Complete genome of a nonphotosynthetic cyanobacterium in a diatom reveals recent adaptations to an intracellular lifestyle

Takuro Nakayama[a,b,c,1], Ryoma Kamikawa[d,e], Goro Tanifuji[b,d], Yuichiro Kashiyama[f,g,h], Naohiko Ohkouchi[f], John M. Archibald[b], and Yuji Inagaki[a,d]

[a]Center for Computational Sciences, University of Tsukuba, Tsukuba 305-8577, Japan; [b]Department of Biochemistry and Molecular Biology, Canadian Institute for Advanced Research, Program in Integrated Microbial Biodiversity, Dalhousie University, Halifax, NS, Canada B3H 4R2; [c]National Institute for Environmental Studies, Tsukuba 305-8506, Japan; [d]Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba 305-8572, Japan; [e]Graduate School of Global Environmental Studies, Graduate School of Human and Environmental Studies, Kyoto University, Kyoto 606-8501, Japan; [f]Japan Agency for Marine-Earth Science and Technology, Yokosuka 237-0061, Japan; [g]Department of Environmental and Biological Chemistry, Fukui University of Technology, Fukui 910-8505, Japan; and [h]Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency, Chiyoda 153-8902, Japan

Edited by Robert Haselkorn, University of Chicago, Chicago, IL, and approved June 24, 2014 (received for review March 21, 2014)

The evolution of mitochondria and plastids from bacterial endosymbionts were key events in the origin and diversification of eukaryotic cells. Although the ancient nature of these organelles makes it difficult to understand the earliest events that led to their establishment, the study of eukaryotic cells with recently evolved obligate endosymbiotic bacteria has the potential to provide important insight into the transformation of endosymbionts into organelles. Diatoms belonging to the family Rhopalodiaceae and their endosymbionts of cyanobacterial origin (i.e., "spheroid bodies") are emerging as a useful model system in this regard. The spheroid bodies, which appear to enable rhopalodiacean diatoms to use gaseous nitrogen, became established after the divergence of extant diatom families. Here we report what is, to our knowledge, the first complete genome sequence of a spheroid body, that of the rhopalodiacean diatom *Epithemia turgida*. The *E. turgida* spheroid body (*Et*SB) genome was found to possess a gene set for nitrogen fixation, as anticipated, but is reduced in size and gene repertoire compared with the genomes of their closest known free-living relatives. The presence of numerous pseudogenes in the *Et*SB genome suggests that genome reduction is ongoing. Most strikingly, our genomic data convincingly show that the *Et*SB has lost photosynthetic ability and is metabolically dependent on its host cell, unprecedented characteristics among cyanobacteria, and cyanobacterial symbionts. The diatom–spheroid body endosymbiosis is thus a unique system for investigating the processes underlying the integration of a bacterial endosymbiont into eukaryotic cells.

photosynthesis | organelle evolution | pseudogenization

The establishment of two bacterium-derived organelles, mitochondria and plastids, triggered drastic changes in the metabolic capacity, genome architecture, and lifestyle of eukaryotic cells. Mitochondria, energy-producing organelles ubiquitously found in eukaryotic cells, can be traced back to a single endosymbiosis between the ancestral (i.e., amitochondrial) eukaryote and an α-proteobacterium. The plastids of photosynthetic eukaryotes are the result of an endosymbiosis between a heterotrophic eukaryote and a cyanobacterium, an event that took place after the mitochondrial endosymbiosis (1–4). Determining the evolutionary events that led to the origins of these organelles is key to understanding the evolution of eukaryotic cells and their genomes. However, these two endosymbioses occurred more than 1 billion years ago (2–6), and present-day organisms provide limited information with which to explore the early stages of the transformation of an endosymbiotic bacterium into a fully integrated eukaryotic organelle.

Diatom species belonging to the family Rhopalodiaceae are unique among eukaryotes in possessing cyanobacterium-derived intracellular structures termed "spheroid bodies" (7), in addition to plastids and mitochondria. Phylogenetic analyses have shown that the spheroid bodies of rhopalodiacean diatoms originated from a single nitrogen-fixing cyanobacterium closely related to *Cyanothece* spp., which is distinct from the cyanobacterium that gave rise to the ancestral plastid (8, 9). Significantly, the spheroid bodies are considered obligate endosymbionts, as these structures are inseparable from the host (i.e., diatom) cells, and passed to daughter cells during host cell division (8, 10). The establishment of this diatom–cyanobacterium endosymbiosis can be traced back to the middle Miocene epoch ∼12 Ma (9), which is much more recent than the birth of mitochondria or plastids. Nitrogen fixation is predicted to be an important cellular function carried out by the spheroid bodies, as rhopalodiacean diatom cells can grow in media containing no nitrogen source (7). Indeed, previous studies experimentally confirmed nitrogenase activity in *Rhopalodia gibba* (8, 11), and genes involved in nitrogen fixation were found in partial genomic data from the spheroid body of *R. gibba* (*Rg*SB) (12). Interestingly, despite its cyanobacterial ancestry, the spheroid body lacks chlorophyll autofluorescence (13), and photosynthetic genes in *Rg*SB were predicted to be nonfunctional (12), suggesting that the symbiotic cyanobacteria are no longer capable of carrying out photosynthesis

Significance

Members of the diatom family Rhopalodiaceae possess a cyanobacterial endosymbiont called a "spheroid body." The spheroid body evolved much more recently than did mitochondria or plastids and is predicted to fix nitrogen. Here we present what is, to our knowledge, the first completely sequenced spheroid body genome from a rhopalodiacean diatom. Comparative analyses revealed that the endosymbiont is metabolically reduced, confirming its status as an obligate endosymbiont. The genome possesses genes for nitrogen fixation, and, to our surprise, no essential genes for photosynthesis. Thus, the spheroid body is, to our knowledge, the first known example of a nonphotosynthetic cyanobacterium, free-living or symbiotic. Rhopalodiacean diatoms have the potential to provide unique insight into the evolution of bacterial endosymbionts and their hosts.

