Diversity and Evolution of Protist Mitochondria: Introns, Gene Content and Genome Architecture

A Dissertation Submitted to
the Graduate School of Life and Environmental Sciences,
the University of Tsukuba
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Science
(Doctral Program in Biologial Sciences)

Yuki NISHIMURA

Table of Contents

Abstract

Mitochondria are the organelles that were derived from an endosymbiotic α-proteobacterium captured by the ancestral eukaryotic cell. In addition to ATP synthesis through aerobic respiration, various essential metabolisms, such as amino acid synthesis, β-oxidation of fatty acid, and formation of FeS cluster, are catalyzed in mitochondria. As α-proteobacterium-derived organelles, mitochondria usually contain bacteria-type genomes (mtDNAs). Compared with the genomes of the extant α-proteobacteria, the mtDNAs determined so far are highly reduced in both genome size and gene content, suggesting severe reductive pressure worked during the process transforming the α-protebacterial endosymbiont to the organelle in early eukaryotic evolution. Nevertheless, size, gene content, and structure of mtDNAs are known to vary amongst eukaryotes, as the genomes have been evolved independently and continuously on branches of the tree of eukaryotes. As mtDNA data have been sampled from phylogenetically restricted lineages, such as metazoa, fungi, land plants, and green algae, our current knowledge regarding mtDNA diveristy is highly likely biased. In this study, I sequenced the mtDNAs of phylogenetic relatives of cryptophytes, as well as those of which are potentially related to cryptophytes, to revise our view on mtDNA diversity. Recent studies indicated cryptophytes are related to diverse non-photosynthetic lineages, such as *Palpitomonas bilix*, goniomonads, and kathablepharids, and these lineages are assembly called Cryptista. Although genomic and/or transcriptomic data are available from at least a single species in each subgroup of Cryptista, mtDNA data were available only from two cryptophytes prior to my study. Thus, I sequenced the mtDNAs of *P. bilix* and the katablepharid *Leucocryptos marina* to model the mtDNA evolution in Cryptista. I also determined the complete mtDNA of the haptophyte *Chrysochromulina* sp. NIES-1333, and an as-yet-to-be-described centroheliozoan strain SRT127, which represent two

lineages being proposed to be related to the Cryptista.

Abbreviation list

Genes encoded in mitochondrial genomes of eukaryotes

Functional Categories	Genes
Electron transport and ATP synthesis	
NADH dehydrogenase subunits	*nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, *nad8*, *nad9*, *nad10* , *nad11*
Succinate dehydrogenase subunits	*sdh2*, *sdh3*, *shd4*
Cytochrome *bc*1 complex subunits	*cob*
Cytochrome c oxidase subunits	*cox1*, *cox2*, *cox3*
ATP synthase subunits	*atp1*, *atp3*, *atp4*, *atp6*, *atp8*, *atp9*
Translation	
Small subunit ribosomal proteins	*rps1*, *rps2*, *rps3*, *rps4*, *rps7*, *rps8*, *rps10*, *rps11*, *rps12*, *rps13*, *rps14*, *rps16*, *rps19*
Large subunit ribosomal proteins	*rpl1*, *rpl2*, *rpl5*, *rpl6*, *rpl10*, *rpl11,* *rpl14*, *rpl16*, *rpl18*, *rpl19*, *rpl20*, *rpl27*, *rpl31*, *rpl32*, *rpl34*, *rpl35*, *rpl36*
Elongation factor	*tufA*
Ribosomal RNAs	*rnl*, *rns*, *rrn*5
transfer RNAs	*trnA..Y*
tmRNA	*ssrA*
Transcription	
Core RNA polymerase	*rpoA*, *rpoB*, *rpoC*
Sigma-like factor	*rpoD*
Protein import	
SecY-type transporter	*secY*
SecY-independent transporters	*tatA*, *tatC*
Protein maturation	
Cytochrome *c* oxidase assembly	*cox11*, *cox15*
Cytochrome *c* maturation	*ccmA*, *ccmB*, *ccmC*, *ccmD*
RNA processing	
RNase P	*rnpB*

The functionally assignable genes which were vertically inherited from α-porteobacteria are shown. All genes in the table are present in at least one mitochondrial genome. Modified from Burger et al. (2013)

Terminology

Abbreviation	Description	Abbreiation	Description
BP	Bootstrap probability	**ML**	Maximum likelihood
BPP	Bayesian posterior probability	**MMETSP**	Marine Microbial Eukaryotes Transcriptome Sequencing Project
DIG	Diagoxigenin	**MP**	Maximum parsimony
En	Endonuclease	**nr database**	non-redundant database
gI intron	Group I intron	**mtDNA**	Mitochondrial genome
gII intron	Group II intron	**ORF**	Open reading frame
HE	Homing endonuclease	**PFGE**	Pulsed-field gel electorophoresis
IEP	Intron encoded protein	**RNP**	Ribonucleoprotein
Ma	Maturase	**RT**	Reverse transcriptase

Chapter 1. General introduction

The origin and evolution of mitochondria

Mitochondria, double membrane-bound organelles, are ubiquitously found across the tree of eukaryotes, and responsible for various biological processes, such as oxidative respiration, amino acid metabolism, fatty acid metabolism, FeS cluster assembly, and apoptosis. It is widely accepted that mitochondria can be traced back to a single α-proteobacterial endosymbiont in the common ancestor of extant eukaryotes (Gray et al. 1999; Gray et al. 2001), because all extant eukaryotes harbor the mitochondria or mitochondrion-derived organelles, which are mainly found in eukaryote lineages adapted to microaerophilic/anaerobic environments (Tovar et al. 2003; Embley and Martin 2006; Müller et al. 2012; Makiuchi and Nozaki 2014). The endosymbiotic origin of mitochondria is consistent with the fact that the aforementioned organelles contain bacterial-type genomes. Comparing to α-proteobacterial genomes, mitochondrial genomes (mtDNAs) are drastically reduced in terms of genome size and gene content, suggesting that a large portion of the genes was discarded or transferred to the host nuclear genome during the transition from a bacterial endosymbiont to an organelle (Gray 1993, Gray et al. 1999).

Although all mtDNAs can be traced back to a single α-proteobacterial genome, significant variation in genome size, genome structure, and gene content has been observed among the mtDNAs studied to date (Burger et al. 2003a; Gray et al. 2004). To explain such diversity in mtDNAs, different tempo and mode of mtDNA evolution need to be postulated for different branches in the tree of eukaryotes. For example, human mtDNA is a circular molecule of 16 Kbp in length, which contain only 13 genes encoding proteins involved in electron transfer chain (Boore 1999). However, mtDNAs can be further reduced, as those of malaria parasites (e.g., *Plasmodium*

falciparum) are approximately 6 Kbp in length containing only three protein-coding genes (Ji et al. 1996). In sharp contrast, the most gene-rich mtDNA is that of the jakobid *Andalucia godoyi*: The *A. godoyi* mtDNA possesses 66 genes encoding functionally annotated proteins involved in electron transport chain, as well as translation, transcription, and protein import/maturation (Burger et al. 2013). Interestingly, functionally annotated proteins found in any mtDNAs known to date can be found in the protein set in the *A. godoyi* mtDNA, suggesting that *A. godoyi* possesses the most ancestral mtDNA.

Mitochondrial genomes also vary in terms of architecture. Majority of mtDNAs comprises a single circular molecule. In addition, single-linear, multi-linear, and multi-circular types of mtDNAs have been found in phylogenetically diverse eukaryotes (Burger et al. 2003). Malaria parasites (e.g., *P. falciparum*) and ciliates are known to possess single-linear mtDNAs (Burger et al. 2000; de Graaf et al. 2009). The mesomycetozoean *Amoebidium parasiticum* posesses the mtDNA comprisng a set of hundred distinct linear molecules (Burger et al. 2003b). The mtDNAs of kinetoplastids form the network structure consisting of two types of circular DNA moleculses, (i) 'maxicircles' carrying protein-coding genes, and (ii) 'minicircles' carrying guide RNA-coding genes of which transcripts assist the maturation of mRNAs transcribed from maxicircles (Liu et al. 2005).

Mobile introns in mitochondrial genome

Mitochondria are apparently not immune to lateral transmission of genetic materials, as a large number of mobile introns have been found in mtDNAs. In mtDNAs, two distinct types of mobile introns are found so far; group I (gI) introns and group II (gII) introns (Lang et al. 2007). These introns often harbor open reading frames (ORFs) encoding putative proteins with site-specific

endonuclease activity (intron encoded proteins or IEPs). It is generally believed that IEPs mediate invasions of their host introns into intron-less loci. The intron mobility beyond genomes is one of the major explanations for a complex pattern of intron distribution within closely related genomes (Lambowits 2004; Haugen et al. 2005).

Group I introns found in the nuclear and organellar genomes in eukaryotes, and bacterial genomes, but not in archaeal genomes (Saldanha et al. 1993; Sandegren and Sjöberg 2004). In eukaryotic nuclear genomes, gI introns have been found only in ribosomal RNA (rRNA) genes (Haugen et al. 2004). A typical secondary structure of gI intron RNAs contains 10 helical structures designated as P1-P10 (Haugen et al. 2005, Fig. 1A). Many gI introns were found to harbor ORFs encoding homing endonuclease (HE). Most of the HEs harbored in mitochondrial gI introns belong to LAGLIDADG or GIY-YIG-type endonuclease family, each of the aforementioned protein families possesses characteristic sequence motifs (Stoddard 2005). A HE protein recognizes and introduces double-strand break to the DNA sequence in an intron-less locus, which is identical to the homing site of the host intron (Fig. 2A). Then, the gI intron invades into the cleaved DNA through homologues recombination between the inron-bearing and intron-less loci (Nielsen and Johansen 2009, Fig. 2A).

Group II introns are also identified in various genomes including those of viruses and phages, but not in any eukaryotic genomes (Michel 1982; Toro 2003). Typical gII intron RNAs need to be folded into a characteristic secondary structure consisting of six domains, which are crucial for splicing reaction (Bonen and Vogel 2001, Fig. 1B). The majority of the IEPs harbored in gII introns is composed of three domains, namely (i) reverse transcriptase (RT) domain, (ii) X or sometimes referred to as maturase (Ma) domain, and (iii) endonuclease (En) domain (San Filippo and

Lambowits 2002). Although some gII introns can be self-spliced *in vitro*, IEPs are believed to be indispensable for *in vivo* splicing reaction (Lehmann and Schmidt 2003). An IEP binds to the corresponding unspliced intron RNA, forming a ribonucleoprotein (RNP). Splicing reaction is principally catalyzed by the intron RNA, but X domain of the binding IEP assists the reaction. En and RT domains are likely not used for splicing reaction but intron transmission from an intron-bearing locus to an intron-less locus. For instance, Cousineau et al. (1998) hypothesized the process of lateral transmission of gII intron as follows (Fig. 2B). The spliced RNP recognizes the specific DNA sequence, and inserts the intron RNA into the top DNA strand. En domain cleaves the bottom DNA strand to initiate reverse transcription by RT domain. Finally, the host DNA repair system removes the intron RNA and fills the resultant gap in the top DNA strand (Fig. 2B).

The organisms which are lacking in mitochondrial genome data

The mtDNA diversity described above is mainly based on the data from phylogenetically limited lineages. According to the NCBI Organelle Genome Resource (http://www.ncbi.nlm.nih.gov/genome/browse/?report=5, last accessed at 17th Oct 2015), there are 6,346 completely sequenced mtDNAs, but 6186 (97.4%) are occupied by metazoans, fungi, and land plants. Thus, to understand the true diversity and evolution of mtDNAs, we need to accumulate the mtDNA data from phylogenetically broad lineages (the current consensus knowledge of eukaryotic taxonomy are reviewed in Adl et al 2012; Fig. 3A), in addition to the three branches of the eukaryotic evolutionary tree. Here, I focus on members of a eukaryotic assemblage called Hacrobia (Okamoto et al. 2009; Fig. 3B), of which mtDNAs have not been studied well prior to this study.

Hacrobia was proposed to comprise diverse unicellular eukaryotes (protists) including

both phototrophic and heterotrophic species, namely cryptophytes, goniomonads, *Palpitomonas bilix*, kathablepharids, haptophytes, telonemids, centrohelids, picozoans, and rappemonads (Fig. 3B). Cryptophytes and haptophytes contain plastids that are remnants of red algal endosymbionts (Lane and Archibald 2008). Goniomonads and kathablepharids are heterotrophic, but show clear phylogenetic affinities to cryptophytes (Martin-Cereceda et al. 2010). *P. bilix* is a recently described heterotrophic species, and appeared to be a basal branch of the clade comprising cryptophytes, gonimonads, and kathablepharids (Yabuki et al 2010; Yabuki et al. 2014). Thus, the clade of cryptophytes, gonimonads, kathablepharids, and *P. bilix* is now called as Cryptista (Yabuki et al. 2014). Both telonemids and centrohelids are heterotrophic organisms (Smith and Patterson 1986; Klaveness et al. 2005;). Although the two lineages have been known for a long time, their phylogenetic positions are not resolved yet (Shalchian-Tabrizi et al. 2006; Sakaguchi et al. 2007). Picozoans was firstly recognized by environmental surveys of small subunit ribosomal RNA (SSU rRNA) sequences in sea water, and proposed as potential relatives of cryptophytes and kathablepharids (Not et al. 2007). Initially, picozoans was considered to be photosynthetic, as the picozooan cells often associated with an organelle-like structure with the potential autofluorescence of phycobiliproteins, which are found exclusively in red algal and cryptophyte plastids. However, latter studies consistently and clearly indicated that picozoans are heterotrophic (Seenivasan et al. 2012; Moreira and López-García 2014). Rappemonads are uncultured eukaryotes with no cellular identity, but known only from environmental plastid rRNA sequences (Kim et al. 2011). The putative plastid rRNA sequences of rappemonads and with those of red algal and red alga-derived plastids grouped together with high statistical support (Kim et al. 2011).

In this study, I investigated the mtDNAs of four species belonging to Hacrobia, the

kathablepharid *Leucocryptos marina*, the haptophyte *Chrysochromulina* sp. NIES-1333, *Palpitomonas bilix*, and an undescribed centrohelid strain SRT127. I here discuss the evolution of mtDNA with special emphases on gene content and genome structure, as well as the origins of mobile introns. I further constructed the novel database of alignments for phylogenetic analyses, and applied to the multi-gene phylogeny to inspect the monophyly of Hacrobia based on the mtDNA data including those I sequenced in this study.

Chapter 2. Lateral transfers of mobile introns among distantly related mitochondrial genomes

Summary

So far, only group I (gI) and group II (gII) introns are found in mitochondrial genomes (mtDNAs). Both introns are self-splicing ribozymes which need to be folded into proper secondary and tertiary structures. The distribution of gI/gII introns is often contradicted to the organismal phylogeny, as these introns are mobile genetic elements that can be laterally transferred between distantly related mtDNAs. Intron mobility is likely facilitated by proteins encoded in introns (intron encoded proteins or IEPs), which bind to and cleave double strand DNAs. As sequence specificity varies amongst IEPs, the intron hosting a certain IEP can be inserted into the specific position in a foreign genome. Thus, if introns are found in the homologous positions of different genomes, their IEPs are considered to be homologous to each other. In other words, if IEPs share an intimate evolutionary affinity, their host introns can be homologous to each other.

In this chapter, I discuss the evolutionary origins of introns found in a 12-Kbp fragment of the mtDNA of a kathablepharid *Leucocryptos marina* (2-1), and the complete mtDNA of a haptophyte *Chrysochromulin*a sp. NIES-1333 (2-2), both of which were determined in this study. The mtDNA of *L. marina* contains two gI introns. Comparisons of intron insertion sites and phylogenetic analyses of two IEPs in the *L. marina* introns suggested that the two introns are evolutionarily distinctive to one another; one is homologous to introns in green algal mtDNAs, and the other is to those in fungal mtDNAs. In the organismal phylogeny, neither green algae nor fungi are closely related to kathablepharids, suggesting that intron transfer occurred between distantly related organisms.

The mtDNA of *Chrysochromulin*a sp. NIES-1333 harbors three gII introns, and only one of the three introns was found to encode an IEP. Curiously, an open reading frame (ORF), of which amino acid sequence shows a strong similarity to those of gII IEPs, was found outside of introns (henceforth here designated as IEP-like ORF). Based on a phylogenetic analysis of IEP sequences, I concluded that the IEP-containing intron shares the origin with an intron found in the mtDNA of diatom. I also recovered a strong affinity between the IEP-like ORF in the *Chrysochromulina* mtDNA and an IEP detected in the mtDNA of the diatom *Phaeodactylum tricornutum*. Thus, I propose that (i) two out of the three introns in the *Chrysochromulina* mtDNA shared the origins with the introns in diatom mtDNAs, and (ii) the IEP-like ORF in the *Chrysochromulina* mtDNA was used to be 'intron-hosted.'

