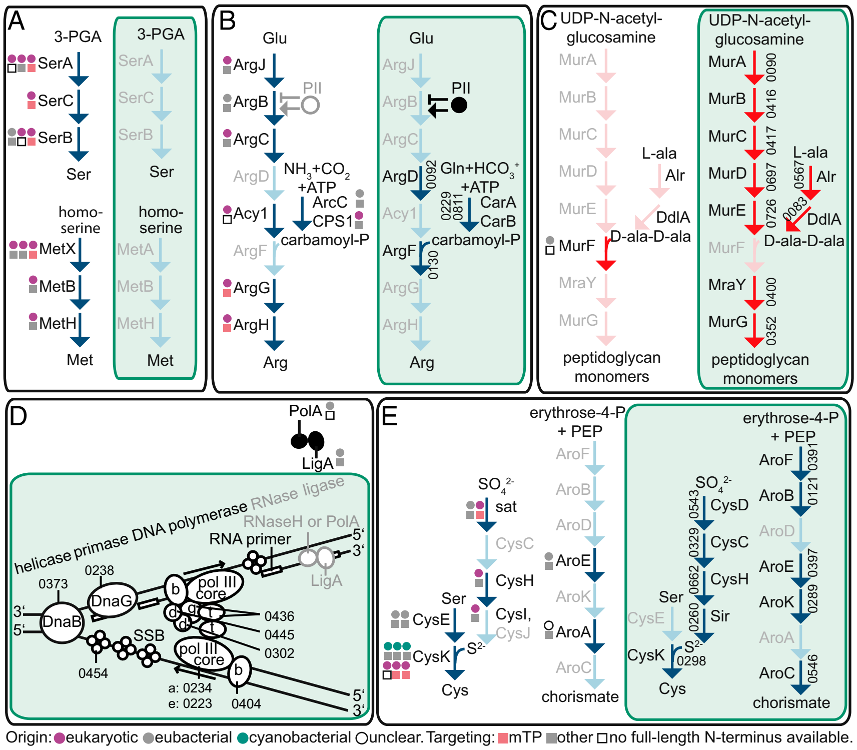

Fig. 1. Metabolic pathways and DNA replication in *P. chromatophora*. The distribution of chromatophore-encoded (within green rectangles) and nuclear-encoded genes is shown, although the subcellular localization of the gene products is unknown. Numbers associated with chromatophore-encoded enzymes are locus tags for the respective genes (e.g., 1234 represents PCC_1234). Pale lettering/arrows indicate that the gene is missing from the chromatophore genome or absent in nuclear transcriptome data. Circles and rectangles adjacent to the enzymes indicate their phylogenetic origin and targeting prediction (TargetP prediction; mTP and SP predictions with a reliability class <3 are shown), respectively; they are defined immediately below the figure. Multiples of the individual symbols represent the presence of multiple protein versions encoded by the transcript dataset. 3-PGA, 3-phosphoglycerate; PII, the PII nitrogen-sensing protein (see text); PEP, phosphoenolpyruvate; SSB, single-strand binding protein. The pathways shown are for the synthesis of serine and methionine (Ser, Met, *A*), arginine (Arg, *B*), peptidoglycan (*C*) and the precursor of aromatic amino acids (chorismate) and cysteine (Cys, *E*) as well as for DNA replication (*D*).

"lost" gene encodes UDP-N-acetylmuramoyl-tripeptide:D-Ala-D-Ala ligase (MurF), which ligates the dipeptide D-Ala-D-Ala to the growing peptide side chain of peptidoglycan monomers (Fig. 1*C*). All remaining steps of peptidoglycan biosynthesis are encoded on the chromatophore genome. Intriguingly, analysis of the *P. chromatophora* transcriptome dataset revealed the presence of a nuclear-encoded MurF of β-proteobacterial origin (Figs. 1*C* and 2*A*).

Predominance of HGT in the Evolution of *P. chromatophora*. The finding that a β-proteobacterial MurF was encoded on the *P. chromatophora* nuclear genome prompted us to search for additional bacterial genes on this genome. Based on phylogenetic analysis of proteins encoded by *P. chromatophora* nuclear transcripts, there are at least 150 independent bacterial gene acquisitions that are often followed by gene family expansions, resulting