Author contributions: T.N. and Y.I. designed research; T.N., R.K., G.T., Y.K., and N.O. performed research; T.N., R.K., G.T., Y.K., N.O., J.M.A., and Y.I. analyzed data; and T.N., R.K., G.T., Y.K., N.O., J.M.A., and Y.I. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. AP012549).

[1]To whom correspondence should be addressed. Email: ntakuro@ccs.tsukuba.ac.jp.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1405222111/-/DCSupplemental.

and are metabolically dependent on their host cells. Nonetheless, spheroid bodies still retain thylakoid membranes (7), the function of which is unclear, and there remains a possibility of retention of photosynthetic electron transport-related metabolism.

Present knowledge supports the idea that spheroid bodies depend on, and are controlled by, their host cells. The endosymbiotic relationship between rhopalodiacean diatoms and their spheroid bodies, as well as that between the testate amoeba *Paulinella chromatophora* and its cyanobacterium-derived photosynthetic organelle (14–16), thus provides a unique opportunity to investigate how an endosymbiotic bacterium integrates with a eukaryotic cell at the biochemical and genetic level (9, 12, 17, 18). Toward the goal of understanding how, and the extent to which, spheroid bodies are integrated into their diatom hosts, we determined the complete genome sequence of the spheroid body in the rhopalodiacean diatom *Epithemia turgida*. The spheroid body genome was found to be considerably reduced compared with its most closely related free-living cyanobacteria, and, remarkably, has already lost key metabolic capacities, including photosynthesis.

Results and Discussion

The Spheroid Body Genome Is Reduced in Size and Gene Content. The spheroid bodies in a clonal culture of *E. turgida* were isolated by discontinuous density Percoll gradient centrifugation. DNA extracted from the spheroid body-rich fraction was subjected to whole-genome amplification followed by 454 and Illumina sequencing. Sequence assembly was carried out by a combination of Velvet (19) and GS de novo assembler software, followed by PCR-based contig gap filling (Sanger sequencing). We obtained a single circular chromosome of 2.79 Mbp with a G+C content of 33.4% (Fig. 1). The genome has 39 tRNA genes (which can translate all 61 aa codons), two identical rRNA gene operons, and 1,720 protein-coding genes (Table 1). We also detected 225 pseudogenes, and no transposable elements were identified. The *E. turgida* spheroid body (*Et*SB) genome shares two major syntenic gene clusters with other cyanobacterial genomes—one for nitrogen fixation genes (*nif* genes) and the other for genes encoding ribosomal proteins. In addition to the *Et*SB genome, we assembled a 5.7-Kbp fragment with a low G+C content (Dataset S1). Most of the protein-coding genes on this fragment show no specific affinity to homologs of *Cyanothece* spp., which bear a close affinity to the spheroid bodies in cyanobacterial phylogeny (as detailed later). It is at present unclear whether this 5.7-Kbp fragment is an extrachromosomal element (e.g., plasmid) of *Et*SB origin; we did not analyze it further.

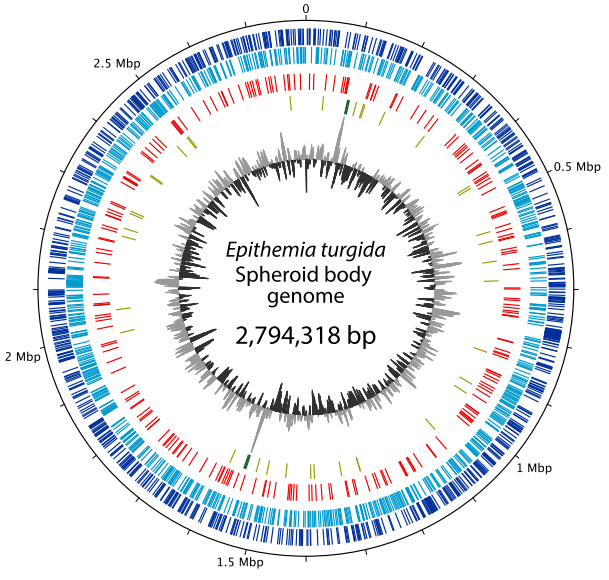

Fig. 1. Map of the circular chromosome of the spheroid body of the diatom *E. turgida*. Outer two circles (dark and light blue) show positions of protein-coding genes on plus and minus strands. The red bars on the third circle indicate pseudogenes. tRNA genes (yellow bars) and rRNA genes (green bars) are displayed on the fourth circle. The innermost circle shows G+C content (window size: 5,000 bp).

The spheroid bodies of rhopalodiacean diatoms have been shown to be specifically related to members of the cyanobacterial genus *Cyanothece* (9) (Fig. S1). Among cyanobacteria most closely related to spheroid bodies, complete genome data are available for three free-living *Cyanothece* species, strains PCC 8801 (National Center for Biotechnology Information BioProject ID PRJNA59027), PCC 8802 (PRJNA59143), and ATCC 51142 (PRJNA20319) (20, 21), as well as an uncultured oceanic species living symbiotically with a haptophyte alga (UCYN-A) (22, 23). The *Et*SB genome was found to be smaller in size and lower in G+C content than the free-living *Cyanothece* strains, which possess genomes of 4.68–5.36 Mbp in size and 37.9–39.8% G+C content (Table 1). Consistent with the observed differences in genome size, we found a marked difference in the number of ORFs between the *Et*SB genome and its free-living relatives. Whereas only 1,720 ORFs were identified in the *Et*SB genome, the genomes of the three free-living *Cyanothece* strains possess 4,367–5,304 ORFs. Moreover, 54.5% of the ORFs on the *Et*SB genome could be assigned to functional categories in Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology (KO), whereas only 32.6–38.0% of the ORFs in the free-living *Cyanothece* strains could be given a functional designation. Overall, these data suggest that the *Et*SB genome has undergone significant reductive evolution, as seen in other obligate bacterial symbionts (15, 24–26). Furthermore, the presence of numerous pseudogenes implies that genome reduction is still ongoing in the genome.

