Multi-omics Analysis of the Symbiotic Green Algae, Chlorella Variabilis, Revealing the Genetic Basis of the Obligate Endosymbiotic Lifestyle

Ryuhei Minei
Nagahama Institute of Bio-Science and Technology
Ryo Hoshina
Nagahama Institute of Bio-Science and Technology
Rina Higuchi
Kobe University
Lin Chen
Kobe University
Yuki Akizuki
Nagahama Institute of Bio-Science and Technology
Yasunobu Terabayashi
Takara Bio Inc
Satoshi Kira
Takara Bio Inc
Masanari Kitagawa
Takara Bio Inc
Toshinobu Suzaki
Kobe University
Atsushi Ogura (aogu@whelix.info)
Nagahama Institute of Bio-Science and Technology

Research Article

Keywords: Plastid, Secondary endosymbiosis, Paramecium bursaria, Genome, Transcriptome

Posted Date: September 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-713024/v1

Abstract

Background: Photosynthetic eukaryotes have evolved through the acquisition of plastids by secondary endosymbiosis, a process that requires several steps. Immediately before plastid acquisition, the genome of the symbiont is known to be dramatically reduced, but few studies have focused on the genomic changes in the symbiont at the early stages of secondary endosymbiosis.

Methods: To investigate the genetic basis of the transition from facultative to obligate endosymbiosis, we compared the genomes of *Chlorella variabilis*, a representative symbiotic alga, with that of *Paramecium bursaria*, to compare closely related free-living species and transcriptomes between organisms in symbiotic and non-symbiotic conditions.

Results: We found that the non-reduced genome of *C. variabilis* and its genes play a crucial role in endosymbiosis, being involved in cell wall biogenesis and degradation, and metabolic exchanges with the host. Our results suggest that the genetic mechanism underlying the enhancement of photosynthesis under symbiosis is the increasing light absorption efficiency and carbon fixation capacity of the endosymbiont, resulting in an increase in the supply of maltose to *P. bursaria*.

Introduction

Endosymbiosis is a driving force of evolution, facilitating the diversification of organisms. Mitochondria and plastids originated from endosymbiotic alpha-proteobacteria and cyanobacteria, respectively (Gray 1999). Secondary endosymbiosis leads to the diversification of the algal lineage. Plastid acquisition originated from endosymbiosis with cyanobacteria, a primary endosymbiosis whose descendants form the monophyletic supergroup Archaeplastida, composed of three lineages: green algae and land plants, red algae, and glaucophytes (Adl et al. 2019). Green and red algae have been repeatedly taken up and integrated as plastids by other eukaryotes, a process known as secondary endosymbiosis, whereas primary endosymbiosis seldom occurs. Besides Archaeplastida, secondary endosymbiosis has led to the establishment of broad diversified algal lineages: the Chlorarachniophyta and Euglenozoa descended from green algae and the Alveolata, Crypthphyta, Haptophyta, and Stramenopile from red algae (Ponce-Toledo et al. 2019).

Inouye and Okamoto (2005) have proposed that secondary endosymbiosis requires many steps for plastid establishment. Previously, several organisms at various intermediate stages from endosymbiosis to plastid have been reported and broadly classified into three stages. Early in secondary endosymbiosis, in the first stage, the relationship between host and symbiont is facultative endosymbiosis. A phagotrophic protist engulfs algae as prey and retains the undigested algae as a temporary symbiont. A host needs to take up of new algae repeatedly, because cell division does not become synchronized between host and symbiont. In the second stage, the facultative endosymbiotic relationship progresses to obligate endosymbiosis. Synchronized cell division between host and symbiont makes the relationship permanent, without the need for the incorporation of new algae. During the third stage, just before plastid

establishment, many genes of the symbiont are horizontally transferred to the host genome, so that the nucleus of the symbiont markedly shrinks, becoming a vestigial nucleus, referred to as a nucleomorph. The mitochondria of the symbiont disappear. Finally, the nucleus of the symbiont disappears and the symbiont becomes a plastid.

Many studies have investigated the genetic changes from symbiont to the plastid in the third stage of secondary endosymbiosis. Nucleomorphs have been found in Chlorarachniophyta, Crypthphyta, and some dinoflagellates (Sarai et al. 2020). Several nucleomorphs from Chlorarachniophyta and Crypthphyta have been sequenced. The genome size of the nucleomorphs ranges from 373 to 703 kbp (Suzuki et al. 2015), whereas that of green and red algae ranges from 12 to 343 Mbp and 11 to 105 Mbp, respectively (Blaby-Haas & Merchant 2019). The genome of the symbiont at the third stage is therefore significantly reduced. In a nucleomorph, 284–610 genes are housekeeping genes, such as those coding for rRNA and proteins involved in translation and transcription. The nuclear genomes of Chlorarachniophyta *Bigelowiella natans* and the cryptophyte *Guillardia theta* have been sequenced and were found to contain 353 and 508 genes derived from the symbiont, respectively, reflecting gene transfer from the symbiont to host (Curtis et al. 2012). Despite these studies, little is known about genetic changes involved in the transition from facultative to obligate endosymbiont in the second stage.

Among the various organisms in the second stage of secondary endosymbiosis, we focused on the well-known endosymbiotic relationship between the ciliate *Paramecium bursaria* and the green algae *Chlorella variabilis*. Their cell division is synchronized (Kadono et al. 2004), indicating that they have established obligate endosymbiosis. In the natural environment, the aposymbiotic *P. bursaria* is rarely found (Tonooka & Watanabe 2002), and non-symbiotic *C. variabilis* has never been found (Hoshina et al. 2010). In the laboratory, methods of culturing *P. bursaria* and *C. variabilis* separately have been established (Siegel 1960), so various interactions between them have been reported. For instance, *P. bursaria* supplies nitrogen to *C. variabilis* (Kamako et al. 2005), while *C. variabilis* supplies a large amount of photosynthetic product in the form of maltose to *P. bursaria* (Muscatine et al. 1967). The genetic mechanisms underlying these transfers remain unclear.

As an initial approach to the elucidation of the genetic basis of the second stage of secondary symbiosis, comprehensive studies using data from next-generation sequencings, such as comparative genomics and transcriptome analysis, are powerful. Genomic and transcriptomic studies of *P. bursaria* have shed light on the genetic mechanisms of symbiosis with *C. variabilis* from the perspective of the host (Kodama et al. 2014; He et al. 2019). From the perspective of the symbiont, the genomes of *C. variabilis* (Blanc et al. 2010) and *Micractinium conductrix* (Arriola et al. 2018), another symbiont of *P. bursaria*, have been reported, but no comparative genomic studies have been performed to clarify the genetic changes from facultative to obligate endosymbionts in the second stage. Transcriptome studies under symbiotic conditions are limited (Quispe et al. 2016), and no comprehensive data are available to clarify the genetic mechanism of symbiosis.

In this study, to reveal the genetic basis of obligate endosymbiosis, we prepared genome datasets of *C. variabilis*, *C. vulgaris*, and *C. sorokiniana* and transcriptomes of *C. variabilis*, under symbiotic and non-symbiotic conditions, and analyzed them. Since the taxonomy of the genus *Chlorella* remains controversial (Heeg & Wolf 2015), we selected *C. vulgaris*, the type species of the genus *Chlorella*, as the non-symbiotic free-living species, and *C. sorokiniana* as the outgroup, for genome comparison with *C. variabilis*. Although we expected to observe a presage of large-scale genome reduction in the third stage of secondary endosymbiosis, the genome of *C. variabilis* did not shrink but rather expanded, compared to that of *C. vulgaris*. Some of the *C. variabilis*-specific expanded genes were significantly upregulated under symbiotic conditions compared to non-symbiotic conditions, suggesting that they play an important role in endosymbiosis. They have been found to contain genes possibly involved in cell wall biogenesis and degradation and metabolic exchange with the host. We also investigated the gene networks and pathways that play a central role in host-symbiont interactions: chlorophyl synthesis, the light-harvesting complex (LHC), the Calvin cycle, and the D-fructose-6P to maltose pathways. Enhanced photosynthesis caused by the increase of amount of chlorophyl, LHC, and enzymes comprising the Calvin cycle is likely to enable the transfer of a large supply of maltose to the host.

Materials And Methods

Read data preparation

Read data of *Chlorella sorokiniana* UTEX 1602 were downloaded from the NCBI BioProject PRJNA290386 (Arriola et al. 2018). In *Chlorella variabilis* NC64A (ATCC 50528), read data of a genome were downloaded from NCBI BioProject PRJDB7392 (Minei et al. 2018), and RNA-seq data were generated in this study. Details of the RNA-seq are described in the transcriptome analysis section. *Chlorella vulgaris* NIES-686 (= SAG 211-11b; CCAP 211/11b; UTEX 259) was cultured in G medium (Hoshina et al. 2018) under 14-h light and 10-h dark conditions at 25°C. Genomic DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). Samples of 50–100 µg of cells were incubated for 5 min at 65℃ in 400 µl Buffer AP1 and 4 µl RNase A, and 400 µl glass beads (0.1 mm) were added. Cells were disrupted using Bead Smash 12 (Wakenyaku Co. Ltd., Kyoto, Japan) at 5,000 rpm for 30 s. This disruption was repeated five times, and then the samples were incubated for 10 min at 65℃. Thereafter, we followed the manufacturer's protocol. Before total RNA extraction using RNeasy Mini Kit (Qiagen), collected cells were resuspended in ten times the amount of RNAlater® (Thermo Fisher Scientific, Waltham, MA, USA) and disrupted in the same way as for DNA extraction. After adding 500 µl Buffer RLT, cell lysates were vortexed and incubated for 3 min at 56℃. The supernatants were recovered by centrifugation at 13,000 g and used as input to the manufacturer's protocol. Extracted genomic DNA and total RNA were shipped to BGI for library preparation and sequencing by PacBio® RS II (Pacific Biosciences, Menlo Park, CA) and HiSeq X Ten (Illumina, San Diego, CA, USA), following the manufacturer's protocol. Details on the read data are summarized in Table S1.

