Hyper-Eccentric Structural Genes in the Mitochondrial Genome of the Algal Parasite Hemistasia phaeocysticola

Akinori Yabuki1,*, Goro Tanifuji2,3, Chiho Kusaka1, Kiyotaka Takishita1, and Katsunori Fujikura1

1Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokosuka, Kanagawa, Japan
2Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Ibaraki, Japan
3Present address: Department of Zoology, National Museum of Nature and Science, Tsukuba, Ibaraki, Japan

*Corresponding author: E-mail: yabukia@jamstec.go.jp.

Accepted: August 20, 2016

Data deposition: This project has been deposited at GenBank/EMBL/DDBJ under the accession DRA003989 (PacBio sequencing data for the mtDNA of Hemistasia phaeocysticola) and LC114081-LC114090 & LC120367-LC120387 (the mitochondrial genomic DNA and cDNA sequences obtained by the Sanger method).

Abstract

Diplonemid mitochondria are considered to have very eccentric structural genes. Coding regions of individual diplonemid mitochondrial genes are fragmented into small pieces and found on different circular DNAs. Short RNAs transcribed from each DNA molecule mature through a unique RNA maturation process involving assembly and three types of RNA editing (i.e., U insertion and A-to-I and C-to-U substitutions), although the molecular mechanism(s) of RNA maturation and the evolutionary history of these eccentric structural genes still remain to be understood. Since the gene fragmentation pattern is generally conserved among the diplonemid species studied to date, it was considered that their structural complexity has plateaued and further gene fragmentation could not occur. Here, we show the mitochondrial gene structure of Hemistasia phaeocysticola, which was recently identified as a member of a novel lineage in diplonemids, by comparison of the mitochondrial DNA sequences with cDNA sequences synthesized from mature mRNA. The genes of H. phaeocysticola are fragmented much more finely than those of other diplonemids studied to date. Furthermore, in addition to all known types of RNA editing, it is suggested that a novel processing step (i.e., secondary RNA insertion) is involved in the RNA maturation in the mitochondria of H. phaeocysticola. Our findings demonstrate the tremendous plasticity of mitochondrial gene structures.

Key words: gene fragmentation, RNA processing, mitochondrial genome, diplonemids, Euglenozoa.

Introduction

Gene coding regions are often interrupted. Introns are the primary structural interrupters of coding region, such as spliceosomal introns, group I and II introns, and archael introns (Moreira et al. 2012). On the other hand, some genes of microbial eukaryotes, such as ciliates and apicomplexans, are further fragmented and found at multiple loci in genomes. These fragmented genes may result in discontinuous products that work as subunits or split proteins (Heinonen et al. 1987; Edqvist et al. 2000; Feagin et al. 2012). Among the known gene structures, the most complex and unusual are considered to be those of the mitochondrial genome of diplonemids (Euglenozoa), a sister group of kinetoplastid flagellates (Marande et al. 2005; Vlček et al. 2011).

Diplonemids are unicellular flagellates that are free-living or parasitize algae and invertebrates (Lukes et al. 2015). While the diversity of diplonemids remains incompletely understood (Lara et al. 2009; de Vargas et al. 2015), the fragmented genes found in the diplonemid mitochondrial genomes have been studied intensively; their complex and unusual structures have attracted attention in the field of not only organelle genomics, but also RNA processing. The mitochondrial genomes of diplonemids studied to date, namely Diplonema spp. and Rhynchopus euleides, comprise multiple mini-circular chromosomes ~5–12 kb (Burger et al. 2012). Although it still remains unknown how many genes there are in the mitochondrial genomes of diplonemids, the coding regions of all genes studied to date including the rRNA genes are...
fragmented into small pieces called “modules”; no nonfragmented version of the gene has been found (Vlcek et al. 2011; Valach et al. 2014). For example, the protein-coding region of cox1 is fragmented into nine pieces, and its fragmentation pattern is generally conserved among the diplonemid species studied to date (Marande and Burger 2007; Kiethega et al. 2011); the number of modules and fragmented positions found in cox1 are generally the same among four diplonemid species sequenced to date. Each module is independently transcribed, and primary transcripts undergo secondary assembly into a single RNA (Marande et al. 2005; Marande and Burger 2007; Kiethega et al. 2013).

