13764–13777 *Nucleic Acids Research, 2014, Vol. 42, No. 22* *Published online 27 November 2014*
doi: 10.1093/nar/gku1266

Widespread occurrence of organelle genome-encoded 5S rRNAs including permuted molecules

Matus Valach[1,*], Gertraud Burger[1], Michael W. Gray[2] and B. Franz Lang[1,*]

[1]Department of Biochemistry and Robert-Cedergren Centre of Bioinformatics and Genomics, Université de Montréal, Montréal, QC, H3C 3J7, Canada and [2]Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS, B3H 4B2, Canada

Received September 12, 2014; Revised November 14, 2014; Accepted November 17, 2014

ABSTRACT

5S Ribosomal RNA (5S rRNA) is a universal component of ribosomes, and the corresponding gene is easily identified in archaeal, bacterial and nuclear genome sequences. However, organelle gene homologs (*rrn5*) appear to be absent from most mitochondrial and several chloroplast genomes. Here, we re-examine the distribution of organelle *rrn5* by building mitochondrion- and plastid-specific covariance models (CMs) with which we screened organelle genome sequences. We not only recover all organelle *rrn5* genes annotated in GenBank records, but also identify more than 50 previously unrecognized homologs in mitochondrial genomes of various stramenopiles, red algae, cryptomonads, malawimonads and apusozoans, and surprisingly, in the apicoplast (highly derived plastid) genomes of the coccidian pathogens *Toxoplasma gondii* and *Eimeria tenella*. Comparative modeling of RNA secondary structure reveals that mitochondrial 5S rRNAs from brown algae adopt a permuted triskelion shape that has not been seen elsewhere. Expression of the newly predicted *rrn5* genes is confirmed experimentally in 10 instances, based on our own and published RNA-Seq data. This study establishes that particularly mitochondrial 5S rRNA has a much broader taxonomic distribution and a much larger structural variability than previously thought. The newly developed CMs will be made available via the Rfam database and the MFannot organelle genome annotator.

INTRODUCTION

5S Ribosomal RNA (5S rRNA), a universal component of 50S prokaryotic (bacterial and archaeal) and 60S eukaryotic cytosol ribosomes, is highly conserved in sequence and secondary structure, comprising a triskelion-like ('three-legged') configuration. Accordingly, 5S rRNA genes are readily recognized in the genome sequences of these organismal groups. 3D reconstructions based on X-ray crystallography show that in the archaeon *Haloarcula marismortui*, 5S rRNA constitutes one of the main structural components of the central protuberance (CP) of the 50S large subunit (LSU) (1). Within the CP, protein-mediated interactions between 5S and 23S rRNA are particularly frequent, stabilizing contacts with domains II and V of the 23S rRNA.

Mitochondria and chloroplasts harbor translation systems that share (sometimes remote) similarities with their symbiotic, evolutionary bacterial antecedents (α-Proteobacteria and Cyanobacteria, respectively). Few organelle ribosomes have been isolated and directly examined for the presence of 5S rRNA, but the respective genes (pt-*rrn5*) are identified readily in most chloroplast genomes, based on sequence similarity alone. Exceptions are in the highly derived plastids of Alveolata, notably coccidian apicomplexans and many dinoflagellates (2) (Figure 1). Mitochondria, on the other hand, present a quite different picture. A clearly recognizable 5S rRNA, encoded by the mitochondrial genome and distinct from that specified by the nuclear and chloroplast genomes, has so far been identified only in ribosomes of angiosperms (Streptophyta) and in select protist (protozoan) groups (3), implying the presence of 5S rRNA-containing mitochondrial ribosomes in these cases, too. However, in other major eukaryotic lineages [i.e. animals and fungi (Opisthokonta), ciliates and apicomplexans (Alveolata) and kinetoplastids (Euglenozoa)], where complete mitochondrial genome sequences are available, no evidence has been found to date of a mtDNA-encoded 5S rRNA (mt-*rrn5*) (Figure 1). Moreover, where mitochondrial ribosomes have been isolated from such lineages and directly analyzed, a 5S rRNA species has not been identified (4–8). Indeed, a 3D cryo-electron microscopic (EM) map of the mammalian mitochondrial (55S) ribosome has revealed that in an expanded LSU CP, the stabilizing interactions normally provided by 5S rRNA have largely been assumed by proteins (9). In addition, much of the CP mass

[*]To whom correspondence should be addressed. Tel: +1 514 343 6111 (Ext 5172); Fax: +1 514 343 2210; Email: matus.a.valach@gmail.com
Correspondence may also be addressed to B. Franz Lang. Tel: +1 514 343 5842; Fax: +1 514 343 2210; Email: franz.lang@umontreal.ca

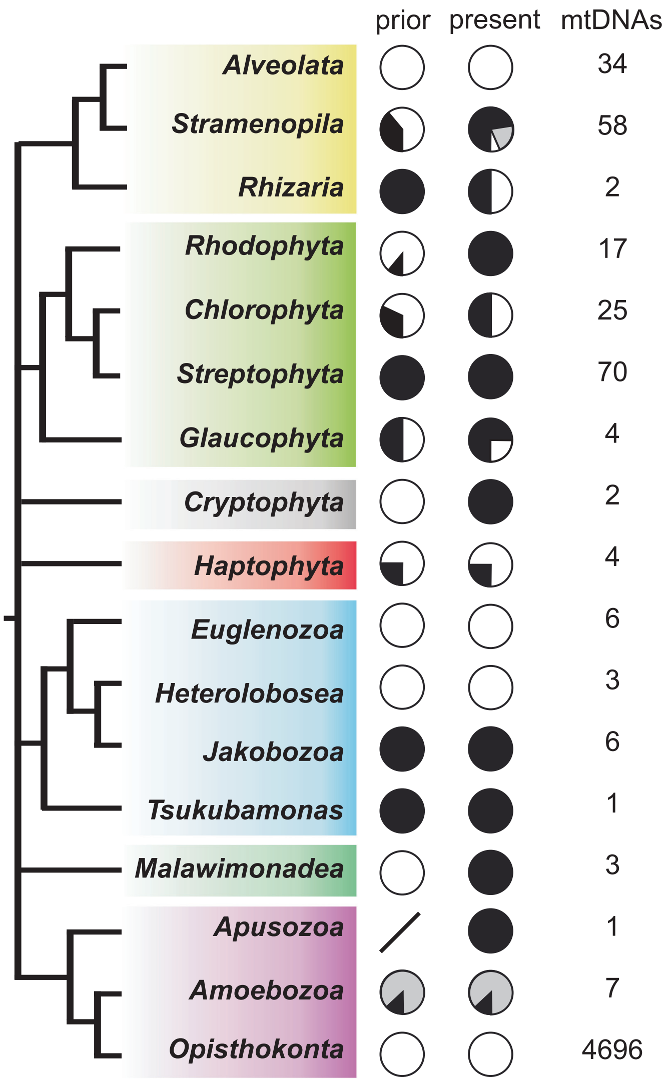

Figure 1. Taxonomic distribution of mtDNA-encoded 5S rRNA across eukaryotes. The schematic tree is based on our current understanding of phylogenetic relationships [for example, see (50)]. The column 'mtDNAs' specifies the number of complete or almost complete mtDNAs that were searched with the newly developed CM models. In the present study, we added one complete mtDNA sequence to Rhizaria, Streptophyta, Heterolobosea and Apusozoa, as well as two to Malawimonadea. The percentage of mtDNAs that contain an *rrn5* gene is depicted as a pie chart (filled black, *rrn5*; gray, *rrn5*-like). Specific information on the presence/absence of mtDNA-encoded *rrn5* is as follows: Stramenopila, absent in *Cafeteria roenbergensis* and *Proteromonas lacertae*, *rrn5*-like genes in 11 species; Rhizaria, present in *Bigelowiella natans*; Chlorophyta, present in 12 out of 25 species; Haptophyta, present in *Pavlova lutheri*; Apusozoa, present in *Thecamonas trahens*; Amoebozoa, present in *Acanthamoeba castellanii*, *rrn5*-like genes in all other amoebozoans. For a detailed listing of species and corresponding search results, see Supplementary Table S1.

of the yeast mitoribosome LSU is composed of the LSU mt-rRNA expansion segments, which (together with additional proteins) spatially replace 5S rRNA without adopting a 5S-like structure (7). Recent high-resolution cryo-EM structures of the mammalian mitoribosome LSU (39S) have shown hints of a short RNA at a position similar to that of a conventional 5S rRNA (5), which most likely corresponds to a stably incorporated tRNA (6,8). In contrast, several previous reports indicated that mammalian mitochondria take up the nucleus-encoded, cytosolic 5S rRNA (10,11) and incorporate it into the mitoribosome through interaction with the protein L18 (12). Thus, even in some model organisms, the evolutionary fate of *rrn5* genes that are considered to have been lost from organelle genomes, and structural consequences for the mitochondrial ribosome in these cases, are not entirely clear.