Genome dataset construction

Prepared read data were checked using SeqKit (Shen et al. 2016) and Seqtk (Li), and the statistics are summarized in Table S1. Long reads from RS II were assembled using SMART analysis software (Pacific Biosciences) with the parameters: genome size, 40 M; target coverage, 25; and polymerase read quality, 0.75. In hybrid assembly, all short and long reads were assembled using MaSuRCA (ver. 3.3.1) (Zimin et al. 2017) and evaluated using QUAST (Gurevich et al. 2013) and BUSCO (ver. 3.1.0) (Simão et al. 2015) using chlorophyta_odb10. After removing chloroplast and mitochondrial genomes from the assembled genomes, using BLAST, repeat sequences were predicted using RepeatModeler (Flynn et al. 2020) and then softmasked; that is, repeat sequences of genomes were converted to lower case bases, using RepeatMasker. After quality control of the RNA-seq data using AfterQC (Chen et al. 2017), the filtered reads were mapped to organelle-eliminated genomes using HISAT2 (ver. 2.1.0) (Kim et al. 2019). Mapped SAM format data were sorted and converted to BAM format. Softmasked genome and sorted BAM files were used as inputs to gene prediction using BRAKER2 (ver. 2.1.0) (Brůna et al. 2021), and predicted genes were evaluated using Python scripts and BUSCO. Representative amino acid sequences of each gene were extracted from results of BRAKER2 using GffRead (Pertea & Pertea 2020) and Python scripts. The extracted proteome data were annotated using BLAST against the UniProt Reviewed (SwissProt) and *Arabidopsis thaliana* proteome (UP000006548), DIAMOND (Buchfink et al. 2021) search against the NCBI nr database, and InterProScan. Taxonomies of the top hit genes in DIAMOND against the nr database were summarized and visualized in a pie chart using Krona (Ondov et al. 2011).

Comparative genomics

Gene Ontology (GO) terms were linked to each gene by integrating the results of the annotation against SwissProt and InterProScan using Blast2GO and summarized and visualized in a bar chart using WEGO (Ye et al. 2018). For synteny analysis, whole-genome pairwise alignments of *C. variabilis*, *C. vulgaris*, and *C. sorokiniana* were performed and visualized as dot plots using D-GENIES (Cabanettes & Klopp 2018). Ortholog groups (OGs) were identified from all protein sequences of *C. variabilis*, *C. vulgaris*, and *C. sorokiniana* using OrthoFinder (ver. 2.3.3) (Emms & Kelly 2019). These OGs were visualized using Venn diagram produced with the R package VennDiagram, and the *C. variabilis*-specific lost, gained, and duplicated genes were extracted. To perform gene set enrichment analysis (GSEA) of *C. variabilis*-specific lost OGs, conserved only in *C. vulgaris* and *C. sorokiniana*, *A. thaliana* UniProt ids linked corresponding *C. vulgaris* representative genes were used as input for Metascape (Zhou et al. 2019) using GO Molecular Function, Biological Process, and Cellular Component annotations. *C. variabilis*-specific duplication OG was defined as the case where the copy number of a gene of *C. variabilis* was greater than that of both *C. sorokiniana* and *C. vulgaris* and the copy number of a gene of either *C. sorokiniana* or *C. vulgaris* was non-zero.

Phylogenetic tree construction

Proteome sequence data of *Auxenochlorella protothecoides* UTEX 25, *Coccomyxa subellipsoidea* C-169 (Blanc et al. 2012), and *Micractinium conductrix* (Arriola et al. 2018) were downloaded from NCBI BioProject PRJNA261964, PRJNA221161, and PRJNA290385, respectively. Proteomes of *Chlorella* sp. A99 (Hamada et al. 2018) and *Helicosporidium* sp. were downloaded from the website of the Okinawa

Institute of Science and Technology (https://www.oist.jp/) and Supporting Information Data S13 of Pombert et al. (2014), respectively. The remaining proteome data were predicted from genomes using Augustus (ver. 3.3.2) (Stanke & Waack 2003) with *C. vulgaris* as training data, generated from gene prediction using BRAKER2. The genome of *Chlorella* sp. ArM0029B, *Auxenochlorella pyrenoidosa* (Fan et al. 2015), *Parachlorella kessleri* (Ota et al. 2016), *Prototheca cutis*, *Prototheca stagnorum* (Suzuki et al. 2018), and *Prototheca zopfii* (Severgnini et al. 2018) were downloaded from NCBI BioProject PRJNA214256, PRJNA171991, PRJDB3487, PRJDB3669, PRJDB3715, and PRJNA388740, respectively. Single-copy orthologs were identified from the proteomes of 14 species using OrthoFinder (ver. 2.2.3). In the protein sequence of each ortholog, non-homologous regions were detected and masked using PREQUAL (Whelan et al. 2018). They were aligned using MAFFT (Katoh & Standley 2013), and a maximum-likelihood tree was constructed using IQ-TREE (ver. 1.6.5) (Minh et al. 2020), adapting the LG + F + G4 protein substitution model.

Molecular evolution analysis

Single-copy orthologs were identified from proteome sequence data of *C. variabilis*, *C. vulgaris*, and *C. sorokiniana* as the outgroup. Their protein-coding DNA sequences and aligned protein sequences using MAFFT were prepared as input for PAL2NAL (Suyama et al. 2006). This software converted multiple sequence alignments of proteins and their corresponding DNA sequences into codon alignments. The Ds, Dn, and Dn/Ds values of each branch were calculated using the CODEML included PAML package (Yang 2007) with the free ratio model. The genes with alignment length shorter than 150 bp, with Ds or Dn values higher than 2, or with Dn/Ds higher than 5 were excluded. These values were visualized in violin plots and two-dimensional density distributions using the R package ggplot. To detect positively selected genes, the branch-site model of CODEML was used, setting *C. variabilis* as the foreground branch and *C. vulgaris* and *C. sorokiniana* as the background branches. Maximum-likelihood ratio tests were performed comparing two models: the null model assuming all codons in all branches with $Dn/Ds \leq 1$ and the alternative model assuming some codons in the foreground branch with $Dn/Ds > 1$. The p-value was calculated using chi-square statistics and adjusted for multiple testing using the false discovery rate (FDR) method. Genes with adjusted p-value < 0.05 were identified as candidate positively selected genes.

Transcriptome analysis

In a preparation of free-living conditions (FL), *C. variabilis* strain Kb1, previously isolated and cloned (Higuchi et al., 2018), was cultured in C medium (Ichimura, 1971) supplemented with 0.03% (w/v) L-serine under 12-h light and 12-h dark conditions at 25℃ (Kato & Imamura 2008). For symbiotic conditions, *Paramecium bursaria* strain Pb-Kb1, having *C. variabilis* strain Kb1 inside the cells, were cultured monoxenically with *Chlorogonium capillatum* (NIES-3374) as the sole food source, in sterilized Volvic® supplemented with 1g/L sodium acetate and 5 g/L yeast extract under 12-h light and 12-h dark conditions at 25℃. The *C. variabilis* cells were isolated by gently disrupting the cells of *P. bursaria* using an ultrasonic disrupter Handy Sonic UR-21P (Tomy Seiko Co. Ltd., Tokyo, Japan). After four and eight days of incubation, collected cells were used as symbiotic log and stationary phase conditions (SL and SS), respectively. All experiments were performed in duplicate. Filtered reads were mapped to the genome

of *C. variabilis* using HISAT2 (ver. 2.1.0) and counted per gene using featureCounts included in the Subread package (Liao et al. 2014). Using the R package edgeR (Robinson et al. 2010), raw count data were filtered using the filterByExpr function and normalized using calcNormFactors. Genes with FDR-adjusted p-value < 0.05 were extracted as significantly differentially expressed genes (DEGs) using the glmLRT function. Heat maps, dendrograms, and trend lines were generated by the pheatmap function included in the R package gplots and ggplot, respectively. GSEA was performed for genes belonging to each Cluster, with corresponding *A. thaliana* UniProt id by Metascape using GO Molecular Function, Biological Process, and Cellular Component annotations.

Results

De novo **genome assembly and annotation**

To assemble unfragmented and non-redundant genomes of *C. variabilis*, *C. vulgaris*, and *C. sorokiniana*, we adopted a hybrid assembly technique using short- and long-read data. The read data of *C. variabilis* and *C. sorokiniana* were acquired from the public database. We sequenced the genomic DNA of *C. vulgaris* using PacBio RS II and Hiseq X Ten. RS II produced approximately 2.6 Gbp long-read data with an average length of 6.7 kbp, and HiSeq X produced 7.5 Gbp short-read data (Table S1). We assembled the hybrid data of the three species and evaluated the degree of fragmentation of these assemblies using indexes such as the number of contigs, the largest contig size, and N50. In the assembly of *C. vulgaris*, the number of contigs decreased from 237 to 91, the largest contig size increased from approximately 1.7 to 4.1 Mbp, and N50 increased from approximately 0.6 to 1.7 Mbp in the assembly using hybrid data, compared with the assembly using only long-read data. To connect the contigs and reduce the number of redundant contigs, we scaffolded each assembly of the three species and evaluated the efficiency of assembly by measuring the preservation rate of Chlorophyta universal single-copy orthologs using BUSCO (Seppey et al. 2019). The preservation rates of *C. variabilis*, *C. vulgaris*, and *C. sorokiniana* were 93.3%, 93.1%, and 92.8%, respectively, indicating that these genomes correctly reflect the original gene repertoire of each species. All proportions of redundantly predicted single-copy orthologs were less than 1% for the three species, indicating that these genomes correctly reflected the copy number of each gene (Table S2). Assembly of the genomes showed that the genome of *C. variabilis* is about 6 Mbp larger than that of *C. vulgaris*.