Three types of RNA editing [i.e., uridine (U) insertion and A-to-I and C-to-U substitutions] are also involved in the mRNA maturation process before the end products of assembled RNA sequences are translated to the proper amino acid sequences (Marande and Burger 2007; Kiethega et al. 2013; Moreira et al. 2016). As misassembled mRNA, in which the short RNAs are assembled in incorrect order, has not been documented, this RNA maturation process is considered to progress quite accurately (Burger et al. 2012). However, the underlying molecular mechanism(s) remain(s) unknown. Although it is unclear how and why these unusual gene structures were established, they have been suggested to be a result of constructive neutral evolution; it has also been suggested that there are no functional advantages of gene fragmentation (Flegontov et al. 2011). Furthermore, because of the shared manner of gene fragmentation among diplonemid species examined to date, the complexity of their gene structures may have plateaued during evolution and further gene fragmentation may have not occurred in this lineage (Flegontov et al. 2011).

*Hemistasia phaeocysticola* was recently identified as a representative of a large environmental clade within diplonemids, and its phylogenetic position is distinct from that of other diplonemids whose mitochondrial gene structures have been analyzed (Yabuki and Tame 2015). In the present study, we show the hyper-eccentric degree of the mitochondrial gene structure of *H. phaeocysticola* with comparison to mitochondrial genes in other diplonemids. We also discuss the possibility of a unique RNA maturation process and the evolutionary constraint on the gene structure in *H. phaeocysticola* mtDNA.

### Materials and Methods

#### DNA and RNA Extraction and cDNA Preparation

Total genomic DNA and RNA were extracted from ~100 ml samples of 2-week-old cultures (NIES-3356) using the DNeasy Plant mini kit (Qiagen) and TRIzol (Life Technologies), respectively. cDNA was synthesized from DNase-treated RNA by SuperScript II (Life Technologies) using random hexamers or by the GeneRacer Kit (Life Technologies). As the RNA transcribed in mitochondria does not have a 5’ cap structure, the steps for the removal of the 5’ phosphates and the 5’ cap structure were skipped in cDNA synthesis using the GeneRacer Kit.

#### Sequencing of cox1 cDNA and Chromosomes Possessing cox1 Modules

Polymerase chain reaction (PCR) was performed using HotStar Taq (Qiagen), and PCR products were cloned into the pCR2.1 vector using a TOPO TA Cloning Kit (Life Technologies). The clones were completely sequenced on both strands. The partial sequence of cox1 cDNA was amplified from the synthesized cDNA using the dip cox1 F and dip cox1 R primer set (supplementary table S1, Supplementary Material online). Amplification consisted of 35 cycles of denaturing at 94°C for 30 s, annealing at 52°C for 30 s, and extension at 72°C for 1 min. Based on this partial sequence, we designed six primers (p1–p6) (supplementary table S1, Supplementary Material online), and the complete cDNA sequence of mature cox1 mRNA was amplified using these primers and adapter primers from the cDNA synthesized by the GeneRacer kit. The PCR program for full-length cox1 cDNA consisted of 35 cycles of denaturing at 94°C for 30 s, annealing at 55°C for 30 s, and extension at 72°C for 2 min. The PCR conducted with genomic DNA and three combinations of outward primer sets (i.e., p1 and p2, p3 and p4, and p5 and p6) for the amplification of chromosomal sequences consisted of two continuous steps: (1) three cycles of denaturing at 94°C for 30 s, annealing at 58°C for 30 s, and extension at 72°C for 2.5 min; (2) 32 cycles of denaturing at 94°C for 30 s, annealing at 55°C for 30 s, and extension at 72°C for 2.5 min. The PCR to confirm the sequence variation and existence of immature mRNA was conducted with the primer set, “cox1 FULL F” and “cox1 FULL R”; the program included 25 cycles of denaturing at 94°C for 30 s, annealing at 55°C for 30 s, and extension at 72°C for 2 min. The acquired cDNA sequences were manually aligned and the structure of the cDNA sequences that were derived from possible immature and misassembled cox1 mRNA (i.e., LC114084–LC114087) predicted by comparison with the cDNA sequence derived from mature mRNA (i.e., LC114081).