Although replacement of 5S rRNA by protein in some mitochondrial (and perhaps plastid) systems can account for the absence of an organelle 5S rRNA, it is equally plausible that current search regimes simply fail to identify highly derived *rrn5* genes in certain other organelle genomes. This latter possibility most probably applies to species, whose organelle-encoded LSU rRNA retains a high degree of 'typical' sequence and secondary structure and lacks yeast-like expansion segments. An example in this regard is the protist *Acanthamoeba castellanii*, where a mtDNA-encoded 5S rRNA was missed during the initial annotation of the complete mitochondrial genome sequence (13), and only later recognized through biochemical characterization (3).

For predicting non-coding RNA genes, computational searches that make use of secondary structure information are more accurate than methods based solely on sequence conservation. Most powerful are covariance models (CMs), i.e. profiles that define a specific structural RNA family by its sequence features plus the covariance of base-paired residues (14,15). The Rfam database (16) is a comprehensive repository of CMs for more than 2000 RNA families. The current Rfam CM for the 5S rRNA family (RF00001 version 11.0) was derived from bacterial, archaeal and eukaryotic nuclear counterparts. Since this model finds only a minor fraction of annotated organelle *rrn5* genes, we set out to develop dedicated mitochondrion- and plastid-specific 5S rRNA CMs. As we show here, the new models are highly sensitive and specific, detecting a number of previously unrecognized *rrn5* genes in published organelle genome sequences. The most intriguing findings are a class of 5S rRNAs with permuted secondary structure encoded in brown algal mitochondria, and the identification of two apicoplast 5S rRNAs.

MATERIALS AND METHODS

Development of CMs

Annotated mitochondrial and plastid *rrn5* genes (alternatively designated *rrf* in plastid genomes) were retrieved from GenBank (complete mitochondrial and plastid genome sections: http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=2759&opt=organelle and http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=2759&opt=plastid, version 22 July 2014). The *Chondrus crispus* mitochondrial *rrn5* (NCBI Gene ID 7020988) was removed from the downloaded sequences, because the authors' gene assignment has been disputed (17). Similarly, the *Bryopsis hypnoides* plastid *rrn5* (Gene ID 8463250) was removed due to the evidently incorrect gene annotation [both Basic Local Alignment Search Tool (BLAST) and CM searches identify a different locus in the plastid genome as *bona fide rrn5*]. Mitochondrial and plastid gene sequences were aligned separately with MUSCLE v3.6 (18) and incorporated into the Genetic Data Environment (GDE) sequence editor (19). Multiple alignments were then inspected by eye and manually adjusted in a few regions to improve primary sequence plus secondary structure fit, the latter assisted by minimum energy secondary structure predictions with RNAalifold (20). The verified annotated sequences include 108 mtDNA-encoded and 500 ptDNA-encoded *rrn5* genes (Supplementary Table S1; marked by '+' in the 'Annotation' column). These data sets, referred to as the mt-gene test set and the pt-gene test set, were used for developing and testing CMs. For building the models, sequence alignments of test set *rrn5* sequences served as input for the Cmbuild and Cmcalibrate programs of Infernal v1.1, after masking columns that are not reliably aligned (15). The Cmbuild option '- -hand' ensures that only the confidently aligned sequence positions are used for building mitochondrion- and plastid-specific CMs (referred to as mt-5S and pt-5S models, respectively). Use of the tree weighting option '--wgsc' (21) increases the chance of detecting sequences in an organismal group that is less well represented in the seed alignment. With these two basic CMs, we searched for *rrn5* genes in individual organelle genome sequences by employing Cmsearch with default settings, i.e. local alignment, an inclusion (significance) *E*-value threshold of 10^{-02} and a reporting *E*-value threshold of 10.

Organelle *rrn5* sequences discovered and validated in the course of our analyses (see below) were included in an additional CM (mtAT-5S) based on a wide taxonomic sampling and a focus on derived and A+T-rich 5S rRNAs that are less effectively identified with the basic mt-5S model. A fourth model has been developed (mtPerm-5S) based on the permuted 5S rRNAs encoded by mtDNAs from brown algae and potentially several other stramenopiles. All models will be made available (together with the seed sequence alignments) in the Rfam database. They will be also included in our automated organelle genome annotation tool MFannot (http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl).

Identification of additional 5S rRNA candidates and secondary structure analyses

The above-described models were used to identify previously unknown 5S rRNA genes, in complete organelle genome sequences retrieved from the NCBI Organelle Genome Resources website (as of 22 July 2014). Homologs of newly detected, divergent *rrn5* candidates were further sought by BLAST searches (22) in GenBank *nr*, in the database of the 1000-plants genome initiative (http://onekp.com), and the Joint Genome Institute's genome portal (23). Multiple sequence alignments including candidates were then analyzed with RNAalifold to estimate the plausibility of folding into a 5S rRNA (triskelion) structure. Thermodynamic folding of single sequences was predicted with RNAfold 2.0 (24), sequence alignments were visualized with either GDE (19) or R-CHIE (25) and secondary structures were drawn with R2R v1.0.3 (26).

Evaluation of Cmsearch performance

We compared the performance of the three models, Rfam model RF00001 (v11.0; referred to as RF-5S), mt-5S and pt-5S in their ability to detect *rrn5* in our organelle genome test sets. The mitochondrial genome (mt-genome) test set is composed of all complete mtDNA sequences with (reliably) annotated *rrn5* (i.e. the genomes corresponding to the mt-gene test set), plus all complete mtDNAs from animals and fungi (which in our view all lack *rrn5*). The plastid genome (pt-genome) test set is composed of all complete ptDNA sequences with annotated *rrn5*. A true positive hit is considered one that matches at least the most highly conserved β and γ domains of the annotated *rrn5* gene (Figure 2A). Deviations in the start and end positions of the gene are allowed. For evaluating a model's performance in finding, for example, mt-*rrn5*, the true positive rate (TPR) is calculated by the formula:

$$TPR = No.\ of\ found\ mt\text{-}genes/$$
$$(No.\ of\ all\ mt\text{-}genes\ in\ the\ mt\text{-}genome\ test\ set);$$

and the false discovery rate (FDR) by:

$$FDR = No.\ of\ wrongly\ assigned\ mt\text{-}genes/$$
$$(No.\ of\ all\ mt\text{-}genes\ in\ the\ mt\text{-}genome\ test\ set).$$

Strains and cultures

Andalucia godoyi PRA-185, *Malawimonas jakobiformis* (ATCC 50310), *M. californiana* (ATCC 50740), *Malawimonas* sp. (kindly provided by Alastair Simpson, Dalhousie University, Halifax, Canada), *Klebsormidium flaccidum* (UTEX 321), *Thecamonas trahens* (ATCC 50062) and *Paracercomonas marina* (ATCC 50344) were grown in liquid WCL medium, *Jakoba bahamiensis* (ATCC 50695) in liquid F/2 medium, and *Stachyamoeba lipophora* (ATCC 50324) in liquid PAS medium. Media composition can be found at http://megasun.bch.umontreal.ca/People/lang/FMGP/methods.html. Cultures were supplemented with live *Enterobacter* essentially as described (27), except for *Klebsormidium flaccidum*, which grows on synthetic media.