2-1. *Leucocryptos marina*

Introduction

Group I introns are one of the major classes of introns found in bacterial genomes, mitochondrial and plastid genomes, and eukaryotic nuclear genomes (Saldanha et al. 1993; Bhattacharya 1998), as well as genomes of viruses/phages (Sandegren and Sjöberg 2004). In eukaryotes, gI introns in nuclear genomes are exclusively inserted in ribosomal RNA (rRNA) genes, whereas the introns reside in genes for both rRNAs and proteins in organellar genomes (Cannone et al. 2002). Group I introns need to be spliced by folding a characteristic secondary and tertiary structures. The typical secondary structure of gI introns consists of approximately 10 double helical elements designated as P1-P10 (Haugen et al. 2005; Edgell et al. 2011). These helical elements are organized into three domains at the tertiary structural level, which are important for efficient splicing of this class of introns (Adams et al. 2004). Many gI introns host ORFs for homing endonucleases (HEs) (Belfort

and Roberts 1997), which may facilitate intron invasion into the intron-less alleles within a population of the same species, as well as those in different species (Sellem et al. 1996; Johansen et al. 1997; Sanchez-Puerta et al. 2011). Intron-encoded (intronic) HEs in mtDNAs are divided into two types, such as LAGLIDADG and GIY-YIG superfamilies, on the basis of highly conserved motifs (Stoddard 2005).

Kathablepharida is a group of heterotrophic eukaryotes distributed in diverse aquatic environments (Auer and Arndt 2001). A phylogenetic analysis using a multi-gene dataset strongly suggests that katablepharids, goniomonads, cryptophytes and *Palpitomonas* together form a monophyletic clade, Cryptista (Yabuki et al. 2014). Here, I report two gI introns hosting LAGLIDADG-type HEs in the mtDNA of kathablepharid *Leucocryptos marina*, and explored the evolutionary histories of these introns by combining their putative secondary structures, the intron positions, and the phylogenetic affinities of the intronic HEs.

Material & Methods

Cell culture

The cultures of the kathablepharid *L. marina* NIES-1335 and the haptophyte *Chrysochromulina* sp. NIES-1333 were purchased from the National Institutes for Environmental Study (NIES). *L. marina* was cultured in f/2 medium (http://mcc.nies.go.jp/02medium.html#2) with *Chyrsochromulina* sp. NIES-1333 as a prey at 20 ºC under 14 h light/10 h dark cycles.

Extraction of DNA and RNA and preparation of cDNA

The *L. marina* cells, together with the prey (*Chrysochromulina*) cells, were harvested by centrifugation and then subjected to DNA and RNA extractions by using Plant DNA Isolation

Reagent (TaKaRa) and RNeasy Plant Minit Kit (QIAGEN), respectively. Complementary DNA was synthesized from total RNA by Superscript II reverse transcriptase (Invitorogen) with random hexamers. These experiments mentioned above were conducted by following the manufactures' instructions. The DNA and cDNA were used as the templates for polymerase chain reactions (PCR), intending to amplification of mtDNA fragments and gene transcripts respectively.

Amplification of mitochondrial genes and sequencing

Six mitochondrial gene transcripts, namely *cob*, *cox1*, *cox3*, *nad1*, *nad7*, and *nad11*, were amplified by reverse transcriptase PCR (RT-PCR) with the primer sets shown in Table 1. PCR products were cloned into pGEM-T Easy vector (Promega). For each gene transcript, 8 clones were completely sequenced and compared to confirm no sequence heterogeneity among clones, except the *cob* and *cox3* transcripts. The *cob* and *cox3* PCR amplicons appeared to consist of two distinctive types, one with and the other without in-frame TGA stop codons (no in-frame TGA codon was found in the *cox1*, *nad1*, *nad7* or *nad11*). The amplicons with in-frame TGA codon were considered to be from the haptophyte prey cells for two reasons. (i) Phylogenetic analyses indicated that the two amplicons were distantly related to each other, and only the one with in-frame TGA stop codons showed a close affinity to the haptophyte homologues (Fig. 4). (ii) The genus *Chrysochromulina* belongs to one of the two classes in Haptophyta, Prymnesiophyceae, whose mtDNAs assign TGA to tryptophan (Hayashi-Ishimaru et al. 1997; Inagaki et al. 1998; Puerta et al. 2004). Based on the phylogenetic analyses and non-standard usage of TGA codon, I concluded that the *cob* and *cox3* transcripts with in-frame TGA codons were likely to originate from the haptophyte cells, and were not considered in the following experiments.

The intergenic spacer regions between *nad11* and *nad1*, *nad1* and *nad7*, *nad7* and *cox3*, *cox3* and *cob*, and *cob* and *cox1* were also amplified with outwarded exact match primers designed based on the six mitochondrial gene sequences determined beforehand (see the above paragraph). These PCR were performed as described in Masuda et al. (2011) and Kamikawa et al. (2009). Cloning and sequencing of PCR products were conducted as described above. The partial mtDNA sequence was deposited to DNA Data Bank of Japan (GenBank/EMBL/DDBJ accession no. AB63966).

Identification of gI introns and secondary structure prediction

Each of *cob* and *cox1* genes in the *L. marina* mtDNA appeared to be interrupted by a single gI intron with a HE. The boundaries of exon and intron were determined by comparing the corresponding cDNA and genomic sequences. The intron secondary structures were predicted using MOLD (Zuker 2003), followed by manual modification by referring the general structures of gI introns presented in GOBASE (O'Brien et al. 2009).

Phylogenetic analysis of intronic HEs

The HE encoded in the *L. marina cob* intron ($HE^{Lm\text{-}cob}$) was aligned with 29 HEs belonging to the LAGALIDADG_2 superfamily, which showed significant similarity to $HE^{Lm\text{-}cob}$ in TBLASTN search against the GenBank non-redundant (nr) database (*E* value < 10^{-10}). The alignments from the BLAST search were carefully assessed, and excluded redundant sequences and the sequences which produced very short alignments with $HE^{Lm\text{-}cob}$. After manual refinement followed by the exclusion of ambiguously aligned positions, 183 amino acid (aa) positions remained in the final 'LAGLIDADG_2' alignment. The same procedure described above was repeated to prepare a

'LAGLIDADG_1' alignment including the HE encoded in *Leucocryptos cox1* intron ($HE^{Lm\text{-}cox1}$). The final LAGLIDADG_1 alignments contain 25 HEs and 191 unambiguously aligned aa positions.

The two HE alignments were separately subjected to maximum likelihood (ML) analysis. The LG model (Le and Gascuel 2008) incorporating empirical aa frequencies and among-site rate variation approximated by a discrete gamma (Γ) distribution with four categories (LG+Γ+F) was selected as the most appropriate model for the aa substitutions in the LAGLIDADG_1 alignment by the program Aminosan (Tanabe 2011) under the Akaike information criterion. Similarly, the VT model (Müller and Vingron 2000) incorporating empirical aa frequencies and among-site rate variation approximated by a discrete Γ distribution with four categories (VT+Γ+F) was selected as the most appropriate model for the LAGLIDADG_2 alignment. The ML analyses were performed using RAxML 7.2.1 (Stamatakis 2006) with the selected models described above. The ML tree was heuristically searched from 10 distinct maximum-parsimony (MP) starting trees. In RAxML bootstrap analyses (100 replicates), the heuristic tree search was performed from a single MP starting tree per replicate.

The two HE alignments were also analyzed by Bayesian inference with the LG+Γ+F model using PhyloBayes v. 3.2 (Lartillot et al. 2009). As VT model is not available in PhyloBayes, the LG+Γ+F model was instead applied to the LAGLIDADG_2 alignments. Two independent Markov Chain Monte Carlo chains (MCMC) were run for 72,000-78,000 generations. The first 100 generations were discarded as 'burn-in' on the basis of the log-likelihood plots (data not shown). For each analysis, the frequencies of all bipartitions observed in the two independent MCMC runs were compared, and the convergence between the two chains were checked by the 'maxdiff' value being smaller than that recommended in the PhyloBayes manual (i.e, maxdiff < 0.1). Subsequently, the

consensus trees with branch lengths and Bayesian posterior probabilities (BPPs) were calculated from the rest of the sampling trees.

Results & Discussion

Overview of the 12 kbp mitochondrial genome fragment of *L.marina* and group I introns

I successfully sequenced an approximately 12 Kbp-long fragment of *L. marina* mtDNA including 9 genes (*nad11*, *nad1*, *nad6*, *atp6*, *nad7*, *cox2*, *cox3*, *cob* and *cox1* in this order, Fig. 5). The intergenic spacer regions are short, ranging from 4-64 bp in length. Neither tRNA nor rRNA gene was identified.

By the comparison between the cDNA and genomic sequences, two introns, one in the *cob* gene and the other in the *cox1* gene, were detected. No sign of RNA editing was found so far. Both two introns are likely of gI, as these sequences can be folded into typical secondary structures comprising of 11-12 double helical domains referring to P1-P10 (Fig. 6). In our BLASTN survey, the putative core regions of the *cob* intron showed similarity to the *cob* intron in a green alga *Chaetosphaeridium globsum* (GenBank accession: AF494279), which is classified as a member of group ID, with an *E* value of 10^{-13}. On the other hand, the putative core region of the *cox1* intron appeared to share sequence similarity with those of group IA1 introns (e.g., the one lying in the large subunit of mitochondrial rRNA gene of a green alga *Scenedesmus obliqus*, GenBank accession: AF202057, with an *E* value of 2×10^{-6}). The two introns are distinguishable from one another by the two following features: (i) The *cox1* intron has two extra stems, P7.1 and P9.1, which are absent in the *cob* intron, and (ii) The *cob* intron harbors an ORF between P1 and P2, while the *cox1* intron has an ORF between P1 and P10 (see Fig. 6).

The ORFs hosted in the *cob* and *cox1* introns encode 217 aa residue-long and 267 aa residue-long polypeptides, respectively. The two intronic ORFs likely encode LAGLIDADG-type HEs, but no significant similarity was detected between their putative aa sequences by a BLASTP search (bl2seq) with default parameters. Henceforth here, the HE hosted in the *cob* and *cox1* genes are designated as $HE^{Lm\text{-}cob}$ and $HE^{Lm\text{-}cox1}$, respectively. $HE^{Lm\text{-}cob}$ appeared to belong to LAGLIDADG_2 superfamily (pfam031611), while $HE^{Lm\text{-}cox1}$ shows affinity to superfamily LAGLIDADG_1 (pfam00961).

Origin of the *Leucocryptos cob* intron

The alignment consisting of 30 LAGLIDADG_2 HEs sequences including $HE^{Lm\text{-}cob}$ was prepared and subjected to the ML and Bayesian phylogenetic analyses. In the unrooted ML tree of this alignment, $HE^{Lm\text{-}cob}$ and two HEs encoded in the *cob* of two green algae, *Nephroselmis olivacea* and *Chlorokybus atmophyticus*, grouped together with a BP of 98% and a BPP of 1.00, suggesting that $HE^{Lm\text{-}cob}$ and green algal HEs evolved from a single ancestral protein (Fig. 7A). The ancestral intron most likely (i) lied at the phase-0 position of the codon corresponding to Gln138 in the *Saccharomyces cerevisiae cob* gene (GenBank accession: NC_001224), and (ii) hosted a LAGDLIDADG_2 HE in the loop region between P1 and P2 as shown in Fig. 8A. Unfortunately, it is difficult to retrieve deeper insights for the origin of the *Leucocryptos cob* intron by inserted positions, as the HEs hosted by the introns lying in the homologous positions were sporadically distributed in the LAGLIDADG_2 phylogeny.

Origin of the *Leucocryptos cox1* intron

A 'LAGLIDADG_1' alignment comprising the aa sequences of $HE^{Lm\text{-}cox1}$ and 24 members of

LAGLIDADG_1 superfamily was subjected to the ML and Bayesian phylogenetic analyses. The unrooted LAGLIDADG_1 ML phylogeny united $HE^{Lm\text{-}cox1}$ and the HE hosted in the fourth out of 15 *cox1* introns in the fungus *Rhizophydium* sp. with a BP of 71% and a BPP of 1.00 (Fig. 7B). Although the statistical support for this clade was inconclusive, the introns hosting these HEs described above exclusively share the homing position; phase-0 of the codon corresponding to Thr93 in the *S. cerevisiae cox1* gene (GenBank accession: NC_001224). Thus, the *Leucocryptos cox1* intron and the fourth intron in *Rhizophydium cox1* gene likely derived from a single ancestral intron, which lied at phase-0 of the codon corresponding to Thr93 in the *S. cerevisiae* homologue, and hosted a LAGLIDADG_1 HE in the loop region between P1 and P10 (Fig. 8B).

The clan of $HE^{Lm\text{-}cox1}$ and the HE in the fourth intron of *Rhizophidium cox1* was further connected to the HE encoded in the first out of 16 *cox1* introns of the fungal *Podospora anserina*, and that encoded in a single intron of the mycetozoan *Dictyostelium fasciculatam* (BP of 70% and BPP of 0.99). Both *Podospora* and *Dictyostelium* introns lie at the phase-1 of the codon corresponding to Ala94 in the *S. cerevisiae cox1* gene, being in close proximity but apparently distinct from the homing position of the *Leucocryptos* and *Rhizophidum* introns. One possibility is that $HE^{Lm\text{-}cox1}$ and the *Rhizophidum* HE, and the *Podospora* and *Dictyostelium* HEs have evolved from a single ancestral HE and still recognize the identical (or very similar) nucleotide sequences, but the cleavage position altered after the separation of two HE pairs. In any case, the evolutionary link between the *cox1* introns in *Leucocryptos* and *Rhizophidium*, and those in *Podospora* and *Dictyostelium* can be assessed only after the enzymatic properties of the HEs hosted in the four *cox1* introns are characterized.

Intron evolution in the *Leucocryptos* mitochondrial genome

Introns in organellar genomes are generally regarded as mobile genetic elements powered by intronic HEs, as 'trans-genomic' invasion have been accumulated in the literature (Sanchez-Puerta et al. 2011). In the global eukaryotic phylogeny inferred in the recent study, kathablepharids highly likely from a clade with goniomonads, cyrptomonads, and *Palpitomons bilix*, referred to as Cryptista (Yabuki et al. 2014), but are closely related to neither green algae nor fungi (Okamoto and Inouye 2005; Kim et al. 2006). Thus, the evolutionarily homologous introns resides in distantly related mtDNAs can be rationalized by lateral transfer events. Nevertheless, considering the cyclic model for gain and loss of selfish genetic elements including gI introns (Goddard and Burt 1999), we cannot exclude the alternative scenario which assumes that (i) the two introns in *cob* and *cox*1 genes discussed above have been vertically inherited from the common ancestor of a large taxonomic assemblage including kathablepharids, green algae and fungi, but (ii) secondary intron loss occurred in other descendant lineages. Nevertheless, HE sequences considered here highly unlikely represent the true diversity of LAGLIDADG_2 and/or LAGLIDADG_1 HE superfamilies. Thus, the origins and evolutions of the two gI introns found in the *Leucocyptos* mtDNA should be revisited after in-depth surveying introns and intronic HEs in the mtDNAs of phylogenetically broad eukaryotic lineages, particularly those of close relatives of kathablepharids, such as goniomonads and cyrptophytes.

2-2. *Chrysochromulina* sp. NIES-1333

Introduction

Group II introns are one of the major classes of introns, and found in the genomes of prokaryotes (bacteria and archaea) (Toro 2003), mitochondria and plastids (Michel et al. 1982) which are derived

from an α-proteobacterium and a cyanobacterium, respectively (Gray et al. 1999; Gould et al .2008). So far, gII introns have been identified in mtDNAs from members of phylogenetically diverse eukaryotic groups such as metazoans (Dellaporta et al. 2006; Vallès et al 2008), jakobids (Lang et al 1997; Burger et al. 2013), members of Archaeplastida (Bégu et al. 2009; Turmel et al. 2007; Mao et al. 2012), fungi (Paquin and Lang 1996; Foury et al. 1998), cryptophytes (Hauth et al. 2005), haptophytes (Ehara et al. 2000; Smith et al. 2014) and stramenopiles (Oudot-Le Secq et al. 2001; Kamikawa et al. 2009; Ravin et al. 2010). These gII introns possess features at both primary and secondary structural levels. At the primary structural level, gII introns possess highly conserved sequence motif at the 5′ and 3′ ends (i.e., 5′-GTGYG…AY-3′; Y for T or C) (Bonen and Vogel 2001). At the secondary structural level, we anticipate the transcripts of typical gII introns (intron RNAs) to form a characteristic bulge structure with six stems, so-called domains I to VI (Toor et al. 2001). Both primary and secondary structures of gII intron RNAs are generally believed to be critical for splicing reaction (Lambowitz and Zimmerly 2004). 5′

Group II introns can be regarded as mobile genetic elements, which are transmittable between an inron-bearing and intron-less alleles (intron homing). The mobility of gII introns are most likely conferred by IEPs. Typical IEPs comprise three functionary distinct domains, namely (i) reverse transcriptase (RT) domain, (ii) X (or maturase) domain, and (iii) endonuclease (En) domain (San Filippo and Lambowitz 2002), although some IEPs were reported to lack RT and/or En domains (Bonen and Vogel 2001; Lambowitz and Zimmerly 2011; Zimmerly and Semper 2015). Among the three domains in IEPs, RT and En domains are predicted to catalyze reverse transcription of intron RNA and cleave the target (intron-less) allele, respectively (Bonen and Vogel 2001; San Filippo and Lambowitz 2002). Domain X may not be responsible for intron mobility, but assists

splicing by stabilizing the conformation of intron RNA. Nevertheless, we have known of many 'IEP-free' gII intron, and it is difficult to predict the protein factors, which cooperate with a particular IEP-free intron in *trans*. As far as I know, this is the first report that predicted as the mtDNA-encoded *trans* factor involved in the splicing of IEP-free introns in mtDNA.