We predicted the gene structures—exons, introns, and repeat sequences—from the assembled genomes of three species and added their functional annotations. To obtain accurate gene structures, we predicted the structures based on expression data from RNA-sEq. The average lengths of Chlorophyta universal single-copy orthologs obtained from the genomes of the three species were more than 3 kbp, the number of exons was 8, and the preservation rate was 93%, showing that the structure of each gene was accurately predicted (Table S3). We found that *C. variabilis* carried 1,475 more genes than *C. vulgaris*. These predicted genes were functionally annotated using the UniProtKB, *Arabidopsis thaliana* proteome, NCBI non-redundant (nr), and InterProScan databases for gene names, GSEA, contamination check, and domain search, respectively. In the homology search against the nr database, the proportions of genes

originating from bacteria and viruses were less than 1%, showing that the genomes of the three species were free from contamination (Fig. S1). Next, we identified repeat sequences in the genomes of the three species using a *de novo* approach. We found that *C. variabilis* possessed 5% more repeats than the others (Table S4). The genome of *C. variabilis* was 6 Mbp larger than that of *C. vulgaris*, probably due to the presence of repeat sequences. The total length of the repeat sequences in the *C. variabilis* genome was 3 Mbp larger than that of *C. vulgaris*. *C. variabilis* possessed 1,475 more genes than *C. vulgaris*, and multiplying 1,475 by the average length of the *C. variabilis* gene region (2,954) yields about 4.5 Mbp. The statistics of the three genome datasets generated are summarized in Table 1.

Phylogeny and molecular evolution of C. variabilis

To confirm a phylogenetic relationship between *C. variabilis* and *C. vulgaris*, we constructed a tree using 449 single-copy orthologs from 14 Trebouxiophyceae species including *Coccomyxa subellipsoidea* as the outgroup. The tree generated indicated that *C. vulgaris* is most closely related to *C. variabilis* (Fig. 1A). The terminal branch lengths of *C. variabilis* and *C. vulgaris* were 0.077 and 0.123, respectively, indicating that the mutation rate of *C. variabilis* was lower than that of *C. vulgaris*. The Ds value of *C. variabilis*, representing the number of mutations per synonymous site, was lower than that of *C. vulgaris*. The two species had similar Dn values (the number of mutation per non-synonymous site) and Dn/Ds ratios, so positively selected genes were not detected using the likelihood ratio test (Fig. 1B).

Comparative genomics of C. variabilis **and** C. vulgaris

To analyze the synteny between the three species at the whole-genome level, we produced dot plots. In comparisons between *C. variabilis* and *C. vulgaris*, each dot representing a homologous match between the two sequences was located on the diagonal more closely than the other comparison. These results showed that syntenies between *C. variabilis* and *C. vulgaris* are conserved throughout their genomes without large-scale insertions, deletions, inversions, or duplications (Fig. 1B). This finding also indicates that *C. variabilis* and *C. vulgaris* are closely related.

To identify the genetic features of *C. variabilis*, we compared the profiles of their GO terms and inferred orthologous gene groups among the three species. GO terms for each gene were derived from domain annotation, and these results were summarized in a bar chart. In the bar chart, most GO terms had the same patterns, but the “extracellular region part” term was presented only in *C. variabilis* (Fig. 2A). In ortholog analysis, we identified OGs, which are defined as groups of genes descended from a single ancestral gene (Emms & Kelly 2019). A total of 8,931 OGs were identified, containing 85.9% of all predicted genes among 3 species, and the remaining genes were not assigned to any OGs and were hence designated as singletons. These OGs were classified by conserved patterns among the three species and visualized in a Venn diagram (Fig. 2B). In the Venn diagram, 7,412 OGs containing 83.0% of all identified OGs were conserved among the 3 species, suggesting that the species are closely related. We then focused on three groups composed of OGs and singletons: *C. variabilis*-specific lost, gained, and duplicated genes.

The *C. variabilis*-specific lost genes were contained in the 287 OGs conserved only in *C. vulgaris* and *C. sorokiniana* (Fig. 2B). To estimate the functions included in these OGs, we conducted GSEA, which identifies significantly enriched biological terms using GO terms, resulting in the identification of “cellular response to reactive oxygen species (ROS)” being the most enriched (Fig. S2A). Superoxide dismutase (SOD, OG0002817), which catalyzes the dismutation of superoxide radicals into oxygen and hydrogen peroxide; monodehydroascorbate reductase (MDR, OG0008335), which catalyzes the conversion of monodehydroascorbate to ascorbate of major antioxidant; and peptide-methionine (R)-S-oxide reductase (OG0008244), which protects against oxidation of proteins, were lost (Table S5). The *C. variabilis*-specific gained genes were contained in 10 OGs with 39 genes and 1,426 singletons conserved only in *C. variabilis* (Fig. 2B). In these OGs, “protein phosphatase 1 binding” was the most enriched term, as identified using GSEA (Fig. S2B). The *C. variabilis*-specific duplicated genes were contained in 297 OGs with 978 genes extracted from the 3 categories of OGs in the Venn diagrams: 7,412 OGs with 8,655 genes conserved among 3 species, 394 OGs with 510 genes conserved only in *C. variabilis* and *C. vulgaris*, and 805 OGs with 1,004 genes conserved only in *C. variabilis* and *C. sorokiniana* (Fig. 2C). In these OGs, “calmodulin-dependent protein kinase activity” was the most enriched, as identified using GSEA (Fig. S2C). Among *C. variabilis*-specific gained and duplicated genes, we analyzed the transcriptome under symbiotic conditions to identify the genes involved in the symbiotic lifestyle.

Transcriptome dynamics of C. variabilis under symbiosis

To identify DEGs under symbiotic conditions, we compared RNA-seq data derived from the three conditions illustrated in Fig. S3, namely, *C. variabilis*, cultured under non-symbiotic conditions without host *P. bursaria* (the free-living condition; FL) as a control, cultured under symbiotic conditions in the *P. bursaria* log phase (the symbiotic and log phase condition; SL), and cultured under symbiotic conditions in the *P. bursaria* stationary phase (the symbiotic and stationary phase condition; SS). More than 98% of reads remained after filtering, and more than 86% of reads mapped to the genome, showing that the sequencing process was effective, and six samples had no contamination. These mapped reads were counted by gene, and after filtering and normalization, we extracted 4,401 significantly DEGs from the pairwise comparison among the 3 conditions, using the generalized likelihood ratio test.

We calculated the relative expression values for each DEG, performed hierarchical clustering based on the expression patterns of each gene, and visualized these results with a heat map and dendrogram (Fig. 3A). The 4,401 DEGs were classified into four Clusters with similar expression patterns (Fig. 3B). Under symbiotic conditions, 2,085 genes were downregulated (Cluster 1) and 1,038 genes were upregulated (Cluster 2). In the SL condition, 1,114 genes were upregulated (Cluster 3), and in SS conditions, 164 genes were upregulated (Cluster 4). Cluster 1 included genes involved in cellular quality control of DNA and proteins, with GO terms such as “cellular response to DNA damage stimulus,” “damaged DNA binding,” and “proteolysis” being strongly enriched in GSEA (Fig. S4A). Cluster 2 included genes related to rRNA and tRNA, with GO terms “ncRNA metabolic process” identified by GSEA (Fig. S4B), suggesting activation of protein synthesis. Cluster 3 included genes linked to GO terms such as “cell cycle,” “DNA replication,” and “chromosome” (Fig. S4C), showing that *C. variabilis* proliferates in synchronization with host cell

division, a finding that is consistent with previous reports. The genes localized in the chloroplast were contained in Clusters 2, 3, and 4 (Fig. S4B, C, and D). Genes related to chlorophyl belonged to Cluster 4 (Fig. S4D), showing the specific activation of photosynthesis under symbiotic conditions.

Multi-omics analysis of C. variabilis

To identify genes playing a crucial role under symbiotic conditions, we combined the results of ortholog and transcriptome analysis and listed them in the Supplemental Data. *C. variabilis*-specific gained or duplicated OGs and DEGs were extracted. Among the *C. variabilis*-specific 1,465 genes gained, 468 DEGs were included, composed of 237, 106, 102, and 23 DEGs belonging to Clusters 1, 2, 3, and 4, respectively. Only 95 of the 468 DEGs were annotated against a gene in the BLAST search of UniProtKB, including chitosanase (OG0011577; g11572), which degrades the chitosan constituting the cell wall of *Chlorella*, included in Cluster 2, chloride channel (OG0011693; g572) included in Cluster 3, and permease (OG0012137; g2843), a membrane transport protein included in Cluster 4 (Table S6). Unannotated OG0000793 was composed of five genes, and this OG was specifically gained and duplicated in *C. variabilis*. OG0000793 contained genes with different expression patterns, with three genes belonging to Clusters 2, 3, and 4, respectively (g5461, g6542, and g11141), and the other two genes (g3072 and g3111) were not DEGs, suggesting that they play different roles, or subfunctionalization (Table S6).