Nucleic acid purification, construction of libraries and sequencing

For sequencing of the mitochondrial genomes reported here, mtDNAs were isolated via CsCl-bisbenzimide equilibrium gradient centrifugation, sequenced by the Sanger technology, assembled with Phred/Phrap (28,29) and annotated with the MFannot tool (http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl), essentially as described earlier (27,30–31). Complete mtDNA sequences have been deposited in GenBank under the accession numbers KP165385-KP165391. For transcriptome sequencing, total RNA including small RNAs was extracted from cells using the RNeasy Plus Universal Kit (Qiagen). RNA-Seq libraries were constructed using the TruSeq Small RNA Sample Prep kit (Illumina) following the supplier's instructions, except that total RNA was not size-fractionated. Both the library preparation and the paired-end Illumina sequencing were outsourced to the technology platform of the Genome

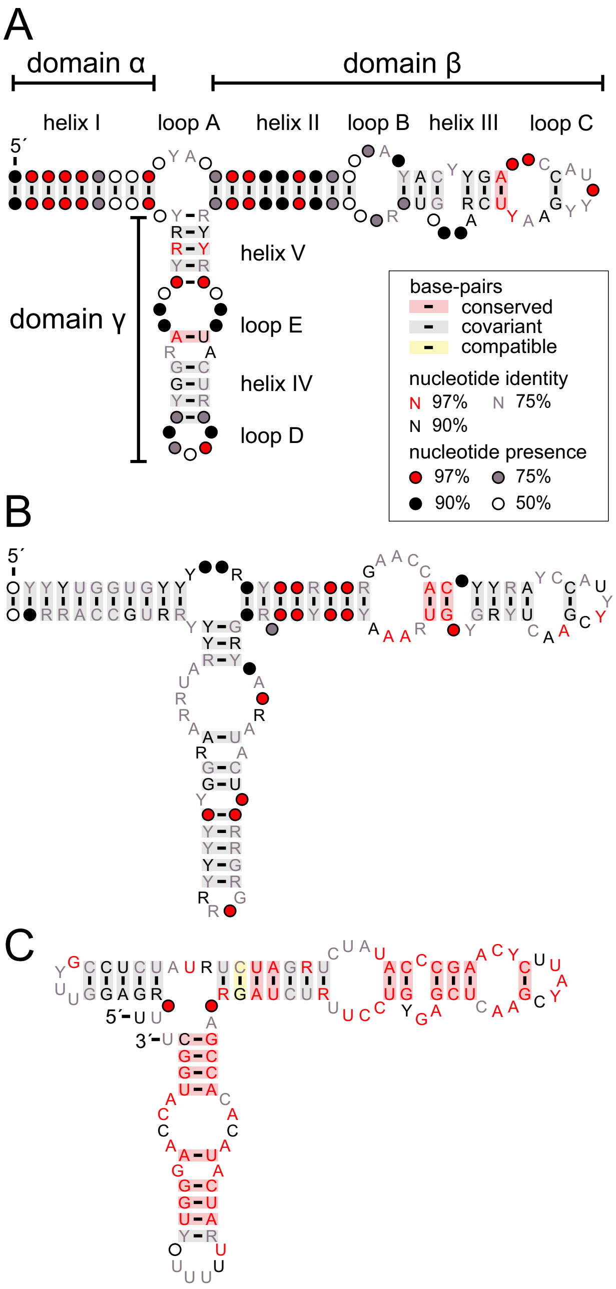

Figure 2. Consensus secondary structure models of organelle 5S rRNA. Sequences were weighted using the GSC algorithm (21). Nucleotides in IUB code are conserved. Circles indicate positions with variable nucleotide identity (below 75% conservation). Conserved, covariant or one-sided compatible substitutions in canonical (Watson–Crick) base-pairs are shaded. See the inset box for details. (**A**) mt-5S rRNA (based on 94 distinct sequences, i.e. identical sequences were excluded). Domains, helices and loops are annotated according to (1). Brown algal mitochondrial sequences were omitted from the mitochondrial consensus (see Results for details). (**B**) pt-5S rRNA (based on 189 distinct sequences). Where a plastid genome contained several non-identical *rrn5* genes, only the highest-scoring one has been included. (**C**) Permuted mt-5S rRNAs from brown algae (based on 23 distinct sequences). Used were not only *rrn5* loci from complete mitochondrial genome sequences available in GenBank (Supplementary Table S1), but also from 10 partial mitochondrial sequences obtained from the 1000-plant genome database (*Laminaria japonica*, *Petalonia fascia*, *Punctaria latifolia*, *Sargassum hemiphyllum*, *Sa. henslowianum*, *Sa. integerrimum*, *Sa. thunbergii*, *Sa. vachellianum*, *Scytosiphon dotyi* and *Sc. lomentaria*; http://onekp.com).

Quebec Innovation Center in Montreal, generating 18–39 million Illumina HiSeq2500 (up to 150 nt read length for *A. godoyi*, *M. jakobiformis* and *M. californiana*) and 6.8 million Illumina MiSeq paired-end reads (up to 250 nt read length for *J. bahamiensis*). The applied procedure yields RNA-Seq data that allow accurate mapping of 5S rRNA termini and precise determination of RNA steady-state levels.

Transcriptome data sets, read processing and gene expression analyses

Transcriptome data for *Pyropia yezoensis* (32), *Phaeodactylum tricornutum* (33), *Thalassiosira pseudonana* (34), *Ectocarpus siliculosus* (35) and *Toxoplasma gondii* (36) were downloaded from the NCBI Sequence Read Archive v1.0 at http://www.ncbi.nlm.nih.gov/Traces/sra. According to the available information, these libraries were made from size-fractionated small or micro-RNAs, with a (single) read length of ∼35–50 nt (for SRA accession numbers, see Supplementary Table S1). Small RNA-Seq data from *Phytophthora sojae* were kindly provided by Mark Gijzen (Agriculture and Agri-Food Canada) (37). Note that RNA-Seq data from standard mRNA or total RNA libraries are not useful for 5S rRNA identification, because they are strongly biased against transcripts that are not poly-adenylated, or because short (<200 nt) transcripts are often eliminated during library construction. Reads were processed with Cutadapt v1.2.1 (38) to remove adapter sequences and to trim low quality positions (quality base 20, error rate 0.1 and minimum length 18). Reads were then mapped onto the organelle genome with bowtie2 (39) using the option end-to-end and otherwise default settings. Mapping was visualized with Geneious R7 (Biomatters, New Zealand). RNA expression levels for individual genes were calculated as RPKM with Artemis v16.0.0 (40,41) based on BAM alignments of mapped reads. 5S rRNA expression was compared with that of tRNAs, small subunit (SSU) and LSU rRNAs, and to intergenic regions as a baseline. Termini were mapped at single nucleotide resolution by parsing the bowtie2-generated SAM file. To verify co-linearity of transcript and gene sequence, bowtie2 mapping was performed in local mode. Soft-clipped portions of reads consisted mostly of adapter fragments that were not removed in the clipping step; there was no indication for sequence rearrangement of transcripts compared to the corresponding genes. For *Toxoplasma gondii*, we also used the read coverage tracks of small ncRNA-seq for the strain ME49 (tachyozoites) displayed on the ToxoDB 9.0 website (http://toxodb.org/toxo) (42) by the genome browser v2.48. The assembly scaffold tgme49_asmbl.1944 represents the apicoplast genome.

RESULTS

The universal Rfam CM misses many annotated mt-*rrn5* genes

The Rfam model RF00001 (v11.0; referred to in the following as RF-5S) was evaluated for its performance in detecting organelle *rrn5* when using Cmsearch with default parameters. In complete mitochondrial genomes, this model finds only ∼70% of the known genes in the mt-gene test set

Table 1. Performance of *rrn5* CMs on organelle genome test data sets[a]

Model	RF-5S		mt-5S	pt-5S
Test dataset	mt-genomes	pt-genomes	mt-genomes	pt-genomes
True positives	76	497	108	500
False positives	3	0	4	1
False negatives	32	3	0	0
TPR	70.4%	99.4%	100%	100%
FDR	2.8%	0%	3.7%	0.2%

[a] Hits within the inclusion threshold as reported by Cmsearch used with default settings. The mt-genome test data set includes 4804 complete mitochondrial genomes (108 sequences with annotated mt-*rrn5* and 4696 metazoan and fungal sequences lacking mt-*rrn5*) and the pt-genome test data set includes 500 complete plastid genomes with annotated pt-*rrn5*. Note that the number of false positives reported by mt-5S and pt-5S drops to zero when filtering out hits with scores below 30 (see Figure 3).

(see Materials and Methods), with *E*-values between 10^{-9} and 10^{-2} and scores from 43 to 18 (Table 1). In some instances, the RF-5S model recognizes exclusively the highly conserved portions of domain β (helix III and loop C; for nomenclature, see Figure 2A). The identified genes include the most bacteria-like members in jakobids, green algae and plants. The three lowest-scoring hits are false positives (e.g. in mtDNA of the animal *Ursus thibetanus*), based on the following reasoning. The corresponding sequences do not align continuously with mt-*rrn5* but introduce indels longer than 20 nt; they often overlap neighboring structural RNA genes (e.g. LSU rRNA, SSU rRNA, tRNAs); and their secondary structures lack helices or loops that are typical for 5S rRNA. Supplementary Table S1 compiles *E*-values and scores for the various search models, as well as taxonomic information and GenBank accession numbers of genome records.