Here, I completely sequenced the mtDNA of the haptophyte *Chrysochromulina* sp. NIES-1333, and identified three introns in total, two of those are found in *cox1* gene and the other is found in *rnl* gene. Analyses of the intron sequences suggest that the three introns in the *Chrysochromulina* mtDNA belonged to gII. I identified two ORFs encoding putative IEPs. Both showed significant similarity to gII intron-hosted IEPs in the mtDNAs; one is *orf627* encoded in the second *cox1* intron, and the other is *orf584*, which is free-standing (i.e. not hosted by any introns). Phylogenetic analyses of IEPs and comparisons of intron position across phylogenetically diverse mtDNAs revealed that the *Chrysochromulina* mtDNA shares homologous introns with those of distantly related species.

Material & Methods

Cell culture, extraction of DNA/RNA, and preparing cDNA

The haptophyte *Chrysochromulina* sp. NIES-1333 was purchased from NIES. The haptophyte cells were grown in f/2 medium at 20 ºC under 14 h light/10 h dark cycles. The cells were harvested by centrifugation. Total DNA and total RNA were extracted from the harvested cells by CTAB buffer as described in Kamikawa et al. (2005) and TRIzol (Invitrogen), respectively. Total RNA was used to synthesize cDNA with random hexamers and Superscript II reverse transcriptase (Invitrogen). RNA extraction and cDNA synthesis were performed following manufactures' protocols.

Amplification of mitochondrial genes and sequencing

The entire mtDNA was amplified by combination of PCR with LA *Taq* DNA polymerase (TaKaRa), genome walking with Genome Walker Universal kit (Clonthech). Amplified DNA fragment < 3 Kbp and those of < 10 Kbp were cloned into pGEM-T Easy vector (Promega) and pCR-XL-TOPO vector, respectively. The short PCR products (< 3 Kbp) were sequenced by the Sanger method using ABI 3130 (Applied Biosystems). 454 pyro-sequencing by the Gs Junior system was performed on the long (> 10 Kbp) amplicons. Newbler (454 Sequencing, Roche) was applied to assemble the pyro-sequencing reads. The DNA amplification and sequencing described above were conducted as described in manufactures' instructions. The *Chrysochromulina* mtDNA was finally assembled into a circular molecule, with an approximate 34 Kbp in length. The complete mtDNA sequence is available in DDBL/EMBL/GenBank under the accession number AB930144.

Genome annotation

Genes encoding proteins and rRNA were identified by BLASTX and BLASTN searches against the NCBI nr database, respectively (Altschul et al. 1990). Transfer RNA genes were identified by using tRNAscan-SE (Schattner et al. 2005). Independent from the analyses described above, we re-annotated the genome by MFannot (http://megasun.bch.umontreal.ca/RNAweasel/).

Both *cox1* and *rnl* genes in the *Chryshochromulina* mtDNA appeared to be intervened by introns. The precise intron-exon boundaries were determined by sequencing the corresponding transcripts (cDNAs) of the intron-containing genes. The secondary structures of the introns identified in the mtDNA were predicted by MFOLD (Zuker 2003), followed by manual refinement by referring to Toor et al. (2001) and GOBASE database (O'Brien et al. 2009).

Phylogenetic analysis

The two ORFs (*orf627* and *orf584*), of which conceptual aa sequences show high similarity to the IEPs of gII intorns, were found in the *Chrysochromulina* mtDNA. The conceptual aa sequences of the two ORFs were aligned with those of 46 IEPs encoded in other mtDNAs and 4 bacterial homologues by Muscle (Edgar 2004). The IEP sequences were retrieved from the GenBank database by referring to previous studies (Kamikawa et al. 2009; Ravin et al. 2010). After manual refinements and exclusion of ambiguously aligned positions, the final alignment includes 52 IEPs and 453 aa positions.

The alignment described above was subjected to the ML and Bayesian methods to infer phylogenetic relationship using RAxML 7.2.6 (Stamatakis 2006) and PhyloBayes 3.3 (Lartillot et al. 2009), respectively. The LG+Γ+F model (Le and Gascuel 2008) was applied for both ML and Bayesian inferences. The ML tree was selected by heuristic searches from 10 randomized MP starting trees. In RAxML bootstrap analyses (100 replicates), the heuristic tree search was performed from a single MP starting tree per replicate. In Bayesian analysis, two independent MCMC chains were run for 5,800-5,850 cycles, reaching maxdiff value of 0.08353 (Manual suggests that maxdiff is smaller than 0.1 when chains reach to convergence). The first 100 cycles were discarded as 'burn-in'; the consensus tree, branch length, and BPPs were calculated from the remaining trees.

Results & Discussion

Overview of the *Chrysochromulina* mitochondrial genome

The mtDNA of *Chrysochromulina* sp. NIES-1333 was assembled into a circular molecule of 34,291 bp in length with an A+T content of 70.0% (Fig. 9). We identified 16 functionally assignable ORFs (including those for two IEPs; see below). UGA codons are most likely assigned for tryptophan

instead of terminal signal, as reported in other member of Prymnesiophyceae, one of the two classes of Haptophyta (Hayashi-Ishimaru et al. 1997; Inagaki et al. 1998; Puerta et al. 2004). We detected 26 tRNA genes and a set of small and large subunits of rRNAs; No 5S rRNA gene was identified. A set of tRNA genes identified in the mtDNA is sufficient to translate all aa codons except for GGN (N = A, G, C, or U) codon for glycine (Table 2). All genes mentioned above were encoded on a single strand. A region with an approximate length of 1.6 Kbp, which contained a single tRNA gene for isoleucine, was found to be duplicated (arrows in Fig. 9).

In terms of gene repertory, the *Chrysochromulina* mtDNA in fundamentally similar to those of other haptophytes, namely *Emiliania huxleyi* (Puerta et al. 2004; Smith and Keeling et al. 2012), *Chrychochromulina tobin* (Hovde et al. 2014), *Diacronema lutheri* (www.bch.umontreal.ca/ogmp/projects/pluthgen.html), *Phaeocystis* spp. (Smith et al. 2014), as shown in Table 3.

General features of the intron in the *Chrysochromulina rnl* gene

The *rnl* gene hosts a single intron with no apparent ORF (designated as Ch_rnli). The intron was found to be inserted at the position between the 837th and 838th bases in the *Homo sapiens* homolog (GeneID: 4550 in NC_012920). Ch_rnli starts with 5′-GTGCG… and ends with …CT-3’, which is similar to the consensus motifs shared amongst typical gII introns (5′-GTGYG…AY-37, Y for T or C). Although the intron sequence was too divergent to predict the entire secondary structure, the domains V and VI, the typical secondary structures of gII introns, were successfully identified with the aid of MFannot (Fig. 10). Thus, I characterized Ch_rnli as a gII intron.

The homing position of Ch_rnli was found to be identical to those of *rnl* introns found in

a member of Archaeplastida (the red alga *Pyropia haitanesis*; NC_017751), two members of Stramenopiles (the blown alga *Pylaiell alittorali*s, and the diatom *Phaeodactylum tricornutum*; NC_003055 and HQ840789, respectively), and a member of Fungi (*Gigaspora rose*a; NC_016985) (Fig. 10). Note that none of the *rnl* intron in the mtDNAs of other haptophytes, *D. lutheri* and *Phaeocystis globosa*, shares the insertion positions with that of *Chrysochromulina* sp NIES-1333. The secondary structures of domains V and VI are predicted in Ch_rnli and the four introns described above, but detected no apparent homology at the nucleotide sequence and secondary structure level among them (Fig. 10). Furthermore, Ch_rnli has no IEPs, which is a key aspect to inferring intron evolution (Lambowitz and Zimmerly 2011). Thus, I avoid discussing the evolutionary relationship among Ch_rnli and the introns listed above, solely based on their homing positions.

General features of the two introns in the *Chrysochromulina cox1* gene

Two introns were identified in the *Chrysochromulina cox1* gene. I designated the first and second introns in *cox1* gene as Ch_cox1i1 and Ch_cox1i2, respectively. Both introns starts with 5′-GTGCG… and ends with …AC-3', being consistent with the consensus motifs of gII introns (5′-GTGYG…AY-3'). Both Ch_cox1i1 and Ch_cox1i12 can be folded into the characteristic secondary structures shared among gII introns, albeit with some ambiguity remaining in domain I (indicated as 'DI' in Fig. 11 A & B). Altogether, the two *cox1* introns were considered to belong to gII. Ch_cox1i1 is inserted in phase-2 of the codon corresponding to Phe68 in the *Homo sapiens cox1* gene, sharing the homing position with the gII intron in the *cox1* genes of the cryptophyte *Rhodomonas salina*, and the diatom *Phaeodactylum tricornatum* (Fig. 11). Ch_cox1li2 was found at

phase-2 of the codon corresponding to Phe237 in the *H. sapiens cox1* gene, being homologous to those of the gII intron in the *cox1* genes of the haptophyte *D. lutheri* and the diatom *Ulnaria acus* (Fig. 11D). Ch_cox1i2 hosts an IEP, while Ch_cox1i1 encodes no apparent ORF.

Evolution of Ch_cox1i2 and its IEP

The IEP encoded in Ch_cox1i2, ORF627, most likely facilitates splicing of the host intron. The ORF627 aa sequence showed apparent similarity to other gII intron-hosted IEP sequences deposited in the GenBank database; the top BLAST hit was an IEP encoded in the first gII intron of the *cox1* gene in the haptophyte *D. lutheri* (Dl_cox1i) with a 49% sequence similarity and an *E*-value of 0.0. In both ML and Bayesian analysis of the IEP alignment, *Chrysochromulina* ORF627 formed a clade with two IEPs in *cox1* gII introns, namely Dl_cox1i and that of the diatom *U. acus* (Ua_cox1i) with a ML bootstrap value (MLBP) of 96% and BPP of 1.00 (Fig. 12). As we generally believe that gII introns and their IEPs have coevolved (Lambowitz and Zimmerly 2011), the intimate relationship among the IEPs encoded in Ch_cox1i2, Dl_cox1i and Ua_cox1i suggests that their host introns are derived from a single ancestral intron bearing an IEP. The single origin of Ch_cox1i2, Dl_cox1i, and Ua_cox1i discussed above is consistent with the fact that the three introns share a homing position (Fig 11).

The ancestral haptophyte species likely possesses a *cox1* gene with a particular gII intron, as *Chrysochromulin*a sp. and *D. lutheri* are representatives of two major classes in Haptophyta, Prymnesiophyceae and Pavlovaphyceae, respectively. The scenario demands that multiple intron losses occurred in the *cox1* genes of *Emiliania huxleyi*, members of the genus *Phaeocyctis*, *Isochrysis galbana* and *Chrysochromulina tobin*.

The IEP phylogeny and comparison of homing positions imply that the homologues introns are present in two distantly related branches (haptophytes and diatoms) in the tree of eukaryotes. This sporadic intron distribution can be explained by a scenario incorporating lateral intron transfer. There is an alternative, but less plausible scenario assuming that the *cox1* gene in the ancestral organism, which has existed to prior to the divergence of major eukaryotic assemblages including diatoms and haptophytes, may have already possessed a gII intron at phase-2 of the codon corresponding to Phe237 in the *H. sapiens cox1* gene, and would have been (secondarily) lost in multiple descendants (i.e., ancestral co-occurrence followed by multiple secondary losses). I prefer the scenario incorporating lateral intron transfer to the alternative scenario, but these scenarios should be reexamined by future studies based on broader diversity of gII introns (and their IEPs) compared with those considered in this study.

Link between a free-standing *orf584* and an IEP-free Ch_cox1i1

Most IEPs in mtDNAs are encoded in intronic ORFs (as observed in Ch_cox1i2; see above), but a few of those are free-standing (e.g., OFR732 in the liverworts *Marchantia polymorpha*; highlighted by a star in Fig. 12). The BLAST search showed that *Chrysochromulina* ORF584, which is free-standing in the genome, bore a significant sequence similarity to gII intron-hosted IEPs; The top BLAST hit of ORF584 aa sequence was the IEP (OER724) encoded in a first intron of *cox1* gene in the diatom *P. tricornutum* (Pt_cox1i1) with a 53% sequence similarity and *E*-value of 0.0. ORF584 equips RT, maturase/X, and En domains, implying that this protein assists intron splicing. The phylogenetic analysis recovered a robust affinity between *Chrysochromulina* ORF584 and ORF724 encoded in Pt_cox1i1 with a MLBP of 100% and a BPP of 1.00 (Fig. 12). This indicates that the two

proteins were derived from the single ancestral IEP encoded in a gII intron, which is homologous to Pt_cox1i1, the host intron of ORF724. Curiously, Pt_cox1i1 and Ch_cox1i1 appear to share a homing position (Fig. 11D). I also noticed that the nucleotide sequence of domain VI in Ch_cox1i1 and that in Pt_cox1i1 are similar to one another (Fig. 11C), although this domain sequences are generally variable among gII introns (Toor et al. 2001). The homing position and sequence similarity in domain VI between Ch_cox1i1 and Pt_cox1i1 suggest that the two introns are homologous to each other. Altogether, I here propose that ORF584 used to be encoded in Ch_cox1i1, and still assists the splicing of the host intron even after being free-standing secondarily in the current *Chrysochromulina* mtDNA. To the best of my knowledge, this is the first report of co-relation between free-standing 'IEP' and IEP-free intron in a mtDNA [see Zoschke et al. (2010) for a similar case but in a plastid genome].

The first intron in the *R. salina cox1* gene (Rs_cox1i1) is unlikely to be homologous to Pt_cox1i1 or Ch_cox1i1, although the three introns share the homing position (Fig 11). The IEP phylogeny placed the IEP encoded in Rs_cox1i1 in a remote position from the clade of ORF584 and ORF724 (Fig. 12), strongly arguing against the homology between Rs_cox1i1 and Pt_cox1i1/Ch_cox1i1. The homology between Pt_cox1i1 and Ch_cox1i1, which were found in two phylogenetically distant related species (i.e., a haptophyte and a diatom), can be explained by lateral intron transfer. Nonetheless, we cannot exclude the alternative possibility assuming ancestral co-occurrence followed by multiple secondary losses. I prefer the simplicity of the first scenario incorporating lateral intron transfer, but the alternative scenario should not be ignored before mtDNA diversity is sufficiently covered.

Acknowledgement

First of all, I am deeply grateful to my supervisors, Associate Prof. Yuji Inagaki and Prof. Tetsuo Hashimoto (University of Tsukuba) for their helpful discussion and warm encouragement during the PhD program. I would like to express my gratitude to Assistant Prof. Ryoma Kamikawa (Kyoto University) for eager training in the earliest stage of my research carrier. I cannot finish this study without extraordinarily tolerant and supports of them three. Associate Prof. Toshiyuki Amagasa accepted me as a graduate student in dual degree programs and greatly contributed to construct my alignment database. Assistant Prof. Goro Tanifuji (University of Tsukuba) gave me a lot of technical advice for pulsed-field gel electrophoresis, southern blot analysis and suggested using useful programs such as SPAdes, Botie2 and so on. Dr. Takuro Nayakama also gave salutary suggestions about using and writing computer programs with Ruby. The culture of *Palpitomonas bilix* was kindly provided by Dr. Akinori Yabuki (Japan Agency for Marine-Earth Science and Technology, JAMSTEC). Mr. Takashi Shiratori (University of Tsukuba) kindly provided the culture of SRT127. I received technical supports for 454 pyrosequencig from Ms. Hiroko Yuki (Okinawa Institute of Science and Technology, OIST). Finally, I thank all the lab member of Molecular Evolution of Microbes (Hashi & Yuji lab at University of Tsukuba), Plant Diversity and Evolutionary Cell Biology (Ishida lab at University of Tsukuba), and Kitagawa Data Engineering (University of Tsukuba) for their helps.

I am supported by a Japan Society for the Promotions of Science (JSPS) Research Fellowship for Young Scientists and this study was funded in part of KAKENHI (No. 25789)

References

Adams PL, Stahley MR, Kosek AB, Wang J, Strobel SA. 2004. Crystal structure of a self-splicing group I intron with both exons. Nature. 430:45–50.

Adl SM, Simpson AG, Lane CE, et al. 2012. The revised classification of eukaryotes. J. Eukaryot. Microbiol. 59:429–493.

Aguileta G, de Vienne D, Ross O, Hood M, Giraud T, Petit E, Gabaldón T. 2014. High Variability of Mitochondrial Gene Order among Fungi. Genome Biol Evol. 6:451–465.

Allen JW, Ginger ML, Ferguson SJ. 2004. Maturation of the unusual single-cysteine (XXXCH) mitochondrial *c*-type cytochromes found in trypanosomatids must occur through a novel biogenesis pathway. Biochem. J. 383:537–542.

Allen JW, Jackson AP, Rigden DJ, Willis AC, Ferguson SJ, Ginger ML. 2008. Order within a mosaic distribution of mitochondrial *c*-type cytochrome biogenesis systems? FEBS J. 275:2385–2402.

Allen JW. 2011. Cytochrome *c* biogenesis in mitochondria--Systems III and V. FEBS J. 278:4198–4216.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Auer B, Arndt H. 2001. Taxonomic composition and biomass of heterotrophic flagellates in relation to lake trophy and season. Freshwater Biol. 46:959–972.

Babbitt SE, Sutherland MC, Francisco BS, Mendez DL, Kranz RG. 2015. Mitochondrial cytochrome *c* biogenesis: no longer an enigma. Trends Biochem. Sci. 40:446–455.

Bégu D, Araya A. 2009. The horsetail *Equisetum arvense* mitochondria share two group I introns with the liverwort *Marchantia*, acquired a novel group II intron but lost intron-encoded ORFs. Curr. Genet. 55:69–79.