Among the *C. variabilis*-specific 978 duplicated genes in the 297 OGs, 406 DEGs of 204 OGs were included, composed of 203 genes from 130 OGs, 81 from 68 OGs, 102 from 68 OGs, and 19 DEGs from 16 OGs belonging to Clusters 1, 2, 3, and 4, respectively. We focused on 152 OGs including 202 DEGs belonging to Clusters 2, 3, and 4, because the expression of these genes was upregulated under symbiotic conditions, indicating a high likelihood of involvement in symbiosis. In the 152 OGs, the genes were related to cell wall biogenesis or degradation and metabolic exchange with the host. The former included alpha-galactosidase (OG0000160), arabinosyltransferase (OG0000063), chitosanase (OG0000466), and beta-glucan synthesis-associated protein (OG0000291). The latter facilitates mutual transport of certain metabolites between *C. variabilis* and *P. bursaria*, including efflux transmembrane transporter (OG0000073) transferring a specific substance from the inside of the cell to the outside, proton/sulfate cotransporter (OG0000123), sugar-phosphate/phosphate translocator (OG0001113), transporter of a specific amino acid (OG0001180), ammonium transporter (OG0000350), and glutamate dehydrogenase (OG0001044), which assimilates ammonium to glutamate.

Symbiosis gene network and pathways

To investigate the pathways participating in symbiosis, we automatically extracted the genes corresponding to each component constituting pathways from integrated data of ortholog and transcriptome analysis (Supplemental Data 1) and plotted their expression patterns on the map from the KEGG Pathway database. These genes were manually curated and are summarized in Table S7. In the pathway “porphyrin and chlorophyl metabolism,” the expression of genes encoding 16 enzymes constituting the biosynthetic pathway of chlorophyl a and b from glutamate tended to increase in SS-

specific conditions (Fig. 4A). Among 30 genes corresponding to 16 enzymatic reactions, 16 genes were DEGs, including 9, 2, and 5 genes belonging to Clusters 2, 3, and 4, respectively. OG0000400 and OG0000788, corresponding to coproporphyrinogen III oxidase (1.3.3.3) and magnesium-protoporphyrin IX monomethyl ester (oxidative) cyclase (1.14.13.81), respectively, were *C. variabilis*-specific duplicated and contained DEGs belonging to Clusters 2 and 3. In addition to chlorophyl biosynthesis, all genes encoding chlorophyl-binding subunits of LHC, which enables efficient absorption of light energy, were highly expressed in SS-specific conditions (Fig. 4B). These results suggest that the increase in chlorophyl and LHC content enhances light absorption and conversion into energy, which then promotes photosynthesis.

In the pathway “carbon fixation in photosynthetic organisms,” the gene expression of 13 enzymes constituting the Calvin-Benson cycle tended to increase under both SL and SS conditions, compared with FL conditions (Fig. 4C). Among 25 genes of 20 OGs corresponding to 13 enzymatic reactions, 17 genes were DEGs, including 2, 12, and 3 genes belonging to Clusters 1, 2, and 3, respectively. The two copies of the ribulose-bisphosphate carboxylase, or Rubisco (4.1.1.39), gene were both DEGs in Cluster 2. The upregulation of each gene in the Calvin cycle, such as Rubisco, leads to the synthesis of a large number of corresponding enzymes, suggesting enhancement of carbon fixation, which then promotes photosynthesis. In “starch and sucrose metabolism,” the expression patterns of the genes encoding eight enzymes constituting the biosynthetic pathway of maltose from D-fructose-6P tended to increase in SL-specific conditions (Fig. 4D). Among 21 genes of 15 OGs corresponding to 8 enzymatic reactions, 13 genes were DEGs, including the 2, 3, and 8 genes belonging to Clusters 1, 2, and 3, respectively. The genes encoding endoglucanase (3.2.1.4) and beta-glucosidase (3.2.1.4) were DEGs in Cluster 3, suggesting acceleration of the degradation from cellulose to glucose under SL-specific conditions. The gene encoding glucose-6-phosphate isomerase (5.3.1.9) was a DEG in Cluster 2, indicating that maltose is produced from D-fructose-6P in the Calvin cycle via starch. The enzymes degrading starch into maltose include alpha-amylase (3.2.1.1), beta-amylase (3.2.1.2), and isoamylase (3.2.1.68). All the genes of alpha-amylase were DEGs and may play a particularly crucial role in supplying maltose to *P. bursaria*. The gene encoding alpha-glucosidase (3.2.1.20) degrading maltose is a DEG in Cluster 1, indicating the suppression of maltose degradation under endosymbiosis.

Discussion

In this study, we investigated the genetic changes occurring in the second stage of the move from facultative to obligate endosymbiosis and the genetic mechanisms of the endosymbiotic lifestyle, using *C. variabilis*. The genetic change, contrary to our hypothesis, did not involve the reduction of the size of the genome of *C. variabilis*. The genes involved in the biosynthesis and degradation of the cell wall and the metabolic exchange with the host play crucial roles in endosymbiosis. The genetic mechanism underlying the enhancement of photosynthetic capacity during symbiosis increases light absorption efficiency and carbon fixation capacity, resulting in a massive supply of maltose to the *P. bursaria*.

Quality of the genome datasets

For comparative genomics, we prepared high-quality genome datasets of *C. variabilis*, *C. vulgaris*, and *C. sorokiniana*, organisms that are closely related, using the same pipeline. This approach ensured the accuracy of the findings in this study. The three gene models predicted from the unfragmented and non-redundant genomes (Table S2) correctly reflected the original gene repertoire and copy number (Table S3), and contamination was not detected (Fig. S1). Although the taxonomy of the genus *Chlorella* is unresolved (Heeg & Wolf 2015), the results of the phylogenetic tree (Fig. 1A), synteny (Fig. 2A), and ortholog inference (Fig. 2C) confirm the close relatedness of the three species. Several genomes of other strains of *C. vulgaris* have been reported (Guarnieri et al. 2018; Cecchin et al. 2019), but we sequenced NIES-686 (= SAG 211-11b; CCAP 211/11b; UTEX 259), the authentic strain of *C. vulgaris*.

Non-reduction of the genome of C. variabilis

Genome comparisons showed that *C. variabilis* has a larger genome size (6 Mbp), more genes (1,475) (Table 1), and no higher mutation rate (Fig. 1A) than *C. vulgaris*, and contrary to expectations, we did not observe a presage of large-scale genome reduction in the third stage of secondary endosymbiosis. However, *C. variabilis* lost some ROS-related genes (Fig. S2A and Table S5), and the expressions of genes involved in cellular quality control of DNA and protein were downregulated under symbiotic conditions compared to non-symbiotic conditions (Figs. 3 and S4A), which may presage genome reduction. The fact that *C. variabilis* and *P. bursaria* can be artificially cultured separately, and their symbiotic associations can be reconstructed (Siegel 1960), suggests that their endosymbiosis is in the relatively early part of stage 2. The larger number of genes in *C. variabilis* probably derives from gene gain or duplication (Fig. 2B), which may have arisen before the establishment of endosymbiosis with *P. bursaria* rather than after it. As a result, the adaptive capacity of *C. variabilis* has probably been enhanced, enabling endosymbiotic life. *C. variabilis*-specific gained and duplicated genes were upregulated under symbiosis (Table S6 and Supplemental Data), suggesting that they play an important role in endosymbiosis. It has been reported that the properties of *C. variabilis* are variable and the organism is capable of environmental adaptation (Krauss & Shihira 1965).

C. variabilis **-specific gained and duplicated DEGs**

Ortholog and transcriptome analyzes identified *C. variabilis*-specific gained and duplicated genes and OGs, some of which were upregulated under symbiotic conditions compared to non-symbiotic conditions (Supplemental Data). These genes included some related to cell wall biosynthesis and degradation and metabolic exchange with *P. bursaria* (Table S6), an observation that suggests that they play an important role in endosymbiosis. The structure of the cell wall may have required adaptation to the intracellular environment of *P. bursaria*, which is different from that under FL conditions. In this study, *C. variabilis* possessed the highest number of genes associated with the GO term “extracellular region part,” compared to *C. vulgaris* and *C. sorokiniana* (Fig. 2B). Pathway analysis indicated that cellulose is degraded into glucose under symbiosis (Fig. 4D and Table S7). It has been reported that the composition of the cell wall is different between symbiotic and free-living *Chlorella* (Takeda 1995) and the thickness of the cell wall of *C. variabilis* under symbiotic conditions is about half compared with that in non-symbiotic conditions and

has a different composition (Higuchi et al. 2018). We investigated the metabolites exchanged using transfer proteins, based on the results of the homology search (Table S6). The supplied ammonium (OG0000350 and OG0001044) and amino acids (OG0001180) are consistent with a previous study showing that *C. variabilis* can utilize them as a nitrogen source (Kamako et al. 2005). To the best of our knowledge, this is the first time that the exchange of chloride (OG0011693), sugar-phosphate/phosphate (OG0001113), and sulfate (OG0000123) has been reported. Sulfurate exchange has been reported in symbiotic relationships between corals-algae (Yuyama & Watanabe 2008) and hydra-algae (Cook 1976). The *C. variabilis*-specific gained/duplicated DEGs whose functions were not predictable using homology search (e.g., OG0000793) are expected to reveal important findings through further analysis using other methods.

Genetic mechanism of the enhancement of photosynthesis

The pathway analysis found that the expression of *C. variabilis* genes corresponding to components the chlorophyl synthesis, LHC, Calvin cycle, and D-fructose-6P to maltose pathways tended to be upregulated under endosymbiosis (Fig. 4 and Table S7). The increase in chlorophyl (Fig. 4A) and LHC (Fig. 4B) content probably enhances light absorption and conversion into energy. The upregulated expression of genes encoding all of the enzymes composing the Calvin cycle (Fig. 4C) probably enhances carbon fixation. These results suggest that the enhancement of photosynthesis by increasing light absorption efficiency and carbon fixation capacity underlies the massive supply of maltose to *P. bursaria*. Muscatine, Karakashian, and Karakashian (1967) reported that *C. variabilis* excretes about half of its total photosynthate as maltose outside the cell, for which the enhancement of photosynthesis is probably required. *P. bursaria* exhibits positive phototaxis (Reisser & Häder 1984) and provides CO_2 by respiration (Reisser 1980), showing properties corresponding to the enhancement of light absorption and carbon fixation of *C. variabilis*. An increase in chlorophyl content and carbon fixation rate in *C. variabilis* under endosymbiosis with *P. bursaria* has been reported (Liang et al. 2020).