In complete plastid genomes, RF-5S identifies 99% of the 500 known genes in the pt-gene test (see Materials and Methods), with *E*-values between 10^{-18} and 10^{-3} and scores between 82 and 22 (Table 1). However, annotated genes from *Euglena longa* (euglenid, Euglenozoa), as well as *Gnetum parvifolium* and *Ephedra equisetina* (gnetophytes, Streptophyta), remain undetected. Nevertheless, no false positives are reported.

Taken together, the RF-5S model performs well on pt-*rrn5*, but poorly on mt-*rrn5*. Although α-proteobacterial genes were included in building this model (16), their sequences are apparently too distant from the mitochondrial counterparts to allow effective mt-*rrn5* identification.

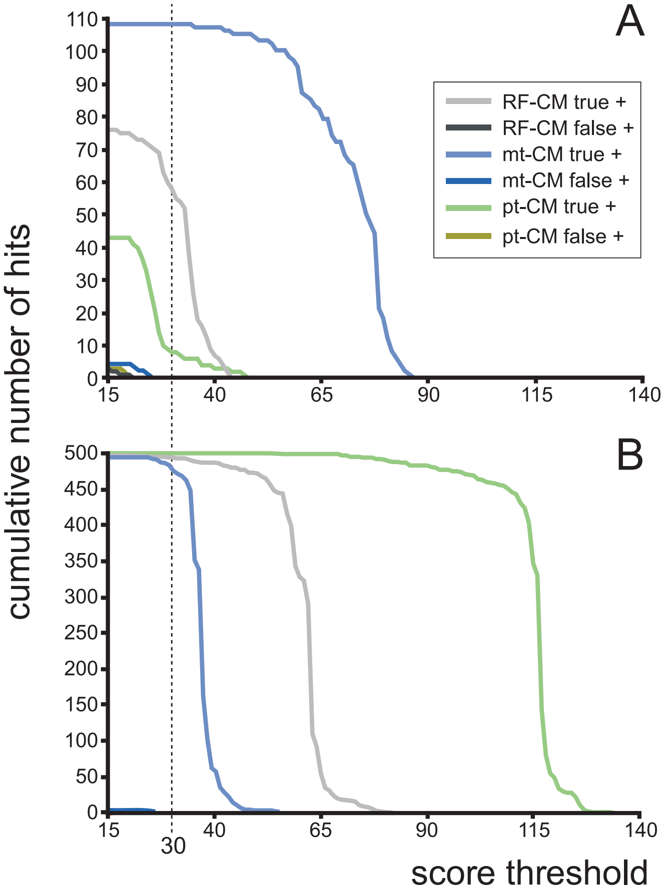

Figure 3. Performance of *rrn5* CMs on organelle genome test data sets. Horizontal axis, the score threshold of hits reported by Cmsearch (default settings). Vertical axis, the number of hits with a score above the threshold. Gray, blue and green continuous lines, hits obtained with the CMs RF-5S, mt-5S and pt-5S, respectively. Light line shading, true positive hits; dark line shading, false positive hits. Dotted line, the score threshold considered biologically meaningful. (**A**) Mitochondrial genome test data set. (**B**) Plastid genome test data set (see Materials and Methods for details).

The mt-5S and pt-5S models detect all known organelle *rrn5* genes and distinguish mt-*rrn5* from pt-*rrn5*

Employing the Infernal package (15), we have built mitochondrion-specific (mt-5S) and plastid-specific (pt-5S) search models. Figure 2A and B show the common secondary structure, conservation and covariance of loci detected by the mt-5S and pt-5S models in the mitochondrial and plastid genomes, respectively. In screens of complete mtDNA sequences with mt-5S, the model finds all members of the mt-gene test set (Table 1) with an *E*-value range from 10^{-17} to 10^{-5} and corresponding scores between 85 and 35 (Supplementary Table S1). In most instances, the positions of the predicted loci precisely match those of GenBank annotations or slightly deviate (by at most 5 nt); exceptionally large discrepancies (up to 25 nt) occur in brown algal mtDNAs as discussed in a separate section below. Notably, the model also recognizes the biochemically characterized yet highly divergent 5S rRNA in *Acanthamoeba castellanii* (3), which has been notoriously refractory to computational detection, including searches with RF-5S CM. The four hits in animal and fungal mtDNAs with scores of 24 and below are regarded as false positives because canonical elements of mt-*rrn5* are conspicuously lacking in the folded sequences.

When screening complete ptDNAs with pt-5S, the model detects all members of the pt-gene test set (Table 1), with *E*-values ranging from 10^{-32} to 10^{-11} and scores between 134 and 54 (Supplementary Table S1). A single false positive was retrieved with a score of 23.4, which is substantially below those of the test set. Based on the results above, hits with scores >30 obtained with the mt-5S and pt-5S models can be considered reliable (Figure 3). Note, however, that hits with lower values may be correct as well, but need to be validated by additional structural, comparative and/or biochemical information (see below).

To determine across-genome specificity, we tested the mt-5S model on ptDNAs and the pt-5S model on mtDNAs. The mt-5S model reports only 11% pt-*rrn5* genes with a score above 40, and for all these hits, the score is significantly lower than that assigned by the pt-5S model (Figure 3). The pt-5S model retrieves nine mitochondrial hits, all in plant mtDNAs (Supplementary Table S1), which are also retrieved by the mt-5S model. In these cases, however, the scores with pt-5S ($E < 10^{-20}$, score >90) are much higher than those with mt-5S. These mitochondrial loci most likely originate from ptDNA, because their sequences are 96–100% identical to pt-*rrn5* genes of the corresponding species (Supplementary Figure S1). Inter-organelle gene transfer is frequent in plants (43,44).

In sum, the two organelle CMs exhibit excellent genome specificity, readily discriminating between genuine and transferred plant *rrn5* genes, and probably also between mt-*rrn5* and pt-*rrn5* in data sets for which the source of the sequence is unknown (e.g. metagenomic or total DNA data sets).

Mitochondrion- and plastid-specific CMs reveal numerous, previously unrecognized *rrn5* genes

With the mt-5S model, we scanned a taxonomically broad collection of mtDNAs including some genome sequences that are partial, plus new sequences generated by us from six protist species. These latter are *Klebsormidium flaccidum*, *Malawimonas californiana*, *Malawimonas* sp., *Paracercomonas marina* (ATCC 50344, earlier misidentified as *Cercomonas longicauda* (45)), *Stachyamoeba lipophora* and *Thecamonas trahens*.

In addition to the known mt-*rrn5* genes, the model returned 40 new hits across most eukaryotic groups, now extended to jakobids, malawimonads, plants, green algae, red algae, glaucophytes, stramenopiles (brown algae, diatoms, raphidophytes, eustigmatophytes, pelagophytes), cryptophytes, haptophytes, apusozoans and amoebozoans (Figure 1). Out of these 40 new genes, only 11 (the most bacteria-like) are also retrieved by the RF-5S model, generally with very low scores (Supplementary Table S1).

Several lines of evidence support the authenticity of the new hits. First, all newly detected genes fall in unassigned, intergenic genome regions; only a few overlap minimally (2–6 nt) the upstream neighboring gene [e.g. in mtDNAs of the diatom *Ulnaria* (*Synedra*) *acus* and the eustigmatophyte *Nannochloropsis oceanica*, and in ptDNA of the rhodophyte *Cyanidioschyzon merolae*]. Such short overlaps are fairly common in organelle genomes (46,47). Second, newly detected genes are located on the same strand as neighboring genes, with the exception of mt-*rrn5* from the green algae *Bathycoccus prasinos* and *Helicosporidium* sp., which are encoded on the opposite strand within a 10-kbp-long, densely packed coding region, and from the land plant *Asclepias syriaca*, located within a 10-kbp-long non-coding region that is delimited by other ribosomal RNA genes. Third, mt-*rrn5* genes tend to reside adjacent to other rRNA-specifying genes, forming rRNA operon-like arrangements that probably allow balanced co-transcription. For instance, the newly detected mt-*rrn5* in the cryptophyte *Rhodomonas salina* is flanked by *rns* and *rnl* (encoding the SSU and LSU rRNAs), all located on the same DNA strand. Further, in the excavates *Reclinomonas* and *Tsukubamonas*, and in *Malawimonas californiana*, mt-*rrn5* is located directly upstream of *rns* (for further examples, see Supplementary Table S1).