The partial nature of the *R. gibba* spheroid body genomic data obtained by Kneip et al. (12) precluded whole genome-scale comparisons between the spheroid body and free-living cyanobacteria, or between the spheroid body and other symbiotic cyanobacteria. Our KO-based analysis of the ORFs in the complete *Et*SB genome revealed that the spheroid body has already lost a large part of the metabolic functions presumed to have been present in its free-living ancestors. The numbers of KO IDs assigned to each predicted protein set for *Cyanothece* spp. ATCC 51142, PCC 8801, and PCC 8802 are 1,309, 1,325, and 1,318, respectively. In sharp contrast, only 849 KO IDs could be assigned to *Et*SB proteins, and most of these (842 KO IDs) were shared with at least one of the *Cyanothece* reference genomes (Fig. 2*A*). Large differences in the number of KO IDs between the *Et*SB and its free-living relatives were observed in several KEGG functional categories, most notably "energy metabolism," "metabolism of cofactors and vitamins," "carbohydrate metabolism," "membrane transport," and "signal transduction" (Fig. 2*B*). For instance, the *Et*SB genome possesses only half of the IDs present in sequenced *Cyanothece* genomes in the category of energy metabolism, mostly the result of a reduction in energy production abilities (e.g., photosynthesis, as detailed later). To the extent that the KO ID repertoire represents the metabolism of an organism, the *Et*SB genome has significantly reduced its metabolic capacity from its cyanobacterial progenitor, which is likely to have had a similar set of genes to strain PCC 8801, PCC 8802, or ATCC 51142.

In addition to the *Et*SB, two complete genome sequences are available for cyanobacterial nitrogen-fixing symbionts: '*Nostoc azollae*,' which is a vertically inherited but extracellular symbiont of the small heterosporous water fern *Azolla filiculoides* (27), and the uncultured cyanobacterium UCYN-A, which is thought to have an extra- and/or intracellular symbiotic relationship with haptophyte algae (22, 23, 28). Additionally, there is a near-complete genome sequence for another cyanobacterial symbiont, *Richelia intracellularis*, which has a mutually dependent relationship with a diatom distantly related to rhopalodiacean species (29).

Table 1. Genome overview of the spheroid body of *E. turgida* and two closely related cyanobacteria

Detail	Spheroid body of *E. turgida*	*Cyanothece* sp. PCC 8801	*Cyanothece* sp. ATCC 51142
Genome size, Mbp*	2.79	4.68	5.36
G+C, %			
Total	33.4	39.8	37.9
Noncoding regions*	26.4	34.6	32.3
rRNA operons	2	2	2
tRNA genes	39	43	43
Protein-coding genes	1,720	4,367	5,304
With functional annotation†	937 (54.5%)	1,659 (38.0%)	1,727 (32.6%)
With ambiguous function†	783 (45.5%)	2,708 (62.0%)	3,577 (67.4%)
Pseudogenes	225	199	6

*Values for main chromosomes.
†Predicted from KO assignment. Values in parentheses indicate percentages among the total protein-coding genes of each genome.

Whereas the genomes of '*N. azollae*' and *R. intracellularis* are larger than the *Et*SB genome (5.49 Mbp and 3.24 Mbp, respectively; Table S1), UCYN-A has a severely reduced genome of 1.44 Mbp, nearly half the size of the *Et*SB genome. As in *Et*SB, the UCYN-A, *R. intracellularis*, and '*N. azollae*' genomes have smaller KO ID repertoires (683, 965, and 1,084, respectively; Fig. 2*C* and Table S1) than the free-living *Cyanothece* spp., which have >1,300 KO IDs (Fig. 2*A*). In the functional categories of energy metabolism and metabolism of cofactors and vitamins, UCYN-A and *Et*SB exhibit similar reductive trends compared with '*N. azollae*' and *R. intracellularis* (Fig. 2*D*). Nevertheless, UCYN-A has fewer KO IDs than does *Et*SB in other functional categories, such as "carbohydrate metabolism," "amino acid metabolism," and "nucleotide metabolism," suggesting that the genome of this uncultured cyanobacterial symbiont has experienced more severe reduction in terms of genome size and metabolic capacity than that of *Et*SB, *R. intracellularis*, and '*N. azollae*' (Fig. 2 *C* and *D* and Table S1).

Spheroid Bodies Are Nonphotosynthetic Cyanobacterial Obligate Endosymbionts. To our knowledge, the complete *Et*SB genome provides us with the first opportunity to predict the entire set of