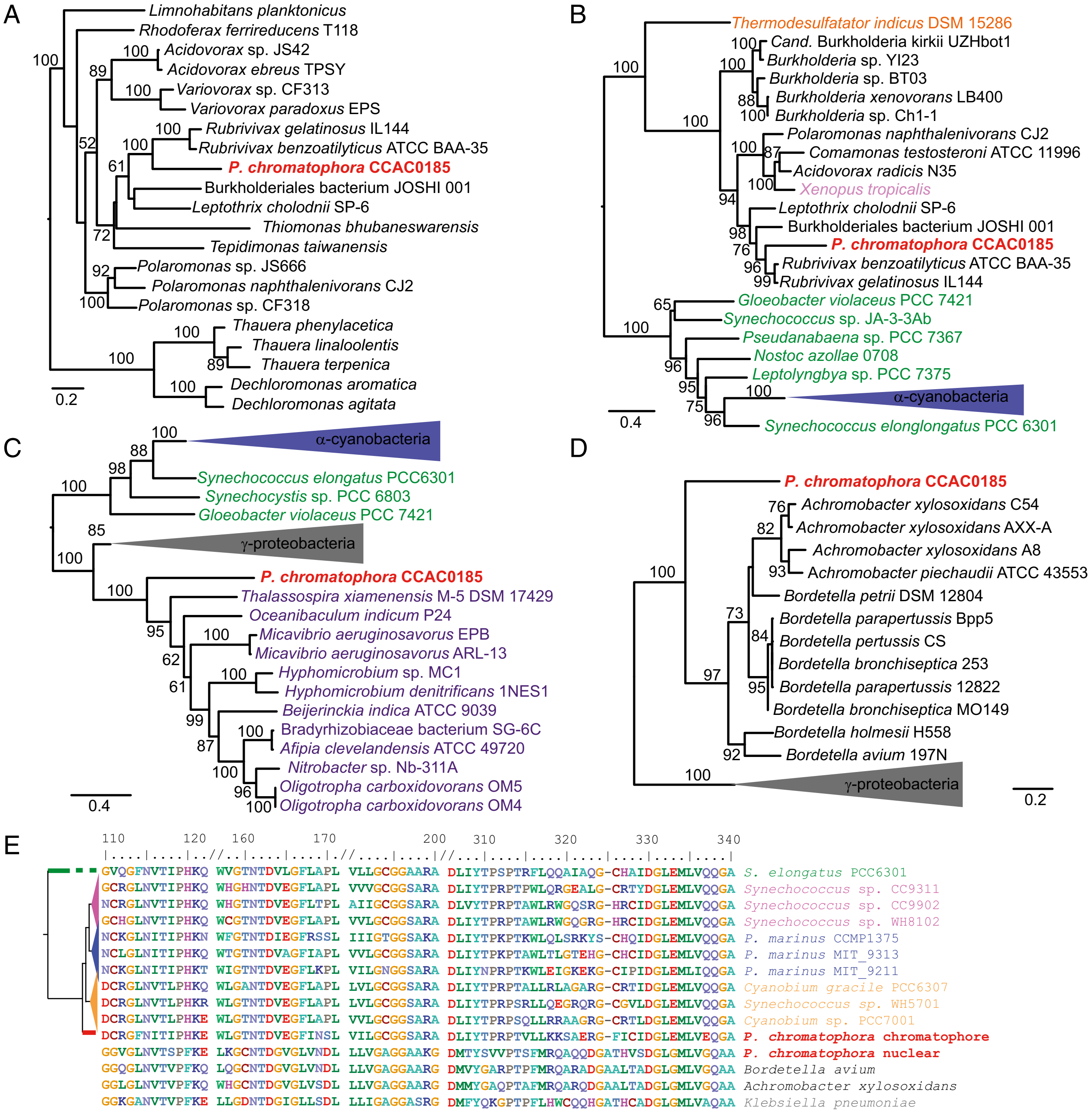
A
Limnohabitans planktonicus
Rhodoferax ferrireducens T118
Acidovorax sp. JS42
Acidovorax ebreus TPSY
Variovorax sp. CF313
Variovorax paradoxus EPS
Rubrivivax gelatinosus IL144
Rubrivivax benzoatilyticus ATCC BAA-35
P. chromatophora CCAC0185
Burkholderiales bacterium JOSHI 001
Leptothrix cholodnii SP-6
Thiomonas bhubaneswarensis
Tepidimonas taiwanensis
Polaromonas sp. JS666
Polaromonas naphthalenivorans CJ2
Polaromonas sp. CF318
Thauera phenylacetica
Thauera linaloolentis
Thauera terpenica
Dechloromonas aromatica
Dechloromonas agitata
100
89
100
100
100
52
100
61
72
92
100
100
100
89
100
0.2
B
Thermodesulfatator indicus DSM 15286
Cand. Burkholderia kirkii UZHbot1
Burkholderia sp. YI23
Burkholderia sp. BT03
Burkholderia xenovorans LB400
Burkholderia sp. Ch1-1
Polaromonas naphthalenivorans CJ2
Comamonas testosteroni ATCC 11996
Acidovorax radicis N35
Xenopus tropicalis
Leptothrix cholodnii SP-6
Burkholderiales bacterium JOSHI 001
P. chromatophora CCAC0185
Rubrivivax benzoatilyticus ATCC BAA-35
Rubrivivax gelatinosus IL144
Gloeobacter violaceus PCC 7421
Synechococcus sp. JA-3-3Ab
Pseudanabaena sp. PCC 7367
Nostoc azollae 0708
Leptolyngbya sp. PCC 7375
α-cyanobacteria
Synechococcus elonglongatus PCC 6301
100
100
100
100
88
100
100
87
100
94
98
76
96
99
65
100
96
95
75
100
96
0.4
C
α-cyanobacteria
Synechococcus elongatus PCC6301
Synechocystis sp. PCC 6803
Gloeobacter violaceus PCC 7421
γ-proteobacteria
P. chromatophora CCAC0185
Thalassospira xiamenensis M-5 DSM 17429
Oceanibaculum indicum P24
Micavibrio aeruginosavorus EPB
Micavibrio aeruginosavorus ARL-13
Hyphomicrobium sp. MC1
Hyphomicrobium denitrificans 1NES1
Beijerinckia indica ATCC 9039
Bradyrhizobiaceae bacterium SG-6C
Afipia clevelandensis ATCC 49720
Nitrobacter sp. Nb-311A
Oligotropha carboxidovorans OM5
Oligotropha carboxidovorans OM4
100
88
98
100
85
100
100
95
100
62
61
100
99
87
100
100
96
100
0.4
D
P. chromatophora CCAC0185
Achromobacter xylosoxidans C54
