Current Biology 23, 1399–1408, August 5, 2013 ©2013 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2013.05.062

Article

Draft Assembly of the *Symbiodinium minutum* Nuclear Genome Reveals Dinoflagellate Gene Structure

Eiichi Shoguchi,[1,10,*] Chuya Shinzato,[1,10]
Takeshi Kawashima,[1,10] Fuki Gyoja,[1] Sutada Mungpakdee,[1]
Ryo Koyanagi,[1] Takeshi Takeuchi,[1] Kanako Hisata,[1]
Makiko Tanaka,[1] Mayuki Fujiwara,[1] Mayuko Hamada,[1]
Azadeh Seidi,[2] Manabu Fujie,[2] Takeshi Usami,[2] Hiroki Goto,[2]
Shinichi Yamasaki,[2] Nana Arakaki,[2] Yutaka Suzuki,[3]
Sumio Sugano,[3] Atsushi Toyoda,[4] Yoko Kuroki,[5]
Asao Fujiyama,[4,6] Mónica Medina,[7] Mary Alice Coffroth,[8]
Debashish Bhattacharya,[9] and Nori Satoh[1,*]

[1]Marine Genomics Unit
[2]DNA Sequencing Section
Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan
[3]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8568, Japan
[4]National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan
[5]RIKEN Research Center for Allergy and Immunology, Yokohama, Kanagawa 230-0045, Japan
[6]National Institute of Informatics, Tokyo 101-8430, Japan
[7]Department of Biology, Penn State University, University Park, PA 16802, USA
[8]Department of Geology, State University of New York at Buffalo, Buffalo, NY 14260, USA
[9]Department of Ecology, Evolution, and Natural Resources, Rutgers University, New Brunswick, NJ 08901-8520, USA

Summary

Background: Dinoflagellates are known for their capacity to form harmful blooms (e.g., “red tides”) and as symbiotic, photosynthetic partners for corals. These unicellular eukaryotes have permanently condensed, liquid-crystalline chromosomes and immense nuclear genome sizes, often several times the size of the human genome. Here we describe the first draft assembly of a dinoflagellate nuclear genome, providing insights into its genome organization and gene inventory.

Results: Sequencing reads from *Symbiodinium minutum* were assembled into 616 Mbp gene-rich DNA regions that represented roughly half of the estimated 1,500 Mbp genome of this species. The assembly encoded ∼42,000 protein-coding genes, consistent with previous dinoflagellate gene number estimates using transcriptomic data. The *Symbiodinium* genome contains duplicated genes for regulator of chromosome condensation proteins, nearly one-third of which have eukaryotic orthologs, whereas the remainder have most likely been acquired through bacterial horizontal gene transfers. *Symbiodinium* genes are enriched in spliceosomal introns (mean = 18.6 introns/gene). Donor and acceptor splice sites are unique, with 5′ sites utilizing not only GT but also GC and GA, whereas at 3′ sites, a conserved G is present after AG. All spliceosomal snRNA genes (*U1–U6*) are clustered in the genome. Surprisingly, the *Symbiodinium* genome displays unidirectionally aligned genes throughout the genome, forming a cluster-like gene arrangement.

Conclusions: We show here that a dinoflagellate genome exhibits unique and divergent characteristics when compared to those of other eukaryotes. Our data elucidate the organization and gene inventory of dinoflagellates and lay the foundation for future studies of this remarkable group of eukaryotes.

Introduction

Dinoflagellates are ecologically and economically important, unicellular eukaryotes that inhabit both marine and freshwater habitats [1]. Most dinoflagellates are 10–100 μm in diameter and are characterized by two flagella, with a unique cell covering referred to as the theca (Figure 1A). Many marine dinoflagellates exist as free-living phytoplankton and are the most important eukaryotic primary producers after diatoms, whereas others are heterotrophs and mixotrophs that are important grazers of plankton. Photosynthetic symbionts such as *Symbiodinium* are essential to reef-building corals [2], and taxa such as *Alexandrium* form harmful algal blooms (“red tides”). Over 60 species produce toxins and have profound impacts on the fishing industry, the recreational values of coastal zones, and public health [3]. Dinoflagellates also produce a wide variety of secondary metabolites. In addition, dinoflagellates have one of the most extensive fossil records among microbial eukaryotes, because of their ability to form resistant cysts [4].

Dinoflagellates are members of the well-supported Alveolata, which also includes ciliates and apicomplexans (Figure S1A available online) [5]. Ciliates include common aquatic protists, such as *Tetrahymena thermophila* and *Paramecium tetraurelia*, which lack plastids. With a few exceptions, most of these taxa are important heterotrophs (micrograzers) or parasites. Apicomplexans are exclusively animal parasites, exemplified by the malarial parasite, *Plasmodium falciparum*. Each alveolate lineage has had a distinct evolutionary trajectory with regard to nuclear genome organization, resulting in three divergent outcomes [6, 7]. Ciliates contain two nuclei, a somatic macronucleus, and a micronucleus for reproduction. In contrast, apicomplexans, due to their obligate parasitic life style, have substantially reduced genomes, with highly degenerate plastids referred to as apicoplasts [8]. The most intriguing alveolates, however, are the dinoflagellates [9, 10], the nucleus of which is characterized by permanently condensed liquid-crystalline chromosomes (Figures 1B and 1C) [11, 12] that lack nucleosomes. Nuclear DNA is associated with basic nuclear proteins or histone-like proteins (HLPs). These have a secondary structure similar to that of bacterial HLPs [13, 14]. In addition, recent studies of dinoflagellate genome data show repeated gene copies that are arranged in tandem arrays [15], *trans*-splicing of messenger RNAs [16, 17], and a reduced role for transcriptional regulation, compared to other eukaryotes [18, 19].

Given these remarkable characteristics, elucidation of the structure and composition of dinoflagellate genomes is essential to understanding their packaging of chromosomal DNA

[10]These authors contributed equally to this work

*Correspondence: eiichi@oist.jp (E.S.), norisky@oist.jp (N.S.)

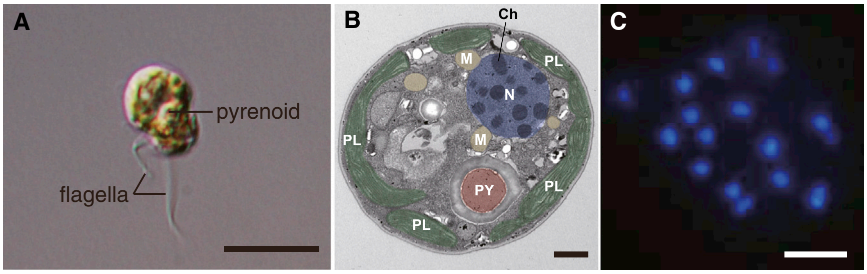

Figure 1. A Dinoflagellate, *Symbiodinium minutum*

(A) *S. minutum* zoospore. A short, transverse flagellum originating from the cingulum and a long, longitudinal flagellum originating from the sulcus are evident in the zoospore. A pyrenoid is also visible. The scale bar represents 10 μm.

(B) Electron micrograph showing permanently condensed chromosomes (Ch) of *S. minutum*. The nucleus (N) is shown in purple, plastids (PL) in green, mitochondria (M) in orange, and pyrenoid (PY) in brown. The scale bar represents 1 μm.

(C) DAPI staining of the nucleus showing permanently condensed chromosomes of *S. minutum*. The scale bar represents 1 μm.

and expression of encoded genes. However, dinoflagellates possess some of the largest nuclear genomes known among eukaryotes (1,500–245,000 megabases [Mbp]), which have previously thwarted whole-genome sequencing of members of this lineage [9, 10]. We selected a culturable dinoflagellate, *Symbiodinium minutum* (Figure 1) [20], that has one of the smallest reported dinoflagellate genomes (~1,500 Mb) [21], and here we report the results of our analysis of its nuclear genome.