The gene expression pattern of each pathway was different. Genes constituting the Calvin cycle tended to be upregulated in both SS and SL conditions (Fig. 4C), genes constituting the pathway of chlorophyl synthesis (Fig. 4A) and LHC (Fig. 4C) tended to be upregulated specifically in SS conditions, and genes constituting the pathway from D-fructose-6P to maltose tended to be upregulated specifically in SL phase (Fig. 4D). These results suggest that *P. bursaria* suppresses the synthesis of chlorophyl and LHC in *C. variabilis* by decreasing the supply of the nitrogen source, glutamine, to *C. variabilis*, and promoting the supply of maltose to *P. bursaria* in *C. variabilis* during their proliferation. Chlorophyll synthesis is closely linked to nitrogen content (Liang et al. 2020), but *C. variabilis* can only utilize NH_3 and amino acids as nitrogen sources (Kamako et al. 2005). It has been reported that *C. variabilis* can utilize glutamine among amino acids (Hamada et al. 2018). Glutamate, which is the starting material for chlorophyl synthesis, can be synthesized from glutamine. He et al. (2019) found that the RNAi knockdown of glutamate synthase in *P. bursaria* reduced the number of symbiotic *C. variabilis*. Therefore, *P. bursaria* may control chlorophyl synthesis in *C. variabilis* according to the level of glutamine supplied to *C. variabilis*.

In conclusion, we found a non-reduced genome of *C. variabilis*, indicating that the symbiotic relationship between *C. variabilis* and *P. bursaria* is relatively early in stage 2 of secondary endosymbiosis. We identified the genes and pathways playing crucial roles under conditions of endosymbiosis. However, this study has several limitations. To characterize the stage of symbiosis between *C. variabilis* and *P. bursaria*, further comparative analysis with other symbiotic endosymbiotic relations is needed. Algae closely related to *C. variabilis* include several species that, as in *C. variabilis*, engage in endosymbiosis with specific protists, which have arisen as phylogenetically independent organisms (Hoshina & Kusuoka 2016). Additional functional analyzes are needed, using techniques such as genome editing of genes related to symbiosis, as identified in this study. However, our study sheds new light on the second stage of secondary endosymbiosis, using genomic comparisons between closely related symbiotic and non-symbiotic species and their transcriptomes under endosymbiosis. We believe that our findings will play a leading role in providing a basis for uncovering the evolution of chloroplast acquisition by secondary symbiosis.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and analysed during the current study are available in the DDBJ Sequence Read Archive, with the BioProject Accession ID: PRJDB12190.

Competing interests

The authors declare that they have no potential conflict of interest. Dr. Ogura, associate editor for BMC genomics, was not involved in the editorial review of or decision to publish this article.

Funding

This work was supported by Grant-in-Aid for Challenging Exploratory Research 20K21452, Grant-in-Aid for Scientific Research (C) 16K07491 to RH, Grant-in-Aid for Scientific Research (C) 19K06814 to RH, and Grant-in-Aid for JSPS Research Fellow 19J15617 to RM.

Authors' contributions

A.O. and R.Hoshina conceived of the presented idea. R.M., carried out the experiment and analysis. R.M. wrote the manuscript with the support from A.O. and R.Hoshina. R. Higuchi, L.C., T.S contributed to

sample preparation. Y.A., Y.T., S.K., A.Y. and M.K. contributed genome analyses. All authors discussed the results and contributed to the final manuscript.

Acknowledgements

Not applicable

References

1. Adl, Sina M., David Bass, Christopher E. Lane, Julius Lukeš, Conrad L. Schoch, Alexey Smirnov, Sabine Agatha, et al. 2019. “Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes.” *The Journal of Eukaryotic Microbiology* 66 (1): 4–119.
2. Arriola, Matthew B., Natarajan Velmurugan, Ying Zhang, Mary H. Plunkett, Hanna Hondzo, and Brett M. Barney. 2018. “Genome Sequences of Chlorella Sorokiniana UTEX 1602 and Micractinium Conductrix SAG 241.80: Implications to Maltose Excretion by a Green Alga.” *The Plant Journal: For Cell and Molecular Biology* 93 (3): 566–86.
3. Blaby-Haas, Crysten E., and Sabeeha S. Merchant. 2019. “Comparative and Functional Algal Genomics.” *Annual Review of Plant Biology* 70 (April): 605–38.
4. Blanc, Guillaume, Irina Agarkova, Jane Grimwood, Alan Kuo, Andrew Brueggeman, David D. Dunigan, James Gurnon, et al. 2012. “The Genome of the Polar Eukaryotic Microalga Coccomyxa Subellipsoidea Reveals Traits of Cold Adaptation.” *Genome Biology* 13 (5): R39.
5. Blanc, Guillaume, Garry Duncan, Irina Agarkova, Mark Borodovsky, James Gurnon, Alan Kuo, Erika Lindquist, et al. 2010. “The Chlorella Variabilis NC64A Genome Reveals Adaptation to Photosymbiosis, Coevolution with Viruses, and Cryptic Sex.” *The Plant Cell* 22 (9): 2943–55.
6. Brůna, Tomáš, Katharina J. Hoff, Alexandre Lomsadze, Mario Stanke, and Mark Borodovsky. 2021. “BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP + and AUGUSTUS Supported by a Protein Database.” *NAR Genomics and Bioinformatics* 3 (1): lqaa108.
7. Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost. 2021. “Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND.” *Nature Methods* 18 (4): 366–68.
8. Cabanettes, Floréal, and Christophe Klopp. 2018. “D-GENIES: Dot Plot Large Genomes in an Interactive, Efficient and Simple Way.” *PeerJ* 6 (June): e4958.
9. Cecchin, Michela, Luca Marcolungo, Marzia Rossato, Laura Girolomoni, Emanuela Cosentino, Stephan Cuine, Yonghua Li-Beisson, Massimo Delledonne, and Matteo Ballottari. 2019. “Chlorella Vulgaris Genome Assembly and Annotation Reveals the Molecular Basis for Metabolic Acclimation to High Light Conditions.” *The Plant Journal: For Cell and Molecular Biology* 100 (6): 1289–1305.
10. Chen, Shifu, Tanxiao Huang, Yanqing Zhou, Yue Han, Mingyan Xu, and Jia Gu. 2017. “AfterQC: Automatic Filtering, Trimming, Error Removing and Quality Control for Fastq Data.” *BMC Bioinformatics* 18 (Suppl 3): 80.

11. Cook, Clayton B. 1976. “Sulfate Utilization in Green Hydra.” In *Coelenterate Ecology and Behavior*, edited by G. O. Mackie, 415–22. Boston, MA: Springer US.
12. Curtis, Bruce A., Goro Tanifuji, Fabien Burki, Ansgar Gruber, Manuel Irimia, Shinichiro Maruyama, Maria C. Arias, et al. 2012. “Algal Genomes Reveal Evolutionary Mosaicism and the Fate of Nucleomorphs.” *Nature* 492 (7427): 59–65.
13. Emms, David M., and Steven Kelly. 2019. “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics.” *Genome Biology* 20 (1): 238.
14. Fan, Jianhua, Kang Ning, Xiaowei Zeng, Yuanchan Luo, Dongmei Wang, Jianqiang Hu, Jing Li, et al. 2015. “Genomic Foundation of Starch-to-Lipid Switch in Oleaginous Chlorella Spp.” *Plant Physiology* 169 (4): 2444–61.
15. Flynn, Jullien M., Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. 2020. “RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (17): 9451–57.
16. Gray, M. W. 1999. “Evolution of Organellar Genomes.” *Current Opinion in Genetics & Development* 9 (6): 678–87.
17. Guarnieri, Michael T., Jennifer Levering, Calvin A. Henard, Jeffrey L. Boore, Michael J. Betenbaugh, Karsten Zengler, and Eric P. Knoshaug. 2018. “Genome Sequence of the Oleaginous Green Alga, Chlorella Vulgaris UTEX 395.” *Frontiers in Bioengineering and Biotechnology* 6 (April): 37.
18. Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics* 29 (8): 1072–75.
19. Hamada, Mayuko, Katja Schröder, Jay Bathia, Ulrich Kürn, Sebastian Fraune, Mariia Khalturina, Konstantin Khalturin, Chuya Shinzato, Nori Satoh, and Thomas Cg Bosch. 2018. “Metabolic Co-Dependence Drives the Evolutionarily Ancient Hydra-Chlorella Symbiosis.” *eLife* 7 (May). https://doi.org/10.7554/eLife.35122.
20. Heeg, Jaqueline S., and Matthias Wolf. 2015. “ITS2 and 18S rDNA Sequence-Structure Phylogeny of Chlorella and Allies (Chlorophyta, Trebouxiophyceae, Chlorellaceae).” *Plant Gene* 4 (December): 20–28.
21. He, Ming, Jinfeng Wang, Xinpeng Fan, Xiaohui Liu, Wenyu Shi, Ning Huang, Fangqing Zhao, and Miao Miao. 2019. “Genetic Basis for the Establishment of Endosymbiosis in Paramecium.” *The ISME Journal* 13 (5): 1360–69.
22. Higuchi, Rina, Chihong Song, Ryo Hoshina, and Toshinobu Suzaki. 2018. “Endosymbiosis-Related Changes in Ultrastructure and Chemical Composition of Chlorella Variabilis (Archaeplastida, Chlorophyta) Cell Wall in Paramecium Bursaria (Ciliophora, Oligohymenophorea).” *European Journal of Protistology* 66 (October): 149–55.
23. Hoshina, Ryo, and Yuko Fujiwara. 2013. “Molecular Characterization of Chlorella Cultures of the National Institute for Environmental Studies Culture Collection with Description of Micractinium

Inermum Sp. Nov., Didymogenes Sphaerica Sp. Nov., and Didymogenes Soliella Sp. Nov. (Chlorellaceae, Tr." *Phycological Research* 61 (2): 124–32.