#### Southern Hybridization Analysis

Total genomic DNA from *H. phaeocysticola* was prepared using a standard cetyltrimethylammonium bromide (CTAB) protocol (Doyle and Doyle 1990). Southern hybridization probes targeting the shared region, the distinct noncoding region of the chromosomes possessing modules 13-1 and 13-2, were amplified using the following primers: “CONS probe F,” “CONS probe R,” “M13-1 probe F,” “M13-1 probe R,” “M13-2 probe F,” and “M13-2 probe R” (supplementary table S1, Supplementary Material online). The PCR products were cloned into the pGEM T-easy vector (Promega) and used as templates for digoxigenin (DIG)-labeled probe.
synthesis according to the manufacturer’s instructions (PCR DIG labeling kit, Roche Diagnostics). The 200 ng of total DNA per lane were applied for electrophoresis and transferred to a positively charged nylon membrane by capillary transfer. Southern hybridization was carried out in DIG Easy Hyb buffer (Roche Diagnostics) with two μl/ml DIG-labeled probes overnight at 37 °C. The membranes were rinsed twice for 10 min at room temperature in low-stringency buffer (0.1% SDS, 2 x SSC), followed by two 15-min rinses at 65 °C in high-stringency buffer (0.1% SDS, 0.5 x SSC). Hybridization signals were detected according to standard procedures in the DIG detection kit (Roche Diagnostics).

Mitochondrial Genome Sequencing

The mitochondrial chromosomes were independently amplified from the total DNA using two primer sets—"CONS. fr and CONS. rF1" and "CONS. fr and CONS. rF2"—to reduce the risk of biased amplification. Note that these primers were designed to outwardly anneal within the shared region that is possibly conserved among all chromosomes (see below). Each amplification consisted of 40 cycles of denaturing at 94 °C for 30 s, annealing, and extension at 72 °C for 3.5 min. The PCR products contained sequences of various length and more than 100 clones were randomly sequenced to check that the amplification was successful (it should be noted that these clones were not sequenced completely). The PCR products were divided into two groups: 1-2 and 2-20 kb, and then directly utilized as initial material (without any fragmentation process) for the sequence library of PacBio RSII. Hence, two libraries were constructed using the SMRTbell Template kit 1.0 (Pacific Biosciences). The mitochondrial genome library construction and sequencing by PacBio RSII (Pacific Biosciences) were performed at the Dragon Genomics Center (TAKARA Bio). The libraries constructed from the 1-2- and 2-20-kb PCR products were sequenced by one and two SMRT cells, respectively. The circular consensus sequencing (CCS) reads were generated from raw reads by the Dragon Genomics Center.

Complete cDNA Sequencing of cob, cox2, and nad7

The primers for 3' and 5' RACE were designed on the basis of the module sequences of cob, cox2, and nad7 detected in the preliminary mitochondrial genome sequencing (supplementary table S1, Supplementary Material online), and the complete cDNA sequences were amplified. Each PCR run consisted of 35 cycles of denaturing at 94 °C for 30 s, annealing at 55 °C for 30 s, and extension at 72 °C for 2 min.

Identification of Module Regions for Each Mitochondrial Gene

To identify the modules of each mitochondrial gene, BLAST 2.2.28+ (ftp://ftp.ncbi.nih.gov/blast/executables/blast+/2.2.28/, last accessed 9 April 2015) was utilized. A local BLAST database was constructed from the provided CCS reads. Module sequences were searched for using the complete cDNA sequences of each gene as a query using blastn. The boundary positions between neighboring modules were estimated by eye. The coding regions that could not be detected even by blastn with the option for short sequence search (i.e., “-task blastn-short”) were left as missing regions in this study, but the undetected regions composed of only T residues were regarded as the sites for uridine-insertion RNA editing. The G/T content of each missing region was calculated manually. It was here assumed that each missing region exists as a single module in the mitochondrial genome and tentative module numbers were assigned to them, respectively.