The pt-5S model finds 12 previously unrecognized *rrn5* genes in plastid genomes. Of those 12, the RF-5S model detects eight. Interestingly, pt-5S (but not RF-5S) revealed *rrn5* in the apicoplast (plastid-derived organelle) genome of *Toxoplasma gondii* and *Eimeria tenella*. Further, in both cases, the corresponding *E*-values ($<10^{-8}$) and scores (>40) are slightly below the lowest values of the pt-*rrn5* test set (10^{-11} and 54), but well above those that we consider to be false positives ($>10^{-3}$ and <25). In 13 ptDNAs from different eukaryotic groups, we find matches in addition to the annotated *rrn5*, mostly located in repeat regions (Supplementary Table S1). For example, the second locus in *Trebouxiophyceae* sp. MX-AZ01 shares 78% sequence identity with the annotated *rrn5* and the predicted secondary structure deviates from that of the previously annotated *bona fide* 'master' gene. Generally, supernumerary copies of pt-*rrn5* are thermodynamically less stable than their authentic counterpart and likely represent pseudo-genes.

Mitochondrial 5S rRNA in most brown algae adopts a permuted secondary structure

In brown algal mitochondrial genomes, the mt-5S model predicts *rrn5* genes with high scores. However, gene termini differ by 6–25 nt from published annotations. The region common to both our predictions and the original annotations corresponds to the highly conserved β and γ domains of 5S rRNA (i.e. helices II-III and IV-V; Figure 2A). However, close inspection reveals that both alternatives are questionable. Most published annotations overlap the downstream tRNA gene considerably, and the inferred secondary structures have either extra A+U-rich hairpins or a long insertion in the 3′ moiety of loop A. Conversely, the mt-5S-assigned termini overlap the upstream *rns* gene, and the deduced secondary structure has an extended 5′ moiety of loop A and an atypical closing helix I.

To investigate this conundrum in more detail, we collected additional syntenic *rns*-*rrn5*-*trnM* gene regions from partial brown algal mtDNAs, which extended our data set to 23 distinct mt-*rrn5* sequences from phaeophytes. Comparative analysis indicates (except in *Dictyota dichotoma*, see below) a shared, yet unconventional, thermodynamically highly stable hairpin upstream of helix II (Figure 2C). This hairpin does not overlap the upstream *rns* sequence, and the resulting overall secondary structure adopts the conventional triskelion shape of 5S rRNA. Therefore, we posit that the hairpin is an integral part of phaeophyte mt-5S rRNA, replacing the conventional helix I, which otherwise brings together the molecule's 5′ and 3′ termini. In this configuration, the ends of phaeophyte mt-5S rRNAs have shifted positions and are located at the intersection of domains α and γ instead of the distal portion of helix I, an arrangement we refer to as 'permuted'. As documented below, this permuted arrangement is corroborated by transcriptome data from *Ectocarpus*. A search model based on these permuted RNA structures (mtPerm-5S) identifies all brown algal *rrn5* genes with higher scores than with the non-permuted mt-5S model (Supplementary Table S1).

Dictyota, a member of the most deeply branching phaeophyte clade (48), is the only brown alga with low support for the permuted mt-*rrn5* structure (with only three base-pairs in the hairpin), while the conventional shape practically lacks helix I (Supplementary Figure S2O and P). Whether this structure represents the ancestral state of brown algal mt-5S rRNA is unclear. Deeper sampling at the base of the brown algal phylogenetic tree will help to retrace the potential transition from a conventional ancestral to the permuted domain arrangement. To solve this puzzle, transcript data will be required.

Highly divergent mt-5S rRNAs in additional stramenopile lineages

After finding unconventional mt-5S rRNA in phaeophytes, we scrutinized results from searches with the mt-5S model for hits below the default inclusion threshold. Candidate *rrn5* genes were examined if they had a conserved β or γ domain, internal indels of up to 10 nt, and overlaps with neighboring genes for up to 25% of their length. These criteria selected several candidates in mtDNAs from stramenopile species belonging to oomycetes, *Blastocystis*, diatoms, chrysophytes, synurids and labyrinthulomycetes. Table 2 compiles the domain divergences and helix I configurations of these loci.

The highest-scoring below-threshold hit occurs in mtDNAs of the *Phytophthora* genus. A BLAST search with this gene in the NCBI *nr* database retrieved 14 additional *rrn5* candidates in complete and partial oomycete mtDNAs. Comparative analyses show that the sequence of this locus is well conserved across oomycetes, maps to previously unassigned genomic regions and can be folded into a typical 5S rRNA secondary structure as detailed below. These loci have the highest A+T content (86%) among all *bona fide rrn5* genes.

To accommodate the sequence bias in these newly detected 5S rRNAs, we built an additional mitochondrial CM (mtAT-5S) based on divergent, A+T-rich *rrn5* sequences (including those of oomycetes, raphidophytes, glaucophytes, rhodophytes, cryptophytes and amoebozoans; for a full list see Supplementary Table S1, column 'Used for model building'). When searching in all complete mtDNA sequences, mtAT-5S detects above the threshold all loci found by the mt-5S model except two: one in *Jakoba bahamiensis* due to a U-rich γ domain; and the other in *Prasinoderma coloniale*, due to an overall G+C-rich sequence. Derived (non-permuted) mt-*rrn5* obtain much higher scores with mtAT-5S (scores and *E*-values up to 88 and 10^{-14}, respectively) than with mt-5S (Supplementary Table S1). For example, a potential locus in one of the *Blastocystis* mtDNAs was reported by the mt-5S model as a below-threshold hit, but is well within the inclusion values with mtAT-5S (score 42, *E*-value 10^{-6}; see Table 2 for details). Still, for less divergent mt-*rrn5*, the scores are higher with mt-5S than with mtAT-5S, and the latter model misses certain loci readily detected by the former.

Secondary structure modeling of derived (non-permuted) stramenopile mt-*rrn5*

Representative secondary structure models of putative, derived mt-5S rRNAs from non-phaeophyte stramenopiles are depicted in Figure 4A and further models are shown in Supplementary Figure S2. For example, although sequence conservation is very low, the *Blastocystis* sequence can be folded into a typical mt-5S rRNA (Figure 4A), with covariant residues in all five helices of the molecule and a triskelion arrangement (Supplementary Figure S2B and D). Conversely, counterparts from oomycetes (Figure 4A) have covariance support for only two out of the five helices, but the primary sequence is more conserved than in *Blastocystis* (Supplementary Figures S2A and C and S3B). Secondary structure models of derived mt-*rrn5* from four other stramenopile lineages are shown in Supplementary Figure S2E–H.

Among the notable deviations are the size-reduced loop C and extended loop E in mt-*rrn5* candidates from the raphid pennate diatoms (Supplementary Figure S2Q–U). Interestingly, the sequences have the propensity to adopt the regular 5S rRNA shape with a short helix I and an alternative secondary structure with a ∼15-nt-long hairpin in the 5′ region, instead of a conventional (open-ended) helix I, thus forming a permuted α domain as in phaeophytes (Supplementary Figure S2Q–T). The same applies to the mt-*rrn5* candidate of the synurid *Chrysodidymus synuroideus* and the chrysophyte *Ochromonas danica*, where a 5′ permuted configuration has a higher thermodynamic stability than the conventional structure (Supplementary Figure S2K–N). In these cases, as well as in the two diatoms *Phaeodactylum* and *Thalassiosira*, the conventional folding is unusual because of a weak and/or short helix I (Supplementary Figure S2U and V). Transcriptome data are available for two diatoms and one oomycete, corroborating the expression of the proposed deviant mt-5S rRNAs.

RNA-Seq data confirm predicted *rrn5* genes and precisely map 5S rRNA termini

We generated RNA-Seq data for *Andalucia godoyi*, *Jakoba bahamiensis*, *Malawimonas jakobiformis* and *M. californiana* to verify transcription and identify mt-5S rRNA termini. In these libraries, mt-*rrn5* transcripts are abundant and evenly covered by reads, demonstrating expression of these loci. Hundreds or more reads even map across the entire 5S rRNA (Figure 5A, C, E and G; Supplementary Table S1). In *A. godoyi*, the ends match exactly the mt-5S prediction (Figure 5B), whereas in *J. bahamiensis* the ends are slightly shifted (1 nt at the 5′ and 3 nt at the 3′ end; Figure 5D). In malawimonads, both termini are more extended (in *M. californiana*, both ends by 12 nt; in *M. jakobiformis*, the 5′ and 3′ ends by 11 and 8 nt, respectively). Thus, helix I is exceptionally long in these two latter taxa (Figure 5F and H).