24. Hoshina, Ryo, Mitsunori Iwataki, and Nobutaka Imamura. 2010. "Chlorella Variabilis and Micractinium Reisseri Sp. Nov. (Chlorellaceae, Trebouxiophyceae): Redescription of the Endosymbiotic Green Algae of Paramecium Bursaria (Peniculia, Oligohymenophorea) in the 120th Year." *Phycological Research* 58 (3): 188–201.
25. Hoshina, Ryo, and Yasushi Kusuoka. 2016. "DNA Analysis of Algal Endosymbionts of Ciliates Reveals the State of Algal Integration and the Surprising Specificity of the Symbiosis." *Protist* 167 (2): 174–84.
26. Inouye, Isao, and Noriko Okamoto. 2005. "Changing Concepts of a Plant: Current Knowledge on Plant Diversity and Evolution." *Plant Biotechnology* 22 (5): 505–14.
27. Kadono, T., T. Kawano, H. Hosoya, and T. Kosaka. 2004. "Flow Cytometric Studies of the Host-Regulated Cell Cycle in Algae Symbiotic with Green Paramecium." *Protoplasma* 223 (2–4): 133–41.
28. Kamako, Shin-Ichiro, Ryo Hoshina, Seiko Ueno, and Nobutaka Imamura. 2005. "Establishment of Axenic Endosymbiotic Strains of Japanese Paramecium Bursaria and the Utilization of Carbohydrate and Nitrogen Compounds by the Isolated Algae." *European Journal of Protistology* 41 (3): 193–202.
29. Katoh, Kazutaka, and Daron M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4): 772–80.
30. Kato, Yutaka, and Nobutaka Imamura. 2008. "Effect of Sugars on Amino Acid Transport by Symbiotic Chlorella." *Plant Physiology and Biochemistry: PPB / Societe Francaise de Physiologie Vegetale* 46 (10): 911–17.
31. Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. "Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype." *Nature Biotechnology* 37 (8): 907–15.
32. Kodama, Yuuki, and Masahiro Fujishima. 2007. "Infectivity of Chlorella Species for the Ciliate Paramecium Bursaria Is Not Based on Sugar Residues of Their Cell Wall Components, but on Their Ability to Localize beneath the Host Cell Membrane after Escaping from the Host Digestive Vacuole in the Early Infection Process." *Protoplasma* 231 (1–2): 55–63.
33. Kodama, Yuuki, Haruo Suzuki, Hideo Dohra, Manabu Sugii, Tatsuya Kitazume, Katsushi Yamaguchi, Shuji Shigenobu, and Masahiro Fujishima. 2014. "Comparison of Gene Expression of Paramecium Bursaria with and without Chlorella Variabilis Symbionts." *BMC Genomics* 15 (March): 183.
34. Krauss, R. W., and I. Shihira. 1965. "NASA Technical Reports Server (NTRS)." ntrs.nasa.gov. 1965. https://ntrs.nasa.gov/search.jsp?R=19660005349.
35. Liang, Chengwei, Xiao Yang, Lu Wang, Xiao Fan, Xiaowen Zhang, Dong Xu, and Naihao Ye. 2020. "Different Physiological and Molecular Responses of the Green Algae Chlorella Variabilis to Long-Term and Short-Term Elevated CO2." *Journal of Applied Phycology* 32 (2): 951–66.

36. Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30.
37. Li, Heng. n.d. *Seqtk*. Github. Accessed May 20, 2021. https://github.com/lh3/seqtk.
38. Minei, Ryuhei, Ryo Hoshina, and Atsushi Ogura. 2018. “De Novo Assembly of Middle-Sized Genome Using MinION and Illumina Sequencers.” *BMC Genomics* 19 (1): 700.
39. Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.” *Molecular Biology and Evolution* 37 (5): 1530–34.
40. Muscatine, Leonard, Stephen J. Karakashian, and Marlene W. Karakashian. 1967. “Soluble Extracellular Products of Algae Symbiotic with a Ciliate, a Sponge and a Mutant Hydra.” *Comparative Biochemistry and Physiology* 20 (1): 1–12.
41. Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy. 2011. “Interactive Metagenomic Visualization in a Web Browser.” *BMC Bioinformatics* 12 (September): 385.
42. Ota, Shuhei, Kenshiro Oshima, Tomokazu Yamazaki, Sangwan Kim, Zhe Yu, Mai Yoshihara, Kohei Takeda, et al. 2016. “Highly Efficient Lipid Production in the Green Alga Parachlorella Kessleri: Draft Genome and Transcriptome Endorsed by Whole-Cell 3D Ultrastructure.” *Biotechnology for Biofuels* 9 (January): 13.
43. Pertea, Geo, and Mihaela Pertea. 2020. “GFF Utilities: GffRead and GffCompare.” *F1000Research* 9 (April). https://doi.org/10.12688/f1000research.23297.2.
44. Pombert, Jean-François, Nicolas Achille Blouin, Chris Lane, Drion Boucias, and Patrick J. Keeling. 2014. “A Lack of Parasitic Reduction in the Obligate Parasitic Green Alga Helicosporidium.” *PLoS Genetics* 10 (5): e1004355.
45. Ponce-Toledo, Rafael I., Purificación López-García, and David Moreira. 2019. “Horizontal and Endosymbiotic Gene Transfer in Early Plastid Evolution.” *The New Phytologist* 224 (2): 618–24.
46. Quispe, Cristian F., Olivia Sonderman, Maya Khasin, Wayne R. Riekhof, James L. Van Etten, and Kenneth W. Nickerson. 2016. “Comparative Genomics, Transcriptomics, and Physiology Distinguish Symbiotic from Free-Living Chlorella Strains.” *Algal Research* 18 (September): 332–40.
47. Reisser, Werner. 1980. “The Metabolic Interactions between Paramecium Bursaria Ehrbg. and Chlorella Spec. in the Paramecium Bursaria-Symbiosis.” *Archives of Microbiology* 125 (3): 291–93.
48. Reisser, Werner, and Donat-P Häder. 1984. “Role of Endosymbiotic Algae in Photokinesis and Photophobic Responses of Ciliates.” *Photochemistry and Photobiology* 39 (5): 673–78.
49. Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40.
50. Sarai, Chihiro, Goro Tanifuji, Takuro Nakayama, Ryoma Kamikawa, Kazuya Takahashi, Euki Yazaki, Eriko Matsuo, et al. 2020. “Dinoflagellates with Relic Endosymbiont Nuclei as Models for Elucidating Organellogenesis.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (10): 5364–75.

51. Seppey, Mathieu, Mosè Manni, and Evgeny M. Zdobnov. 2019. “BUSCO: Assessing Genome Assembly and Annotation Completeness.” In *Gene Prediction: Methods and Protocols*, edited by Martin Kollmar, 227–45. New York, NY: Springer New York.
52. Severgnini, Marco, Barbara Lazzari, Emanuele Capra, Stefania Chessa, Mario Luini, Roberta Bordoni, Bianca Castiglioni, Matteo Ricchi, and Paola Cremonesi. 2018. “Genome Sequencing of Prototheca Zopfii Genotypes 1 and 2 Provides Evidence of a Severe Reduction in Organellar Genomes.” *Scientific Reports* 8 (1): 14637.
53. Shen, Wei, Shuai Le, Yan Li, and Fuquan Hu. 2016. “SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation.” *PloS One* 11 (10): e0163962.
54. Siegel, R. W. 1960. “Hereditary Endosymbiosis in Paramecium Bursaria.” *Experimental Cell Research* 19 (March): 239–52.
55. Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” *Bioinformatics* 31 (19): 3210–12.
56. Stanke, Mario, and Stephan Waack. 2003. “Gene Prediction with a Hidden Markov Model and a New Intron Submodel.” *Bioinformatics* 19 Suppl 2 (October): ii215–25.
57. Suyama, Mikita, David Torrents, and Peer Bork. 2006. “PAL2NAL: Robust Conversion of Protein Sequence Alignments into the Corresponding Codon Alignments.” *Nucleic Acids Research* 34 (Web Server issue): W609–12.
58. Suzuki, Shigekatsu, Rikiya Endoh, Ri-Ichiroh Manabe, Moriya Ohkuma, and Yoshihisa Hirakawa. 2018. “Multiple Losses of Photosynthesis and Convergent Reductive Genome Evolution in the Colourless Green Algae Prototheca.” *Scientific Reports* 8 (1): 940.
59. Suzuki, Shigekatsu, Shu Shirato, Yoshihisa Hirakawa, and Ken-Ichiro Ishida. 2015. “Nucleomorph Genome Sequences of Two Chlorarachniophytes, Amorphochlora Amoebiformis and Lotharella Vacuolata.” *Genome Biology and Evolution* 7 (6): 1533–45.
60. Takeda, Hiroshi. 1995. “Cell Wall Composition and Taxonomy of Symbiotic Chlorella from Paramecium and Acanthocystis.” *Phytochemistry* 40 (2): 457–59.
61. Tonooka, Yuki, and Tsuyoshi Watanabe. 2002. “A Natural Strain of Paramecium Bursaria Lacking Symbiotic Algae.” *European Journal of Protistology* 38 (1): 55–58.
62. Whelan, Simon, Iker Irisarri, and Fabien Burki. 2018. “PREQUAL: Detecting Non-Homologous Characters in Sets of Unaligned Homologous Sequences.” *Bioinformatics* 34 (22): 3929–30.
63. Yang, Ziheng. 2007. “PAML 4: Phylogenetic Analysis by Maximum Likelihood.” *Molecular Biology and Evolution* 24 (8): 1586–91.
64. Ye, Jia, Yong Zhang, Huihai Cui, Jiawei Liu, Yuqing Wu, Yun Cheng, Huixing Xu, et al. 2018. “WEGO 2.0: A Web Tool for Analyzing and Plotting GO Annotations, 2018 Update.” *Nucleic Acids Research* 46 (W1): W71–75.
65. Yuyama, Ikuko, and Toshiki Watanabe. 2008. “Molecular Characterization of Coral Sulfate Transporter Homolog That Is up-Regulated by the Presence of Symbiotic Algae.” *Fisheries Science:*

FS 74 (6): 1269–76.