Results and Discussion

Analysis of 616 Mbp of Gene-Rich Genome Regions in *Symbiodinium minutum*

A single clone of *S. minutum* was cultured under unialgal conditions and used for DNA extraction. The haploid nucleus in this species consists of approximately 18 permanently condensed chromosomes (Figure 1C). Genome size was determined with flow cytometry and K-mer analysis [22] and was estimated to be ~1,500 Mbp (Figures S1B and S1C). Approximately 37-fold coverage of the genome was achieved with Roche 454 GS-FLX [23] and Illumina Genome Analyzer IIx (GAIIx) next-generation sequencers [24]. These data yielded an assembly with a contig N50 = 34.6 Kbp and scaffold N50 = 126.2 Kbp (Table S1, parts A–C). The assembly was substantiated by bacterial artificial chromosome (BAC) and fosmid end sequencing (Figures S1D–S1F). The contigs (and scaffolds) totaled ~616 Mbp (Table 1). Approximately 26,300 Mb of RNA sequencing (RNA-seq) data obtained under two different culture conditions (seven libraries) were assembled into 63,104 unique transcriptomes (N50 = 1,684 nucleotides), 26,691 of which encoded complete open reading frames (Table 1 and Table S1, parts D and E).

Gene prediction yielded 41,925 protein models in the ~616 Mbp assembly, 77.2% of which (32,366 gene models) were supported by RNA-seq data (blastn, e value $< 1 \times 10^{-10}$; Table 1). The vast majority of the transcriptome was encoded in the 616 Mbp draft assembly, suggesting that these contigs represent the euchromatin-like region of the *Symbiodinium* genome. DNA transposons, retrotransposons, and tandem repeats comprised 0.5%, 1.1%, and 4.6%, respectively, of the assembled genome (Table S1, parts F–H). The following genome browser provides access to the assembled data: http://marinegenomics.oist.jp/genomes/gallery/ (Figure S1G) [25].

Our approach then left approximately 884 Mbp of unassembled sequence in *S. minutum*. Given the high copy number of organelle DNA (i.e., plastid DNA minicircles [26] and mitochondrial DNA), we expect that >10% of the unassembled data is likely to be derived from this fraction (data not shown). A recent study reported that >50% of ~110 Gbp of the *Alexandrium ostenfeldii* (dinoflagellate) genome comprised tandem repeats [27]. Similarly, we examined the frequency of tandem repeats in the unassembled *Symbiodinium* data by calculating their ratios in end sequences of BAC and fosmid clones. DNA transposons, retrotransposons, and tandem repeats constituted approximately 0.2% and 0.6%, 0.7% and 1.0%, and 3.8% and 5.8% of BAC and fosmid end sequences, respectively (Table S1, parts F–H). These results suggest that <10% of the *Symbiodinium* genome is comprised of transposons and tandem repeats. Therefore, these complex DNA sequences

Table 1. Genomic Composition of Three Groups of Alveolates, the Dinoflagellate *Symbiodinium minutum*, the Apicomplexan *Plasmodinium falciparum*, and the Ciliate *Tetrahymena thermophila*

	Symbiodinium minutum	*Plasmodium falciparum*[a]	*Tetrahymena thermophila*[b]
A total assembled length of assembly (bp)	615,520,517	22,853,764	103,927,049
G + C content (%)	43.6	19.4	22.0
Genes			
No. of genes	41,925	5,268	24,725
Average length of genes (bp)	11,959	–	2,579
Average length of transcripts (nt)	2,067	2,283	1,994
Gene models supported by EST (%)	77.2	70.0	–
Exons			
No. of exons per gene	19.6	2.4	4.6
Average length (bp)	99.8	949.0	430.6
Total length (Mb)	82.1	12.0	49.2
Introns			
No. of genes with introns (%)	95.3	53.9	71.4
Average length (bp)	499	178.7	161.8
First two nucleotides at 5′ splice sites	GT/GC/GA	GT	GT
Total length (Mb)	331.5	1.3	14.5
Intergenic Regions			
Average length (bp)	2,064	1,694	1,423
Unidirectional arrangement of genes	Yes	No	No

See also Figure S1 and Table S1.

[a] From Gardner et al. [7].

[b] From Eisen et al. [6].

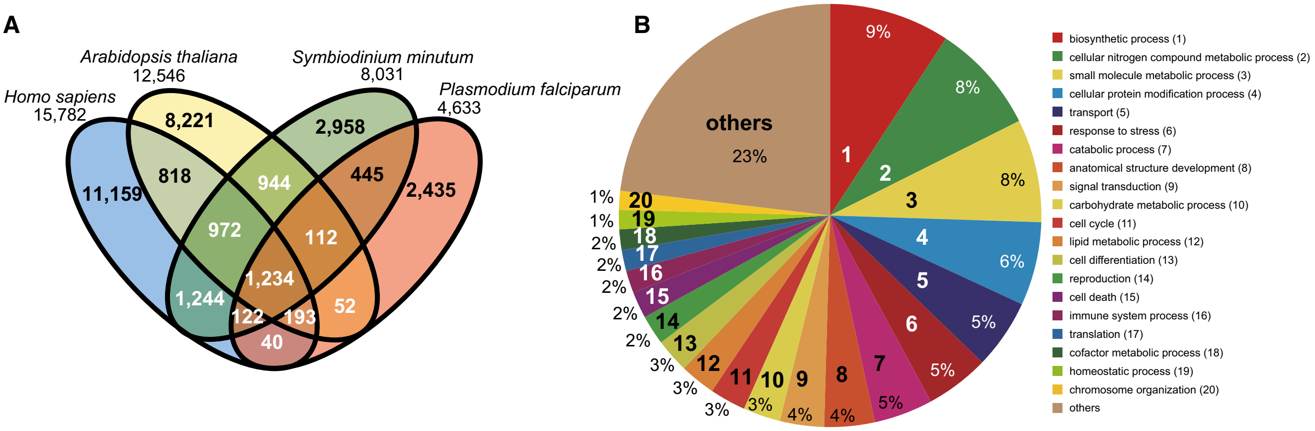

Figure 2. Venn Diagrams of Gene Content of the *Symbiodinium minutum* Genome

(A) Eight thousand thirty-one nuclear gene groups of *Symbiodinium minutum* (shown in green) are compared with 15,782 *Homo sapiens* gene groups (blue), 12,546 *Arabidopsis thaliana* groups (green plant; yellow), and 4,633 *Plasmodium falciparum* groups (pink).

(B) Functional classifications of gene family groups based on biological process of Gene Ontology (GO). The top twenty classifications on 8,031 nuclear gene groups of *Symbiodinium minutum* are shown.

See also Figure S2 and Table S2.

cannot explain our inability to assemble a large fraction of this dinoflagellate nuclear genome. On the other hand, unassembled reads (75 bp in length) from the Illumina sequencer represented 15.6 million (17%) of the total 91 million reads (see the Supplemental Experimental Procedures), and their compositions were similar to that of the assembled genome (data not shown). Although this suggests that much of the fragmented composition of the unassembled genomic region is similar to those of assembled regions, the unassembled sequences should be analyzed further by future studies.

The GC content of the *Symbiodinium* nuclear genome was 44% overall (Table 1), with 51%, 42%, and 43% in exon, intron, and intergenic regions, respectively (Figure S1H). This ratio is comparable to those of metazoans and green plants, but it stands in strong contrast to the AT-rich genomes of other alveolates, such as apicomplexans (19% GC of *Plasmodium falciparum* [7]) and ciliates (28% and 22% GC of *Paramecium tetraurelia* [28] and *Tetrahymena thermophila* [6], respectively) (Table 1).

Although we attempted to avoid microbial contamination in the *Symbiodinium* culture, we found evidence of an α-proteobacterium (Figures S1I–S1N). Scanning-electron microscopy showed the presence of bacteria on the surfaces of *Symbiodinium* cells (Figures S1M and S1N). The largest scaffold in the assembly appears to form a circular DNA of 3.8 Mbp (Figure S1I). The GC content of the scaffold was 53%, compared to 44% in *S. minutum*. In addition, no expressed sequence tag (EST) data were mapped onto the putative bacterial genome fragment (Figures S1J and S1K), indicating that the source of this DNA was quite different from *Symbiodinium*. Phylogenetic analysis with the contaminant 16S ribosomal RNA sequence indicated that this organism shows the closest match to *Parvibaculum lavamentivorans* DS-1 (Figure S1L). *P. lavamentivorans* DS-1^T is the type species of the novel genus *Parvibaculum* in the novel family *Rhodobiaceae* (formerly *Phyllobacteriaceae*) of the order *Rhizobiales* of α-proteobacteria [29].