Similarly, transcriptome data from a red alga (*Pyropia*), two diatoms (*Phaeodactylum* and *Thalassiosira*), an

Table 2. Predicted mt-*rrn5* of stramenopiles (based on combined evidence from covariance analysis, synteny and thermodynamic stability of RNA folding)

Lineage	Number of species[a]	α domain[b]	β domain	γ domain	mt-5S score (*E*-value) range[c]	mtAT-5S score (*E*-value) range[c]	mtPerm-5S score (*E*-value) range[d]
Bacillariophyta (araphid, pennate diatoms)	1	conventional	conserved	conserved	25.8 (1.3×10^{-03})	36.7 (2.3×10^{-04})	n.a.
Bacillariophyta (raphid, pennate diatoms)	5	5′ permuted or conventional	divergent	moderately conserved	11–18.1 (3.4–5.5×10^{-2})	34.4 (1.1×10^{-03})	21.7–59.7 (2.5×10^{-02}–2.5×10^{-10})
Bacillariophyta (centric diatoms)	1	conventional	moderately conserved	divergent	not predicted	23.1 (1×10^{-1})	n.a.
Blastocystis	5	conventional	moderately conserved	divergent	11.3 (1.4)	18.7–42.9 (0.47–8.4×10^{-06})	n.a.
Chrysophyceae	1	5′ permuted or short conventional	divergent	moderately conserved	8.9 (7)	16.9 (1.5)	not predicted
Eustigmatophytes	5	short conventional	conserved	conserved	41–48.5 (4.6×10^{-07}–1×10^{-08})	45.5–63.1 (3.6×10^{-06}–1.4×10^{-09})	n.a.
Labyrinthulomycetes	1	short conventional	moderately conserved	divergent	14.6 (2.9×10^{-1})	20.9 (2×10^{-1})	n.a.
Oomycetes	14	conventional	moderately conserved	moderately conserved	26.7 (1×10^{-04})	30.3–50.1 (4.3×10^{-03}–6.4×10^{-08})	n.a.
Pelagophytes	1	5′ permuted	conserved	conserved	59 (4.5×10^{-11})	58.2 (1.2×10^{-08})	56.9 (1×10^{-09})
Phaeophytes	31	5′ permuted	conserved	conserved	49.4–64.3 (9.1×10^{-09}–2.3×10^{-12})	41.6–55.1 (3.2×10^{-05}–4.6×10^{-08})	69.1–99.4 (2.6×10^{-12}–2.1×10^{-18})
Raphidophytes	2	extended conventional	conserved	conserved	62.9–69.1 (5.7×10^{-12}–2.8×10^{-13})	78.2–88.3 (1.4×10^{-12}–1.7×10^{-14})	n.a.
Synurophyceae	1	5′ permuted	divergent	moderately conserved	20.3 (1.7×10^{-2})	23.9 (5.5×10^{-2})	63.1 (4.8×10^{-11})

[a] Species are listed in Supplementary Table S1. (For phaeophytes, see also legend to Figure 2C.)

[b] α domain arrangement. For detailed secondary structure models, see Supplementary Figure S2.

[c] If only a single score and *E*-value are shown for a lineage with multiple representatives, the model detected mt-*rrn5* in a single species.

[d] mtPerm-5S model has been optimized to detect a 5′ permuted α domain, but it also recognizes conserved β and γ domains often leading to detection of mt-*rrn5* genes with a conventional α domain. n.a., not applicable.

oomycete (*Phytophthora sojae*), a brown alga (*Ectocarpus*) and a coccidian (*Toxoplasma*) confirm that organelle 5S rRNAs predicted by the mt-5S or pt-5S models are expressed (Supplementary Table S1). However, in these libraries, read coverage is highly biased due to the library construction procedure (selection of very short RNAs, targeting miRNAs), with excessive read enrichment at either the 5′ or 3′ ends of 5S rRNA (Figure 6). Still, the analysis of overrepresented read starts and ends corroborates the predicted organelle 5S rRNA termini in the red alga, the oomycete and the coccidian (Supplementary Figures S3A and B and S4E). For *Ectocarpus*, analysis of read termini substantiates convincingly the 3′-end of the predicted permuted secondary structure, while support for the 5′ terminus is weak (among the four reads that cover the 5′ region, two coincide with the predicted terminus) (Supplementary Figure S3C). There is no evidence for splicing or RNA-level rearrangements that reverts the transcript to a conventional structure.

DISCUSSION

Specialized CMs considerably improve detection of organelle *rrn5* genes

In contrast to prokaryotic and nuclear 5S rRNA genes, those encoded by organelle genomes can be difficult to recognize due to sequence divergence, compositional bias and/or structural deviation. This constraint applies particularly to mt-*rrn5*. Accelerated sequence evolution in certain mtDNAs, in conjunction with the older evolutionary age of mitochondria (compared to plastids), exacerbates sequence and secondary structure deviations to a degree that renders a large number of genes unrecognizable by the universal RF-5S model.

The organelle-specific CMs presented here have a high true positive rate and low false discovery rate, significantly outperforming RF-5S in detecting *rrn5* in mitochondrial and plastid genomes. In particular, mt-5S not only reports 25% more true positives in the test set than RF-5S does, but it also revealed 40 previously unrecognized mt-*rrn5* genes (Supplementary Table S1). Note, however, that mt-5S does not recognize the expressed loci referred to as ‘5S-like RNA’ genes from six amoebozoan mitochondrial genomes. These genes have virtually none of the conserved sequence positions that otherwise characterize mt-5S rRNA, although they share a common shape with mt-*rrn5* from *Acanthamoeba* (49). The latter is the only amoebozoan whose mt-*rrn5* gene is confidently identified by mt-5S (score 35, *E*-value 10^{-6}).

The new CMs discriminate surprisingly well between mt-*rrn5* and pt-*rrn5* (based on differences in score or *E*-value). In fact, pt-5S readily recognized mtDNA-located *rrn5* genes of plastid origin, with a 4-fold score difference between genuine and plastid-derived mt-*rrn5*. Vis-a-vis the rampant DNA transfer from chloroplasts to mitochondria in plants (43), the new models will preclude future confusion between

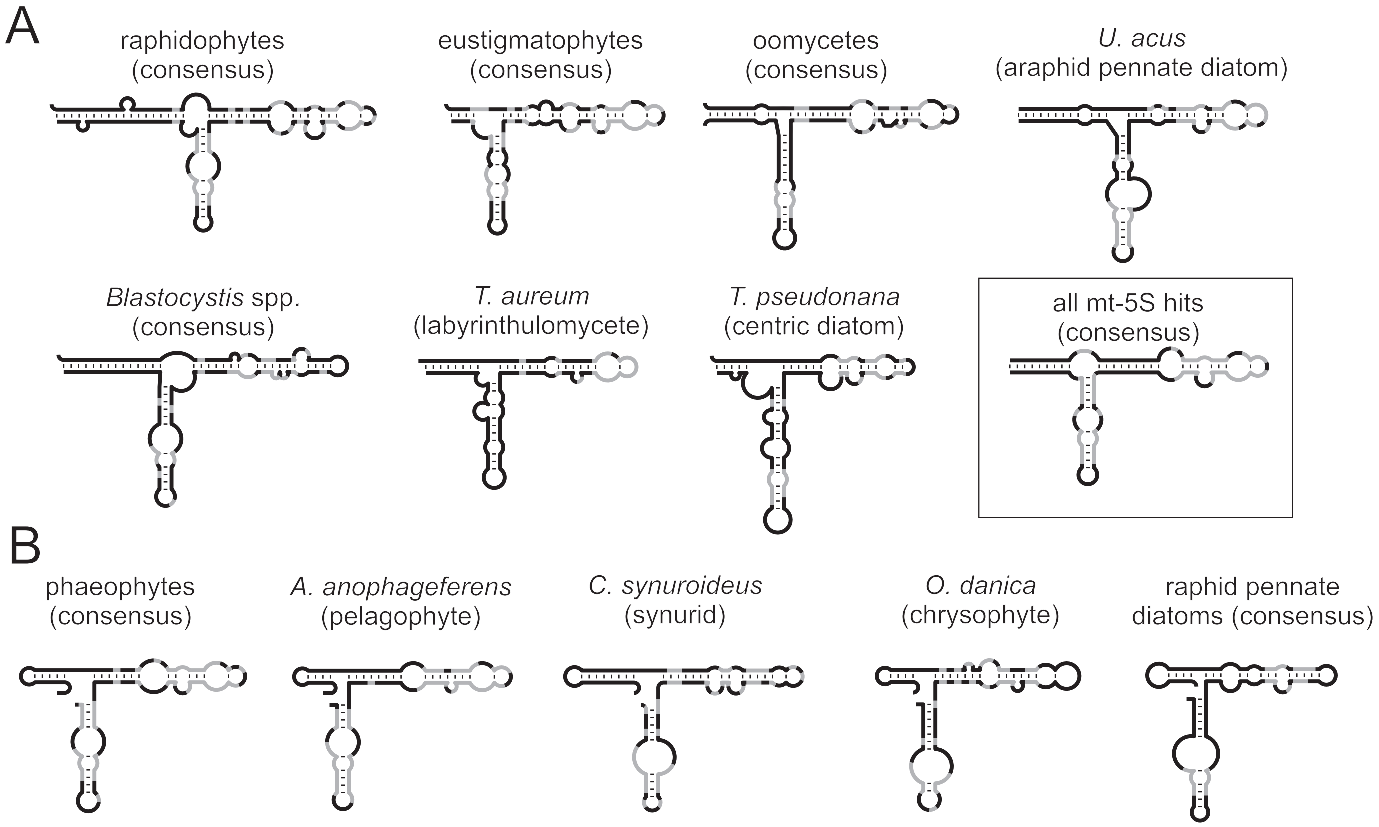
A
raphidophytes (consensus)
eustigmatophytes (consensus)
oomycetes (consensus)
U. acus (araphid pennate diatom)
Blastocystis spp. (consensus)
T. aureum (labyrinthulomycete)
T. pseudonana (centric diatom)
all mt-5S hits (consensus)
B
phaeophytes (consensus)
A. anophageferens (pelagophyte)
C. synuroideus (synurid)
O. danica (chrysophyte)
raphid pennate diatoms (consensus)