Results

Preliminary Survey of cox1-Coding Chromosomes and Detection of the General “Core” of Mitochondrial Chromosomes

The partial cDNA of the mature cox1 mRNA of H. phaeocystica was obtained by reverse-transcription polymerase chain reaction (RT-PCR) using “dip cox1 F” and “dip cox1 R” (supplementary table S1, Supplementary Material online). No product with the same primers was amplified from genomic DNA. The complete cDNA sequence of mature cox1 mRNA was then determined by 5' and 3' RACE (deposited in Genbank as LC114081) (fig. 1A). The amino acid sequence was translated from this cDNA sequence under the rule of genetic code table 4 (i.e., The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/ Spiroplasma Code), but not under the standard code table (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi, last accessed 7 September 2016). It was also confirmed that the amino acid sequences of cox1 in this study cluster with other diplonemid cox1 sequences with 100% support in the phylogenetic tree (supplementary material, Supplementary Material online).

Five mitochondrial chromosomes that possess modules numbered 7, 12 (two types), and 13 (two types) were acquired by the genomic PCR conducted using outward primers annealing to various fragments of the deduced coding region (fig. 1A and B). The mitochondrial chromosome sequences contained homo-polymorphic repeats at numerous sites. Two types of mitochondrial chromosomes possessing module 13, which were 2,682 and 3,185 bp, were completely sequenced by primer walking and deposited in Genbank as LC114082 and LC114083, respectively. The circular architecture of the module 13 chromosome was confirmed by PCR with the inward primer set (i.e., fig. 1B, inv. “m13 F and inv. m13 R” in supplementary table S1, Supplementary Material online). All mitochondrial chromosomes sequenced herein contained nearly identical 500-bp sequences at the 5' upstream region of each module; these sequences were
designated “shared regions (SRs)” (fig. 1B). Southern hybridization analysis using the shared regions (without any restriction enzyme digestion) showed multiple signals (fig. 1C). While the band patterns and intensities of the M13-1 and M13-2 probe signals were different from one another, those signals corresponded to the portion of the signals of shared region probe (fig. 1C).

Amplification and Sequencing of Mitochondrial DNA

Sequences amplified by PCR with the outward primers of the shared region (i.e., “CONS. fR and CONS. rF1” and “CONS. fR and CONS. rF2” in supplementary table S1, Supplementary Material online) included partial sequences of cox1, cox2, cox3, cob, atp6, and nad7 (data not shown). PCR products obtained using the shared region primers were divided into two fractions on the basis of length, for example, 1-2 and 2-20 kb, to generate two different DNA libraries to obtain comprehensive sequence data (see below).

The library constructed from 1- to 2-kb PCR products with shared region primers was sequenced by one SMRT cell, and 227,054 subreads yielding 12,992 CCS reads were obtained. The second library consisting of 2-20-kb PCR products was sequenced by two SMRT cells, and 127,303 and 157,557 subreads, yielding 9,581 and 11,542 CCS reads were obtained, respectively. The raw sequence data were deposited in Genbank as DRA003989. The CCS reads were analyzed in the subsequent analyses (CCS reads are provided upon request). Homo-polymeric repeats were also recognized in each CCS read at numerous sites.