Gene Content of the Dinoflagellate Genome

A total of 41,925 gene models (including 47,014 transcript variants) were estimated for the *S. minutum* genome. We first carried out a pfam domain search to identify frequently occurring domain-containing proteins to gain insight into gene family evolution in the dinoflagellate genome [30]. We found that 20,983 of the 41,925 gene models (50%) encode proteins with known domains, whereas the remaining models lack identifiable domains.

Results of the pfam analysis are shown in Table S2. One of the largest families was the EF hand family, with 1,052 genes (including 4,204 domains). The EF hand is a helix-loop-helix structural domain in a large family of calcium-binding proteins. The second largest family with 901 genes (including 4,108 domains) contained ankyrin repeats. The ankyrin repeat is one of the most common protein-protein interaction motifs in nature. Protein kinases, pentatricopeptide repeat (PPR), and zinc finger proteins accounted for 707 (1,349), 623 (2,001), and 1,025 genes (1,957 domains), respectively. PPRs are likely involved in RNA editing [31]. The fact that RNA editing was found in the organellar genomes of dinoflagellates [32] including *Symbiodinium* (unpublished data) can be explained by the presence of a large number of PPR-related genes in the genome. In addition, 649 genes (707 domains) for ion transporters were identified. It is highly likely that these gene families enable dinoflagellates to fulfill the diverse niches that they have successfully invaded [33].

In order to determine the number of unique and shared genes in *S. minutum*, we examined how many *Symbiodinium* gene families are shared with other eukaryotes. For this analysis, small proteins ($\leq$40 aa) and transcript variants were removed, and the remaining 41,740 proteins were clustered using OrthoMCl [34] with default settings (e value cutoff 1×10^{-5}, protein percent identity $\geq$50%, and MCL inflation of 1.5). When they were assigned to ortholog groups against 150 genomes in the OrthoMCL database (Version 5), 22,220 proteins (53%) were assigned into 8,031 OrthoMCL groups, whereas 6,679 proteins (16%) were clustered into 1,267 *S. minutum*-specific groups. *Symbiodinium* shares a considerable number of gene groups (3,572 and 3,262: 44% and 41%) of orthologs with *Homo* [35] and *Arabidopsis* [36], respectively (Figure 2A). On the other hand, only a few *Plasmodium* gene groups (1,589 and 1,591; 34% and 34%) were shared with *Homo*, and *Arabidopsis*, respectively (Figure 2A). This may provide further evidence that *Plasmodium* has lost many human orthologs during its evolution as an obligate parasite. Similarity searches (tblastn, e value < 10^{-5}) to dinoflagellate ESTs at NCBI also indicated ~46% of predicted proteins are novel or *Symbiodinium*-specific (data not shown), corresponding to ratios of unassigned proteins in the OrthoMCL database (47%). Therefore, this analysis confirmed that many highly divergent or dinoflagellate-specific proteins are encoded in the *Symbiodinium* genome (Figure 2A) [37].

Gene ontology (GO) is a useful method for categorizing putative gene functions [38]. GO analyses of *Symbiodinium* genes shared and unshared with three other organisms show that the GO terms associated with biosynthetic process, cellular nitrogen compound metabolic process, small nucleotide metabolic process, cellular protein modification process, and transport are highly represented in all categories in the Venn diagram (Figures 2B and S2). In contrast, secondary metabolic process is shared by *Symbiodinium*, *Plasmodium*, and *Arabidopsis*, but not by humans (Figure S2). The increased sharing of metabolic pathways in *Symbiodinium*, *Plasmodium*, and *Arabidopsis* was also clarified in the categories of small molecule metabolic process and cofactor metabolic process. *Symbiodinium* gene groups classified by GOs as neurological system process and cytoskeleton organization were shared with humans (Figure S2) in contrast to genes with GOs of cell motility and cell morphogenesis, which were shared with *Symbiodinium*, *Plasmodium*, and humans. Thus, GO classifications clarified not only the patterns of shared proteins, but also those of unshared genes, including proteins with GOs for generation of precursor metabolites and energy (Figure S2).

Specific Gene Expansion in the *Symbiodinium* Genome

Dinoflagellates have been predicted to contain 38,000–87,000 protein-coding genes [30]. The presence of a remarkably larger number of genes in the S. *minutum* genome (41,925) is likely caused by lineage-specific expansion of genes by duplication [30]. A study using massively parallel signature sequencing (MPSS) reported the presence of 40,029 unique EST tags (i.e., inferred to be different genes) in *Alexandrium tamarense*, consistent with this study [19]. It has been argued that tandem duplication of certain genes or gene families, including those for actin, luciferase, and form II Rubisco, occurred in the dinoflagellate lineage [10]. Lineage-specific gene family expansions were defined as orthologous groups with multiple copies in *S. minutum*, whose numbers are significantly greater than those of various other organisms, including *Plasmodium falciparum*, *Tetrahymena thermophila*, *Toxoplasma gondii*, *Phytophthora ramorum*, *Cyanidioschyzon merolae*, *E. coli*, *D. discoideum*, *A. thaliana*, *O. sativa*, and *H. sapiens*.

To identify gene expansions in *S. minutum*, we carried out an analysis based on 2,000 random permutations of exact probability (Table S3, part A) [33]. Proteins containing transposable element-related Pfam domains were removed from the calculation. It was found that 1,064 groups (10,912 genes) are likely to be expanded in the *Symbiodinium* genome. Together with *Symbiodinium*-specific orthologous group genes, it is predicted that a total of 17,703 genes (42.4% of the proteome) might have originated by gene duplication in this dinoflagellate. The possibility of gene expansions by the recycling of processed complementary DNA from messenger RNA (mRNA) is likely to be lower [39] because there are few intronless genes (discussed later). Table S3, part A, summarizes the top 50 expanded gene families in the *Symbiodinium* genome. Chlorophyll a/b-binding proteins and ion channel proteins are highly expanded in the *Symbiodinium* genome. One striking finding is that the regulator of chromosome condensation family protein (RCC1) is highly expanded (Table S3, part A) (discussed below). Calcium channel and calmodulin families are also expanded. Because the largest domain was the EF hand subgroup of calcium-binding proteins, Ca^{2+} metabolism is assumed to have great importance in *Symbiodinium*.

Molecular Basis of Permanently Condensed Chromatin

Dinoflagellate nuclei are characterized by permanently condensed, liquid-crystalline chromosomes (Figures 1B and 1C). The question of how dinoflagellate chromosomes are organized is one of the fundamental issues that remain to be fully understood [9, 10]. In eukaryotes, histone proteins are involved in chromatin modulation, whereas in prokaryotes, histone-like proteins play a role in chromatin modulation. Although it was previously thought that dinoflagellates lacked histone proteins [40] and used histone-like proteins for DNA organization [41], recent studies have revealed the presence of genes that encode four core nucleosomal histones, H2A, H2B, H3, and H4 [42–44]. Histone deacetylase [45], nucleosome assembly protein [45], and an H1-type linker histone-like protein [46] have also been found in dinoflagellates.

Our genome-wide survey revealed the presence of both eukaryotic histone genes and prokaryotic histone-like genes, although orthologs of histone H1 were not found (Table S3, part B, and Figure S3). All four, core histone genes (H2A, H2B, H3, and H4) are duplicated (Figure S3A). In addition, there were 15 histone-like proteins similar to those found in bacteria (Table S3, part B, and Figure S3B).

A recent study revealed that, in addition to enlargement of the genome, a dinoflagellate, *Hermatodinium* sp., gained a novel family of nucleoproteins from an algal virus, termed *d*inoflagellate /viral *n*ucleoprotein (DVNP) [47]. We searched for the presence of DVNP in the *Symbiodinium* genome and identified 19 genes putatively homologous to DVNPs in the genome, suggesting an involvement of this type of protein in chromosome structure in *Symbiodinium* (Table S3, part C, and Figure S3C). Because 11 *Symbiodinium* DVNP-like proteins had additional domains that were not found in the *Hermatodinium* DVNPs, it is likely that DVNP function has diverged in these two dinoflagellates.