Belfort M, Roberts RJ. 1997. Homing endonucleases: keeping the house in order. Nucleic Acids Res. 25:3379–3388.

Bhattacharya D. 1998. The origin and evolution of protist group I introns. Protist. 149:113–122.

Bonen L, Vogel J. 2001. The ins and outs of group II introns. Trends Genet. 17:322–331.

Boore JL. 1999. Animal mitochondrial genomes. Nucleic Acids Res. 27:1767–1780.

Brown MW, Kolisko M, Silberman JD, Roger AJ. 2012. Aggregative multicellularity evolved independently in the eukaryotic supergroup Rhizaria. Curr. Biol. 22:1123–1127.

Burger G, Zhu Y, Littlejohn TG, Greenwood SJ, Schnare MN, Lang BF, Gray MW. 2000. Complete sequence of the mitochondrial genome of *Tetrahymena pyriformis* and comparison with *Paramecium aureli*a mitochondrial DNA. J. Mol. Biol. 297:365–380.

Burger G, Gray MW, Lang BF. 2003a. Mitochondrial genomes: anything goes. Trends Genet. 19:709–716.

Burger G, Forget L, Zhu Y, Gray MW, Lang BF. 2003b. Unique mitochondrial genome architecture in unicellular relatives of animals. Proc. Natl. Acad. Sci. U.S.A. 100:892–897.

Burger G, Gray MW, Forget L, Lang BF. 2013. Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. Genome Biol Evol. 5:418–438.

Burki F, Inagaki Y, Bråte J, et al. 2009. Large-scale phylogenomic analyses reveal that two enigmatic protist lineages, Telonemia and Centroheliozoa, are related to photosynthetic chromalveolates. Genome Biol Evol. 1:231–238.

Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, Jakobsen KS, Pawlowski J. 2007. Phylogenomics reshuffles the eukaryotic supergroups. PLoS ONE. 2:e790.

Burki F, Okamoto N, Pombert J-FF, Keeling PJ. 2012. The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. Proc. Biol. Sci. 279:2246–2254.

Cannone JJ, Subramanian S, Schnare MN, et al. 2002. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics. 3:2.

Cavalier-Smith T. 2003. Protist phylogeny and the high-level classification of Protozoa. European J. Protistol. 39:338–348.

Cavalier-Smith T, Chao EE. 2006. Phylogeny and megasystematics of phagotrophic heterokonts (kingdom Chromista). J. Mol. Biol. 62:388–420.

Cavalier-Smith T, von der Heyden S. 2007. Molecular phylogeny, scale evolution and taxonomy of centrohelid heliozoa. Mol. Phylogenet. Evol. 44:1186–1203.

Cavalier-Smith T, Chao EE, Lewis R. 2015. Multiple origins of Heliozoa from flagellate ancestors: New cryptist subphylum Corbihelia, superclass Corbistoma, and monophyly of Haptista, Cryptista, Hacrobia and Chromista. Mol. Phylogenet. Evol. 93:331–362.

Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 28:464–469.

Cousineau B, Smith D, Lawrence-Cavanagh S, et al. 1998. Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. Cell. 94:451–462.

Cho Y, Qiu YL, Kuhlman P, Palmer JD. 1998. Explosive invasion of plant mitochondria by a group I intron. Proc. Natl. Acad. Sci. U.S.A. 95:14244–14249.

Dellaporta SL, Xu A, Sagasser S, Jakob W, Moreno MA, Buss LW, Schierwater B. 2006. Mitochondrial genome of *Trichoplax adhaeren*s supports placozoa as the basal lower metazoan phylum. Proc. Natl. Acad. Sci. U.S.A. 103:8751–8756.

Dinouël N, Drissi R, Miyakawa I, Sor F, Rousset S, Fukuhara H. 1993. Linear mitochondrial DNAs of yeasts: closed-loop structure of the termini and possible linear-circular conversion mechanisms. Mol. Cell Biol. 13:2315–2323.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Edgell DR, Chalamcharla VR, Belfort M. 2011. Learning to live together: mutualism between self-splicing introns and their hosts. BMC Biol. 29:22.

Ehara M, Watanabe KI, Ohama T. 2000. Distribution of cognates of group II introns detected in mitochondrial *cox1* genes of a diatom and a haptophyte. Gene. 256:157–167.

Emblem Å, Karlsen BO, Evertsen J, Johansen SD. 2011. Mitogenome rearrangement in the cold-water scleractinian coral *Lophelia pertusa* (Cnidaria, Anthozoa) involves a long-term evolving group I intron. Mol. Phylogenet. Evol. 61:495–503.

Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. Nature. 440:623–630.

Eschbach S, Hofmann CJ, Maier UG, Sitte P, Hansmann P. 1991. A eukaryotic genome of 660 kb: electrophoretic karyotype of nucleomorph and cell nucleus of the cryptomonad alga, *Pyrenomonas salina*. Nucleic Acids Res. 19:1779–1781.

Fan J, Lee RW. 2002. Mitochondrial genome of the colorless green alga *Polytomella parva*: two linear DNA molecules with homologous inverted repeat termini. Mol. Biol. Evol. 19:999–1007

Febvre-Chevalier C . 1990. Phylum Actinopoda Class Heliozoa. In Margulis L, Corliss JO, Melkonian M, Chapman DJ (eds) Handbook of Protoctista. Jones and Bartlett, Boston, 347–362

Foury F, Roganti T, Lecrenier N, Purnelle B. 1998. The complete sequence of the mitochondrial genome of *Saccharomyces cerevisia*e. FEBS Lett. 440:325–331.

Fritz-Laylin L, Ginger M, Walsh C, Dawson S, Fulton C. 2011. The *Naegleria* genome: a free-living microbial eukaryote lends unique insights into core eukaryotic cell biology. Res. Microbiol. 162:607–618.

Giegé P, Grienenberger JM, Bonnard G. 2008. Cytochrome *c* biogenesis in mitochondria. Mitochondrion. 8:61–73.

Goddard MR, Burt A. 1999. Recurrent invasion and extinction of a selfish gene. Proc. Natl. Acad. Sci. U.S.A. 96:13880–13885.

Goddard MR, Leigh J, Roger AJ, Pemberton AJ. 2006. Invasion and persistence of a selfish gene in the Cnidaria. PLoS ONE. 1:e3.

Gould SB, Waller RF, McFadden GI. 2008. Plastid evolution. Annu. Rev. Plant Biol. 59:491–517.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27:221–224.

de Graaf RM, van Alen TA, Dutilh BE, Kuiper JW, van Zoggel HJ, Huynh MB, Görtz H-DD, Huynen MA, Hackstein JH. 2009. The mitochondrial genomes of the ciliates *Euplotes minuta* and *Euplotes crassus*. BMC Genomics. 10:514.

Gray MW. 1993. Origin and evolution of organelle genomes. Curr. Opin. Genet. Dev. 3:884–890.

Gray MW. 1999. Evolution of organellar genomes. Curr. Opin. Genet. Dev. 9:678–687.

Gray MW, Burger G, Lang BF. 1999. Mitochondrial evolution. Science. 283:1476–1481.

Gray MW, Burger G, Lang BF. 2001. The origin and early evolution of mitochondria. Genome Biol. 2:REVIEWS1018.

Gray MW, Lang BF, Burger G. 2004. Mitochondria of protists. Annu. Rev. Genet. 38: 477–524.

Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AG, Roger AJ. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc. Natl. Acad. Sci. U.S.A. 106:3859–3864.

Handa H. 2008. Linear plasmids in plant mitochondria: peaceful coexistences or malicious invasions? Mitochondrion. 8:15–25.

Haugen P, Reeb V, Lutzoni F, Bhattacharya D. 2004. The evolution of homing endonuclease genes and group I introns in nuclear rDNA. Mol. Biol. Evol. 21:129–140.

Haugen P, Simon DM, Bhattacharya D. 2005. The natural history of group I introns. Trends Genet. 21: 111–119

Hauth AM, Maier UG, Lang BF, Burger G. 2005. The *Rhodomonas salin*a mitochondrial genome: bacteria-like operons, compact gene arrangement and complex repeat region. Nucleic Acids Res. 33:4433–4442.

Hayashi-Ishimaru Y, Ehara M, Inagaki Y, Ohama T. 1997. A deviant mitochondrial genetic code in prymnesiophytes (yellow-algae): UGA codon for tryptophan. Curr. Genet. 32:296–299.

Hazle T, Bonen L. 2007. Comparative analysis of sequences preceding protein-coding mitochondrial genes in flowering plants. Mol. Biol. Evol. 24:1101–1112.

Herman EK, Greninger AL, Visvesvara GS, Marciano-Cabral F, Dacks JB, Chiu CY. 2013. The mitochondrial genome and a 60-kb nuclear DNA segment from *Naegleria fowleri*, the causative agent of primary amoebic meningoencephalitis. J. Eukaryot. Microbiol. 60:179–191.

Hikosaka K, Kita K, Tanabe K. 2013. Diversity of mitochondrial genome structure in the phylum Apicomplexa. Mol. Biochem. Parasitol. 188:26–33.

Hovde B, Starkenburg S, Hunsperger H, Mercer L, Deodato C, Jha R, Chertkov O, Monnat R, Cattolico R. 2014. The mitochondrial and chloroplast genomes of the haptophyte *Chrysochromulina tobin* contain unique repeat structures and gene profiles. BMC Genomics. 15:604.

Inagaki Y, Ehara M, Watanabe KI, Hayashi-Ishimaru Y, Ohama T. 1998. Directionally evolving genetic code: the UGA codon from stop to tryptophan in mitochondria. J. Mol. Evol. 47:378–384.

Jackson CJ, Reyes-Prieto A. 2014. The mitochondrial genomes of the glaucophytes *Gloeochaete wittrockiana* and *Cyanoptyche gloeocystis*: multilocus phylogenetics suggests a monophyletic archaeplastida. Genome Biol Evol. 6:2774–2785.

Janouškovec J, Horák A, Oborník M, Lukes J, Keeling PJ. 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proc. Natl. Acad. Sci. U.S.A.107:10949–10954

Janouškovec J, Tikhonenkov DV, Mikhailov KV, Simdyanov TG, Aleoshin VV, Mylnikov AP, Keeling PJ. 2013. Colponemids represent multiple ancient alveolate lineages. Curr Biol. 23:2546–2552.

Johansen S, Elde M, Vader A, Haugen P, Haugli K, Haugli F. 1997. *In vivo* mobility of a group I twintron in nuclear ribosomal DNA of the myxomycete *Didymium iridis*. Mol. Microbiol. 24:737–745.

Ji YE, Mericle BL, Rehkopf DH, Anderson JD, Feagin JE. 1996. The *Plasmodium falciparum* 6 kb element is polycistronically transcribed. Mol. Biochem. Parasitol. 81:211–223.

Kamikawa R, Hosoi‐Tanabe S, Nagai S, Itakura S, Sako Y. 2005. Development of a quantification assay for the cysts of the toxic dinoflagellate *Alexandrium tamarens*e using real‐time polymerase chain reaction. Fisheries Sci. 71:987–991

Kamikawa R, Masuda I, Demura M, Oyama K, Yoshimatsu S, Kawachi M, Sako Y. 2009. Mitochondrial group II introns in the raphidophycean flagellate *Chattonella* spp. suggest a diatom-to-Chattonella lateral group II intron transfer. Protist.160:364–375.

Kamikawa R, Kolisko M, Nishimura Y, Yabuki A, Brown MW, Ishikawa SA, Ishida K, Roger AJ, Hashimoto T, Inagaki Y . 2014. Gene content evolution in Discobid mitochondria deduced from the phylogenetic position and complete mitochondrial genome of *Tsukubamonas globosa*. Genome Biol. Evol. 6:306–315.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

Kim E, Simpson AG, Graham LE. 2006. Evolutionary relationships of apusomonads inferred from taxon-rich analyses of 6 nuclear encoded genes. Mol. Biol. Evol. 23:2455–2466.

Kim E, Lane CE, Curtis GA, Kozera C, Bowman S, Archibald JM. 2008. Complete sequence and analysis of the mitochondrial genome of *Hemiselmis andersenii* CCMP644 (Cryptophyceae). BMC Genomics. 9:215.

Kim E, Harrison JW, Sudek S, Jones MD, Wilcox HM, Richards TA, Worden AZ, Archibald JM. 2011. Newly identified and diverse plastid-bearing branch on the eukaryotic tree of life. Proc. Natl. Acad. Sci. U.S.A. 108:1496–1500.

Klaveness D, Shalchian-Tabrizi K, Thomsen HA, Eikrem W, Jakobsen KS. 2005. *Telonema antarcticum* sp. nov., a common marine phagotrophic flagellate. Int. J. Syst. Evol. Microbiol. 55:2595–2604.

Lambowitz AM, Zimmerly S. 2004. Mobile group II introns. Annu. Rev. Genet. 38:1–35.

Lambowitz AM, Zimmerly S. 2011. Group II introns: mobile ribozymes that invade DNA. Cold Spring Harb. Perspect. Biol. 3:a003616.

Lane CE, Archibald JM. 2008. The eukaryotic tree of life: endosymbiosis takes its TOL. Trends Ecol. Evol. 23:268–275.

Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Gray MW. 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature. 387:493–497.

Lang BF, Laforest MJ, Burger G. 2007. Mitochondrial introns: a critical view. Trends Genet. 23:119–125

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9:357–359.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 25:2286–2288.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.

Lehmann K, Schmidt U. 2003. Group II introns: structure and catalytic versatility of large natural ribozymes. Crit. Rev. Biochem. Mol. Biol. 38:249–303.

Liu B, Liu Y, Motyka SA, Agbo EE, Englund PT. 2005. Fellowship of the rings: the replication of kinetoplast DNA. Trends Parasitol.

Lonergan KM, Gray MW. 1994. The ribosomal RNA gene region in *Acanthamoeba castellanii* mitochondrial DNA. A case of evolutionary transfer of introns between mitochondria and plastids? J. Mol. Biol. 239:476–499.

Maddison DR, Swofford DL, Maddison WP. 1997. NEXUS: an extensible file format for systematic information. Syst. Biol. 46:590–621.

Makiuchi T, Nozaki T. 2014. Highly divergent mitochondrion-related organelles in anaerobic parasitic protozoa. Biochimie. 100:3–17.

Mao Y, Zhang B, Kong F, Wang L. 2012. The complete mitochondrial genome of *Pyropia haitanensis* Chang et Zheng. Mitochondrial DNA. 23:344–346.

Martin-Cereceda M, Roberts EC, Wootton EC, Bonaccorso E, Dyal P, Guinea A, Rogers D, Wright CJ, Novarino G. 2010. Morphology, ultrastructure, and small subunit rDNA phylogeny of the marine heterotrophic flagellate *Goniomonas* aff. *amphinema*. J. Eukaryot. Microbiol. 57:159–170.

Masuda I, Kamikawa R, Ueda M, Oyama K, Yoshimasu S, Inagaki Y, Sako Y. 2011. Mitochondrial genomes from two red tide forming raphidophycean algae *Heterosigma akashiwo* and *Chattonella marina* var. *marina*. Harmful Algae. 10:130–137

Meinhardt F, Kempken F, Kämper J, Esser K. 1990. Linear plasmids among eukaryotes: fundamentals and application. Curr. Genet. 17:89–95.

Meyer EH, Giegé P, Gelhaye E, Rayapuram N, Ahuja U, Thöny-Meyer L, Grienenberger JM, Bonnard G. 2005. AtCCMH, an essential component of the *c*-type cytochrome maturation pathway in *Arabidopsis* mitochondria, interacts with apocytochrome *c*. Proc. Natl. Acad. Sci. U.S.A. 102:16113–16118.

Michel F, Jacquier A, Dujon B. 1982. Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. Biochimie. 64:867–881

Moreira D, Le Guyader H, Philippe H. 2000. The origin of red algae and the evolution of chloroplasts. Nature. 405:69–72.

Moreira D, López-García P. 2014. The rise and fall of Picobiliphytes: how assumed autotrophs turned out to be heterotrophs. Bioessays. 36:468–474.

Morell V. 1996. TreeBASE: the roots of phylogeny. Science. 273:5275.

Müller M, Mentel M, van Hellemond JJ, et al. 2012. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. Microbiol. Mol. Biol. Rev. 76:444–495.

Müller, Vingron. 2000. Modeling amino acid replacement. J Comput Biol. 7:761–776.

Nielsen H, Johansen S. 2009. Group I introns: Moving in new directions. RNA Biol. 6:375–383.

Nurk S, Bankevich A, Antipov D, et al. 2013. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. J. Comput. Biol. 20:714–737.

O'Brien EA, Zhang Y, WangE , Marie V, Badejoko W, Lang BF, Burger G. 2009. GOBASE: an organelle genome database. Nucleic Acids Res. 37:D946–50

Okamoto N, Inouye I. 2005. The katablepharids are a distant sister group of the Cryptophyta: A proposal for Katablepharidophyta divisio nova/ Kathablepharida phylum novum based on SSU rDNA and beta-tubulin phylogeny. Protist. 156:163–179.

Okamoto N, Chantangsi C, Horák A, Leander BS, Keeling PJ. 2009. Molecular phylogeny and description of the novel katablepharid *Roombia truncata* gen. et sp. nov., and establishment of the Hacrobia taxon nov. PLoS ONE. 4:e7080.

Oudot-Le Secq MP, Fontaine JM, Rousvoal S, Kloareg B, Loiseaux-De Goër S. 2001. The complete sequence of a brown algal mitochondrial genome, the ectocarpale *Pylaiella littoralis* (L.) Kjellm. J. Mol. Evol. 53:80–88.