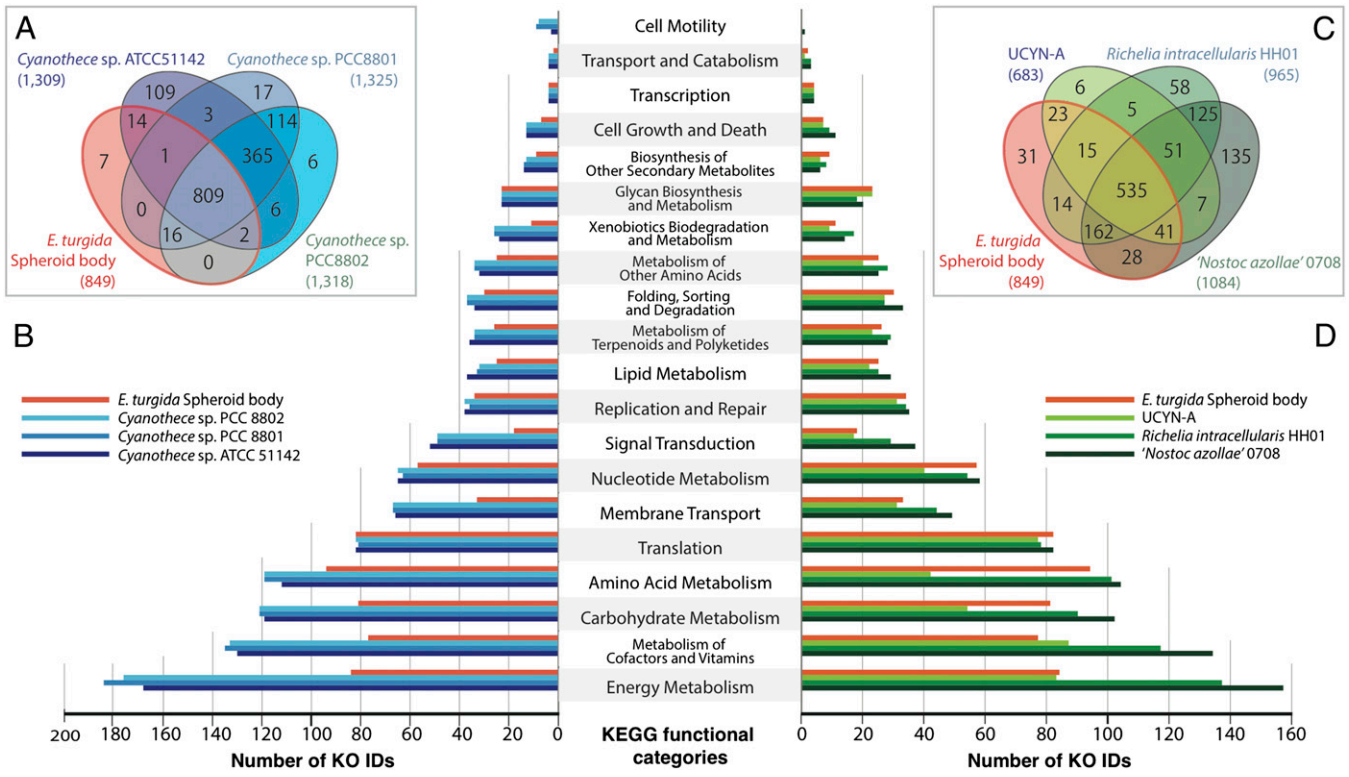

Fig. 2. KO-based comparison between free-living *Cyanothece* strains and cyanobacterial symbionts. (*A*) Venn diagram of KO ID repertoires of the spheroid body of *Epithemia turgida* and *Cyanothece* spp. strain PCC 8801, PCC 8802, and ATCC 51142. The numbers in parentheses indicate the total number of unique KO IDs in each genome. (*B*) Comparison of KO ID numbers between the *Et*SB, PCC 8801, PCC 8802, and ATCC 51142. Each bar indicates numbers of unique KO IDs in each functional category. (*C*) Venn diagram of KO ID repertoires of the *Et*SB, UCYN-A, *R. intracellularis* and '*N. azollae*.' (*D*) Comparison of KO ID numbers between the *Et*SB, UCYN-A, *R. intracellularis*, and '*N. azollae*' in functional categories.

metabolic pathways carried out in the spheroid body (see Fig. 4). Pioneering studies detected no chlorophyll autofluorescence (13), and inactivated photosynthetic genes were found in a partially sequenced spheroid body genome (12). However, given the prominent thylakoid membranes of spheroid bodies (7), it was difficult to exclude the existence of facultative photosynthetic capacity of the endosymbiont based solely on the partial genome sequence of *Rg*SB or microscopic observations. Analysis of the complete genome sequence of *Et*SB presented herein allows us to conclude that the cyanobacterial endosymbionts of rhopalodiacean diatoms truly lack photosynthetic ability. In the *Et*SB genome, genes for photosynthetic components such as photosystems I and II and phycobilisomes (the main peripheral antennae of cyanobacteria) were absent or clearly nonfunctional (Fig. 3). We also found that genes for the entire chlorophyll synthesis pathway, which converts protoporphyrin IX to chlorophyll *a*, appear to be inactivated by pseudogenization (Fig. S2). All things considered, the complete spheroid body genome sequence provides concrete and comprehensive evidence for the loss of photosynthesis. To our knowledge, the *Et*SB is the only known example of a nonphotosynthetic cyanobacterium, free-living or symbiotic.

Despite the absence of a photosynthetic apparatus, a cytochrome b_6f complex, plastocyanin, and ATP synthase are most likely functional on the spheroid body thylakoid membranes, as intact ORFs for such proteins reside in the *Et*SB genome (Fig. 3). In cyanobacteria, the aforementioned set of proteins is involved in respiration together with NADH dehydrogenase and cytochrome *c* oxidase, in addition to photosynthesis (30). Thus, we predict that the *Et*SB still synthesizes ATP by an electrochemical proton gradient. The uptake hydrogenase (Hup, encoded by *hupSL*) (31), which can recycle dihydrogens produced through nitrogen fixation as electron donors, may be involved in the process described earlier as an additional supplier of electrons (Fig. 3). Therefore, although the spheroid body lacks photosynthetic capacity, its thylakoid membranes may be retained as the site of an intact respiratory chain. Furthermore, it is possible that the activity of cytochrome *c* oxidase in the spheroid body serves to protect extremely oxygen-sensitive nitrogenases by reducing oxygen molecules to water.