Achromobacter xylosoxidans AXX-A
Achromobacter xylosoxidans A8
Achromobacter piechaudii ATCC 43553
Bordetella petrii DSM 12804
Bordetella parapertussis Bpp5
Bordetella pertussis CS
Bordetella bronchiseptica 253
Bordetella parapertussis 12822
Bordetella bronchiseptica MO149
Bordetella holmesii H558
Bordetella avium 197N
γ-proteobacteria
76
82
93
100
73
84
95
97
92
100
0.2
E
110 120 160 170 200 310 320 330 340
|....|....|..//..|....|....|.//....|....|//...|....|....|....|....|....|
GVQGFNVTIPHKQ WVGTNTDVLGFLAPL VLLGCGGAARA DLIYTPSPTRFLQQAIAQG-CHAIDGLEMLVQQGA S. elongatus PCC6301
GCRGLNVTIPHKQ WHGHNTDVEGFLAPL VVVGCGGSARA DLIYTPRPTPWLQRGEALG-CRTYDGLEMLVQQGA Synechococcus sp. CC9311
NCRGLNVTIPHKQ WHGTNTDVEGFLTPL AIIGCGGSARA DLVYTPRPTAWLRWGQSRG-HRCIDGLEMLVQQGA Synechococcus sp. CC9902
GCHGLNVTIPHKQ WCGTNTDVEGFLAPL VVLGCGGSARA DLIYTPRPTAWLRWGQGRG-HRCIDGLEMLVQQGA Synechococcus sp. WH8102
NCKGLNITIPHKN WFGTNTDIEGFRSSL IIIGTGGSAKA DLIYTPKPTKWLQLSRKYS-CHQIDGLEMLVQQGA P. marinus CCMP1375
NCLGLNITIPHKQ WTGTNTDVAGFIAPL VVLGCGGSARA DLIYTPKPTAWLTLGTEHG-CHCIDGLEMLVQQGA P. marinus MIT_9313
NCKGLNITIPHKT WIGTNTDVEGFLKPL VILGNGGSARA DLIYNPRPTKWLEIGKEKG-CIPIDGLEMLIQQGA P. marinus MIT_9211
DCRGLNVTLPHKQ WLGANTDVEGFLAPL LVLGCGGSARA DLIYTPRPTALLRLAGARG-CRTIDGLEMLVQQGA Cyanobium gracile PCC6307
DCRGLNVTLPHKR WLGTNTDVEGFLAPL LVLGCGGSARA DLIYTPRPSRLLQEGRQRG-CGVLDGLEMLVQQGA Synechococcus sp. WH5701
DCRGLNVTLPHKE WLGTNTDVEGFLAPL VVLGCGGSARA DLIYTPRPSQLLRRAAGRG-CRTLDGLEMLVQQGA Cyanobium sp. PCC7001
DCRGFNITIPHKE WLGTNTDVEGFINSL VILGCGGSARA DLIYTPRPTVLLKKSAERG-FICIDGLEMLVEQGA P. chromatophora chromatophore
GGVGLNVTSPFKE LKGCNTDGVGLVNDL LLVGAGGAAKG DMTYSVVPTSFMRQAQQDGATHVSDGLGMLVGQAA P. chromatophora nuclear
GGQGLNVTVPFKQ LQGCNTDGVGLVNDL LLVGAGGAARG DMVYGARPTPFMRQARADGAALTADGLGMLVGQAA Bordetella avium
GGLGLNVTVPFKQ WHGCNTDGVGLVSDL LLVGAGGAARG DMMYGAQPTAFMRQARQDGAALTADGLGMLVGQAA Achromobacter xylosoxidans
GGKGANVTVPFKE LLGDNTDGIGLLSDL LLIGAGGASRG DMFYQKGPTPFLHWCQQHGATHCADGLGMLVAQAA Klebsiella pneumoniae