cox1 Gene Structure and Sequence Variations

Seventeen fragments corresponding to most of the protein-coding region of cox1 were retrieved from the CCS data, but seven small regions were missing (fig. 2A). Including these regions, the _H. phaeocysticola_ cox1 protein-coding region was found to be fragmented into at least 24 modules (fig. 2A). The boundary position between the modules could be elucidated clearly, except in a few exceptional cases. For example, as the boundary between modules 7 and 8 overlapped by one T residue, no precise boundary was recognizable (fig. 2B). Five sites for U-insertion type of RNA editing were predicted (i.e., between modules 11 and 12, modules 12 and 13, modules 19 and 20, and modules 20 and 21, and at the end of module 24: fig. 2A, C and D). The exact number of U residues inserted between modules 12 and 13 as well as modules 19 and 20 could not be elucidated, because it was unclear whether the last T residues on modules 12 and 19 were indeed part of modules or not (fig. 2C and D). The five T residues between modules 12 and 13 in _H. phaeocysticola_ were in the same position as an RNA editing site of _Diplonema papillatum_ cox1, in which six uridines are inserted (fig. 2A). The eight boundaries between _D. papillatum_ cox1 modules were exactly shared with cox1 of _H. phaeocysticola_ (fig. 2A).

**Fig. 1.**—Summary of preliminary surveys of cox1-coding chromosomes. (A) Complete cDNA sequence of mature cox1 mRNA. The annealing positions of the primers and the module regions identified in the preliminary surveys are shown with the cDNA sequence. (B) Schematic representation of circular chromosomes possessing modules 7, 12, and 13. Solid lines correspond to the regions that were actually sequenced. The regions shown by dashed-lines still remain to be sequenced. On the physical map of the chromosomes possessing module 13, the positions of probes (i.e., M13-1 and M13-2) that can distinguish each chromosome are also shown. SR, shared region; m7, module 7; m12, module 12; m13, module 13. (C) Southern hybridization results. Left, middle, and right columns correspond to the results obtained using the specific probes for the shared region, M13-1, and M13-2, respectively.
Among the 17 identified modules in *H. phaeocysticola* cox1, sequence variation was found in 12 modules. Among the variations, there were also nonsynonymous substitutions. For example, four different variants of module 20 were identified in the mitochondrial DNA sequences (fig. 2E, supplementary material, Supplementary Material online), three of which are nonsynonymous substitutions. The selected representative sequences of each module in which the variations were found are provided in the supplementary material, Supplementary Material online. The sequence variations were also recognized in the cDNA sequences amplified with the primer set, “cox1 FULL F” and “cox1 FULL R.” Sequencing of 34 clones yielded 25 different sequences (deposited in GenBank as LC114084–LC114087 and LC120367–LC120387). Although eight cDNA sequences could not be translated to amino acid sequences (see below), the other sequences were also different from one another at the amino acid level (supplementary material, Supplementary Material online). Furthermore, several single nucleotide polymorphisms found in cDNA sequences could not be confirmed.

![Figure 2](https://academic.oup.com/gbe/article-abstract/8/9/2870/2669850/2874)
in these genomic sequences (figs. 2 and 3A) and all cDNA sequences corresponding to the module 1 region were inconsistent with the genomic sequences at several sites (fig. 3A). At these inconsistent sites, G and T in cDNA sequences were replaced with A and C in the genomic sequences, respectively.

Seven unidentified regions in cDNA sequence were related to T/G-rich (fig. 3B). In particular, the T/G content of "modules 2, 16, 18, and 22" exceeded 90% (fig. 3B). Sequence variations were also recognized in cDNA data for these missing regions, except "modules 16 and 22" (fig. 3B). Among eight cDNA sequences that could not be translated to amino acid sequences, internal region was missing in three sequences (i.e., LC114084–LC114086: fig. 3C, D and E); for example, the 3’ end of truncated module 2 was directly connected with the 5’ end of module 22 in LC114084 (fig. 3C). While both LC114085 and LC114086 contained truncated module 18, the length of and truncated position in module 18 differed (fig. 3D and E). In one sequence (i.e., LC114087), only module 4 was replaced with the unknown sequence, which could not be identified in the mitochondrial DNA data, and the codon-frame was shifted from this unknown sequence (fig. 3F). A stop codon was found inside LC120371. The three other cDNA sequences (i.e., LC120375, LC120380 and LC120382) lacked or additionally possessed one T residue in continuous T regions.