To gain insight into the energy-metabolic dependence of the spheroid body on its diatom host, we searched the *Et*SB genome for genes encoding enzymes involved in the Calvin and tricarboxylic acid (TCA) cycles. The Calvin cycle appears to be incomplete (Fig. S3). In particular, ORFs for the three subunits of ribulose-1,5-bisphosphate carboxylase/oxygenase (*rbcL*, *rbcS*, and *rbcX*) were found to be disrupted by stop codons, suggesting

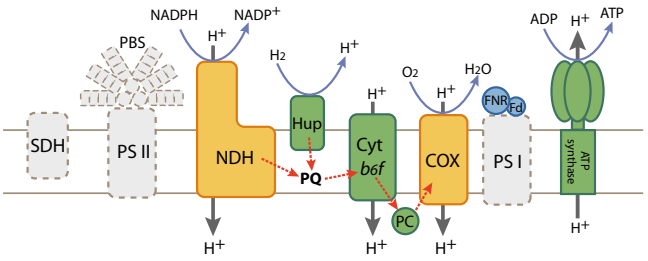

Fig. 3. Schematic representation of the predicted electron flow in the thylakoid membrane of the spheroid body. Dashed red arrows indicate deduced electron flow. Complexes predicted to be inactivated by gene loss and/or pseudogenization are shown by dashed lines. Yellow- and blue-colored complexes are specific to respiration and photosynthesis, respectively, whereas those in green are involved in respiration and photosynthesis. COX, terminal cytochrome *c* oxidase; cyt b_6f, cytochrome b_6f complex; FNR, ferredoxin-NADP$^+$ reductase; Fd, ferredoxin; Hup, uptake hydrogenase; NDH, type 1 NADPH dehydrogenase; PC, plastocyanin; PBS, phycobilisome; PQ, plastoquinone; PS I, photosystem I; PS II, photosystem II; SDH, succinate dehydrogenase.

that the endosymbiont cannot fix carbon dioxide through the Calvin cycle. As the *Et*SB still possesses apparently functional genes for enzymes to catabolize carbohydrates (e.g., enzymes involved in the pentose phosphate pathway and glycolysis, except phosphofructokinase), the endosymbiont likely energetically depends on sugars that are "extracellularly" supplied from the host. This idea is also supported by the presence of a gene encoding an ABC-transporter for sugars in the *Et*SB genome (Fig. 4), despite the fact that the total repertoire of transporter genes is highly reduced compared with its free-living relatives (Fig. 2*B*, "membrane transport"). Additionally, the TCA cycle was found to be incomplete in the *Et*SB (Fig. 4 and Fig. S4); only intact ORFs for citrate synthase, aconitase, isocitrate dehydrogenase, and fumarase were identified. The TCA cycle in the spheroid body is thus unlikely to function in ATP synthesis, but the products from the incomplete cycle are likely used as substrates for the biosynthesis of certain amino acids (oxaloacetate and 2-oxoglutarate are required for biosynthesis of five and four amino acids, respectively; Fig. 4). The incomplete nature of the TCA cycle and the Calvin cycle deduced from the genome data strongly support the obligate endosymbiosis between spheroid bodies and rhopalodiacean diatoms. It is conceivable that some or all of the missing enzymes in the *Et*SB are imported from the host cell. Regardless, the obligate nature of the endosymbiosis would not be changed, as the host cell would still be required to complete the two pathways in the endosymbiont.

Amino acid biosynthetic pathways are often partially or entirely discarded in bacterial endosymbionts (15, 22, 25, 26, 32). In sharp contrast, the *Et*SB genome retains seemingly functional genes for the biosynthesis of all protein amino acids (Fig. 4). Likewise, the biosynthesis of purines, pyrimidines, and most cofactors [e.g., FMN, FAD, NAD(P)$^+$, CoA, heme, and thiamine] is likely to be operating in the *Et*SB.

The key characteristics of the *Et*SB described earlier—(*i*) absence of photosynthesis, (*ii*) incomplete TCA cycle, and (*iii*) intact biosynthetic pathways for protein amino acids, nucleotides, and cofactors—places the spheroid bodies of rhopalodiacean diatoms on a distinct evolutionary trajectory from that of UCYN-A, which discarded the entire TCA cycle, and several biosynthetic pathways for amino acids and purine nucleotides, and indeed any endosymbiotic bacteria known to date.

Is the Spheroid Body Becoming a Nitrogen Fixing Organelle? Nitrogenase activity has been experimentally confirmed in *R. gibba* cells (8, 11). To validate the nitrogen-fixing capacity of *E. turgida* experimentally, *E. turgida* cells were cultured in the presence of $^{15}N_2$ gas, followed by analysis of the isotopic composition of chlorophyll *a*. Incorporation of isotopic nitrogen in chlorophyll *a* was detected (Table S2), indicating that this diatom species can indeed use gaseous nitrogen. A nearly complete set of *nif* genes for nitrogen fixation was identified in the *Et*SB genome, which is homologous to the gene set found in a 51-Kbp genome fragment of the *Rg*SB (12). The composition and order of the *nif* genes were almost identical between the spheroid body genomes of *E. turgida* and *R. gibba*, with the exception of the *nifU* gene encoding a scaffold protein for [Fe_2S_2] cluster assembly. The partial genome of the *Rg*SB harbors an ORF for NifU, which is highly truncated compared with other homologs (12), whereas the orthologous ORF in the *Et*SB genome is disrupted by multiple in-frame stop codons. We predict that the *Et*SB uses a NifU-like protein (ETSB_1030) located in a region distinct from the *nif* gene cluster, instead of the authentic NifU. It is presently not known how the products of nitrogen fixation are transported into the cytosol of the diatom host. Although a gene encoding a putative ammonium transporter persists in the *Et*SB genome (Fig. 4), its precise function is unclear.