Fig. 2. Phylogeny of HGT-derived genes in *P. chromatophora*. Maximum likelihood phylogenetic trees from amino acid alignments of (*A*) MurF, (*B*) PolA, (*C*) LigA, and (*D*) AroE. Numbers at the branches represent bootstrap values. Color code: purple, α-; black, β-; and gray, γ-proteobacteria; blue, α-; and green, β-cyanobacteria; orange, thermodesulfobacteria; pink, Eukarya; and red, *P. chromatophora*. (*E*) Portion of amino acid alignment of nuclear and chromatophore-encoded copies of *P. chromatophora* AroE with proteobacterial and cyanobacterial sequences. The tree (left) represents "species" phylogeny based on the ribosomal operon. The lineages are marked as follows: green, *S. elongatus*; pink, marine *Synechococcus* clade; blue, *Prochlorococcus* clade; orange, *Cyanobium* clade; red, *P. chromatophora* (nuclear and chromatophore genes); black, β-; and gray, γ-proteobacteria.

in at least 229 bacterium-derived genes (*Materials and Methods* and Dataset S1). Only 58 (or 25%) of these genes are of α-cyanobacterial origin, and thus potentially chromatophore-derived, although we cannot exclude the possibility that some may also have arisen via HGT from related cyanobacterial lineages. Most of the remaining 171 HGTs are affiliated with other bacteria, with 64 being confidently assigned to a specific donor bacterial lineage and two for which an HGT or EGT origin could not be unambiguously determined (Fig. S5 *A* and *B* and Dataset S1). For 52 other genes there was not sufficient bootstrap support (i.e., ≥80%) to establish affiliation with a particular bacterial clade, or the sequences originated at the base of a particular lineage, indicating a likely donor group, but with lower confidence. The remaining 53 bacterial genes could not be assigned to a specific clade due to frequent HGTs among these taxa. Nonetheless, these latter genes likely arose via HGT because similar genes are absent in other eukaryotes or α-cyanobacteria (Fig. S5 *A* and *B* and Dataset S1). Therefore, our results suggest a predominance of HGT over EGT in the evolution of the *P. chromatophora* photosynthetic lineage. We hypothesize that this result is explained by the fact that *P. chromatophora* has a phagotrophic ancestry that facilitated the HGT ratchet. Analysis of a partial nuclear genome sequence from wild-caught cells of *Paulinella ovalis*, a phagotrophic sister lineage of *P. chromatophora*, revealed the presence of various bacterial DNA sequences that were likely derived from food vacuoles (18). This partial genome dataset also revealed nuclear genes of α-cyanobacterial origin (e.g., a diaminopimelate epimerase gene), suggesting that in addition to EGT phagotrophy can lead to HGT in the *Paulinella* lineage, as previously hypothesized (19). These results can also be the consequence of the uptake of DNA from the environment by transformation or by viral transduction.

Spliced Leader Sequences and Introns Confirm Nuclear Origin of HGT Genes. Validation of the nuclear origin of *P. chromatophora* HGT candidates is provided by the presence of a conserved 20-nt transspliced leader (SL) sequence on many of these transcripts. The biological function of SLs is not well understood but they are found at the 5′ terminus of mature mRNAs in a phylogenetically diverse group of organisms including euglenozoans, cnidarians, chordates, nematodes, and dinoflagellates (20), but to our knowledge they have not previously been reported from Rhizaria. Of the 17,801 unique nuclear transcripts with an assigned function, 4,649 (26.1%) contained the SL sequence CGGATAWTCCKGCTTTTCTG or a 5′-truncated version of this sequence (but at least CTTTTCTG) within the first 40 nt and usually at the 5′ terminus (Fig. S6). Because RNA sequencing generally results in poor assembly at the 5′ends of transcripts, we expect that the actual fraction of transcripts carrying a SL at their 5′ end is much higher. As expected, SLs were absent from all chromatophore- and mitochondrion-derived transcripts. Of the presumed HGT-derived cDNA contigs, 32% contained an SL (Dataset S1). For the other presumed HGT-derived transcripts, we searched for spliceosomal introns in the corresponding genomic contigs. Using both of these approaches we were able to confirm the nuclear origin for 162 of the 171 genes derived via HGT (Dataset S1).

Chromatophore-Related Functions of HGT Genes. Adaptive HGTs from bacteria have been reported from diverse eukaryotic lineages (e.g., refs. 21–25). Thus, it is likely that some HGT candidates represent ancient transfers to the nuclear genome that are not related to chromatophore function. However, none of the *P. chromatophora* HGTs was present in the partial *P. ovalis* dataset, and many encode proteins that fill specific gaps in chromatophore-encoded metabolic pathways [e.g., D-Ala-D-Ala ligase MurF (Figs. 1*C* and 2*A*), a DNA polymerase I (PolA) responsible for removal of RNA primers and filling in the resulting gaps during DNA replication, a DNA ligase (LigA) that seals DNA nicks (Figs. 1*D* and 2 *B* and *C*), and a serine O-acetyltransferase CysE (Fig. 1*E*)]. Eight bacterial genes of non-α-cyanobacterial provenance function in bacterial cell wall biosynthesis or division, whereas 25 are associated with the processing of genetic information. Twelve HGTs encode transporters that might facilitate metabolite or ion exchange between the chromatophore and the *P. chromatophora* cytoplasm (Fig. S5 *C* and *D* and Dataset S1). For example, a gene encoding a putative Gly/Ala Na^+ symporter may be involved in shuttling cytoplasmically synthesized Gly and Ala into the chromatophore, which lacks genes encoding the pathways for Gly and Ala biosynthesis (Fig. S3).