Gene Structures of cob, cox2, and nad7

The complete cDNA sequences of mature cob, cox2, and nad7 mRNA were determined by RT-PCR and 3’/5’ RACE methods, and their sequences are deposited in Genbank as LC114088, LC114089, and LC114090, respectively (fig. 4). These sequences could also be translated to amino acid sequences by gene code table 4. Most of the modules of each gene were identified in the mitochondrial DNA sequences and the missing regions exhibited high G/T contents as was the case for cox1 (fig. 4). Each gene was fragmented more finely than in D. papillatum, and the boundaries between D. papillatum modules were mostly shared in H. phaeocysticola, except one position in each gene (e.g., the boundary position between modules 5 and 6 in cob of H. phaeocysticola; fig. 4). RNA editing sites for U-insertion were also found in cob, cox2, and nad7
as in cox1. The RNA editing site found at the terminal region of cob in D. papillatum was not found in H. phaeocysticola (fig. 4). The gene fragmentation patterns as well as the position and number of U residues inserted by RNA editing are summarized in figure 4; it should be noted that this schematic summary might not be precisely correct because of overlapping neighbor module sequences at several boundary positions (e.g., between modules 7 and 8 of cob; see supplementary material, Supplementary Material online). Sequence variations in the modules were also recognized by comparison with the genomic sequences of the three genes, as with cox1, and their selected representative sequences are also provided in the supplementary material, Supplementary Material online.

**Discussion**

The four genes examined here are fragmented into small pieces in the mitochondrial genome of H. phaeocysticola, and no nonfragmented versions of the genes were found. These features are consistent with those of other diplonemids. Furthermore, many modules had multiple copies whose sequences differed from one another, even though they correspond to the same coding region. Although module 9 of cox1 of D. papillatum was found in two chromosomes (Vícek et al. 2011), no other multi-copy modules have been reported in other diplonemids. In addition, sequence variation within a given module has not been reported in other diplonemids. Furthermore, the mitochondrial chromosomes of H. phaeocysticola are smaller than those of other diplonemids, although only two chromosomes (i.e., LC114082 and LC114083) were sequenced completely in the present study. Other chromosomes of H. phaeocysticola amplified by module-specific outward primers and incompletely sequenced were also shown to be smaller than 5 kb (fig. 1B). Other chromosomal features—the existence of a shared region and many homo-polymeric repeats—are shared with other diplonemids.

As the existence of such a shared region has also been reported in the chromosomes of other diplonemids (Vícek et al. 2011; Kiethega et al. 2013), this is not a unique feature of H. phaeocysticola. However, this feature is useful for sequencing the mitochondrial genome of H. phaeocysticola, because comprehensive mitochondrial genome amplification can be expected by PCR with outward primers annealing within the shared region. In fact, the modules of other genes (i.e., cob, cox2, nad7, and more) were recovered from the amplified sequences in the preliminary analyses. Thus, this approach allowed us to selectively amplify the
mitochondrial genome of *H. phaeocysticola* and sequence it broadly and deeply using PacBio RS II (Pacific Biosciences).

Module identification based on a BLAST search showed that *cox1* of *H. phaeocysticola* is fragmented into at least 24 pieces (fig. 2A). As three other genes of *H. phaeocysticola* are also fragmented two times finer than the corresponding genes in *D. papillatum*, this hyper-fragmentation is possibly a general feature of the mitochondrial genes of *H. phaeocysticola*. The genes found in the micronuclear genomes of the hypotrichous ciliates are also discontinuous and highly interruped, but these are joined by DNA splicing during macronuclear development and only the noninterrupted genes of the macronuclear genome are transcriptionally active (Landweber et al. 2000; Swart et al. 2013). Hence, the mitochondrial genes found in *H. phaeocysticola* may be the most structurally eccentric transcriptionally active genes known to date.