It is also currently unclear how the spheroid body, which lacks a complete TCA cycle and photosynthetic ability (Fig. 3 and Fig. S4),

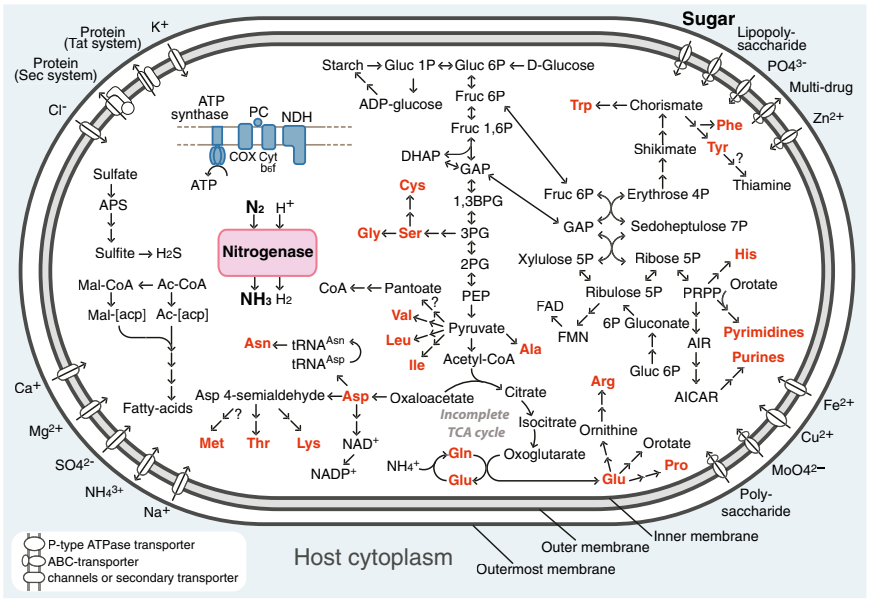

Fig. 4. Overview of the deduced metabolism of the spheroid body in *E. turgida*. Single arrows indicate metabolic reactions and pathways involving multiple steps are represented by double arrows. Amino acids and nucleotides are highlighted by orange. Substances of particular interest are in bold. Question marks highlight pathways that are uncertain even among related *Cyanothece* strains. Transporter types: single oval, P-type ATPase transporter; oval coupled with small oval, ABC-transporters; single rounded rectangle, channels or secondary transporters. Ac-[acp], acetyl-[acyl-carrier protein]; Ac-CoA, acetyl-CoA; 1,3BPG, 1,3-bisphosphoglycerate; AICAR, 5-aminoimidazole-4-carboxamide ribotide; AIR, 5-aminoimidazole ribotide; APS, adenosine 5′-phosphosulfate; COX, terminal cytochrome *c* oxygenase; cyt b_6f, cytochrome b_6f complex; DHAP, dihydroxyacetone phosphate; FAD, flavin adenine dinucleotide; FMN, flavin mononucleotide; Fruc, fructose; GAP, glyceraldehyde-3-phosphate; Gluc, glucose; Mal-[acp], malonyl-[acyl-carrier protein]; Mal-CoA, malonyl CoA; NDH, type 1 NADPH dehydrogenase; PC, plastocyanin; PEP, phosphoenolpyruvate; 2PG, 2-phosphoglycerate; 3PG, 3-phosphoglycerate; PRPP, 5-phosphoribosyl 1-pyrophosphate.

generates sufficient ATP to support an energetically expensive nitrogenase activity (33). If glycolysis and/or the pentose phosphate pathway are the principal electron donors in the spheroid body, sugars or sugar phosphates need to be effectively imported from the host to the endosymbiont. It is also possible that the host supplies ATP to the spheroid bodies, an interesting hypothesis but one that cannot be confirmed or refuted based on the data presently in hand. Characterization of the complete set of spheroid body transporters will be necessary to elucidate the trafficking of metabolites between host and endosymbiont.

Reductive genome evolution has been demonstrated in *Et*SB, UCYN-A, '*N. azollae*,' and *R. intracellularis* (22, 27, 29), but only the *Et*SB has discarded photosynthetic ability; photosynthesis genes (e.g., those for photosynthetic complexes) persist in the genomes of UCYN-A, '*N. azollae*,' and *R. intracellularis* (22, 27, 29). For instance, UCYN-A retains complete gene sets for photosystem I and chlorophyll *a* synthesis, despite its small genome size (22). Altogether, with its lack of photosynthesis, the spheroid bodies of rhopalodiacean diatoms are highly integrated into their diatom host cells, and indeed distinct from all other cyanobacterial symbionts examined to date. However, based on the *Et*SB genome data currently in hand, it is not possible to determine whether the spheroid body is a bona fide organelle in the strictest sense. The evolution of protein import machinery, which enables proteins to move from host to endosymbiont, has been proposed as the critical event in the conversion of endosymbionts into organelles (34, 35). To address this issue, it will be important to survey the host diatom nuclear genome for genes possibly encoding proteins transported to, and functioning in, the spheroid body.

Conclusion

We have sequenced the genome of the *E. turgida* spheroid body, providing important insights into the unique evolutionary status of this nitrogen-fixing intracellular structure. The data strongly suggest that the spheroid bodies of rhopalodiacean diatoms have secondarily lost an important property of cyanobacteria (i.e., photosynthesis), and are unique among prokaryotic symbionts examined to date. However, there is still much to learn about how the diatom cell controls its spheroid bodies. Investigation of this host–symbiont system should provide clues to understanding the general evolutionary processes underlying the conversion of a free-living prokaryote into a fully integrated, host-controlled subcellular entity.

Materials and Methods

Culturing *E. turgida* Cells and Isolation of the Spheroid Bodies. *E. turgida*, isolated from Lake Yunoko, Tochigi, Japan, was grown in CSi-N medium as described by Nakayama et al. (9). Harvested diatom cells were resuspended in PBS solution and mildly broken by vortexing with glass beads. The homogenate containing spheroid bodies was layered onto a discontinuous gradient of 90%, 70%, 60%, and 50% Percoll, and then centrifuged at 12,000 × *g* for 20 min. Intact spheroid bodies were taken from the boundary between the 60% and 70% Percoll fractions. Although the spheroid body-rich fraction also contained contaminating bacteria present in the *E. turgida* culture, DNA-containing organelles from the diatom cells (i.e., nuclei, mitochondria and plastids) were not detected under the light microscope. DNA was extracted from the spheroid body-rich fraction using the DNeasy Plant Mini Kit (Qiagen), followed by whole-genome amplification by using the REPLI-g mini kit (Qiagen).