Paquin B, Lang BF. 1996. The mitochondrial DNA of *Allomyces macrogynus*: the complete genomic sequence from an ancestral fungus. J. Mol. Biol. 255:688–701.

Patron NJ, Inagaki Y, Keeling PJ. 2007. Multiple gene phylogenies support the monophyly of cryptomonad and haptophyte host lineages. Curr. Biol. 17:887–891.

Pérez-Brocal V, Shahar-Golan R, Clark CG. 2010. A linear molecule with two large inverted repeats: the mitochondrial genome of the stramenopile *Proteromonas lacertae*. Genome Biol Evol. 2:257–266.

Philippe H, Snell EA, Bapteste E, Lopez P, Holland PW, Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21:1740–1752.

Ravin NV, Galachyants YP, Mardanov AV, Beletsky AV, Petrova DP, Sherbakova TA, Zakharova YR, Likhoshway YV, Skryabin KG, Grachev MA. 2010. Complete

sequence of the mitochondrial genome of a diatom alga *Synedra acus* and comparative analysis of diatom mitochondrial genomes. Curr. Genet. 56:215–223

Rayapuram N, Hagenmuller J, Grienenberger J-MM, Giegé P, Bonnard G. 2007. AtCCMA interacts with AtCcmB to form a novel mitochondrial ABC transporter involved in cytochrome *c* maturation in *Arabidopsis*. J. Biol. Chem. 282:21015–21023.

Rota-Stabelli O, Yang Z, Telford MJ. 2009. MtZoa: a general mitochondrial amino acid substitutions model for animal evolutionary studies. Mol. Phylogenet. Evol. 52:268–272.

Sakaguchi M, Inagaki I, Hashimoto T. 2007. Centrohelida is still searching for a phylogenetic home: analyses of seven *Raphidiophrys contractilis* genes. Gene. 40:47–54

Saldanha R, Mohr G, Belfort M, Lambowitz AM. 1993. Group I and group II introns. FASEB J. 7:15–24.

San Filippo J, Lambowitz AM. 2002. Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. J. Mol. Biol. 324:933–951.

Sánchez-Puerta MV, Bachvaroff TR, Delwiche CF. 2004. The complete mitochondrial genome sequence of the haptophyte *Emiliania huxleyi* and its relation to heterokonts. DNA Res. 11:1–10.

Sánchez-Puerta MV, Abbona CC, Zhuo S, Tepe EJ, Bohs L, Olmstead RG, Palmer JD. 2011. Multiple recent horizontal transfers of the *cox1* intron in Solanaceae and extended co-conversion of flanking exons. BMC Evol. Biol. 11:277.

Sandegren L, Sjöberg B-MM. 2004. Distribution, sequence homology, and homing of group I introns among T-even-like bacteriophages: evidence for recent transfer of old introns. J. Biol. Chem. 279:22218–22227.

Sanders C, Turkarslan S, Lee DW, Daldal F. 2010. Cytochrome *c* biogenesis: the Ccm system. Trends Microbiol. 18:266–274.

Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 33:W686–9.

Seenivasan R, Sausen N, Medlin LK, Melkonian M. 2013. *Picomonas judraskeda* gen. et sp. nov.: the first identified member of the Picozoa phylum nov., a widespread group of picoeukaryotes, formerly known as “picobiliphytes”. PLoS ONE. 8:e59565.

Sellem CH, d’ Aubenton-Carafa Y, Rossignol M, Belcour L. 1996. Mitochondrial intronic open reading frames in *Podospora*: mobility and consecutive exonic sequence variations. Genetics. 143:777–788.

Shalchian-Tabrizi K, Eikrem W, Klaveness D, et al. 2006. Telonemia, a new protist phylum with affinity to chromist lineages. Proc. Biol. Sci. 273:1833–1842.

Smith DR, Arrigo KR, Alderkamp A-CC, Allen AE. 2014. Massive difference in synonymous substitution rates among mitochondrial, plastid, and nuclear genes of Phaeocystis algae. Mol. Phylogenet. Evol. 71:36–40.

Smith DR, Keeling PJ. 2012. Twenty-fold difference in evolutionary rates between the mitochondrial and plastid genomes of species with secondary red plastids. J. Eukaryot. Microbiol. 59:181-184.

Smith MR, Patterson DJ. 1986. Analyses of heliozoan interrelationships: an example of the potentials and limitations of ultrastructural approaches to the study of protistan phylogeny. Proc. Roy. Soc. B. 227:325–366.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.

Stoddard BL. 2005. Homing endonuclease structure and function. Q. Rev. Biophys. 38:49–95.

Stoltzfus A, O'Meara B, Whitacre J, Mounce R, Gillespie EL, Kumar S, Rosauer DF, Vos RA. 2012. Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis. BMC Res. Notes. 5:574.

Swart EC, Nowacki M, Shum J, et al. 2012. The *Oxytricha trifallax* mitochondrial genome. Genome Biol Evol. 4:136–154.

Szitenberg A, Rot C, Ilan M, Huchon D. 2010. Diversity of sponge mitochondrial introns revealed by *cox1* sequences of Tetillidae. BMC Evol. Biol. 10:288.

Takano H, Kawano S, Kuroiwa T. 1994. Complex terminal structure of a linear mitochondrial plasmid from *Physarum polycephalum*: three terminal inverted repeats and an ORF encoding DNA polymerase. Curr. Genet. 25:252–257.

Tanabe AS. 2011. Kakusan4 and Aminosan: two programs for comparing nonpartitioned, proportional and separate models for combined molecular phylogenetic analyses of multilocus sequence data. Mol Ecol Resour. 11:914–921.

Tanifuji G, Erata M, Ishida K, Onodera N, Hara Y. 2006. Diversity of secondary endosymbiont-derived actin-coding genes in cryptomonads and their evolutionary implications. J. Plant Res. 119:205–215.

Tikhonenkov DV, Janouškovec J, Mylnikov AP, Mikhailov KV, Simdyanov TG, Aleoshin VV, Keeling PJ. 2014. Description of *Colponema vietnamica* sp.n. and *Acavomonas peruviana* n. gen. n. sp., two new alveolate phyla (Colponemidia nom. nov. and Acavomonidia nom. nov.) and their contributions to reconstructing the ancestral state of alveolates and eukaryotes. PLoS ONE. 9:e95467.

Toor N, Hausner G, Zimmerly S. 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. RNA. 7:1142–1152.

Toro N. 2003. Bacteria and Archaea group II introns: additional mobile genetic elements in the environment. Environ. Microbiol. 5:143–151.

Tovar J, León-Avila G, Sánchez LB, Sutak R, Tachezy J, van der Giezen M, Hernández M, Müller M, Lucocq JM. 2003. Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. Nature. 426:172–176.

Turmel M, Otis C, Lemieux C. 2002. The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proc. Natl. Acad. Sci. U.S.A. 99:11275–11280.

Turmel M, Otis C, Lemieux C. 2007. An unexpectedly large and loosely packed mitochondrial genome in the charophycean green alga *Chlorokybus atmophyticus*. BMC Genomics 8:137.

Turmel M, Otis C, Lemieux C. 2013. Tracing the evolution of streptophyte algae and their mitochondrial genome. Genome Biol. Evol. 5:1817-1835

Unseld M, Marienfeld JR, Brandt P, Brennicke A. 1997. The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. Nat. Genet. 15:57–61.

Vahrenholz C, Riemen G, Pratje E, Dujon B, Michaelis G. 1993. Mitochondrial DNA of *Chlamydomonas reinhardtii*: the structure of the ends of the linear 15.8-kb genome suggests mechanisms for DNA replication. Curr. Genet. 24:241–247.

Vallès Y, Halanych KM, Boore JL. 2008. Group II introns break new boundaries: presence in a bilaterian's genome. PLoS ONE. 3:e1488.

Vaughn JC, Mason MT, Sper-Whitis GL, Kuhlman P, Palmer JD. 1995. Fungal origin by horizontal transfer of a plant mitochondrial group I intron in the chimeric *coxI* gene of *Peperomia*. J. Mol. Evol. 41:563–572.

Yabuki A, Inagaki Y, Ishida K. 2010. *Palpitomonas bilix* gen. et sp. nov.: A novel deep-branching heterotroph possibly related to Archaeplastida or Hacrobia. Protist. 161:523–538.

Yabuki A, Chao EE, Ishida KI, Cavalier-Smith T. 2012. *Microheliella maris* (Microhelida ord. n.), an ultrastructurally highly distinctive new axopodial protist species and genus, and the unity of phylum Heliozoa. Protist. 163:356–388.

Yabuki A, Kamikawa R, Ishikawa SA, Kolisko M, Kim E, Tanabe AS, Kume K, Ishida K, Inagki Y. 2014. *Palpitomonas bilix* represents a basal cryptist lineage: insight into the character evolution in Cryptista. Sci. Rep. 4:4641.

Yang EC, Kim KM, Kim SY. Lee J, Boo GH, Lee JH, Nelson WA, Yi G, Schmidt WE, Fredericq S, Boo SM, Bhattacharya D, Yoon HS. 2015. Highly conserved mitochondrial genomes among multicellular red algae of the Florideophyceae. Genome Biol. Evol. 7:2394–2406.

Zhao S, Burki F, Bråte J, Keeling PJ, Klaveness D, Shalchian-Tabrizi K. 2012. *Collodictyon*--an ancient lineage in the tree of eukaryotes. Mol. Biol. Evol. 29:1557–1568.

Zimmerly S, Semper C. 2015. Evolution of group II introns. Mob DNA. 6:7.

Zoschke R, Nakamura M, Liere K, Sugiura M, Börner T, Schmitz-Linneweber C. 2010. An organellar maturase associates with multiple group II introns. Proc. Natl. Acad. Sci. U.S.A. 107:3245–3250.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31:3406–3415.

Tables

Table1 Degenerate primers used for reverse-transcription PCR

Genes	Directions	Sequences(5′-3′)
Cob	Forward	GNGAYGTNAAYAAYGGNTGG
	Reverse	ACDATRTGNGCNGGNGTNACC
Cox1	Forward	ACNAAYCAYAARGAYATHGG
	Reverse	NACNCCNACRAANGTRCACC
Cox3	Forward	CCNTTYCAYTTRGTNGAYCC
	Reverse	NACNACRTCNACRAARTGCC
Nad1	Forward	CGNGGNCCNAAYGTNGTNGG
	Reverse	NARYTCNGCYTCNGCYTCNGG
Nad7	Forward	AAYTTYGGNCCNCARCAYCC
	Reverse	NACNCCRAAYTCNCCYTTNG
Nad11	Forward	GTNGCNGGNAAYTGYKGNATG
	Reverse	NGTNARNGCNCCNACNGGRCA

Table 2. Codon frequency and tRNA anticodon repertoire in the mitochondrial genome of *Chrysochromulina* sp. NIES-1333.

Codons	AA	Frequency	tRNA anticodon	Codons	AA	Frequency	tRNA anticodon	Codons	AA	Frequency	tRNA anticodon	Codons	AA	Frequency	tRNA anticodon
UUU	F	15.32	GAA	UCU	S	17.25	UGA	UAU	Y	3.16	GUA	UGU	C	12.15	GCA
UUC		2.29		UCC		2.29		UAC		14.44		UGC		2.99	
UUA	L	9.16	UAA	UCA		31.34		UAA	[1]	10.39	–[1]	UGA	W	12.5	UCA
UUG		2.82		UCG		6.34		UAG		7.21		UGG		2.82	
CUU		2.99	UAG	CCU	P	8.28	UGG	CAU	H	8.63	GUG	CGU	R	10.56	ACG
CUC		0.53		CCC		2.11		CAC		8.98		CGC		2.64	
CUA		1.76		CCA		18.13		CAA	Q	14.44	UUG	CGA		8.45	
CUG		0.17		CCG		2.99		CAG		3.17		CGG		0.35	
AUU	I	10.21	GAU	ACU	T	22.89	UGU	AAU	N	11.09	GUU	AGU	S	14.6	GCU
AUC		3.35		ACC		3.17		AAC		6.51		AGC		8.10	
AUA		4.05	CAU[2]	ACA		20.95		AAA	K	14.09	UUU	AGA	R	4.93	UCU
AUG	M	17.6	CAU	ACG		8.45		AAG		3.52		AGG		1.23	
GUU	V	11.09	UAC	GCU	A	34.16	UGC	GAU	D	12.68	GUC	GGU	G	41.12	–
GUC		1.41		GCC		3.17		GAC		4.93		GGC		5.46	
GUA		3.17		GCA		17.61		GAA	E	13.73	UUC	GGA		8.45	
GUG		1.94		GCG		7.22		GAG		3.87		GGG		2.64	

[1]No tRNA for stop codons.

[2]The first position of anticodon CAU may be modified to recognize AUA codon.

[3]No tRNA was detected.

Table 3 Gene repertories in haptophye mitochondrial genomes

	Chrysochromulina sp. NIED-1333	*Chrysochromulina tobin*	*Diacronema lutheli*	*Emiliania huxlyei*
rnl	Y[1]	Y	Y[1]	Y
rns	Y	Y	Y[1]	Y
rrn5	N	N	Y	Y
tRNA	23 species	23 species	22 species	23 species
nad1	Y	Y	Y	Y
nad2	Y	Y	Y	Y
nad3	Y	Y	Y	Y
nad4	Y	Y	Y	Y
nad4L	Y	Y	Y	Y
nad5	Y	Y	Y	Y
nad6	Y	Y	Y	Y
cob	Y	Y	Y	Y
cox1	Y[2]	Y	Y[1]	Y
cox2	Y	Y	Y	Y
cox3	Y	Y	Y[1]	Y
atp4	N	Y	Y	Y
atp6	Y	Y	Y[1]	Y
atp8	N	Y	Y	N
atp9	Y	Y	Y[1]	Y
rps3	N	Y	N	Y
rps8	N	Y	N	Y
rps12	Y	Y	Y	Y
rps14	N	Y	Y	Y
rps19	N	N	Y	N
rpl14	N	N	Y	N
rpl16	Y	Y	Y	Y
dam	N	N	N	Y
Others	orf627[a], orf538[b]	orf457	orf636[c], orf105[d]	orf104[d]

Y, yes; N, no. Numbers of introns are shown in brackets.

a: Encoded in the second *cox1* intron

b:Free-standing open reading frame encoding a protein with amino acid sequence simila

II intron-encoded proteins.

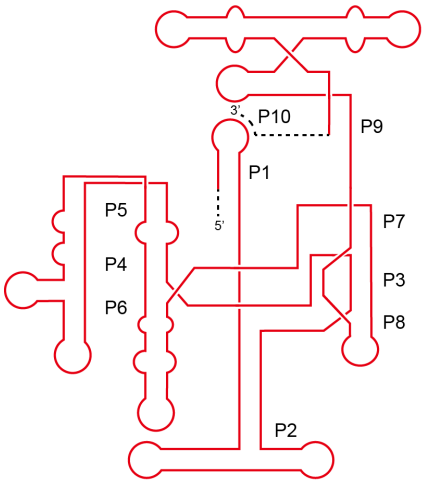
c: Encoded in the *cox1* intron.

d: Encodes an uncharacterized protein

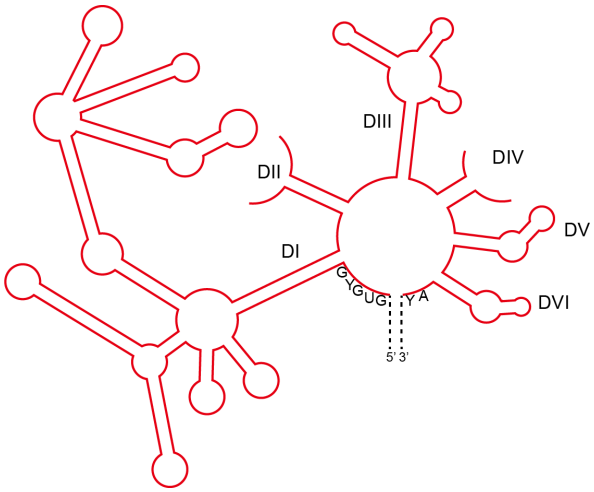
Figures

Fig 1. Model structures of group I and group II RNA

A. Schematic representative of group I intron secondary structures. The conserved stem-loop structures are designated as P1 to P10. B. Schematic representative of group II intron secondary structures. The conserved domain structures are designated as DI to DVI. The consensus intron boundary sequences are shown. In both structures, red and Black lines indicate intron and exon, respectively. These figures are made by referring to Edgell et al. 2011.



A Group I intron



B Group II intron

Fig. 2 Invasion models of group I and group II introns.

A, B Black and gray lines indicate double strand DNAs (dsDNA) with and without intron, respectively. Introns are shown in red. Striped regions are homing positions of introns, which are recognized and cleaved by endonucleases (domain). A. Group I intron homing. Step 1: The homing endonuclease (HE), which is shown as red packman-shape symbol, is expressed from an intronic open reading frame (ORF). Step 2: The HE introduces double strand-break to the recipient (gray) dsDNA. Step 3: The cleaved dsDNA are repaired through homologous recombination with intron-hosting DNA. B. Group II intron homing. Step 1: A spliced intron RNA forms a ribonucleoprotein (RNP; RNA and IEP are indicated by the orange line and the blue circle, respectively) complex with the protein expressed from the ORF hosted in the corresponding intron (intron encoded protein or IEP; step1). Step 2: The RNP complex inserts the intron RNA to the top strand of an intron-lacking dsDNA by reverse-splicing manner. The bottom strand is cleaved by the endonuclease (En) activity in the RNP complex. Step 3: The DNA strand which is complementary to the intron RNA inserted in the top strand, was synthesized by reverse transcription (RT) activity in the RNP complex. Step 4: Intron homing is completed by the host DNA repair system (step 4).