In the gene-group expansion analysis of the *Symbiodinium* genome, genes for RCC1 had the third highest expansion (Table S3, part A). The RCC1 protein binds to chromatin and plays an important role in the regulation of gene expression [48]. In eukaryotes, the RCC1 superfamily has been divided into five subgroups based on the structures of other domains in the proteins [49]. On the other hand, three subfamilies of diverse RCC1-like repeat proteins have been reported in both prokaryotes and eukaryotes [50], namely, proteins related to RCC1 (including five eukaryotic RCC1 subgroups), proteins related to BLIP-II (β-lactamase inhibitor protein-II), and proteins related to the 10-RLR *Shewanella* sequence. Although the function of proteins related to bacterial BLIP-II and eukaryotic RCC1 has been studied [49, 51], that of the 10-RLR *Shewanella* group remains to be elucidated.

We found 189 genes encoding RCC1_2 (PF13540) in the *Symbiodinium* genome at an e value cutoff of 1×10^{-3}. A preliminary phylogenetic analysis of selected sequences showed that *Symbiodinium* proteins with RCC1-like repeats are related to RCC1 or the 10-RLR *Shewanella* sequence, but not BLIP-II (Figure S3D). The *Symbiodinium* protein most similar to

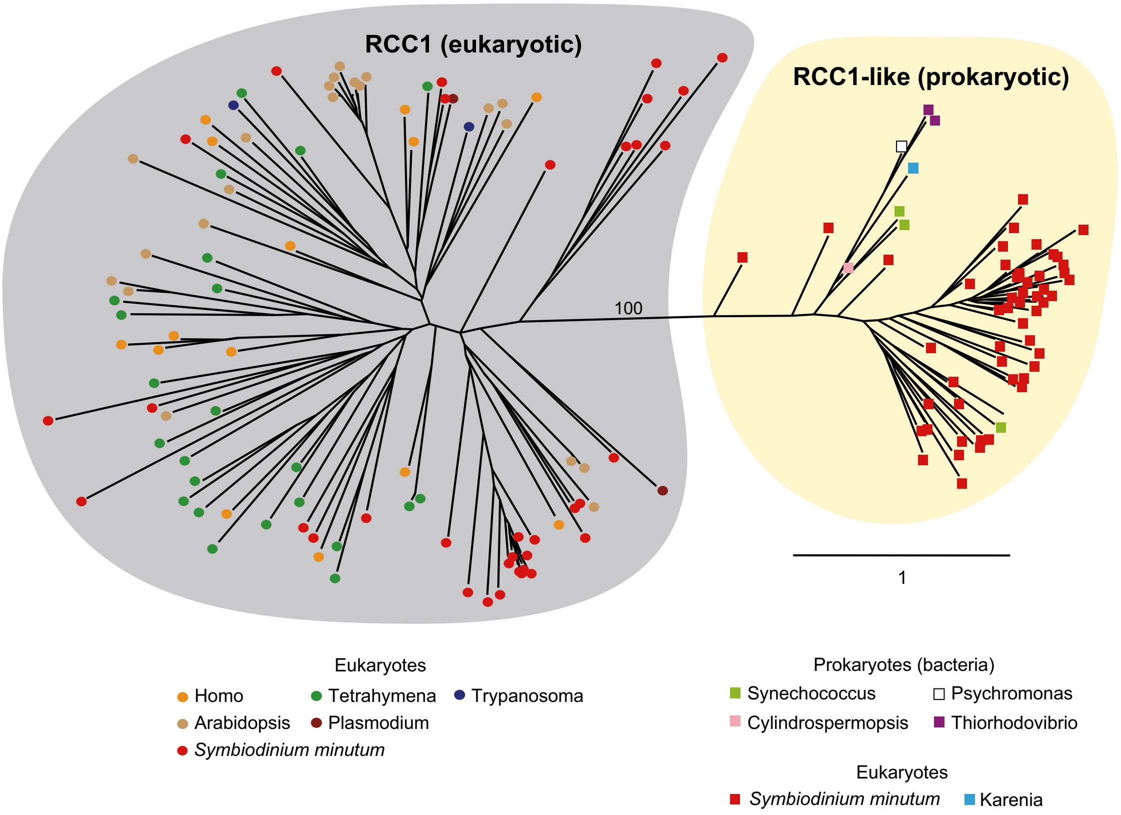

Figure 3. The Presence of Genes for Regulator-of-Chromosome-Condensation Proteins

Regulator of chromosome condensation (RCC1) proteins are eukaryotic proteins that bind to chromatin and play an important role in the regulation of gene expression. A maximum-likelihood phylogeny of 86 RCC1 family proteins encoded in the *S. minutum* genome is shown. The two distinct groupings of eukaryotic RCC1 proteins and prokaryotic RCC1-like proteins are supported by 100% bootstrap duplication. The bar indicates an amino acid substitution per site.

See also Figure S3 and Table S3.

bacterial BLIP-II (SymbB1.v1.2.021769) formed a clade with human RCC1.

To investigate this relationship further, we selected 93 genes at a stringent e value cutoff of 1 × 10^{-20} (Table S3, part D). Of these, amino acid sequences of 86 proteins were manually aligned and used for molecular phylogenic analyses, of which only 14 contained additional domains (Table S3, part D). Maximum likelihood (ML) analysis showed the existence of two distinct clusters among the 86 genes, which were supported by a 100% bootstrap value. One cluster with 34 *Symbiodinium* proteins consisted of those orthologous to eukaryotes, including alveolates, plants, and animals (Figure 3, left), whereas the other included 52 *Symbiodinium* proteins that had similarities to prokaryotes, including cyanobacteria and proteobacteria (Figure 3, right). This result provides a potential explanation for the characteristic architecture of dinoflagellate chromosomes, although the manner in which these proteins interact with each other to establish and maintain the permanently condensed chromosomes remains to be studied.

It has been reported that RCC1 functions as a guanine nucleotide-exchange factor (GEF) of the nuclear G protein, Ran, which is a member of the Ras superfamily [52]. We examined a possible expansion of Ran genes. Although the domain search revealed that the *Symbiodinium* genome contains approximately 90 candidates for Ras superfamily members, molecular phylogenetic analysis showed that only two of the Ras genes are putative Ran orthologs (Figure S3E). Therefore, an expansion of Ran genes is unlikely to have occurred in the *Symbiodinium* genome.

Unique Spliceosomal Splicing

It has been reported that introns are relatively uncommon in dinoflagellate genes [53–56]. For example, a recent survey showed that introns are present in only three of 17 heat shock protein genes sequenced: one canonical intron in *Peridinium willei* and *Thecadiniium yashimaense* and one noncanonical intron in *Polarella glacialis* [56]. Because of difference among dinoflagellate species, the real frequency of introns in dinoflagellate nuclear genes remains to be addressed using more systematic approaches [10]. We determined the number of introns per gene by examining splice sites in both open reading frame (ORF) and 3′ untranslated regions (UTRs) (Table S4). In contrast to previous studies, we found that genes of *S. minutum* are highly intron rich. Of the 41,925 genes, 39,970 (95%) are composed of multiple exons (Table 1), and the average number of exons per gene was 19.6 (Table 1), with some genes containing more than 200 introns (Table S4, part A). Eighty-three percent (646,312/780,874) of their introns were consistent with our transcriptome mapping (Table S4, part B). Approximately 54% and 71% of genes in *Plasmodium falciparum* and *Tetrahymena thermophila* contain introns, with the average number of exons per gene being 2.4 in *Plasmodium*