Figure 4. Secondary structure skeleton models of stramenopile mt-5S rRNAs and mt-5S-like RNAs. Consensus structures are shown if data are available for several members of a group. For the species used to build the consensus, see Table 2. Gray shading indicates nucleotide conservation compared to the mt-5S rRNA secondary structure model. (**A**) Secondary structure of RNAs with conventional folding. Upper row, mt-5S rRNAs. Lower row, mt-5S-like RNAs and the mitochondrial consensus. (**B**) Permuted (5′) secondary structures. For details and alternative foldings into a conventional secondary structure of mt-5S-like RNAs in raphid pennate diatoms, *C. synuroides* and *O. danica*, see Supplementary Figure S2. Candidate 5S-like RNAs are those that a CM reports below the default threshold, but whose sequence has the propensity to fold into a 5S-like triskelion secondary structure; in some cases, the locus is also syntenic with an *rrn5* reported above the threshold.

endogenous and inter-organelle-transferred 5S rRNA sequences.

In contrast to current perception, our study has uncovered mtDNA-encoded *rrn5* in most eukaryotic supergroups for which sequence information is available, notably jakobids, malawimonads, Archaeplastida (plants, red algae and glaucophytes), Stramenopila (brown algae, diatoms, raphidophytes, eustigmatophytes, pelagophytes), Cryptophyta, Haptophyta and Amorphea [apusozoans, amoebozoans; Figure 1; for taxonomy, see (50)]. While it seems as if the bulk of mitochondrial *rrn5* genes resides in Archaeplastida, basal Excavata and Stramenopila, this spotty taxonomic distribution is due to sampling bias (Supplementary Figure S5). The groups where mtDNA-encoded 5S rRNAs is apparently lacking are Opisthokonta (animals, fungi), Alveolata (ciliates, apicomplexans, dinoflagellates), Heterolobosea and Euglenozoa.

Plastid 5S rRNAs are well conserved, except in non-photosynthetic plastids

We found pt-*rrn5* in nearly all plastid genomes for which complete sequences are available. Exceptions are ptDNAs of haemosporidians and piroplasmids, which are non-photosynthetic. The gene either has diverged to a degree that it is unrecognizable or has been lost from apicoplast DNA. In contrast, pt-*rrn5* was detected in the (also non-photosynthetic) coccidians, the sister clade of the two latter taxa (50). The unconventional loop C (Supplementary Figure S4E and F) in these sequences is corroborated by RNA-Seq data (Figure 6F).

In general, pt-*rrn5* sequences are much more highly conserved than their mitochondrial homologs, with rare secondary structure variations in the otherwise ultra-conserved elements of the β and γ domains (Figure 2B). Less drastic deviations include a supernumerary ∼25-nt-long stem-loop at the base of loop C that characterizes pt-5S rRNAs of gnetophytes [the *enfants terribles* in seed plant phylogeny due to fast-evolving plastid gene sequences (51)] (Supplementary Figure S4A). Other deviations involve a shortened γ domain and absent loop E and helix IV as in the chlorarachniophyte *Bigelowiella*, or an A+T-rich γ domain with extended helices IV and V as in *Euglena longa* and *Chromera velia* (Supplementary Figure S4B–D).

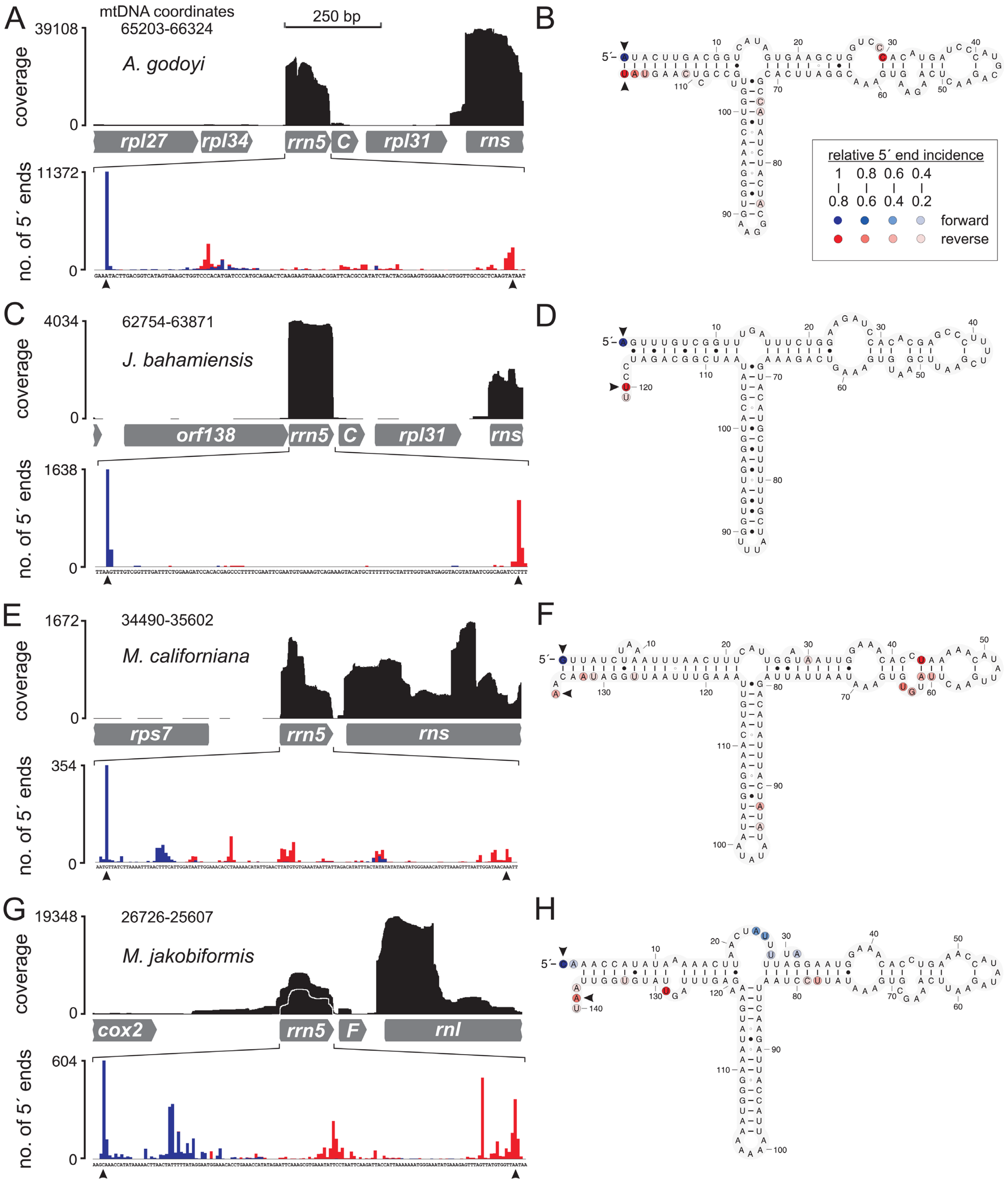

Figure 5. Transcriptome data from jakobid and malawimonad mt-5S rRNAs. (**A** and **B**) *Andalucia godoyi*; (**C** and **D**) *Jakoba bahamiensis*; (**E** and **F**) *Malawimonas californiana*; (**G** and **H**) *Malawimonas jakobiformis*. (**A, C, E** and **G**) RNA-Seq read mapping onto mtDNA sequences. Upper panels, read coverage in linear scale (vertical axis) plotted against the genome region encompassing *rrn5* (500 nt up- and downstream; horizontal axis). Lower panels, count of read 5′ ends (vertical axis) versus mapping position on *rrn5* (horizontal axis); forward reads in blue, reverse reads in red. Black arrowheads indicate experimentally confirmed termini. Only those reads are shown where at least one mate of a pair maps to the *rrn5* locus. The majority of read pairs in (**G**) span either 5S rRNA plus an ∼230-nt-long upstream region, or exclusively 5S rRNA (delimited by the white line). Note that vertical scales vary among samples. (**B, D, F** and **H**) End-mapping results superimposed on secondary structure models. The 5′ ends of forward and reverse reads are indicated by blue and red circles, respectively. The color shades indicate the ratio between the number of reads ending at a given position compared to the number of reads ending at the most frequent position (see inset in **B**).