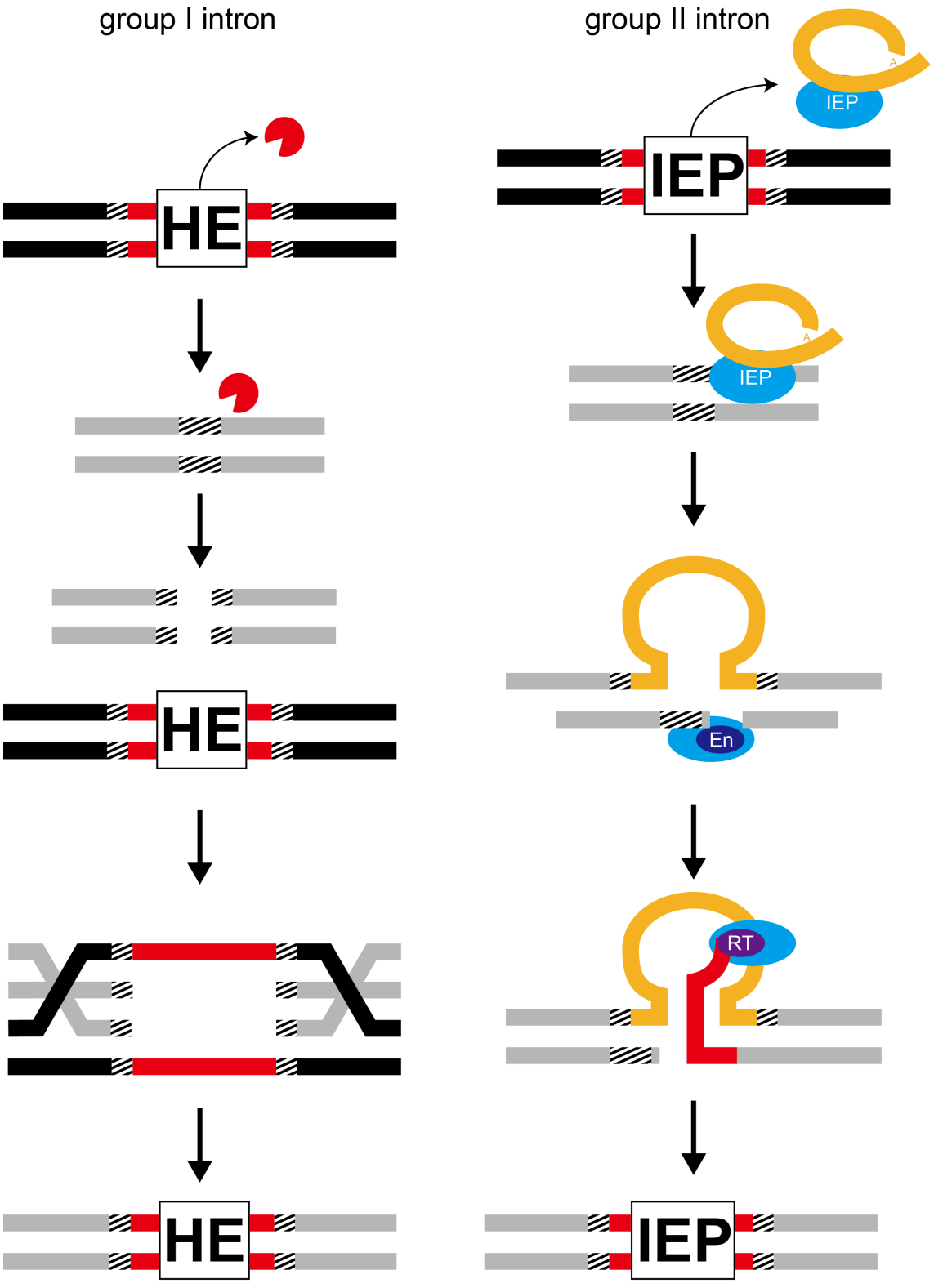
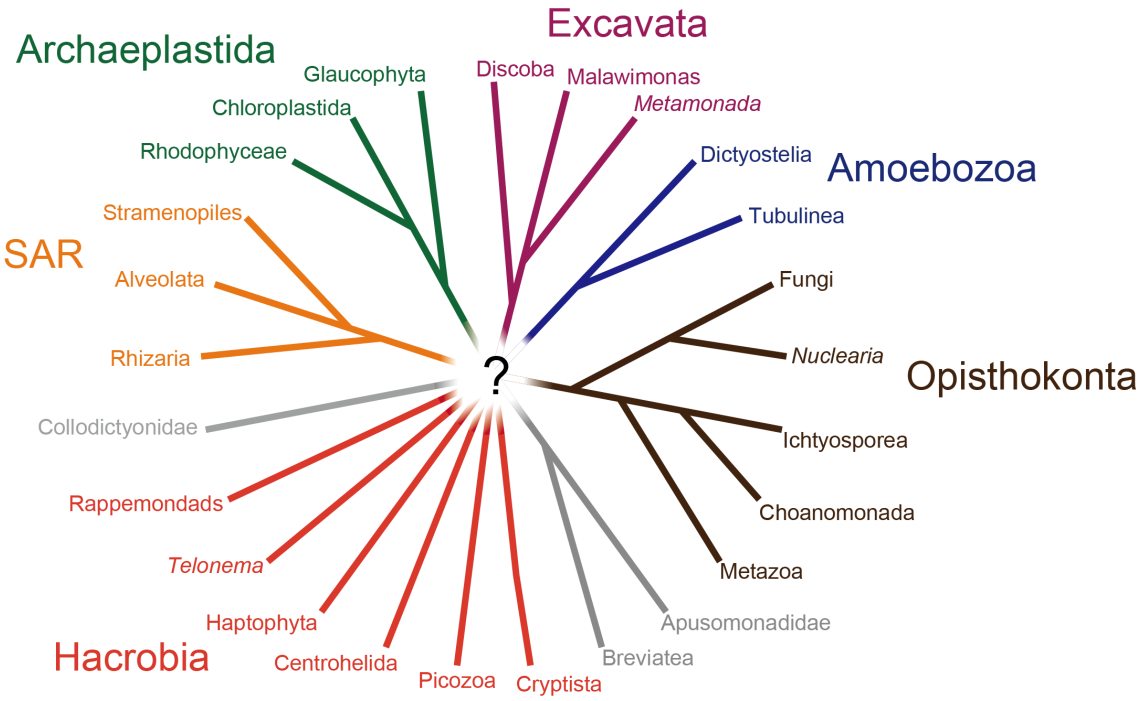

group I intron
group II intron
HE
HE
HE
IEP
IEP
IEP
En
RT
IEP

Fig 3. Tree of eukaryotes.

A. A working hypothesis on the phylogenetic relationship amongst the major eukaryotic lineages mainly based on Adl et al. (2012). Six major assemblages are color-coded, and the branches of 'orphans' lineages, which showed no strong affinity to any of the six assemblages are colored, are colored in gray. As the root of eukaryotes is still controversial, the tree is unrooted. B. The relationship amongst the members of Hacrobia. As it is uncertain whether these lineages are monophyletic, the backbone part of the tree is indicated by dot lines. The relationship among four cryptist lineages were drawn based on Yabuki et al. (2014).

A Grobal tree of eukaryotes



B Tree of Hacrobia

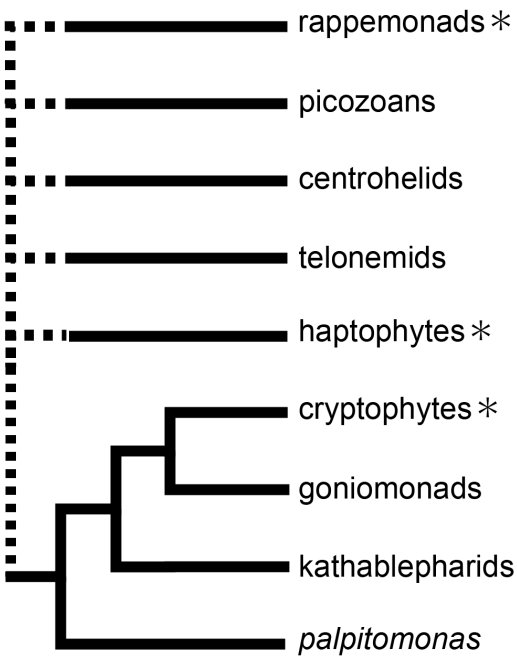
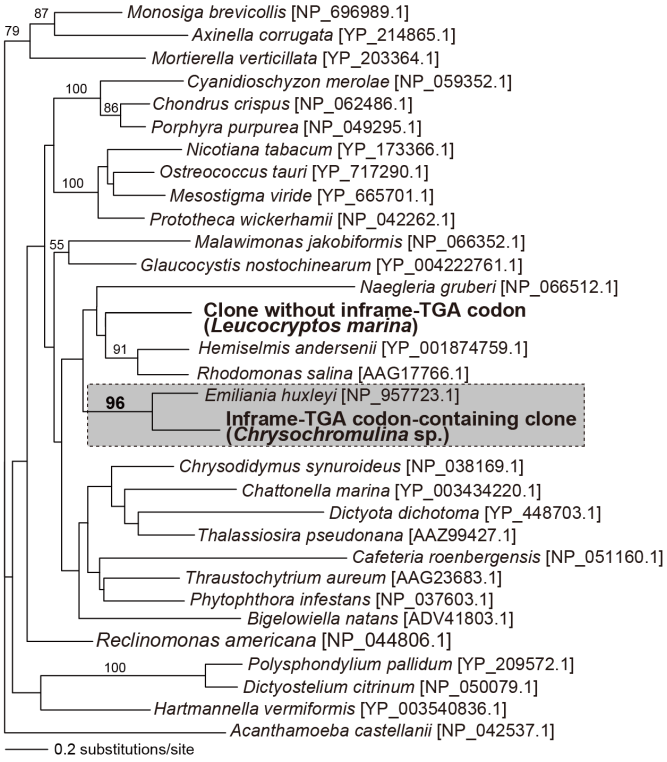


Fig. 4 Maximum-likelihood (ML) analyses of the COB and COX3 amino acid (aa) alignments. A. The ML phylogeny inferred from the COB alignment comprises 31 taxa with 368 unambiguously aligned aa positions B. The ML phylogeny inferred from the COX3 alignment comprising 26 taxa with 218 unambiguously aligned aa positions. Leucocryptos marina and Chrysochromulina sp. are highlighted by bold characters. The haptophyte clade is shades. Only ML bootstrap values equal to or greater than 50% are shown. The two alignments were separately analyzed with LG+Γ+F model by using RAxML ver. 7.2.1. The GenBank accession numbers were given in brackets on the right of species names.

A COB


87
79
Monosiga brevicollis [NP_696989.1]
Axinella corrugata [YP_214865.1]
Mortierella verticillata [YP_203364.1]
100
Cyanidioschyzon merolae [NP_059352.1]
86
Chondrus crispus [NP_062486.1]
Porphyra purpurea [NP_049295.1]
Nicotiana tabacum [YP_173366.1]
100
Ostreococcus tauri [YP_717290.1]
Mesostigma viride [YP_665701.1]
Prototheca wickerhamii [NP_042262.1]
55
Malawimonas jakobiformis [NP_066352.1]
Glaucocystis nostochinearum [YP_004222761.1]
Naegleria gruberi [NP_066512.1]
Clone without inframe-TGA codon
(Leucocryptos marina)
91
Hemiselmis andersenii [YP_001874759.1]
Rhodomonas salina [AAG17766.1]
96
Emiliania huxleyi [NP_957723.1]
Inframe-TGA codon-containing clone
(Chrysochromulina sp.)
Chrysodidymus synuroideus [NP_038169.1]
Chattonella marina [YP_003434220.1]
Dictyota dichotoma [YP_448703.1]
Thalassiosira pseudonana [AAZ99427.1]
Cafeteria roenbergensis [NP_051160.1]
Thraustochytrium aureum [AAG23683.1]
Phytophthora infestans [NP_037603.1]
Bigelowiella natans [ADV41803.1]
Reclinomonas americana [NP_044806.1]
100
Polysphondylium pallidum [YP_209572.1]
Dictyostelium citrinum [NP_050079.1]
Hartmannella vermiformis [YP_003540836.1]
Acanthamoeba castellanii [NP_042537.1]
0.2 substitutions/site

B COX1

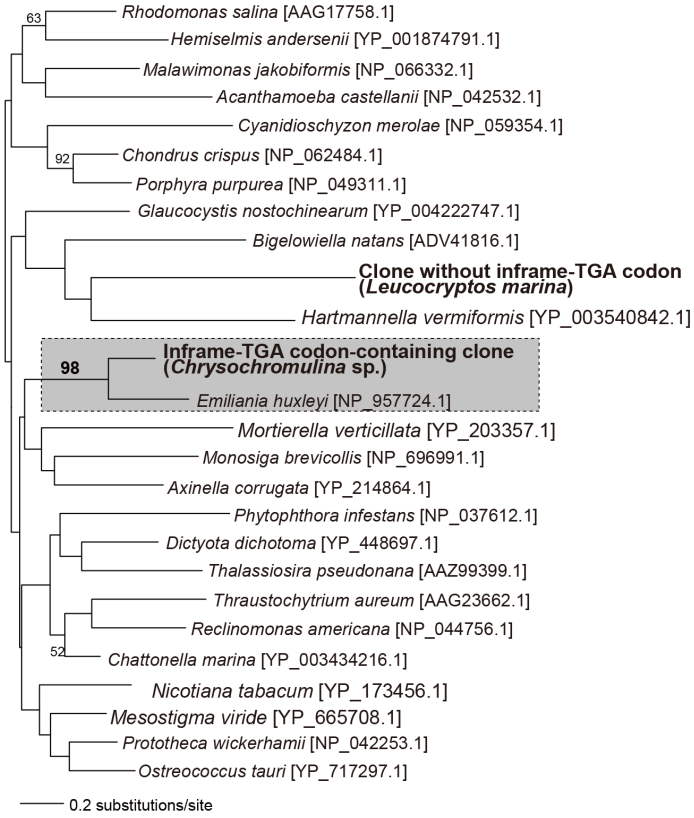

63
Rhodomonas salina [AAG17758.1]
Hemiselmis andersenii [YP_001874791.1]
Malawimonas jakobiformis [NP_066332.1]
Acanthamoeba castellanii [NP_042532.1]
Cyanidioschyzon merolae [NP_059354.1]
92
Chondrus crispus [NP_062484.1]
Porphyra purpurea [NP_049311.1]
Glaucocystis nostochinearum [YP_004222747.1]
Bigelowiella natans [ADV41816.1]
Clone without inframe-TGA codon
(Leucocryptos marina)
Hartmannella vermiformis [YP_003540842.1]
98
Inframe-TGA codon-containing clone
(Chrysochromulina sp.)
Emiliania huxleyi [NP_957724.1]
Mortierella verticillata [YP_203357.1]
Monosiga brevicollis [NP_696991.1]
Axinella corrugata [YP_214864.1]
Phytophthora infestans [NP_037612.1]
Dictyota dichotoma [YP_448697.1]
Thalassiosira pseudonana [AAZ99399.1]
Thraustochytrium aureum [AAG23662.1]
Reclinomonas americana [NP_044756.1]
52
Chattonella marina [YP_003434216.1]
Nicotiana tabacum [YP_173456.1]
Mesostigma viride [YP_665708.1]
Prototheca wickerhamii [NP_042253.1]
Ostreococcus tauri [YP_717297.1]
0.2 substitutions/site

Fig. 5 Primary structure of the partial mitochondrial genome of the katablepharid Leucocryptos marina.

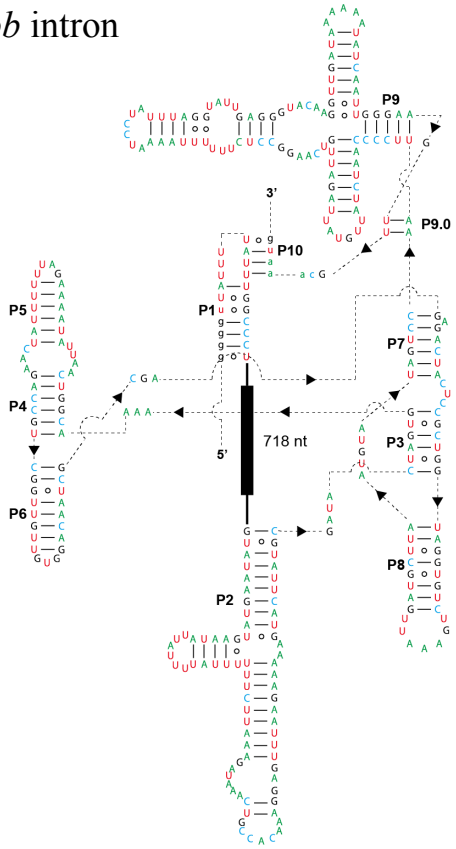
Protein-coding regions are shown by arrows. All the coding regions are located on the same strand (left to right). Unsequenced regions of *nad11* and *cox1* are shown by dot lines. The genes initially amplified by reverse transcriptase PCR are shown in orange, while those amplified from genomic DNA were in green. The IEPs and introns are indicated by black arrows and gray boxes, respectively


nad11
nad1
nad6
atp6
nad7
cox2
cox3
IEP
cob
IEP
cox1
1 kb

Fig 6. Putative secondary structures of the group I intron RNAs in a 12 Kbp mtDNA fragment of *Leucocryptos marina*

A. Putative Watson–Crick and wobble base pairs are shown by lines and open circles, respectively. Capital and small letters represent intron and exon nucleotides, respectively. Stem structures, which are characteristic to group I introns, are labeled as P1–P10. The open reading frame (ORF) for a LAGLIDADG-type homing endonuclease (closed box; 217 amino acid residues) was found in the 718 nucleotide-long loop region between P1 and P2. B. Secondary structure of the *Leucocryptos* cox1 intron. The details of this figure are same as described in A, except the ORF for a LAGLIDADG-type homing endonuclease (closed box; 267 amino acid residues) was found in the 827 nucleotide-long loop region between P1 and P10. P9.1 and P7.1, which are absent in the *Leucocryptos cob* intron, are shaded.