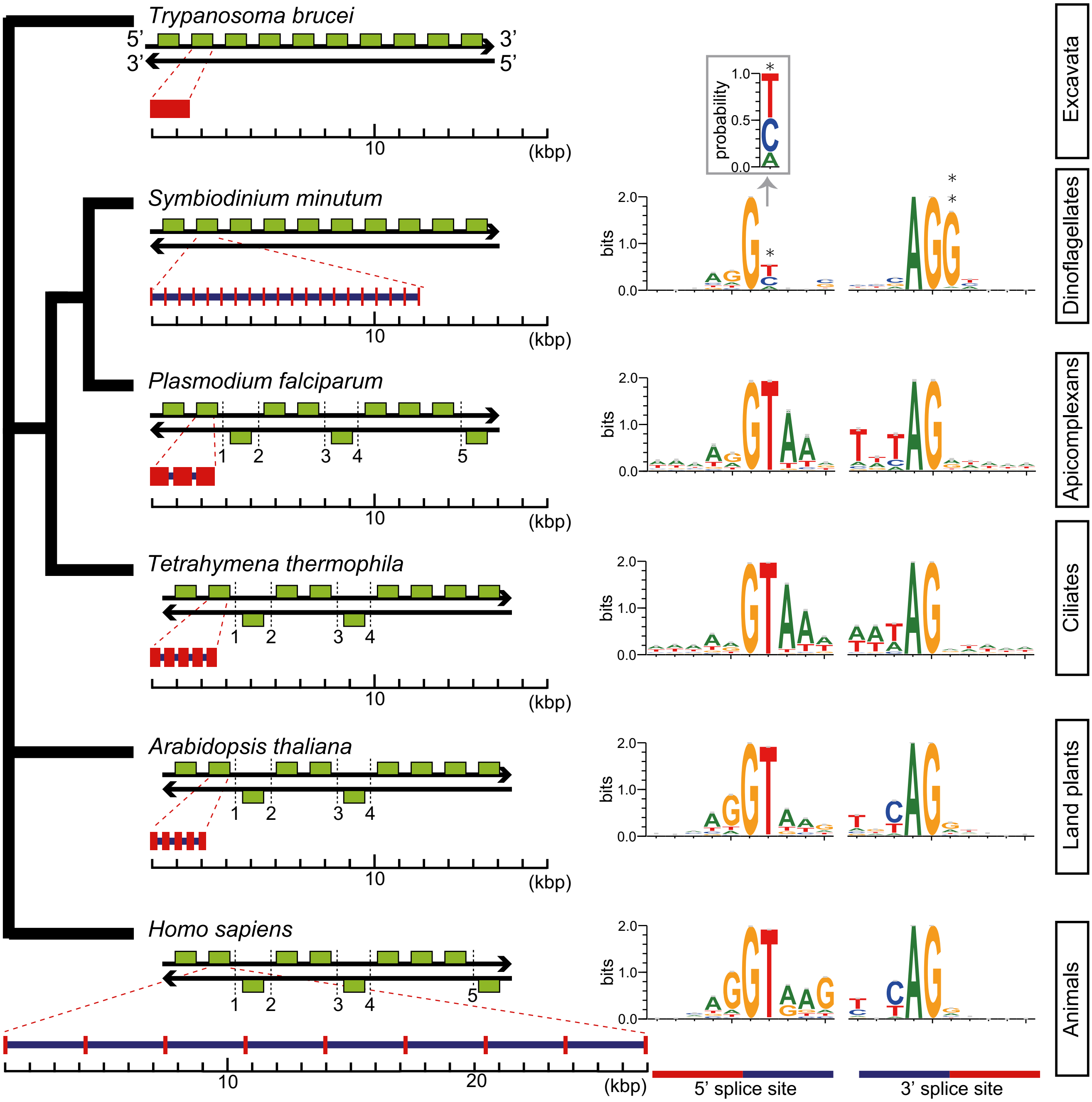

Figure 4. Architecture of Genes and Splice Site Motifs in the Nuclear Genomes of Representative Eukaryotes and Alveolates

Green boxes indicate typical patterns of ten-gene arrangements with the number of strand-switch regions (SSRs), although the SSRs shown here are not always typical. Patterns are based on the analyses shown in Figure 6. Gene architecture shows average gene lengths (exons in red and introns in blue) with the average intron number per gene. The sequence motif of the splice site is illustrated with WebLogo. Only two genes with spliceosomal introns in *Trypanosoma brucei* have been reported, but the motif was not shown. The unusual gene organization on the same strand of DNA shows similarities between *Symbiodinium* and *Trypanosoma*. Additionally, analyses of intron richness and the weakness of 5′ splice site signals (asterisk) indicate that *Symbiodinium* has the most unusual genome organization found in a eukaryote genome to date. The probability of position 2 at the 5′ splice site is shown in inset. A double asterisk shows G conserved at the 3′ splice site.
See also Figure S4 and Table S4.

and 4.6 in *Tetrahymena*, respectively (Table 1). The fact that an average gene length is 11,959 bp, whereas the average length of transcripts was 2,067 nucleotides supports the presence of many introns in *Symbiodinium* genes (Table 1).

In addition, spliceosomal introns of *Symbiodinium* are unique among the sequenced eukaryotic genomes. In *Plasmodium*, *Tetrahymena*, *Arabidopsis*, and *Homo*, introns are excised under the GT-AG rule, wherein GT and AG are used as recognition nucleotides at 5′ and 3′ splice sites, respectively (Figure 4). In contrast, *Symbiodinium* uses GC and GA at the 5′ donor splice site, in addition to GT (Figures 4 and S4 and Table S4, part B). GC usage frequency was nearly

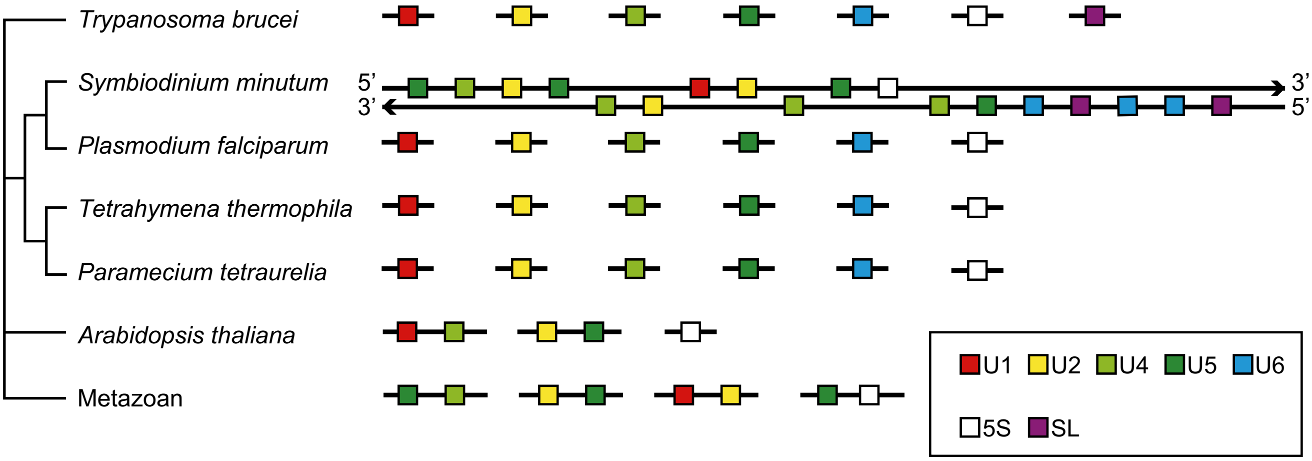

Figure 5. Clustering of Major Spliceosomal RNAs in the *Symbiodinium* Genome

All major spliceosomal RNAs (snRNAs; U1, U2, U4, U5, and U6) are clustered in the *Symbiodinium* genome. This feature is compared with the presence of snRNAs in other eukaryote genomes. The presence of microsynteny is suggested.
See also Figure S5 and Table S5.

equal to that of GT (Figure S4). The presence of these 5′ splice sites provides the first evidence in eukaryotes that the majority of mRNA splicing does not always follow the GT-AG rule. Another feature of *Symbiodinium* splicing is that the 3′ acceptor splice site, AG, is frequently followed by the nucleotide G (Figures 4 and S4), although a similar phenomenon is known in human minor alternative splice sites [57].