A lack of biochemical data makes it impossible to predict the subcellular localization of nuclear-encoded, and in particular, HGT-derived proteins. However, the functional complementarity of nuclear and chromatophore-encoded proteins provides a reasonable basis for our speculation that, similar to the case of the EGT-derived photosynthesis polypeptides PsaE and PsaK, nuclear-encoded HGT-derived proteins are imported into the chromatophore to rescue lost gene functions. In this context it is interesting that a highly conserved *glnB* gene is present on the chromatophore genome (Fig. 1*B* and Fig. S7). This gene encodes the PII nitrogen-sensing protein that regulates arginine biosynthesis through interactions with the N-acetyl glutamate kinase (ArgB) (26), which is encoded on the nuclear genome and derived via HGT from a planctomycete donor. For transcripts that included the full-length N terminus of the encoded protein, as indicated by either the presence of an SL sequence or an in-frame stop codon upstream of the presumable start methionine, the occurrence of potential N-terminal targeting sequences was analyzed using TargetP 1.1 in nonplant mode (27) (Fig. 1 and Figs. S3 and S4). For the enzymes that catalyze the first steps in the arginine biosynthetic pathway (ArgJ, ArgB, and ArgC), as for PsaE and PsaK (17), no N-terminal presequences were predicted. For the last two enzymes of the arginine biosynthetic pathway, ArgG and ArgH, a mitochondrial targeting peptide (mTP) was predicted. TargetP predictions of mTPs and signal peptides (SPs) seem accurate for *P. chromatophora* based on the finding that most enzymes of the TCA cycle and typical ER proteins yield high confidence mTP or SP predictions, respectively (Table S1). Thus, it is likely that some HGT-derived proteins not predicted to contain an mTP or SP are targeted to the chromatophores where they replace lost functions, or play a role in host/chromatophore metabolic integration. However, in other cases the proteins for a given metabolic pathway may be partitioned between the cytoplasm and chromatophore and the connectivity of the pathway established by metabolite exchange between the two compartments.

Presymbiotic Interbacterial HGT vs. Postsymbiotic Bacteria to Eukaryote HGT. Eukaryotic genomes are widely known to contain many genes of bacterial origin (28) that are usually attributed to mitochondrion and plastid endosymbiosis. The diverse phylogenetic origins of these bacterial genes is explained by the fluid genome composition of prokaryotes (8) that resulted in chimeric (presymbiotic) genomes in the donor lineages (29, 30). The observed bursts in HGT frequency that coincide with organelle acquisition in eukaryotes support this interpretation (8). Is this the case in *P. chromatophora*? Do the many bacterial gene transfers we observed have their origins in a highly chimeric α-cyanobacterial genome of the endosymbiont? Alternatively, did these foreign genes arise via EGT from the existing mitochondrial endosymbiont? The second explanation can be largely excluded on two counts: (*i*) The *P. chromatophora* HGT candidates are not found in other eukaryotes, all of which share the same mitochondrion, and (*ii*) many of these HGTs seem to specifically fill gaps in chromatophore pathways. Therefore, it is not reasonable to assume that a mitochondrion-derived *murF* gene was maintained over hundreds of millions of years even though there was no need for peptidoglycan synthesis.

To evaluate the first, more intriguing, scenario we used phylogenomics to determine how many of the 867 protein-coding genes still retained on the chromatophore genome had an HGT (i.e., non-α-cyanobacterial) origin. This analysis demonstrated that 848/867 (97.8%) of the chromatophore-encoded genes are placed unambiguously as sister to, or nested within, the α-cyanobacteria group and therefore are not the result of interphylum HGT. There is a single gene (PCC_0175, a YGGT family membrane protein) for which a noncyanobacterial origin is supported by a bootstrap value ≥80%. This implies that if lineage-specific HGT-derived genes were present in the chromatophore ancestor, they primarily had nonessential functions that did not survive endosymbiosis. Consistent with these findings is the observation that although cyanobacterial genomes are well known to undergo frequent HGTs (29, 31) (*i*) HGT rates in the *Prochlorococcus*/*Synechococcus* clade are the lowest among cyanobacteria (29) and (*ii*) a detailed comparative genomic study of *Prochlorococcus* spp. and marine *Synechococcus* spp. revealed a core set of 1,273 genes present in 12 *Prochlorococcus* species (32). Genes in the core genome encode essential functions including enzymes involved in central carbon metabolism and amino acid and chlorophyll biosynthesis. The larger, less widely distributed component of this pan-genome encodes functions that may relate to niche specificity and that are nonessential under optimal growth conditions. Genes such as *argB*, *murF*, *polA*, *cysE*, and *aroE* (discussed here; Fig. 1) are part of the core genome and are present in all 12 *Prochlorococcus* and 4 *Synechococcus* strains analyzed (32). In addition, the first β-cyanobacterium branching outside of the α-cyanobacteria, *Synechococcus elongatus,* contains the cyanobacterial version of these genes (Fig. 2*E* and Fig. S8), suggesting that the ancestor of the chromatophore also encoded cyanobacterial homologs of these genes. These results support our hypothesis that the *P. chromatophora* host cell acquired the many bacterial genes that we identified primarily through postendosymbiotic HGT, and not EGT from a highly chimeric endosymbiont genome. Finally, we note that insects such as mealybugs that harbor nutritional, bacterial endosymbionts with highly reduced genomes have also gained bacterial genes through HGT. Similar to the situation observed for *P. chromatophora*, the insect HGT-derived genes seem to complement functions lost from the symbiont genome (33).