A *cob* intron



B *cox1* intron

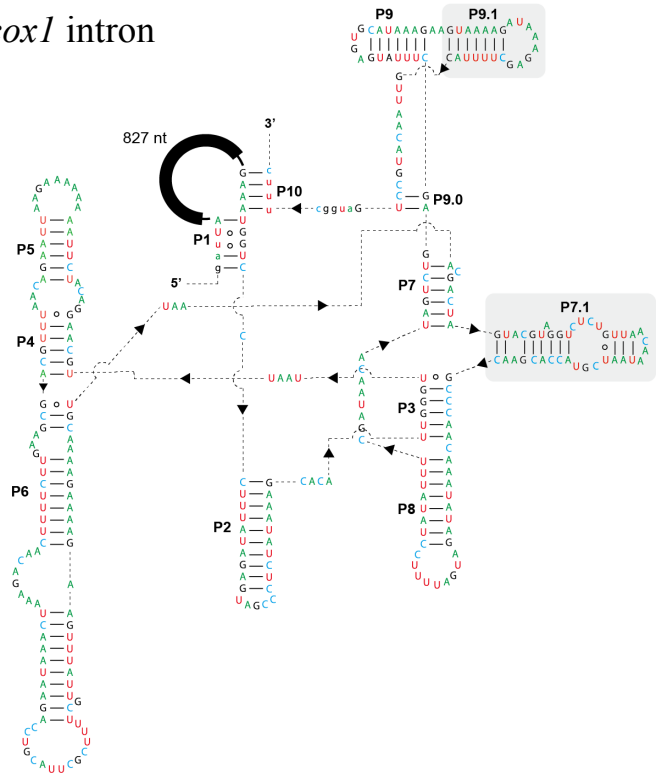
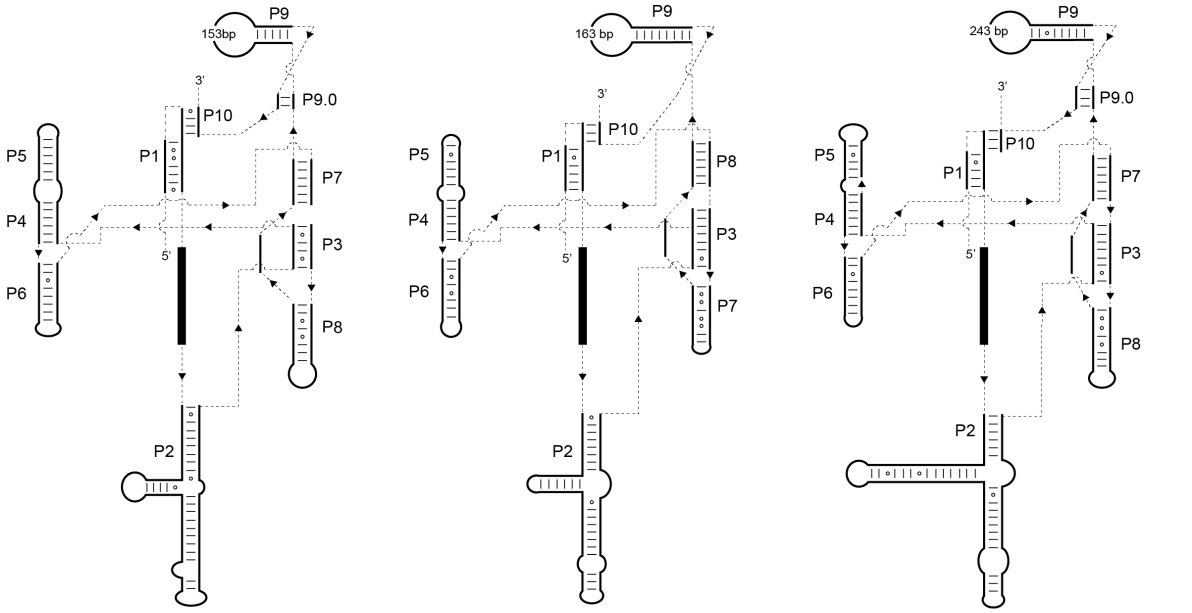


Fig. 7 Comparison of the group I intron RNAs structures.

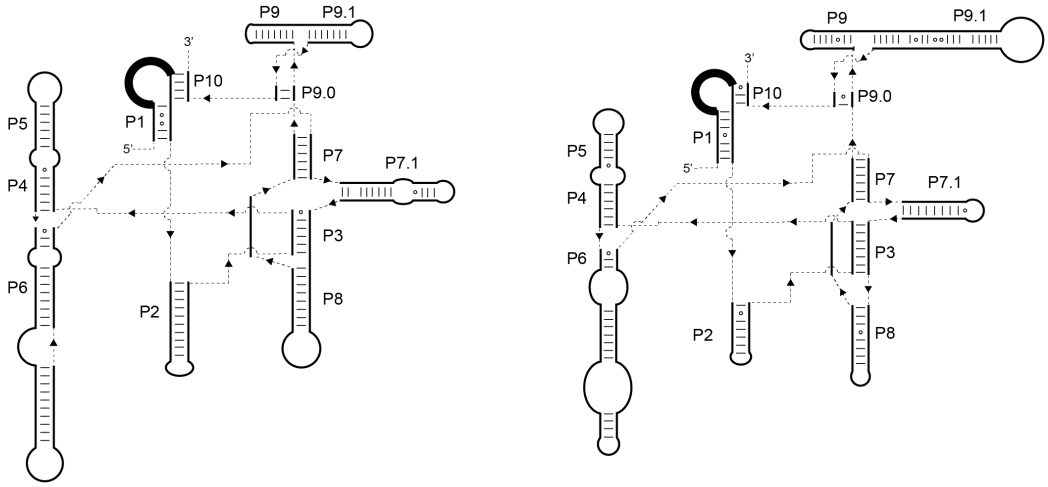
A. Secondary structure of *Leucocryptos*, *Chlorokybus*, and *Nephroselmis cob* intron. LAGLIDADG_2-type homing endonucleases (HEs) are encoded in the region between P1 and P2 in the three introns (shown as closed boxes). Putative Watson-Crick and wobble base pairs are shown by lines and open circles, respectively. Characteristic stem structures for group I introns are indicated (P1-P10). B. Schematic structures of *Leucocryptos* and *Rhizophydium* cox1 introns. Both introns harbor LAGLIDADG_1-type HEs between P1 and P10 (shown in closed boxes). Details are described in A.

A *cob* intron



Lecucocryptos marina *Nephroselmis olivacea* *Chlorokybus atmophyticus*

B *cox1* intron

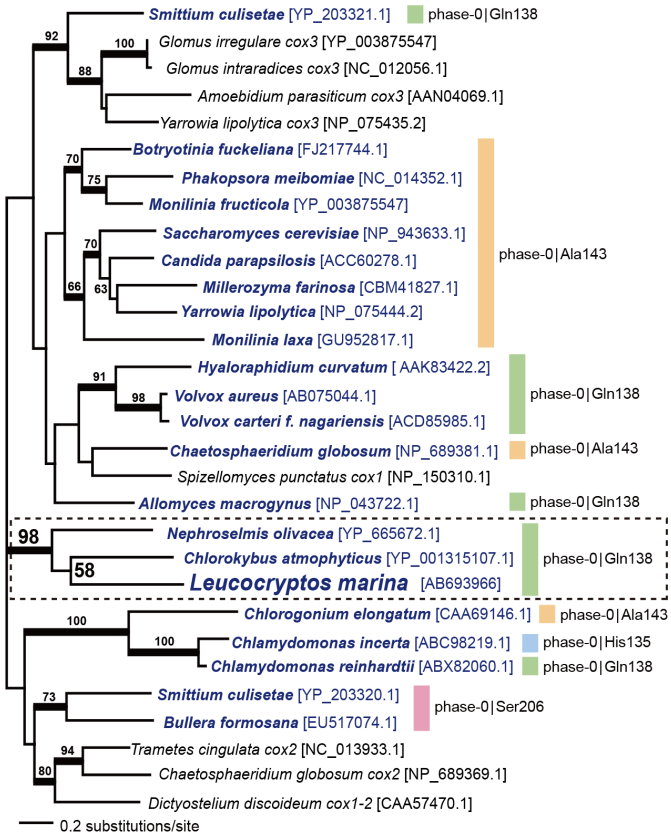


L. marina *Rhizophydium sp.*

Fig. 8 Phylogenetic analysis of group I intron-encoded LAGLIDADG endonucleases.

A. Unrooted ML phylogeny inferred from the LAGLIDADG_2 alignment containing 183 amino acid positions. 30 HEs belonging to LAGLIDADG_2 superfamily were subjected to the ML and Bayesian methods. The HEs hosted in *cob* introns are shown in dark blue. The details of the homing positions of the HE-hosting *cob* introns (phase and codon) are given on the right side of the tree. Codon numbers are based on the *Saccharomyces cerevisiae cob* gene (GenBank accession number NP_009315.1). Only ML bootstrap values equal to or greater than 50 % are shown. The resultant tree inferred from Bayesian analysis was essentially identical to that from the ML analysis (data not shown). The branches supported by Bayesian posterior probabilities (BPPs) equal to or greater than 0.95 were highlighted by thick lines. The GenBank accession number of the HE sequences used in this tree are given in brackets. B. Unrooted ML phylogeny inferred from the LAGLIDADG_1 alignment containing 191 amino acid positions. 25 HEs belonging to LAGLIDADG_1 superfamily were subjected to the ML and Bayesian methods. The HEs hosted in *cox1* introns are shown in dark red. The details of the homing positions of the HE-hosting *cox1* introns (phase and codon) are given on the right side of the tree. Codon numbers are based on the *S. cerevisiae cox1* gene (GenBank accession number NP_009305.1). I am unsure the precise position of the intron identified in the *Flammulina velutipues* cox1 genes, as only HE sequence has been deposited in the GenBank database (labeled with a question mark). Other details are the same as described in A.

A LAGLIDADG_2



B LAGLIDADG_1

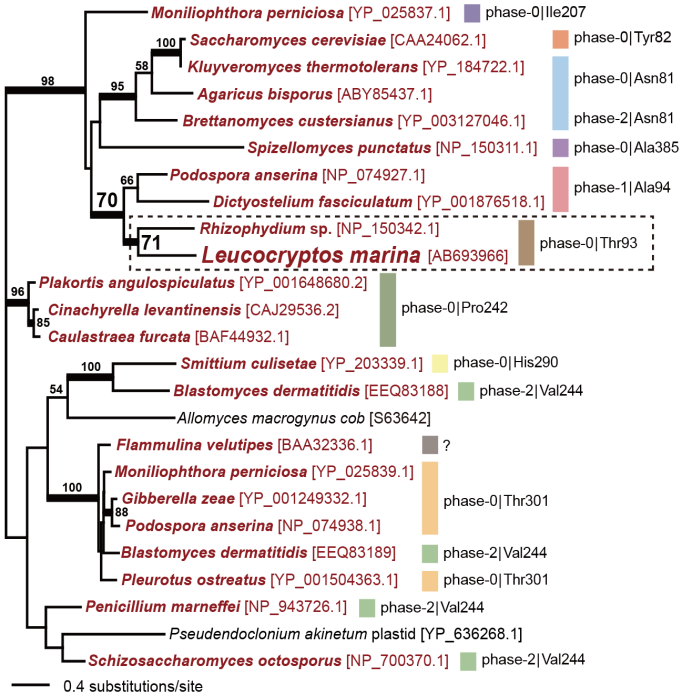


Fig. 9 Mitochondrial genome of *Chrysochromulina* sp. NIES-1333.
Protein-coding genes and rRNA genes are represented by boxes. Gray boxes represent two open reading frames, of which amino acid sequences showed significant sequences similarity to intron-encoded proteins. Transfer RNA genes are represented by lines. Introns are shown in dotted lines. Arrows represent duplicated region.

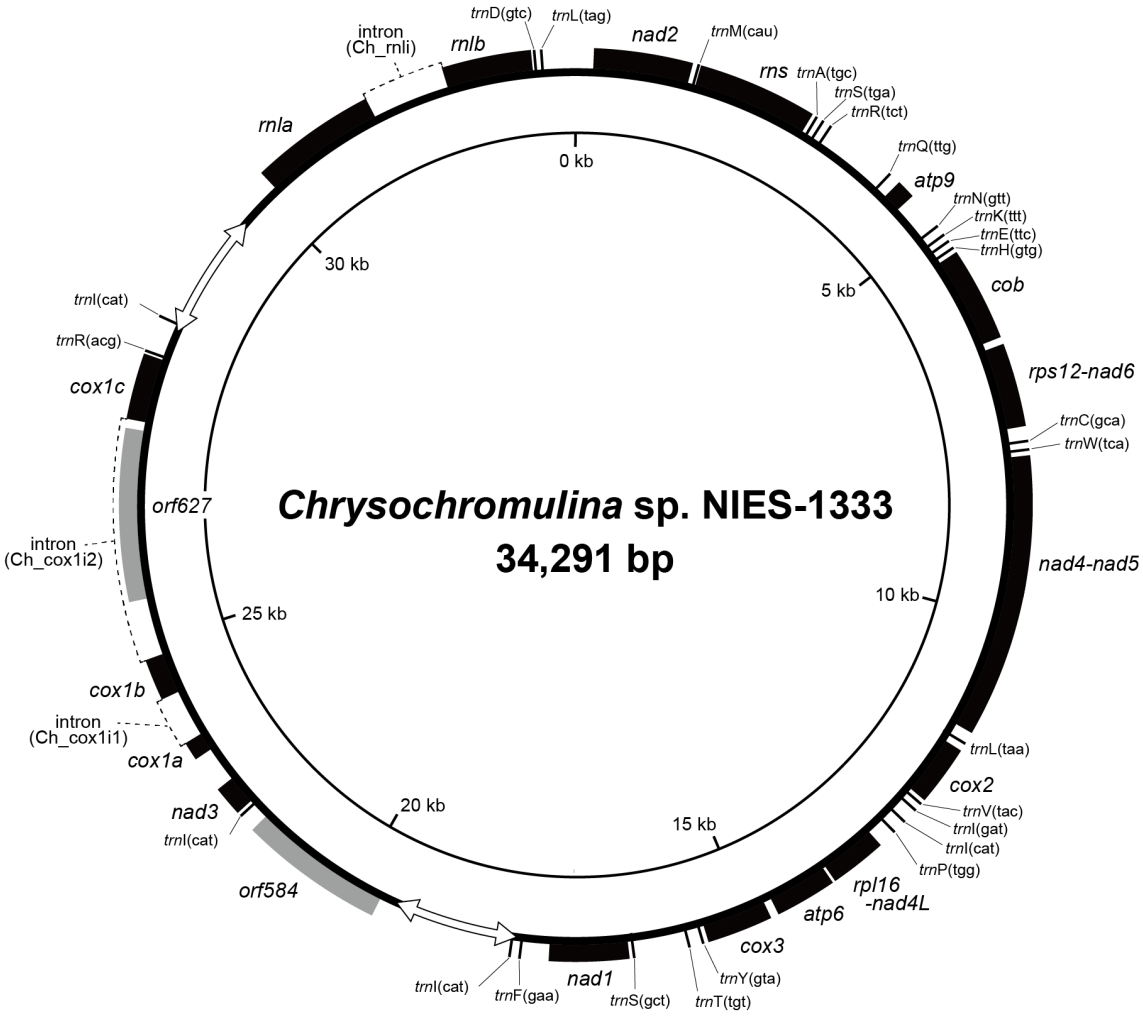
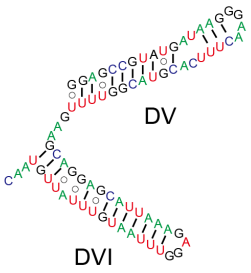

Chrysochromulina sp. NIES-1333
34,291 bp
0 kb
5 kb
10 kb
15 kb
20 kb
25 kb
30 kb
trnD(gtc)
trnL(tag)
nad2
trnM(cau)
rns
trnA(tgc)
trnS(tga)
trnR(tct)
trnQ(ttg)
atp9
trnN(gtt)
trnK(ttt)
trnE(ttc)
trnH(gtg)
cob
rps12-nad6
trnC(gca)
trnW(tca)
nad4-nad5
trnL(taa)
cox2
trnV(tac)
trnI(gat)
trnI(cat)
trnP(tgg)
rpl16
-nad4L
atp6
cox3
trnY(gta)
trnT(tgt)
trnS(gct)
nad1
trnF(gaa)
trnI(cat)
orf584
nad3
trnI(cat)
cox1a
intron
(Ch_cox1i1)
cox1b
intron
(Ch_cox1i2)
orf627
cox1c
trnR(acg)
trnI(cat)
rnla
intron
(Ch_rnli)
rnlb

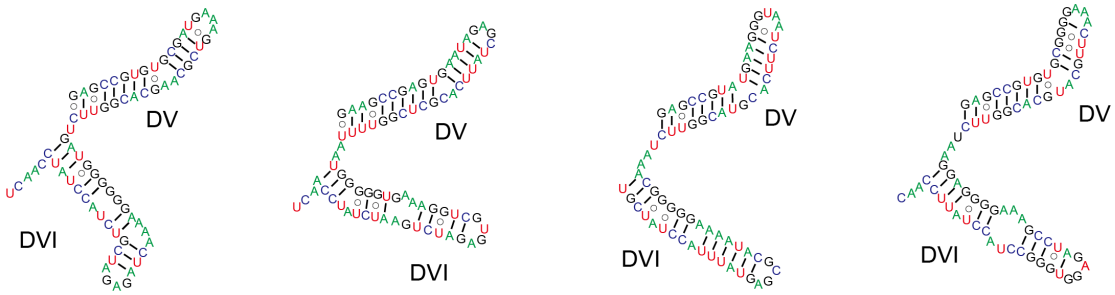
Fig. 10 Comparison of *rnl* intron sharing insertion position with that of *Chrysochromulina* sp.