Key steps in RNA splicing are performed by spliceosomes, including five small nuclear RNA (snRNA) molecules (U1, U2, U4, U5, and U6). The five major snRNAs recognize nucleotide sequences that specify where splicing is to occur, and they participate in spliceosome chemistry [58]. In the *Plasmodium* and *Tetrahymena* genomes, snRNAs are scattered throughout the genome, whereas in metazoans and green plants, two different types of the five major snRNAs are sometimes tandemly aligned (Figure 5) [59, 60]. In contrast, in the *Symbiodinium* genome, all five snRNAs, U1, U2, U4, U5, and U6, occur in a cluster (Figures 5 and S5A), in addition to other snRNAs scattered at about 70 locations (Table S5). This is the first discovery of an snRNA gene cluster in a eukaryote genome. We also confirmed the conservation of base pairing sequences in the snRNAs by aligning these with those of other eukaryotes and the additional unique 5′ sequences of U1, U2, and U5 (Figure S5E). It has been reported that *trans*-splicing of messenger RNAs is common in dinoflagellates [9, 10]. The *Symbiodinium* genome contained spliced-leader (SL) genes with a conserved 5′ SL sequence (Figures 5 and S5).

Unique Arrangement of Genes in the Genome

It has been argued that the nuclear genome of dinoflagellates resembles that of trypanosomes, in spite of their phylogenetically distant relationship (dinoflagellates belong to the Alveolata and trypanosomes belong to the Excavata) [61, 62]. In both groups, genes are often organized in tandem arrays and mRNAs are regulated by *trans*-splicing and polyadenylation [9, 10]. In dinoflagellates, however, the tandem arrays have been reported as only repeats from multiple copies of a single gene [15].

Scaffolds of approximately 200 kbp were compared among *Tetrahymena* (ciliate), *Plasmodium* (aplicomplexa), *Symbiodinium* (dinoflagellate), *Trypanosoma* (euglenozoa) [63], *Arabidopsis* (land plant) [36], and *Homo* (mammalian) [35] (Figure 6A). In contrast to the random arrangement of protein-coding genes in the genomes of *Tetrahymena*, *Plasmodium*, *Arabidopsis*, and *Homo*, those of the *Symbiodinium* and *Trypanosoma* genomes show a clear tendency for tandem and unidirectional gene alignment. The grade of change in gene direction was searched using a ten-gene window (Figure 6B). Graphs of these data for *Plasmodium*, *Tetrahymena*, *Arabidopsis*, and *Homo* show a peak between four or five changes in orientation, indicating the frequency of SSRs between genes in head-to-head or tail-to-tail orientations (Figure 4). In contrast, *Symbiodinium* and *Trypanosoma* show a cluster (Figure 6B). This indicates a strong tendency for tandem alignment of genes or clustering of unidirectionally aligned genes in the *Symbiodinium* and *Trypanosoma* genomes.

In order to gain further information about the unique arrangement of genes, we performed transcription start site (TSS) analysis and determined how many genes would receive spliced leader (SL) sequences during transcription [64]. We found that transcripts of 3,329 genes contained SL sequences (see the Supplemental Experimental Procedures). On the other hand, transcripts of 13,318 genes did not contain SL sequences. Therefore, at least 20% of expressed genes (3,329/16,647) are likely to be SL *trans*-spliced (SLTS). In spite of the conservation of dinoflagellate SL sequences, this frequency is strikingly lower than those predicted by earlier transcriptomic studies of other dinoflagellates [10]. RNA processing may differ in the markedly smaller *S. minutum* genome, and this difference should be explored by further studies.

In trypanosomatids, gene clusters that include several hundred genes are transcribed on a polycistronic mRNA and are *trans*-spliced by the addition of SLs [62, 63]. We examined 117 large, unidirectionally aligned gene clusters that include at least 20 genes and found that, on average, 18% were identified as SLTS genes (Table S6, part A). In *S. minutum*, polycistronic mRNAs are unlikely to be as long as in trypanosomatids, as suggested by recent EST studies from other species [65].

Genes Involved in the Basic Transcriptional Machinery

In relation to the unique features of the S. *minutum* genome (e.g., permanently condensed chromosomes, spliceosomal splicing, and unidirectionally aligned genes), we surveyed

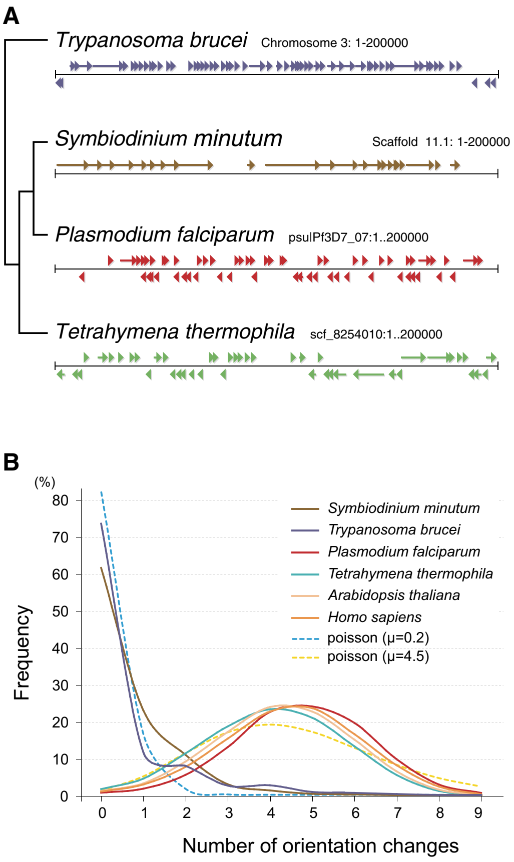

Figure 6. Nuclear Gene Arrangement in the *Symbiodinium minutum* Genome

(A) Examples of gene arrangement in the 200 kbp nuclear genome are shown and compared among *Symbiodinium minutum* (dinoflagellate), *Trypanosoma brucei* (euglenozoan), *Plasmodium falciparum* (apicomplexa), and *Tetrahymena thermophila* (ciliate). In contrast to the genomes of *Tetrahymena*, and *Plasmodium*, which show a random arrangement of protein-coding genes (arrowheads and arrows), the genomes of *S. minutum* and *Trypanosoma* are arranged into a large directional gene cluster in a head-to-tail orientation.

(B) A search of the directional gene cluster using a ten-gene window shows the strong tendency toward unidirectional alignment of genes in the *S. minutum* and *Trypanosoma* genomes. Each line represents a frequency histogram for changes in the gene orientation between successive genes in the genome. The x axis represents the number of orientation changes as one moves through windows of ten genes. For an example, as indicating random orientation, the Poisson distribution with $\mu = 4.5$ (average) is shown. The distribution with $\mu = 0.2$ is shown for comparison to the lines of *Symbiodinium* and of *Trypanosoma*.

See also Figure S6 and Table S6.

genes involved in basic transcriptional machinery. With similarity searches using blastp and tblastn, we found that the *Symbiodinium* genome contains highly conserved basic transcriptional machinery components, including RNA polymerased I, II, and III (Table S6, part B, and Figure S6); basal transcription factors, such as TFIID and TATA-binding protein (TBP; Table S6, part C); and transcription elongation factors (Table S6, part D). Almost all of these genes were substantiated by corresponding mRNAs (Table S6parts B–D). In contrast, we found only a few sequence-specific transcription factors, including 19 gene models with AP2 domain(s), 15 models with HMG_box domain(s), eight models with zf-C2H2 domain(s), and others (Table S6, part E). These results suggest constant, steady transcription of *Symbiodinium* genes with fewer genes under sequence-specific transcriptional control.

Conclusions

Dinoflagellates, a highly diverse group of eukaryotes with respect to lifestyle and phylogenetic breadth, are key players in marine and freshwater ecosystems. We report here that the genome of the dinoflagellate *Symbiodinium minutum* is unique in having divergent characteristics, compared to other "model" eukaryotes. This genome contains almost twice as many gene families as that of its sister group, the apicomplexans. The permanently condensed chromosomes are composed of both prokaryotic and eukaryotic proteins, suggesting a mosaic nuclear structure derived through horizontal gene transfer. Rich spliceosomal introns that use various recognition nucleotides at 5′ splice sites appear to be unique among unicellular eukaryotes. Unidirectional gene alignment across euchromatic genome regions is the second example of this type of gene arrangement, previously known only in the evolutionarily distantly related trypanosomes. Such divergent characters raise questions about how genomes were modified during the evolution of each eukaryote group. For example, what process gave rise to the intron-rich dinoflagellate genes? How is the gene arrangement in large eukaryote genomes controlled? Decoding of other eukaryote genomes as well as organellar genomes is essential to answer these fundamental questions about eukaryote evolution.