66. Zhou, Yingyao, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K. Chanda. 2019. “Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets.” *Nature Communications* 10 (1): 1523.
67. Zimin, Aleksey V., Daniela Puiu, Ming-Cheng Luo, Tingting Zhu, Sergey Koren, Guillaume Marçais, James A. Yorke, Jan Dvořák, and Steven L. Salzberg. 2017. “Hybrid Assembly of the Large and Highly Repetitive Genome of Aegilops Tauschii, a Progenitor of Bread Wheat, with the MaSuRCA Mega-Reads Algorithm.” *Genome Research* 27 (5): 787–92.

Tables

Table1: Summary of genome datasets of *C. variabilis*, *C. vulgaris*, and *C. sorokiniana*.

Species	*C. variabilis*	*C. vulgaris*	*C. sorokiniana*
Lifestyle	Symbiosis	Free-living	Free-living
Sequencer	illumina & Nanopore	illumina & PacBio	illumina & PacBio
Genome size (bp)	44.98 M	38.97 M	58.12 M
GC (%)	67.09	61.50	64.08
Genes	11,634	10,159	13,282
Repeat (%)	14.84	9.47	9.55

Figures

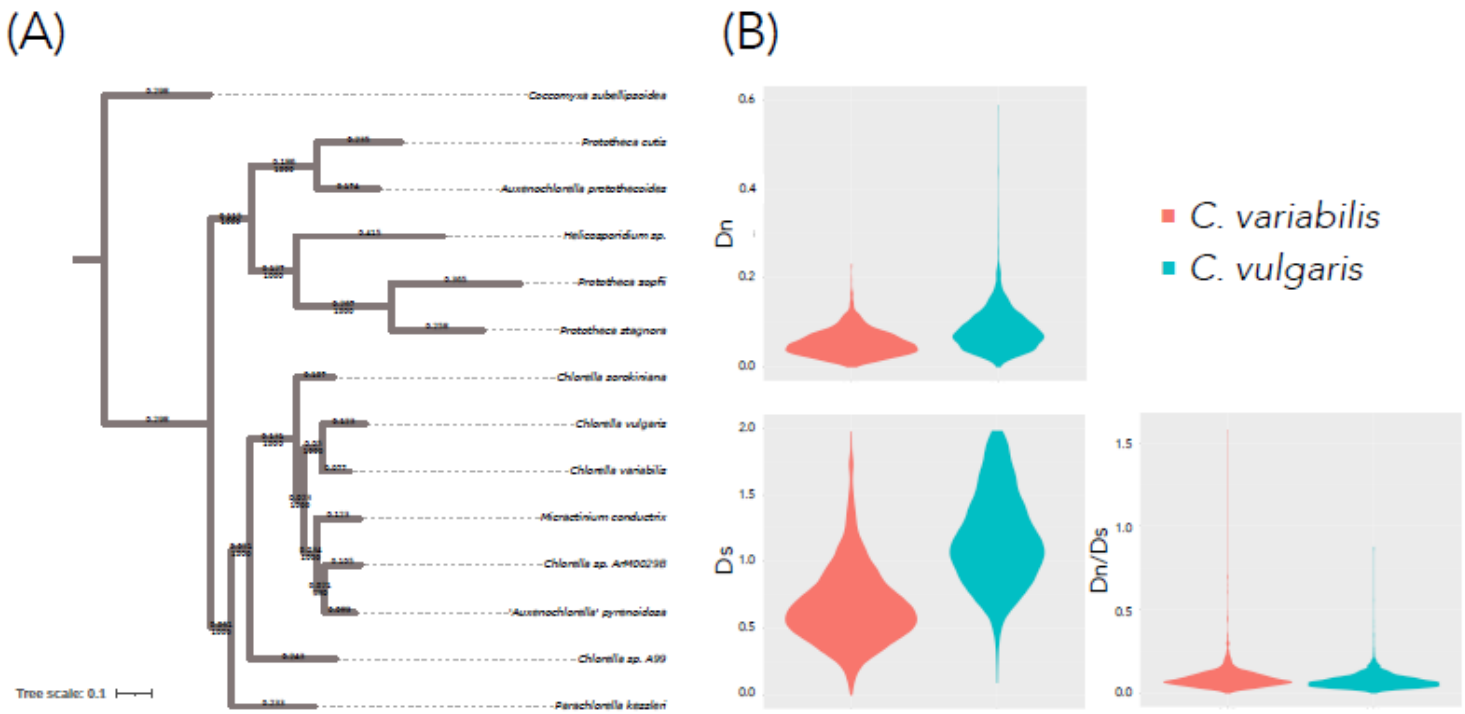

Figure 1

Phylogeny and molecular evolution of C. variabilis. (A) A maximum-likelihood phylogenetic tree using 449 single-copy orthologous genes shared in the 14 species, rooted on Coccomyxa subellipsoidea. The number above and below each branch represents branch length and bootstrap value, respectively. (B) Violin plot indicating distributions of Dn (top left panel), Ds (bottom left panel), and Dn/Ds (right panel) values derived from 6,358 single-copy orthologous genes of C. variabilis, C. vulgaris, with C. sorokiniana as outgroup

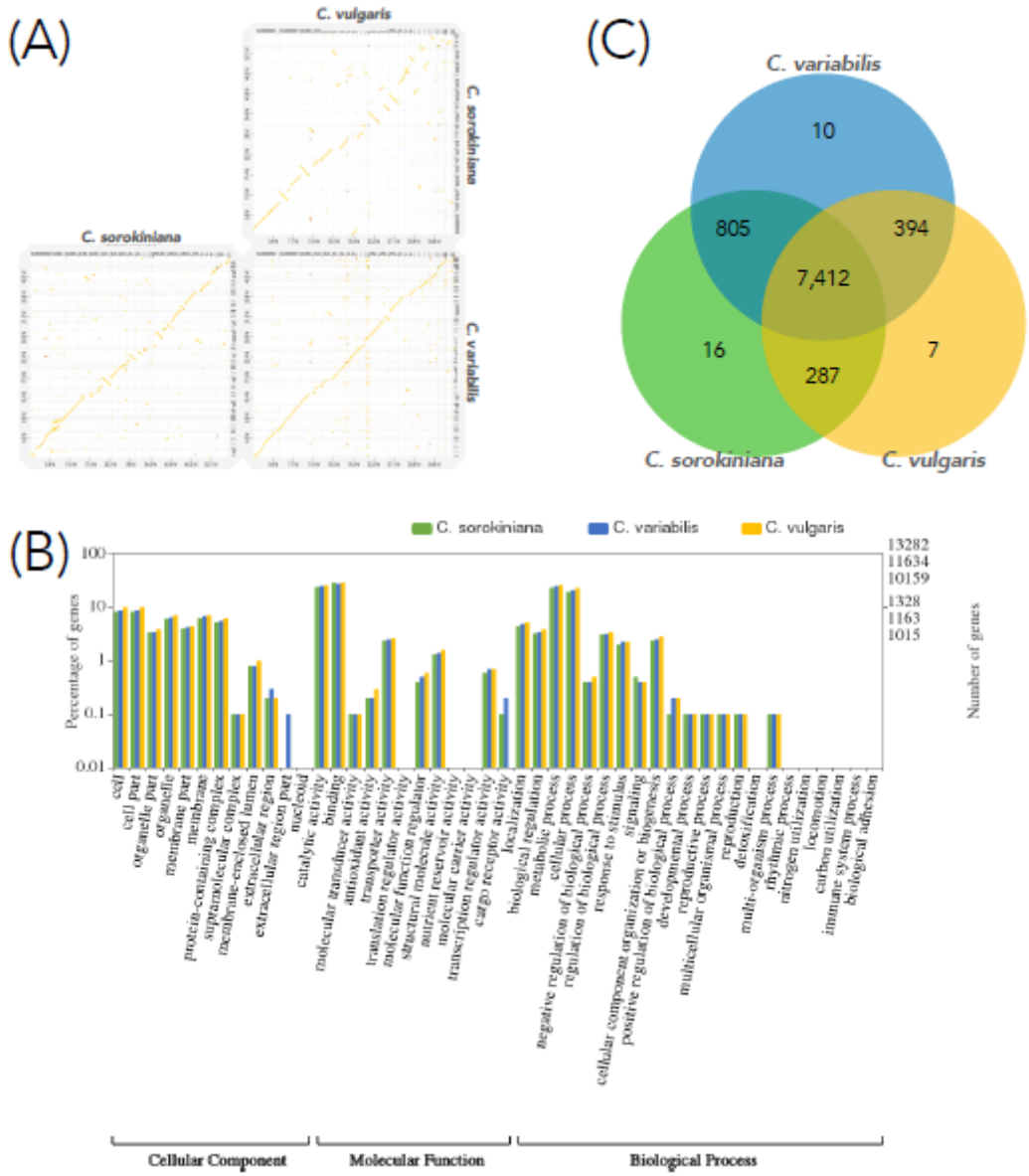

Figure 2

Comparative genomics of C. variabilis, C. vulgaris, and C. sorokiniana. (A) Dot plots indicating whole-genome synteny pairwise comparison among three species. Each axis represents nucleotide sequences of the genome laid end to end, and each yellow dot indicates a homologous match between two genomes. Left panel, C. sorokiniana-C. variabilis comparison; top right panel, C. vulgaris-C. sorokiniana comparison; bottom left, C. vulgaris-C. variabilis comparison. (B) Bar chart indicating classifications of GO terms linked to all genes predicted from genomes of three species. (C) Venn diagram indicating sharing patterns of the 8,931 ortholog groups identified between the 3 species