Intermediates in the Replacement Process. To further test the hypothesis that HGT into the nuclear genome can replace chromatophore genes, we searched for potential intermediates in the replacement process and found full-length chromatophore-encoded genes with bacterial homologs present in the nuclear genome (both copies transcribed). Examples are the shikimate dehydrogenase AroE (Figs. 1*E* and 2 *D* and *E*), an inositol monophosphatase, and the elongation factor leader peptidase A (LepA) (Dataset S1). This potential intermediate replacement state was also identified for EGT-derived genes (15). Once the introduced nuclear gene attains targeting capabilities, the copy of the gene fixed or lost in each case of "gene duplication" via HGT or EGT cannot be predicted with confidence. However, the gene transfer ratchet model described by Doolittle (19) (and our data) predicts that over evolutionary time an increasing number of organelle genes will be lost in favor of nuclear copies. Additional genome data from phagotrophic *Paulinella* species will provide insights into which HGT-derived genes predate endosymbiosis and which may be associated with organelle evolution.

Conclusion

Whereas most eukaryotic genes are vertically inherited, data are accumulating of widespread HGT in eukaryotes that is tied to adaptation (28). The uptake of a bacterial endosymbiont represents a profound change in lifestyle that requires recalibration of the host genetic repertoire, a need that can be partially met via HGT. In addition, the impact of Muller's ratchet that leads to endosymbiont genome reduction seems to drive the fixation of horizontally acquired "compensatory" bacterial genes in the host genome. Thus, similar to EGT, HGT-derived genes may facilitate integration of the endosymbiont by providing the host with transcriptional/translational control over chromatophore metabolic functions, metabolite fluxes between the cytoplasm and chromatophore, and the processing of genetic information. Therefore, like EGT, HGT establishes key connections that enable the host to coordinate host–chromatophore metabolism, growth, and proliferation. We hypothesize that in *P. chromatophora* phagotrophy was initially maintained during chromatophore integration (Fig. 3), with the mixotrophic lifestyle setting the stage for a gene transfer ratchet that facilitated organelle integration by enabling replacement of chromatophore genes with genes derived from either EGT or HGT, in addition to the repurposing of host-derived genes. This is consistent with the observed bursts in HGT frequencies coincident with plastid and mitochondrion acquisition (8).

Materials and Methods

Cultivation of *P. chromatophora* and Generation of Axenic Culture. *P. chromatophora* CCAC0185 was grown as described previously (17). To generate an axenic culture *P. chromatophora* cells were sprayed onto nutrient agar plates. The axenic culture was obtained from a single bacteria-free *P. chromatophora* cell recovered from these plates (for details see *SI Materials and Methods*).

Generation of Sequencing Libraries and Assemblies. Genomic DNA (gDNA) and cDNA derived from axenic *P. chromatophora* cultures were subjected to Illumina and Nextera library generation and sequencing, resulting in

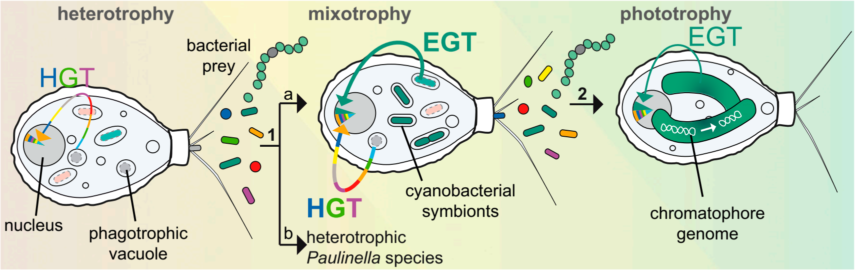

Fig. 3. Evolution of phototrophy from a phagotrophic ancestor in the *Paulinella* clade. In step 1a a mixotrophic cell evolved by maintaining a α-cyanobacterial endosymbiont and exploiting its photosynthetic ability. Over time, the host targeted proteins to the symbiont and inserted membrane transporters to gain control over symbiont growth and division, leading to vertical inheritance of the nascent organelle. Step 1b indicates heterotrophic *Paulinella* species that did not acquire permanent endosymbionts. In step 2 efficient metabolite exchange led to loss of phagotrophy and relaxed functional constraint on many chromatophore genes, leading to massive chromatophore genome reduction. Colored sections represent HGT (multicolor) and EGT (green) components of the nuclear genome; arrow thickness represents prevalence of the particular gene transfer type during different evolutionary stages.