Genome Sequencing and Assembly. Approximately 10 μg of amplified DNA was sequenced on a Roche 454 GS FLX sequencer with Titanium reagents and an Illumina HiSeq 2000 at Hokkaido System Science. 454 and Illumina sequencing generated ~195,000 reads (61 Mbp in total) and ~58 million reads (5.8 Gbp in total), respectively. Illumina reads were assembled into contigs by using Velvet (19), and contigs were combined and assembled together with reads from 454 sequencing using the GS de novo assembler software (Roche Diagnostics). Ten large contigs (>5 Kbp) with high A+T content (>65%) were found in the final contig pool and selected as candidate spheroid body genome sequences. A contig containing the spheroid body rRNA gene operon [identical to a partial sequence already obtained by Nakayama et al. (9)] possessed a relatively low A+T content (~5.2 Kbp, A+T of 49%) and was identified manually. Gaps between the remaining spheroid body-derived contigs were closed by PCR based on information from mate pairs of paired-end Illumina reads in the flanking regions of contigs. A single circular chromosome of 2.79 Mbp with coverage depth of ~145× was obtained.

Genome Annotation. The initial ORF prediction was performed by using GeneMarkS (36). Four short ORFs for genes *rpl36*, *rpmG*, *petL*, and *ndhL* were identified by tBLASTn searches. tRNA genes were detected by tRNAscan-SE (37), and rRNA genes were predicted based on nucleotide similarity. Putative pseudogenes on the spheroid body genome were identified by tBLASTn by using all proteins encoded in the genome of *Cyanothece* sp. PCC 8801 as queries with a threshold of 1.0×10^{-10}. Coding regions interrupted by stop codons and/or disrupted by frame shifts, as well as severely truncated ORFs, were tagged as putative pseudogenes. A total of 176 ORFs initially predicted by GeneMarkS were also found to correspond to pseudogenes. The genome map was generated by using the DNAplotter (38). The fully annotated

genome sequence is available in DNA Data Base in Japan/GenBank/European Molecular Biology Laboratory under accession number AP012549.

KO-Based Analyses. KO ID assignment was performed for all 1,720 ORFs predicted in the spheroid body genome. The initial assignments were performed by using the KEGG Automatic Annotation Server (39) and then manually refined. A complete list of the KO assignments is provided in Dataset S2. All KO IDs assigned for genes from three *Cyanothece* strains (PCC 8801, PCC 8802, and ATCC 51142) and two cyanobacterial symbionts ('*N. azollae*' 0708 and UCYN-A) were retrieved from the KEGG database by using the KEGG API (40). KO IDs for protein sequences of *R. intracellularis* HH01 were assigned by using the KEGG Automatic Annotation Server. KO IDs for pseudogenes in the spheroid body genome were predicted from the KO IDs for their intact homologs in PCC 8801. Metabolic pathways in the *Et*SB were deduced by using KEGG mapper (41).

Phylogenetic Tree Construction. Orthologous proteins for putative spheroid body proteins were searched against genome data from eight cyanobacterial strains by using BLASTp with each spheroid body protein as a query. Only top hits from each cyanobacterial strain with E-values of 0.0 were considered as orthologous proteins. A total of 241 orthologous proteins shared by all eight cyanobacterial strains and the *Et*SB were determined in this manner. The dataset is available upon request. The orthologous proteins were individually aligned by using MUSCLE version 3.6 (42); all positions containing gap(s) were removed. The alignments were concatenated into a single dataset (136,692 sites). Maximum likelihood phylogenetic analysis was performed by using RAxML 8.0.0 (43) under the LG + Γ + F model. Searches for the best trees were conducted starting from 10 random trees, and bootstrap values were obtained with nonparametric bootstrapping by using 100 replicates.

Nitrogen Isotope Tracing Analysis. *E. turgida* cells and two diatom strains without spheroid bodies (*Cyclotella meneghiniana* NIES-805 and *Fragilaria capucina* NIES-391) were cultured in nitrogen-deficient CSi-N medium as described earlier for 4 d. The $^{15}N_2$ gas was added to the gaseous phase of each flask on the first and second day of culturing. All three cultures were then collected through a glass filter (GF/F; Whatman). Extraction and isotopic analysis of chlorophyll *a* from the filters were done by using the procedure reported by Tyler et al. (44).

ACKNOWLEDGMENTS. We thank Dr. Fumie Kasai (National Institute for Environmental Studies) for all her help and cooperation. This work was supported by Japan Society for the Promotion of Sciences Grants 22870037 (to T.N.), 24870004 (to R.K.), and 23117006 (to Y.I.); Ministry of Education, Culture, Sports, Science and Technology of Japan Grant-in-Aid for Scientific Research on Innovative Areas 3308; and a grant from the Institute for Fermentation (Osaka, Japan) (R.K.). T.N. was a JSPS Postdoctoral Fellow for Research Abroad. J.M.A. is a Senior Fellow of the Canadian Institute for Advanced Research, Program in Integrated Microbial Biodiversity.