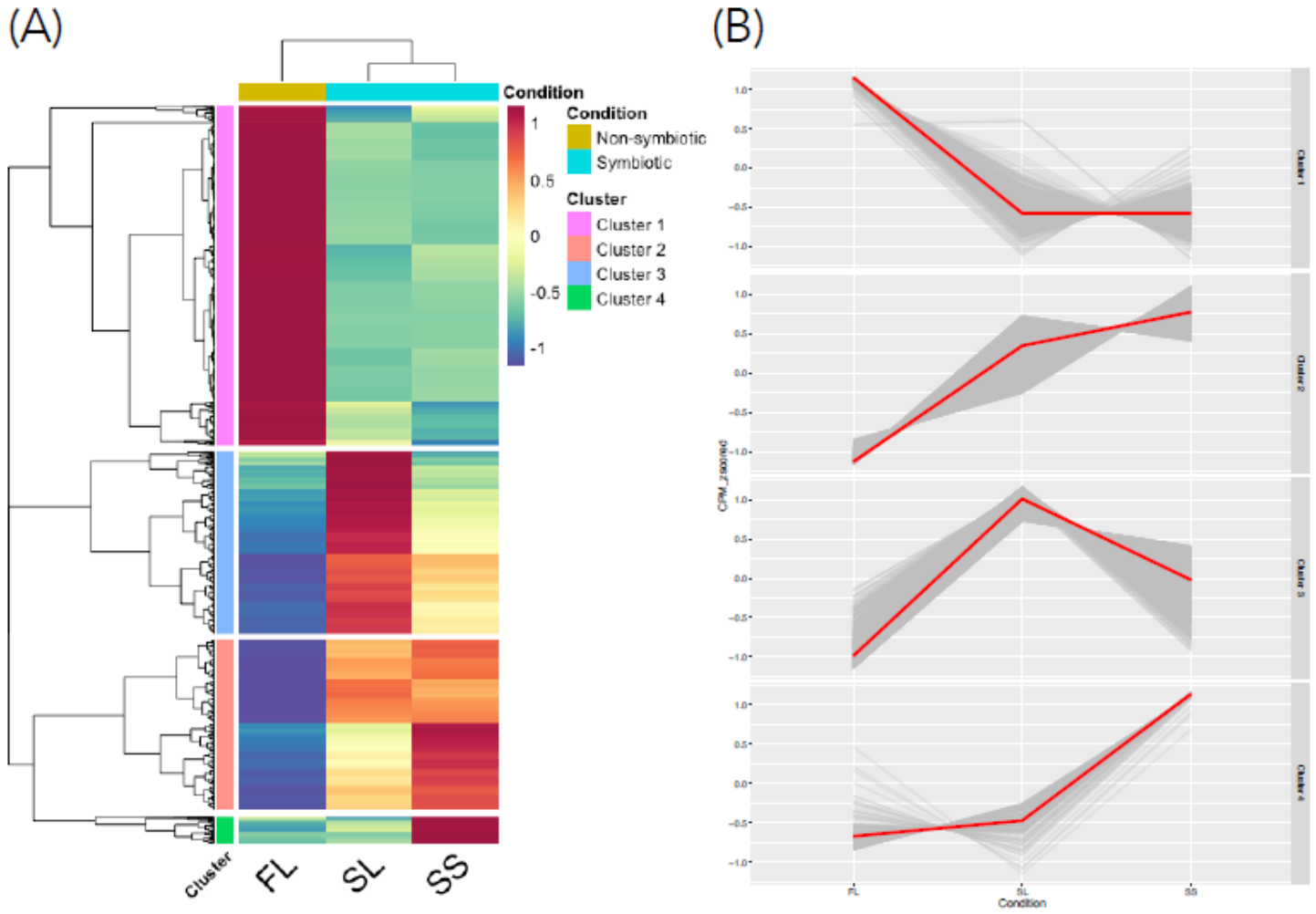

Figure 3

Differentially expressed gene (DEG) analysis under three conditions. (A) Heat map indicating the relative expression pattern of each of the 4401 DEGs, and 2 dendrograms derived from hierarchical clustering of each DEG (left) and condition (top). The DEGs were classified into four Clusters based on the left dendrogram. (B) Trend lines indicating changes in the relative expression levels (“CPM_zscored” of the y-axis label) of DEGs belonging to each Cluster under the three conditions. The gray line represents the expression of each DEG, and the red line shows the median expression of all DEGs in each Cluster. FL, the free-living condition; SL, the symbiotic and log phase condition; SS, the symbiotic and stationary phase condition

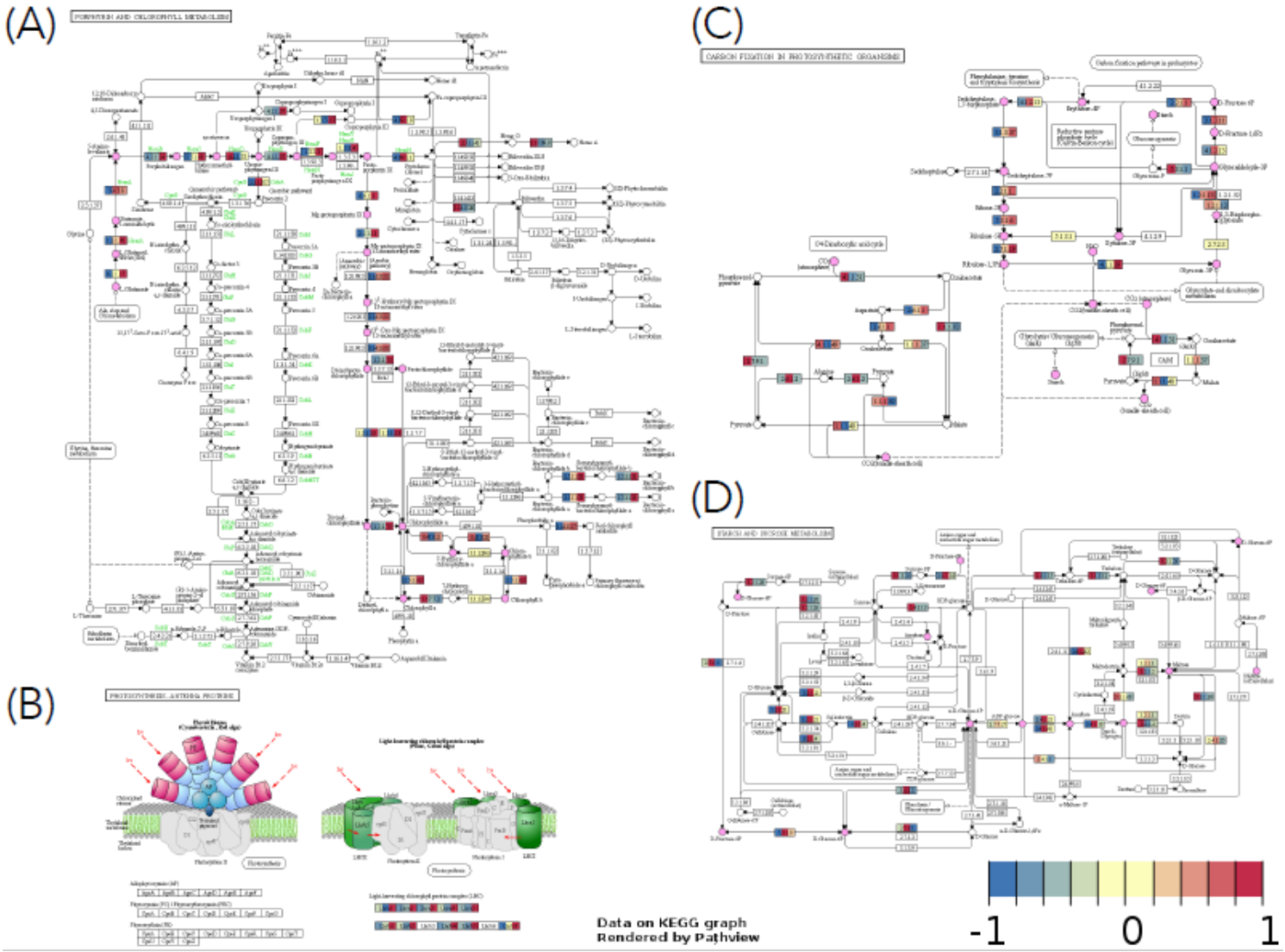

Figure 4

Metabolic pathways, with gene expression patterns mapped: (A) porphyrin and chlorophyl metabolism, (B) photosynthesis-antenna proteins, (C) carbon fixation in photosynthetic organisms, and (D) starch and sucrose metabolism of KEGG Pathway. Square, proteins; circle, metabolites; arrow, enzymatic reaction. Within the squares are the corresponding protein symbols and EC numbers, and their annotations are in Table S7. The square is divided into three zones, each colored based on the median relative expression level of the genes encoding this protein under three conditions: FL, SL, and SS conditions, from left to right. Metabolites expected to increase biosynthesis/degradation are highlighted in pink. FL, the free-living condition; SL, the symbiotic and log phase condition; SS, the symbiotic and stationary phase condition. The uncolored squares, enzymes, indicate that the corresponding genes were not identified

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionaldata.xlsx
- SupFigures.pdf
- SupLegends.docx
- TableS.xlsx