147.4 Gbp gDNA and 4.9 Gbp cDNA raw sequence data. Genome and transcriptome assemblies were generated as detailed in *SI Materials and Methods*. All sequence and assembly data generated in this project can be accessed via NCBI BioProject PRJNA311736.

Estimation of Genome Size. Mapping the Illumina HiSeq data against gDNA contigs that encode a complete or partial CEGMA core eukaryote protein ($n = 78$) resulted in an average coverage of 10.05×. With a total amount of 96.2 Gbp of HiSeq data, we arrived at a genome size estimation of 9.57 Gbp. Mapping the MiSeq data separately (1.81× average coverage of contigs; 17.4 Gbp of data) yields an extremely close estimate of 9.61 Gbp. Estimation of the genome size through *k*-mer counts arrived at a similar result (11.45 Gbp). A histogram of *k*-mer frequency (number of times a particular 31-mer was observed) vs. probability approaches an exponential distribution (Fig. S1), as opposed to an expected normal distribution. This pattern indicates that despite generating 113.6 Gbp of genome data, our sequence coverage was derived predominantly from unique or nonoverlapping DNA amplicons and evidences low-coverage sequencing of an extremely large genome. For more detail see *SI Materials and Methods*.

Phylogenomic Pipeline and Screening for HGT from Bacteria. In brief, an initial phylogenomic analysis was performed as follows. Predicted proteins were queried via BLASTp (*e*-value $\leq 1 \times 10^{-5}$) (34) against a local protein database. A maximum of 12 species from each taxonomic phylum was selected in descending order of blast bitscore from the results to a maximum of 150 total species, and the respective sequences were aligned. Maximum-likelihood phylogenies were generated using RAxML v. 8.2 (35) with 100 bootstrap replicates under the LG+G model. The resulting trees were screened for *P. chromatophora* + prokaryote monophyly with bootstrap support of ≥70% or for trees containing, besides the *P. chromatophora* sequence, sequences solely from prokaryotes. After this initial screening, contigs of potential bacterial origin were manually curated. Curated alignments were then subjected to a second phylogenetic analysis using IQTREE (36) with 2,000 ultrafast bootstrap replicates and automatic model selection. Phylogenetic trees and protein alignments are available at cyanophora.rutgers.edu/paulinella. For more detail see *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank John Coller and Ji Xuhuai (Stanford Functional Genomics Facility), Jane Grimwood and Jeremy Schmutz (HudsonAlpha) for the generation of Illumina sequence data, Karl Forchhammer for advice on the Arg biosynthetic pathway, and Rajat Roy for advice regarding genome size estimation. This study was supported by National Science Foundation Grants MCB-10370 (to A.R.G.) and EF 08-27023 and OCE 11-29203 (to D.B.) and Deutsche Forschungsgemeinschaft Grant NO 9010/1-1 (to E.C.M.N.).