A. Partial secondary structure of the intron designated as Ch_rnli, as it is in the *rnl* gene of *Chrysochromulina* sp. Putative Watson-Crick and wobble base pairs are shown by lines and open circles, respectively. Capital and small letters represent intron and exon nucleotides, respectively. Domains V and VI, which are group II intron-specific stem structures, are indicated as DV and DVI, respectively. B. Partial secondary structure of the intron which sharing the insertion position with Ch_rnli. Details are the same as described in A. Notes that no similarity is found among the five introns at either primary or secondary structural level. C. Intron hosted in rnl genes. Open, gray, and black boxes represent exons, inton, and intron-encoded proteins (IEPs), respectively. Introns, which shares the homing position, are shaded.

A Partial structure of Ch_rnli



B *rnl* intron structures of the orther organisms



Pylaiella littoralis *Phaeodactylum tricornatum* *Pyropia haitanensis* *Gigaspora rosea*

C comparison of insertion positions

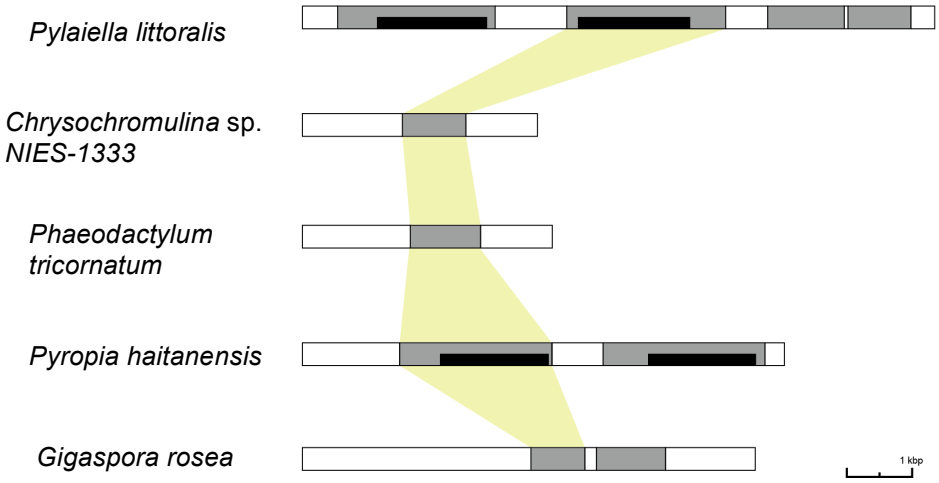
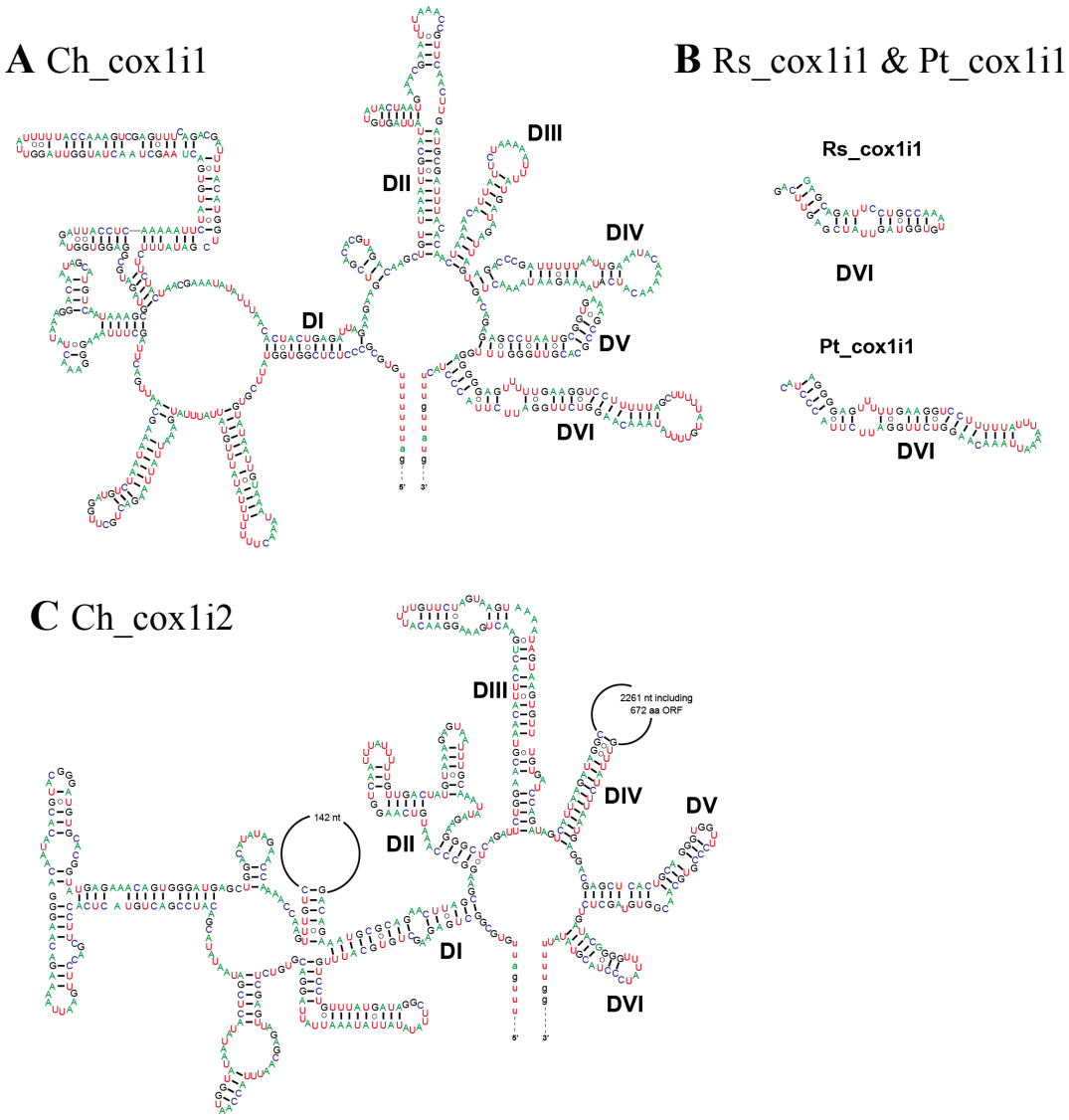


Fig. 11 Putative cox1 introns and comparison among sharing insertion position with that of *Chrysochromulina* sp.

A. Secondary structure of the first *cox1* intron (Ch_cox1i1). The details are the same as described in Fig. 10. B. Partial secondary structure of domain VI (DVI) of the first intron in the cox1 gene of *Phaeodactylum tricornutum* and *Rhosomonas salina* (Pt_cox1i1 and Rs_cox1i1, respectively). Domain VI of Pt_cox1i1 is similar to that of Ch_cox1i1 at both primary and secondary structural levels while there is no similarity between domain VI of Rs_cox11i and that of Ch_cox1i1 or Pt_cox1i1. C. secondary structure of the second cox1 intron (Ch_cox1i2). The details of this figure are the same described above. D. Introns hosted in cox1 genes. Details are the same as described in Fig. 10C.



D Comparison of insertion positions

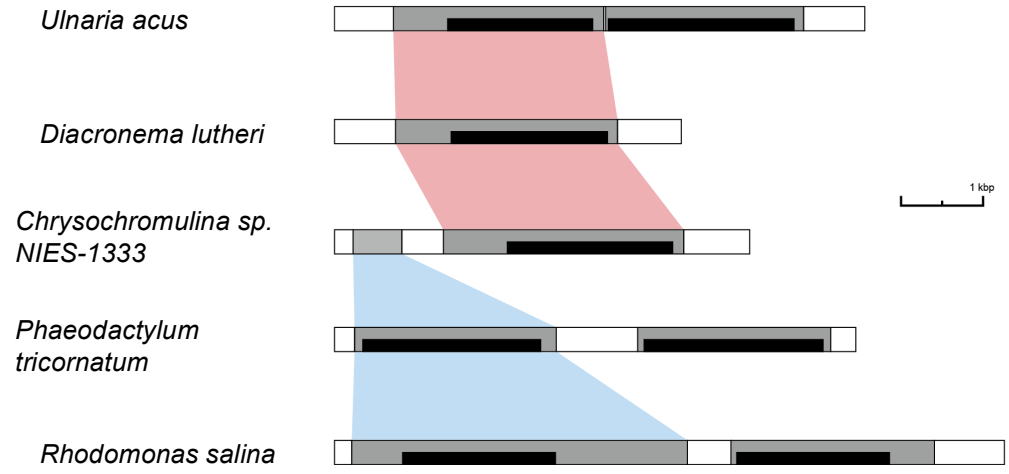
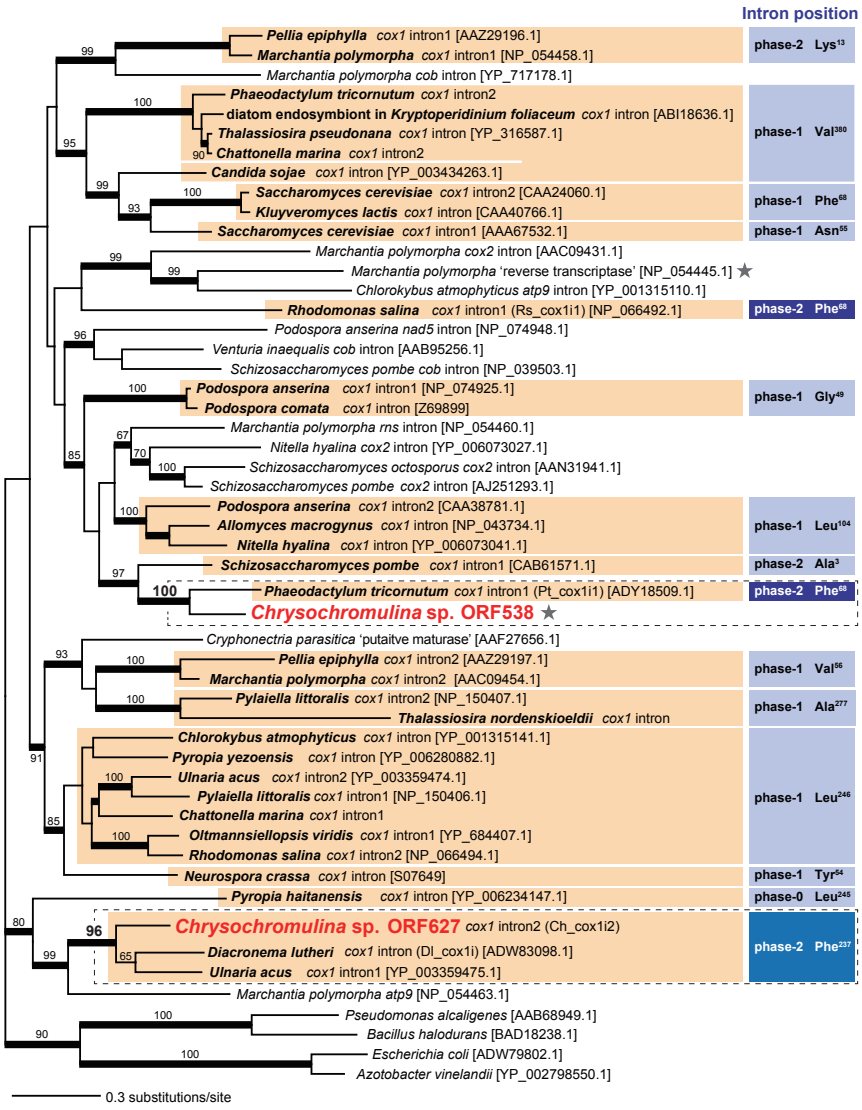


Fig. 12 Phylogeny inferred from 52 Intron-encoded protein (IEP) amino acid sequences.

The IEP alignments were subjected to both maximum-likelihood (ML) and Bayesian methods. As the two methods reconstructed very similar trees, only ML tree is shown here. The tree is rooted by the bacterial sequences. Values at nodes represent ML bootstrap support values greater than 50%. The nodes supported by Bayesian posterior probabilities equal to or greater than 0.95 are highlighted by thick lines. The IPEs in *cox1* introns are shaded in orange. The detailed homing positions of cox1 introns are given on the right side of the tree. Codon numbers are based on the Homo sapiens cox1 gene (GenBank accession number YP_003024028). Free-standing IEPs are highlighted with stars.


Intron position
Pellia epiphylla cox1 intron1 [AAZ29196.1]
Marchantia polymorpha cox1 intron1 [NP_054458.1]
phase-2 Lys13
Marchantia polymorpha cob intron [YP_717178.1]
Phaeodactylum tricornutum cox1 intron2
diatom endosymbiont in Kryptoperidinium foliaceum cox1 intron [ABI18636.1]
Thalassiosira pseudonana cox1 intron [YP_316587.1]
Chattonella marina cox1 intron2
phase-1 Val380
Candida sojae cox1 intron [YP_003434263.1]
Saccharomyces cerevisiae cox1 intron2 [CAA24060.1]
Kluyveromyces lactis cox1 intron [CAA40766.1]
phase-1 Phe68
Saccharomyces cerevisiae cox1 intron1 [AAA67532.1]
phase-1 Asn55
Marchantia polymorpha cox2 intron [AAC09431.1]
Marchantia polymorpha 'reverse transcriptase' [NP_054445.1]
Chlorokybus atmophyticus atp9 intron [YP_001315110.1]
Rhodomonas salina cox1 intron1 (Rs_cox1i1) [NP_066492.1]
phase-2 Phe68
Podospora anserina nad5 intron [NP_074948.1]
Venturia inaequalis cob intron [AAB95256.1]
Schizosaccharomyces pombe cob intron [NP_039503.1]
Podospora anserina cox1 intron1 [NP_074925.1]
Podospora comata cox1 intron [Z69899]
phase-1 Gly49
Marchantia polymorpha rns intron [NP_054460.1]
Nitella hyalina cox2 intron [YP_006073027.1]
Schizosaccharomyces octosporus cox2 intron [AAN31941.1]
Schizosaccharomyces pombe cox2 intron [AJ251293.1]
Podospora anserina cox1 intron2 [CAA38781.1]
Allomyces macrogynus cox1 intron [NP_043734.1]
Nitella hyalina cox1 intron [YP_006073041.1]
phase-1 Leu104
Schizosaccharomyces pombe cox1 intron1 [CAB61571.1]
phase-2 Ala3
Phaeodactylum tricornutum cox1 intron1 (Pt_cox1i1) [ADY18509.1]
phase-2 Phe68
Chrysochromulina sp. ORF538
Cryphonectria parasitica 'putaitve maturase' [AAF27656.1]
Pellia epiphylla cox1 intron2 [AAZ29197.1]
Marchantia polymorpha cox1 intron2 [AAC09454.1]
phase-1 Val56
Pylaiella littoralis cox1 intron2 [NP_150407.1]
Thalassiosira nordenskioeldii cox1 intron
phase-1 Ala277
Chlorokybus atmophyticus cox1 intron [YP_001315141.1]
Pyropia yezoensis cox1 intron [YP_006280882.1]
Ulnaria acus cox1 intron2 [YP_003359474.1]
Pylaiella littoralis cox1 intron1 [NP_150406.1]
Chattonella marina cox1 intron1
Oltmannsiellopsis viridis cox1 intron1 [YP_684407.1]
Rhodomonas salina cox1 intron2 [NP_066494.1]
phase-1 Leu246
Neurospora crassa cox1 intron [S07649]
phase-1 Tyr54
Pyropia haitanensis cox1 intron [YP_006234147.1]
phase-0 Leu245
Chrysochromulina sp. ORF627 cox1 intron2 (Ch_cox1i2)
Diacronema lutheri cox1 intron (Dl_cox1i) [ADW83098.1]
Ulnaria acus cox1 intron1 [YP_003359475.1]
phase-2 Phe237
Marchantia polymorpha atp9 [NP_054463.1]
Pseudomonas alcaligenes [AAB68949.1]
Bacillus halodurans [BAD18238.1]
Escherichia coli [ADW79802.1]
Azotobacter vinelandii [YP_002798550.1]
0.3 substitutions/site