Experimental Procedures

Please see the Supplemental Information for details of the experimental procedures.

DNA obtained from a single clonal culture of the *Symbiodinium minutum* was sequenced with Roche 454 GS-FLX [23] and the Illumina Genome Analyzer IIx (GAIIx) [24]. The 454 shotgun and paired-end reads were assembled de novo by GS De Novo Assembler version 2.3 (Newbler, Roche) [23], and subsequent scaffolding was performed by SOPRA [66] and SSPACE [67] with Illumina mate-pair information. Gaps inside scaffolds were closed with Illumina paired-end data using Gapcloser [68]. RNA-seq analysis was also performed with GAIIx. A set of gene model predictions (the *Symbiodinium minutum* Gene Model version 1.2) was generated mainly with AUGUSTUS, and a genome browser has been established with the Generic Genome Browser 2.17, as employed in our previous studies [25, 69]. Annotation and identification of *Symbiodinium* genes were performed with three approaches, individual methods or combinations of the methods: reciprocal BLAST analyses, screening of the gene models against the Pfam database [70], and phylogenetic analyses. Orthologous gene family annotation was performed with the OrthoMCL database (version 5) (using a Markov Cluster algorithm to group orthologs and paralogs) [34]. A sliding window search was applied to all .gff files to count the frequency of gene orientation changes in each ten successive genes (window size, ten genes; shift size, ten genes). Detail methods are described in the Supplemental Information.

Accession Numbers

The genome assembly has been deposited with DDBJ/EMBL/GenBank as project accession BASF01000001–BASF01033816 (contigs) and DF239013–DF260911 (scaffolds).

Supplemental Information

Supplemental Information includes Supplemental Experimental Procedures, six figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2013.05.062.

Acknowledgments

We thank Steven Aird for his help in preparing the manuscript and Toshio Sasaki for supporting electron microscopy analysis. Supercomputing was supported by the IT Section of OIST. This work was supported in part by Grants-in-Aids from MEXT (no. 23128515 to E.S.) and JSPS (no. 24241071 to N.S.) of Japan and by the National Science Foundation (OCE-09-26822 to M.A.C.).

Received: December 25, 2012
Revised: May 29, 2013
Accepted: May 31, 2013
Published: July 11, 2013

References

1. Graham, L., and Wilcox, L. (2000). Algae (Upper Saddle River: Prentice-Hall).
2. Coffroth, M.A., and Santos, S.R. (2005). Genetic diversity of symbiotic dinoflagellates in the genus *Symbiodinium*. Protist *156*, 19–34.
3. Wang, D.Z. (2008). Neurotoxins from marine dinoflagellates: a brief review. Mar. Drugs *6*, 349–371.
4. Moldowan, J.M., and Talyzina, N.M. (1998). Biogeochemical evidence for dinoflagellate ancestors in the early cambrian. Science *281*, 1168–1170.
5. Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjaeveland, A., Nikolaev, S.I., Jakobsen, K.S., and Pawlowski, J. (2007). Phylogenomics reshuffles the eukaryotic supergroups. PLoS ONE *2*, e790.
6. Eisen, J.A., Coyne, R.S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J.R., Badger, J.H., Ren, Q., Amedeo, P., Jones, K.M., et al. (2006). Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol. *4*, e286.
7. Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature *419*, 498–511.
8. Wilson, R.J., Denny, P.W., Preiser, P.R., Rangachari, K., Roberts, K., Roy, A., Whyte, A., Strath, M., Moore, D.J., Moore, P.W., and Williamson, D.H. (1996). Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. J. Mol. Biol. *261*, 155–172.
9. Wisecaver, J.H., and Hackett, J.D. (2011). Dinoflagellate genome evolution. Annu. Rev. Microbiol. *65*, 369–387.
10. Lin, S. (2011). Genomic understanding of dinoflagellates. Res. Microbiol. *162*, 551–569.
11. Dodge, J.D. (1965). Chromosome structure in the dinoflagellates and the problem of the mesocaryotic cell. Excerpta Med. Int. Congr. Ser. *91*, 339–345.
12. Bouligand, Y., and Norris, V. (2001). Chromosome separation and segregation in dinoflagellates and bacteria may depend on liquid crystalline states. Biochimie *83*, 187–192.
13. Wong, J.T., New, D.C., Wong, J.C., and Hung, V.K. (2003). Histone-like proteins of the dinoflagellate *Crypthecodinium cohnii* have homologies to bacterial DNA-binding proteins. Eukaryot. Cell *2*, 646–650.
14. Chan, Y.H., Kwok, A.C., Tsang, J.S., and Wong, J.T. (2006). Alveolata histone-like proteins have different evolutionary origins. J. Evol. Biol. *19*, 1717–1721.
15. Bachvaroff, T.R., and Place, A.R. (2008). From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate Amphidinium carterae. PLoS ONE *3*, e2929.
16. Lidie, K.B., and van Dolah, F.M. (2007). Spliced leader RNA-mediated trans-splicing in a dinoflagellate, *Karenia brevis*. J. Eukaryot. Microbiol. *54*, 427–435.
17. Zhang, H., Hou, Y., Miranda, L., Campbell, D.A., Sturm, N.R., Gaasterland, T., and Lin, S. (2007). Spliced leader RNA *trans*-splicing in dinoflagellates. Proc. Natl. Acad. Sci. USA *104*, 4618–4623.
18. Erdner, D.L., and Anderson, D.M. (2006). Global transcriptional profiling of the toxic dinoflagellate *Alexandrium fundyense* using Massively Parallel Signature Sequencing. BMC Genomics *7*, 88.
19. Moustafa, A., Evans, A.N., Kulis, D.M., Hackett, J.D., Erdner, D.L., Anderson, D.M., and Bhattacharya, D. (2010). Transcriptome profiling of a toxic dinoflagellate reveals a gene-rich protist and a potential impact on gene expression due to bacterial presence. PLoS ONE *5*, e9688.
20. LaJeunesse, T.C., Parkinson, J.E., and Reimer, J.D. (2012). A genetics-based description of *Symbiodinium minutum* sp. nov. and *S. psygmophilum* sp. nov. (dinophyceae), two dinoflagellates symbiotic with cnidaria. J. Phycol. *48*, 1380–1391.
21. LaJeunesse, T.C., Lambert, G., Andersen, R.A., Coffroth, M.A., and Galbraith, D.W. (2005). *Symbiodinium* (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates. J. Phycol. *41*, 880–886.
22. Chapman, J.A., Ho, I., Sunkara, S., Luo, S., Schroth, G.P., and Rokhsar, D.S. (2011). Meraculous: de novo genome assembly with short paired-end reads. PLoS ONE *6*, e23501.
23. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.
24. Bentley, D.R. (2006). Whole-genome re-sequencing. Curr. Opin. Genet. Dev. *16*, 545–552.
25. Takeuchi, T., Kawashima, T., Koyanagi, R., Gyoja, F., Tanaka, M., Ikuta, T., Shoguchi, E., Fujiwara, M., Shinzato, C., Hisata, K., et al. (2012). Draft genome of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology. DNA Res. *19*, 117–130.
26. Zhang, Z., Cavalier-Smith, T., and Green, B.R. (2002). Evolution of dinoflagellate unigenic minicircles and the partially concerted divergence of their putative replicon origins. Mol. Biol. Evol. *19*, 489–500.
27. Jaeckisch, N., Yang, I., Wohlrab, S., Glöckner, G., Kroymann, J., Vogel, H., Cembella, A., and John, U. (2011). Comparative genomic and transcriptomic characterization of the toxigenic marine dinoflagellate *Alexandrium ostenfeldii*. PLoS ONE *6*, e28012.
28. Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature *444*, 171–178.
29. Schleheck, D., Weiss, M., Pitluck, S., Bruce, D., Land, M.L., Han, S., Saunders, E., Tapia, R., Detter, C., Brettin, T., et al. (2011). Complete genome sequence of *Parvibaculum lavamentivorans* type strain (DS-1(T)). Stand. Genomic Sci. *5*, 298–310.
30. Hou, Y., and Lin, S. (2009). Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. PLoS ONE *4*, e6978.
31. Fujii, S., and Small, I. (2011). The evolution of RNA editing and pentatricopeptide repeat genes. New Phytol. *191*, 37–47.
32. Lin, S., Zhang, H., and Gray, M.W. (2008). RNA editing in dinoflagellates and its implications for the evolutionary history of the editing machinery. In RNA and DNA Editing: Molecular Mechanisms and Their Integration into Biological Systems, H. Smith, ed. (Hoboken: John Wiley & Sons), pp. 280–309.
33. Colbourne, J.K., Pfrender, M.E., Gilbert, D., Thomas, W.K., Tucker, A., Oakley, T.H., Tokishita, S., Aerts, A., Arnold, G.J., Basu, M.K., et al. (2011). The ecoresponsive genome of *Daphnia pulex*. Science *331*, 555–561.
34. Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178–2189.
35. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al.; International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.
36. Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature *408*, 796–815.
37. Chan, C.X., Soares, M.B., Bonaldo, M.F., Wisecaver, J.H., Hackett, J.D., Anderson, D.M., Erdner, D.L., and Bhattacharya, D. (2012). Analysis of *Alexandrium tamarense* (Dinophyceae) genes reveals the complex evolutionary history of a microbial eukaryote. J. Phycol. *48*, 1130–1142.
38. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.
39. Slamovits, C.H., and Keeling, P.J. (2008). Widespread recycling of processed cDNAs in dinoflagellates. Curr. Biol. *18*, R550–R552.
40. Rizzo, P.J. (1987). Biochemistry of the dinoflagellate nucleus. In The Biology of Dinoflagellates, F.J.R. Taylor, ed. (Oxford: Blackwell Scientific), pp. 143–173.
41. Chan, Y.H., and Wong, J.T. (2007). Concentration-dependent organization of DNA by the dinoflagellate histone-like protein HCc3. Nucleic Acids Res. *35*, 2573–2583.
42. Okamoto, O.K., and Hastings, J.W. (2003). Genome-wide analysis of redox-regulated genes in a dinoflagellate. Gene *321*, 73–81.