1. Falkowski PG, et al. (2004) The evolution of modern eukaryotic phytoplankton. *Science* 305(5682):354–360.
2. Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D (2004) A molecular timeline for the origin of photosynthetic eukaryotes. *Mol Biol Evol* 21(5):809–818.
3. Gould SB, Waller RF, McFadden GI (2008) Plastid evolution. *Annu Rev Plant Biol* 59: 491–517.
4. Martin W, Herrmann RG (1998) Gene transfer from organelles to the nucleus: How much, what happens, and Why? *Plant Physiol* 118(1):9–17.
5. Schleiff E, Becker T (2011) Common ground for protein translocation: Access control for mitochondria and chloroplasts. *Nat Rev Mol Cell Biol* 12(1):48–59.
6. Ball SG, et al. (2013) Metabolic effectors secreted by bacterial pathogens: Essential facilitators of plastid endosymbiosis? *Plant Cell* 25(1):7–21.
7. Archibald JM (2015) Evolution: Gene transfer in complex cells. *Nature* 524(7566): 423–424.
8. Ku C, et al. (2015) Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* 524(7566):427–432.
9. Marin B, Nowack ECM, Melkonian M (2005) A plastid in the making: Evidence for a second primary endosymbiosis. *Protist* 156(4):425–432.
10. Nowack ECM (2014) *Paulinella chromatophora*—Rethinking the transition from endosymbiont to organelle. *Acta Soc Bot Pol* 83(4):387–397.
11. Kies L (1974) [Electron microscopical investigations on *Paulinella chromatophora* Lauterborn, a thecamoeba containing blue-green endosymbionts (Cyanelles) (author's transl)]. *Protoplasma* 80(1):69–89.
12. Kies L, Kremer BP (1979) Function of cyanelles in the thecamoeba *Paulinella chromatophora*. *Naturwissenschaften* 66(11):578–579.
13. Nowack ECM, Melkonian M, Glöckner G (2008) Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. *Curr Biol* 18(6): 410–418.
14. Nakayama T, Ishida K (2009) Another acquisition of a primary photosynthetic organelle is underway in *Paulinella chromatophora*. *Curr Biol* 19(7):R284–R285.
15. Nowack ECM, et al. (2011) Endosymbiotic gene transfer and transcriptional regulation of transferred genes in *Paulinella chromatophora*. *Mol Biol Evol* 28(1):407–422.
16. Reyes-Prieto A, et al. (2010) Differential gene retention in plastids of common recent origin. *Mol Biol Evol* 27(7):1530–1537.
17. Nowack ECM, Grossman AR (2012) Trafficking of protein into the recently established photosynthetic organelles of *Paulinella chromatophora*. *Proc Natl Acad Sci USA* 109(14):5340–5345.
18. Bhattacharya D, et al. (2012) Single cell genome analysis supports a link between phagotrophy and primary plastid endosymbiosis. *Sci Rep* 2:356.
19. Doolittle WF (1998) You are what you eat: A gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet* 14(8):307–311.
20. Bitar M, Boroni M, Macedo AM, Machado CR, Franco GR (2013) The spliced leader trans-splicing mechanism in different organisms: Molecular details and possible biological roles. *Front Genet* 4:199 (abstr).
21. Acuña R, et al. (2012) Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proc Natl Acad Sci USA* 109(11):4197–4202.
22. Boto L (2014) Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proc R Soc B* 281(1777):20132450.
23. de Koning AP, Brinkman FSL, Jones SJM, Keeling PJ (2000) Lateral gene transfer and metabolic adaptation in the human parasite *Trichomonas vaginalis*. *Mol Biol Evol* 17(11):1769–1773.
24. Ropars J, et al. (2015) Adaptive horizontal gene transfers between multiple cheese-associated fungi. *Curr Biol* 25(19):2562–2569.
25. Schönknecht G, et al. (2013) Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* 339(6124):1207–1210.
26. Burillo S, Luque I, Fuentes I, Contreras A (2004) Interactions between the nitrogen signal transduction protein PII and N-acetyl glutamate kinase in organisms that perform oxygenic photosynthesis. *J Bacteriol* 186(11):3346–3354.
27. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2(4):953–971.
28. Huang J (2013) Horizontal gene transfer in eukaryotes: The weak-link model. *BioEssays* 35(10):868–875.
29. Dagan T, et al. (2013) Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome Biol Evol* 5(1):31–44.
30. Soucy SM, Huang J, Gogarten JP (2015) Horizontal gene transfer: Building the web of life. *Nat Rev Genet* 16(8):472–482.
31. Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT (2006) Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. *Genome Res* 16(9):1099–1108.
32. Kettler GC, et al. (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* 3(12):e231.
33. Husnik F, et al. (2013) Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell* 153(7):1567–1578.
34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.
35. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
36. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32(1):268–274.
37. Hess S, Suthaus A, Melkonian M (2015) "*Candidatus* Finniella" (*Rickettsiales, Alphaproteobacteria*), novel endosymbionts of viridiraptorid amoeboflagellates (Cercozoa, Rhizaria). *Appl Environ Microbiol* 82(2):659–670.
38. Mahmud P (2014) Reduced representations for efficient analysis of genomic data. PhD thesis (Rutgers, The State University of New Jersey, New Brunswick, NJ).
39. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17(1):10–12.
40. Simpson JT, et al. (2009) ABySS: A parallel assembler for short read sequence data. *Genome Res* 19(6):1117–1123.
41. Pertea G, et al. (2003) TIGR Gene Indices clustering tools (TGICL): A software system for fast clustering of large EST datasets. *Bioinformatics* 19(5):651–652.
42. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Res* 9(9):868–877.
43. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27(4):578–579.
44. Bernt M, et al. (2013) MITOS: Improved de novo metazoan mitochondrial genome annotation. *Mol Phylogenet Evol* 69(2):313–319.
45. Roy RS, Bhattacharya D, Schliep A (2014) Turtle: Identifying frequent k-mers with cache-efficient algorithms. *Bioinformatics* 30(14):1950–1957.
46. Quast C, et al. (2013) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res* 41(Database issue):D590–D596.
47. O'Brien EA, et al. (2007) TBestDB: A taxonomically broad database of expressed sequence tags (ESTs). *Nucleic Acids Res* 35(Database issue):D445–D451.
48. Boguski MS, Lowe TMJ, Tolstoshev CM (1993) dbEST–Database for "expressed sequence tags". *Nat Genet* 4(4):332–333.
49. Katoh K, Asimenos G, Toh H (2009) Multiple alignment of DNA sequences with MAFFT. *Methods Mol Biol* 537:39–64.