Furthermore, the *cox1* cDNA sequences that possess the truncated module sequences (fig. 3C, D and E) suggest that seven small fragments of *cox1* could not be identified in the mitochondrial DNA sequences (fig. 2A). Although the possibility that some chromosomes were not amplified with the PCR conditions used in the present study and/or that the sequencing depth was insufficient cannot be ruled out completely, we are inclined to consider that they may not be encoded directly in the genome and/or may be further fragmented into much smaller pieces. These unidentified regions are predominantly composed of T and G residues (fig. 3B) and several C and A residues in the genomic sequences are replaced with T and G residues in cDNA sequences, respectively (fig. 3A). If these inconsistencies between cDNA and genomic sequences represent the existence of C-to-U and A-to-G (or I, because I’s in RNA are read as G’s by reverse transcriptase) types of RNA editing, the genomic sequences corresponding to some of the unidentified regions are different from the cDNA sequences and therefore they still remain to be identified. Indeed, the existence of C-to-U and A-to-I substitution type of RNA editing has been confirmed in *D. papillatum* recently (Moreira et al. 2016). Hence, it is reasonable to consider that the same RNA editing may also occur in the mitochondrion of *H. phaeocysticola*.

Seven small fragments of *cox1* could not be identified in the mitochondrial DNA sequences (fig. 2A). Although the possibility that some chromosomes were not amplified with the PCR conditions used in the present study and/or that the sequencing depth was insufficient cannot be ruled out completely, we are inclined to consider that they may not be encoded directly in the genome and/or may be further fragmented into much smaller pieces. These unidentified regions are predominantly composed of T and G residues (fig. 3B) and several C and A residues in the genomic sequences are replaced with T and G residues in cDNA sequences, respectively (fig. 3A). If these inconsistencies between cDNA and genomic sequences represent the existence of C-to-U and A-to-G (or I, because I’s in RNA are read as G’s by reverse transcriptase) types of RNA editing, the genomic sequences corresponding to some of the unidentified regions are different from the cDNA sequences and therefore they still remain to be identified. Indeed, the existence of C-to-U and A-to-I substitution type of RNA editing has been confirmed in *D. papillatum* recently (Moreira et al. 2016). Hence, it is reasonable to consider that the same RNA editing may also occur in the mitochondrion of *H. phaeocysticola*.

Furthermore, the *cox1* cDNA sequences that possess the truncated module sequences (fig. 3C, D and E) suggest that
modules 2 and 18 (and some other missing regions) may be further fragmented into much smaller pieces. For example, a fragment of just 7 bp found in truncated module 2 (LC114084 and LC114085) may independently exist in the mitochondrial genome and such very short sequences are very difficult to identify using BLAST. Sequences corresponding to further fragmented modules could actually be found in the mitochondrial DNA sequence data by text search, although it is difficult to distinguish whether these short sequences are indeed modules or just spurious similarity.

The cox1 cDNA sequences lacking internal regions (i.e., LC114084–LC114086: fig. 3C, D and E) may also provide hints about the RNA maturation process in H. phaeocysticola mitochondria. Neighbor modules are assembled accurately step-by-step in Diplonema and Rhynchopus, and this type of immature or misassembled sequence has never been reported (Burger et al. 2012). If the internal region lacking cox1 sequences observed here represent an intermediate stage of RNA maturation, it would be interesting and important to understand the RNA maturation process of H. phaeocysticola. For instance, if some internal regions are inserted secondarily into pre-assembled mRNA, the RNA maturation process may differ at least in part between H. phaeocysticola and other diplonemids. The finding of the sequence in which only module 4 was replaced with the unknown sequence (i.e., LC114087: fig. 3F) may also support the existence of a sequence-insertion step in the RNA maturation process.

Almost all module boundary positions found in Diplonema and Rhynchopus are shared in the homologous genes of H. phaeocysticola. This means that most of the gene structures found in Diplonema and Rhynchopus were established in their common ancestor and that little further gene fragmentation occurred in the lineage leading to Diplonema and Rhynchopus. On the other hand, gene fragmentation continued in the lineage leading to H. phaeocysticola. This indicates that the complexity of the gene structure may have plateaued in the lineage leading to Diplonema and Rhynchopus as previously suggested (Flegontov et al. 2011), but not in the lineage leading to H. phaeocysticola. In other words, further fragmentation was tolerated in the ancestor leading to H. phaeocysticola (fig. 5).