43. Hackett, J.D., Scheetz, T.E., Yoon, H.S., Soares, M.B., Bonaldo, M.F., Casavant, T.L., and Bhattacharya, D. (2005). Insights into a dinoflagellate genome through expressed sequence tag analysis. BMC Genomics *6*, 80.
44. Bayer, T., Aranda, M., Sunagawa, S., Yum, L.K., Desalvo, M.K., Lindquist, E., Coffroth, M.A., Voolstra, C.R., and Medina, M. (2012). *Symbiodinium* transcriptomes: genome insights into the dinoflagellate symbionts of reef-building corals. PLoS ONE *7*, e35269.
45. Lin, S., Zhang, H., Zhuang, Y., Tran, B., and Gill, J. (2010). Spliced leader-based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates. Proc. Natl. Acad. Sci. USA *107*, 20033–20038.
46. Toulza, E., Shin, M.S., Blanc, G., Audic, S., Laabir, M., Collos, Y., Claverie, J.M., and Grzebyk, D. (2010). Gene expression in proliferating cells of the dinoflagellate *Alexandrium catenella* (Dinophyceae). Appl. Environ. Microbiol. *76*, 4521–4529.
47. Gornik, S.G., Ford, K.L., Mulhern, T.D., Bacic, A., McFadden, G.I., and Waller, R.F. (2012). Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. Curr. Biol. *22*, 2303–2312.
48. Dasso, M. (1993). RCC1 in the cell cycle: the regulator of chromosome condensation takes on new roles. Trends Biochem. Sci. *18*, 96–101.
49. Hadjebi, O., Casas-Terradellas, E., Garcia-Gonzalo, F.R., and Rosa, J.L. (2008). The RCC1 superfamily: from genes, to function, to disease. Biochim. Biophys. Acta *1783*, 1467–1479.
50. Stevens, T.J., and Paoli, M. (2008). RCC1-like repeat proteins: a pangenomic, structurally diverse new superfamily of beta-propeller domains. Proteins *70*, 378–387.
51. Lim, D., Park, H.U., De Castro, L., Kang, S.G., Lee, H.S., Jensen, S., Lee, K.J., and Strynadka, N.C. (2001). Crystal structure and kinetic analysis of beta-lactamase inhibitor protein-II in complex with TEM-1 beta-lactamase. Nat. Struct. Biol. *8*, 848–852.
52. Rojas, A.M., Fuentes, G., Rausell, A., and Valencia, A. (2012). The Ras protein superfamily: evolutionary tree and role of conserved amino acids. J. Cell Biol. *196*, 189–201.
53. Okamoto, O.K., Liu, L., Robertson, D.L., and Hastings, J.W. (2001). Members of a dinoflagellate luciferase gene family differ in synonymous substitution rates. Biochemistry *40*, 15862–15868.
54. Zhang, H., and Lin, S. (2003). Complex gene structure of the form II Rubisco in the dinoflagellate *Prorocentrum minimum* (dinophyceae). J. Phycol. *39*, 1160–1171.
55. Rowan, R., Whitney, S.M., Fowler, A., and Yellowlees, D. (1996). Rubisco in marine symbiotic dinoflagellates: form II enzymes in eukaryotic oxygenic phototrophs encoded by a nuclear multigene family. Plant Cell *8*, 539–553.
56. Hoppenrath, M., and Leander, B.S. (2010). Dinoflagellate phylogeny as inferred from heat shock protein 90 and ribosomal gene sequences. PLoS ONE *5*, e13220.
57. Thanaraj, T.A., and Clark, F. (2001). Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. Nucleic Acids Res. *29*, 2581–2593.
58. Rogozin, I.B., Carmel, L., Csuros, M., and Koonin, E.V. (2012). Origin and evolution of spliceosomal introns. Biol. Direct *7*, 11.
59. Marz, M., Kirsten, T., and Stadler, P.F. (2008). Evolution of spliceosomal snRNA genes in metazoan animals. J. Mol. Evol. *67*, 594–607.
60. Wang, B.B., and Brendel, V. (2004). The ASRG database: identification and survey of *Arabidopsis thaliana* genes involved in pre-mRNA splicing. Genome Biol. *5*, R102.
61. Cavalier-Smith, T. (2004). Only six kingdoms of life. Proc. Biol. Sci. *271*, 1251–1262.
62. Lukes, J., Leander, B.S., and Keeling, P.J. (2009). Cascades of convergent evolution: the corresponding evolutionary histories of euglenozoans and dinoflagellates. Proc. Natl. Acad. Sci. USA *106* (*Suppl 1*), 9963–9970.
63. Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B., et al. (2005). The genome of the African trypanosome *Trypanosoma brucei*. Science *309*, 416–422.
64. Yamashita, R., Sathira, N.P., Kanai, A., Tanimoto, K., Arauchi, T., Tanaka, Y., Hashimoto, S., Sugano, S., Nakai, K., and Suzuki, Y. (2011). Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. Genome Res. *21*, 775–789.
65. Beauchemin, M., Roy, S., Daoust, P., Dagenais-Bellefeuille, S., Bertomeu, T., Letourneau, L., Lang, B.F., and Morse, D. (2012). Dinoflagellate tandem array gene transcripts are highly conserved and not polycistronic. Proc. Natl. Acad. Sci. USA *109*, 15793–15798.
66. Dayarian, A., Michael, T.P., and Sengupta, A.M. (2010). SOPRA: scaffolding algorithm for paired reads via statistical optimization. BMC Bioinformatics *11*, 345.
67. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578–579.
68. Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., et al. (2010). The sequence and de novo assembly of the giant panda genome. Nature *463*, 311–317.
69. Shinzato, C., Shoguchi, E., Kawashima, T., Hamada, M., Hisata, K., Tanaka, M., Fujie, M., Fujiwara, M., Koyanagi, R., Ikuta, T., et al. (2011). Using the *Acropora digitifera* genome to understand coral responses to environmental change. Nature *476*, 320–323.
70. Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., et al. (2010). The Pfam protein families database. Nucleic Acids Res. *38* (Database issue), D211–D222.