1. Gray MW, Burger G, Lang BF (1999) Mitochondrial evolution. *Science* 283(5407): 1476–1481.
2. Dyall SD, Brown MT, Johnson PJ (2004) Ancient invasions: From endosymbionts to organelles. *Science* 304(5668):253–257.
3. Gould SB, Waller RF, McFadden GI (2008) Plastid evolution. *Annu Rev Plant Biol* 59: 491–517.
4. Archibald JM (2009) The puzzle of plastid evolution. *Curr Biol* 19(2):R81–R88.
5. 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.
6. Parfrey LW, Lahr DJG, Knoll AH, Katz LA (2011) Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci USA* 108(33):13624–13629.
7. Drum RW, Pankratz S (1965) Fine structure of an unusual cytoplasmic inclusion in the diatom genus, *Rhopalodia*. *Protoplasma* 60:141–149.
8. Prechtl J, Kneip C, Lockhart P, Wenderoth K, Maier U-G (2004) Intracellular spheroid bodies of *Rhopalodia gibba* have nitrogen-fixing apparatus of cyanobacterial origin. *Mol Biol Evol* 21(8):1477–1481.
9. Nakayama T, et al. (2011) Spheroid bodies in rhopalodiacean diatoms were derived from a single endosymbiotic cyanobacterium. *J Plant Res* 124(1):93–97.
10. Geitler L (1977) Zur Entwicklungsgeschichte der Epithemiaceen *Epithemia, Rhopalodia* und *Denticula* (*Diatomophyceae*) und ihre vermutlich symbiotischen Sphäroidkörper. *Plant Syst Evol* 128:259–275.
11. Floener L, Bothe H (1980) Nitrogen fixation in *Rhopalodia gibba*, a diatom containing blue-greenish inclusions symbiotically. *Endocytobiology: Endosymbiosis and Cell Biology, a Synthesis of Recent Research*, eds Schwemmler W, Schenk H (Walter de Gruyter, Berlin), 1st Ed, pp 541–552.
12. Kneip C, Voβ C, Lockhart PJ, Maier U-G (2008) The cyanobacterial endosymbiont of the unicellular algae *Rhopalodia gibba* shows reductive genome evolution. *BMC Evol Biol* 8:30.
13. Kies L (1992) Glaucocystophyceae and other Protists Harbouring Prokaryotic Endocytobionts. *Algae and Symbioses*, ed Reisser W (Biopress, Bristol, UK), pp 353–377.
14. Marin B, Nowack ECM, Melkonian M (2005) A plastid in the making: Evidence for a second primary endosymbiosis. *Protist* 156(4):425–432.
15. Nowack EC, Melkonian M, Glöckner G (2008) Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. *Curr Biol* 18(6): 410–418.
16. Nakayama T, Archibald JM (2012) Evolving a photosynthetic organelle. *BMC Biol* 10: 35.
17. Trapp EM, Adler S, Zauner S, Maier U-G (2012) *Rhopalodia gibba* and its endosymbionts as a model for early steps in a cyanobacterial primary endosymbiosis. *J Endocytobiosis Cell Res* 23:21–24.
18. Adler S, Trapp EM, Dede C, Maier U-G, Zauner S (2014) *Rhopalodia gibba*: The first steps in the birth of a novel organelle? *Endosymbiosis*, ed Löffelhardt W (Springer, Vienna), pp 167–179.
19. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18(5):821–829.
20. Bandyopadhyay A, et al. (2011) Novel metabolic attributes of the genus *Cyanothece*, comprising a group of unicellular nitrogen-fixing cyanobacteria. *MBio* 2(5):e00214-11.
21. Welsh EA, et al. (2008) The genome of *Cyanothece* 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle. *Proc Natl Acad Sci USA* 105(39):15094–15099.
22. Tripp HJ, et al. (2010) Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* 464(7285):90–94.
23. Thompson AW, et al. (2012) Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science* 337(6101):1546–1550.
24. Wernegreen JJ (2002) Genome evolution in bacterial endosymbionts of insects. *Nat Rev Genet* 3(11):850–861.
25. Kuwahara H, et al. (2007) Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, *Calyptogena okutanii*. *Curr Biol* 17(10):881–886.
26. Moran NA, McCutcheon JP, Nakabachi A (2008) Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet* 42:165–190.
27. Ran L, et al. (2010) Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PLoS ONE* 5(7):e11486.
28. Hagino K, Onuma R, Kawachi M, Horiguchi T (2013) Discovery of an endosymbiotic nitrogen-fixing cyanobacterium UCYN-A in *Braarudosphaera bigelowii* (Prymnesiophyceae). *PLoS ONE* 8(12):e81749.
29. Hilton JA, et al. (2013) Genomic deletions disrupt nitrogen metabolism pathways of a cyanobacterial diatom symbiont. *Nat Commun* 4:1767.
30. Vermaas WFJ (2001) Photosynthesis and respiration in cyanobacteria. *eLS* (Wiley, Chichester, UK), www.els.net, 10.1038/npg.els.0001670.
31. Tamagnini P, et al. (2002) Hydrogenases and hydrogen metabolism of cyanobacteria. *Microbiol Mol Biol Rev* 66(1):1–20.
32. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* 407(6800): 81–86.
33. Seefeldt LC, Hoffman BM, Dean DR (2009) Mechanism of Mo-dependent nitrogenase. *Annu Rev Biochem* 78:701–722.
34. Cavalier-Smith T, Lee JJ (1985) Protozoa as hosts for endosymbioses and the conversion of symbionts into organelles. *J Protozool* 32:376–379.
35. Theissen U, Martin W (2006) The difference between organelles and endosymbionts. *Curr Biol* 16(24):R1016–R1017.
36. Besemer J, Lomsadze A, Borodovsky M (2001) GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 29(12):2607–2618.
37. Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* 33(Web server issue):W686-W689.
38. Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J (2009) DNAPlotter: Circular and linear interactive genome visualization. *Bioinformatics* 25(1):119–120.
39. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35(Web server issue):W182-W185.
40. Kawashima S, Katayama T, Sato Y, Kanehisa M (2003) KEGG API: A Web service using SOAP/WSDL to access the KEGG system. *Genome Inform* 14:673–674.
41. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 40(database issue): D109–D114.
42. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
43. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
44. Tyler J, et al. (2010) Tracking aquatic change using chlorine-specific carbon and nitrogen isotopes: The last glacial-interglacial transition at Lake Suigetsu, Japan. *Geochemistry Geophys Geosystems* 11:Q09010.