**Supplementary Material**

Supplementary material and table S1 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org).

**Acknowledgments**

This work was supported by grants from the JSPS awarded to A.Y. (26840133) and K.T. (25650132).

**Literature Cited**

Burger G, Jackson CJ, Waller RF. 2012. Unusual mitochondrial genomes and genes. In: Bullenwell CE, editor. Organelle genetics. Berlin: Springer. p. 41–77.

de Vargas C, et al. 2015. Eukaryotic plankton diversity in the sunlit ocean. Science 348:1261605.

Dobáková E, Flegontov P, Škalický T, Lukeš J. 2015. Unexpectedly streamlined mitochondrial genome of the euglenozoan Euglena gracilis. Genome Biol Evol. 7:3358–3367.

Doyle JJ, Doyle JL. 1990. Isolation of plant DNA from fresh tissue. Focus 12:13–15.

Edqvist J, Burger G, Gray MW. 2000. Expression of mitochondrial protein-coding genes in Tetrahymena pyriformis. J Mol Biol. 297:381–393.

Feagin JE, et al. 2012. The fragmented mitochondrial ribosomal RNAs of Plasmodium falciparum. PLoS One 7:e38320.

Flegontov P, Gray MW, Burger G, Lukeš J. 2011. Gene fragmentation: a key to mitochondrial genome evolution in Euglenozoa? Curr Genet. 57:225–232.

Heinonen TY, Schnare MN, Young PG, Gray MW. 1987. Rearranged coding segments, separated by a transfer RNA gene, specify the two parts of a discontinuous large subunit ribosomal RNA in Tetrahymena pyriformis mitochondria. J Biol Chem. 262:2879–2887.

Kiethega GN, Turcotte M, Burger G. 2011. Evolutionarily conserved cox1 trans-splicing without cis-motifs. Mol Biol Evol. 28:2425–2428.

Kiethega GN, Yan Y, Turcotte M, Burger G. 2013. RNA-level unscrambling of fragmented genes in Diplonema mitochondria. RNA Biol. 10:301–313.

Landweber LF, Kuo TC, Curtis EA. 2000. Evolution and assembly of an extremely scrambled gene. Proc Natl Acad Sci U S A. 97:3298–3303.

Lara E, Moreira D, Vereschchaka A, López-García P. 2009. Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids. Environ Microbiol. 11:47–55.

Lukeš J, Flegontova O, Horák A. 2015. Diplonemids. Curr Biol. 25:R702–R704.

Marande W, Burger G. 2007. Mitochondrial DNA as a genomic jigsaw puzzle. Science 318:415–415.

Marande W, Lukeš J, Burger G. 2005. Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. Eukaryot Cell 4:1137–1146.

Moreira S, Breton S, Burger G. 2012. Unscrambling genetic information at the RNA level. Wiley Interdiscip Rev RNA 3:213–228.

Moreira S, Valach M, Aoulad-Aissa M, Otto C, Burger G. 2016. Novel modes of RNA editing in mitochondria. Nucleic Acids Res. 44:4907–4919. doi: 10.1093/nar/gkw188.

Swart EC, et al. 2013. The Oxytricha trifallax macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. PLoS Biol. 11:e1001473.

Valach M, Moreira S, Kiethega GN, Burger G. 2014. Trans-splicing and RNA editing of LSU rRNA in Diplonema mitochondria. Nucleic Acids Res. 42:2660–2672.

Vlcek C, Marande W, Teijeiro S, Lukeš J, Burger G. 2011. Systematically fragmented genes in a multipartite mitochondrial genome. Nucleic Acids Res. 39:979–988.

Yabuki A, Tame A. 2015. Phylogeny and reclassification of Hemistasia phaeocysticola (Scherfelf) Elbrächter & Schnepf, 1996. J Eukaryot Microbiol. 62:426–429.

**Associate editor:** Sandra Baldauf