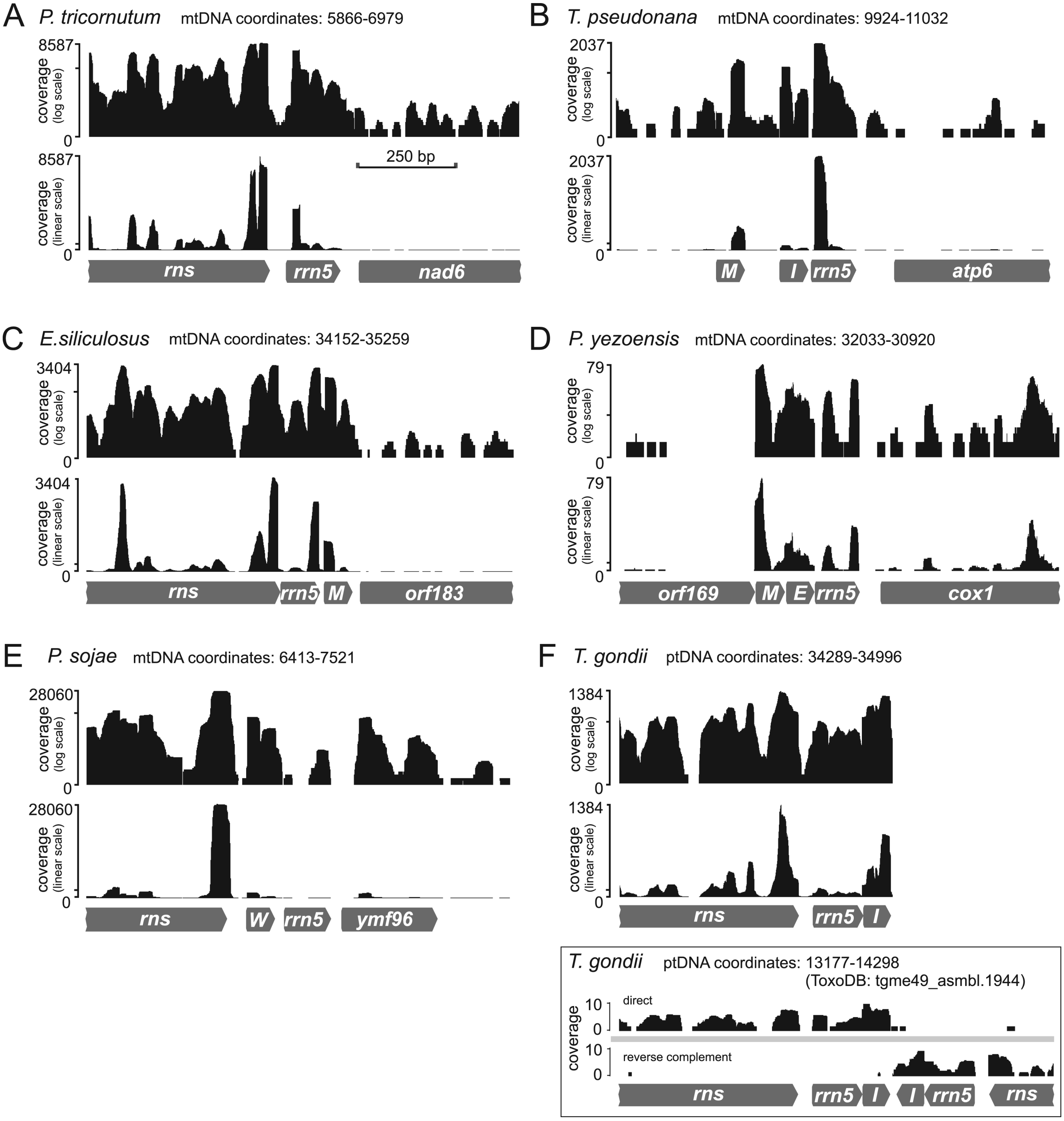

Figure 6. Transcriptome data for non-excavate mt-5S rRNAs. RNA-Seq read mapping onto mtDNA or ptDNA sequences, with coordinates as in GenBank. Upper panels, read coverage in log scale (vertical axis). Lower panels, read coverage in linear scale (vertical axis) plotted against the genome region encompassing *rrn5* (horizontal axis; 500 nt up- and downstream of *rrn5*, except in *T. gondii*, where the genome sequence ends 87 nt downstream of *rrn5*). Note that vertical scales vary among samples, with the tick indicating 10% of the maximal coverage. (**A**) *Phaeodactylum tricornutum* (diatom); (**B**) *Thalassiosira pseudonana* (diatom). Note that for the two diatoms, the available data do not allow precise end-mapping of the predicted loci. (**C**) *Ectocarpus siliculosus* (phaeophyte); (**D**) *Pyropia yezoensis* (rhodophyte); (**E**) *Phytophthora sojae* (oomycete). (**F**) *Toxoplasma gondii* (apicomplexan). Note that the *rrn5* and *trnI* genes abut in *T. gondii*, resulting in a seemingly continuous read coverage. The lowest panel shows the small-transcript mapping data available at the ToxoDB genome browser. The contig 'tgme49_asmbl.1944' represents the apicoplast genome and has a different coordinate system than the corresponding GenBank record (NC_001799).

Moderate deviations of mt-5S rRNAs in jakobids, malawimonads and archaeplastids

As already mentioned, jakobids, malawimonads and archaeplastids are the lineages with the most conservative mt-*rrn5* sequences and structures. In jakobids, not only mt-*rrn5* (17), but also the entire mitochondrial genome is minimally derived (27,52). Malawimonad mt-*rrn5* shows a minor deviation (confirmed by RNA-Seq data)—a helix I whose stem is a dozen residues longer than usual (Figure 5F and H). Most probably, the prolonged helix I does not interfere with the function of 5S rRNA and the mitoribosome as a whole. This view is supported by studies of a *Bacillus subtilis* mutant that is defective in 5S rRNA processing. Despite a substantial α-domain extension in this mutant, the incompletely processed molecule is readily integrated into fully functional ribosomes (53).

Among rhodophyte mt-5S rRNAs, those of the two deeply diverging lineages Cyanidiales and Gigartinales (54) are well conserved and accordingly were reported early on (Supplementary Table S1)(55–57). However, in more derived red algal lineages, mt-*rrn5* has an extended loop B (7–10 nt versus the usual length of 2–5 nt) and an A+T-rich sequence (Supplementary Figure S3A), explaining why the gene has remained unnoticed so far. Experimental support for the expression of mt-5S rRNAs in a derived red alga comes from *Pyropia* RNA-Seq data (Figure 6D).

Deviant mt-5S rRNAs abound in stramenopiles

Contrary to previous views, almost all stramenopile phyla appear to encode *rrn5* in their mitochondrial genomes, with the most sequence-derived, but structurally conserved homologs in *Blastocystis* (Supplementary Figure S2). Only one out of the 13 potential *Blastocystis rrn5* genes is identified with the mtAT-5S model, a situation resembling that of 5S rRNA-like sequences in Amoebozoa (49). Similarly, mt-5S and mtAT-5S models readily detect only a single diatom *rrn5*, that from *Ulnaria* (*Synedra*) *acus* (Table 2 and Supplementary Table S1). Other diatom sequences are even more divergent and at least two appear to be expressed (Figure 6A and B).

The mt-5S rRNA secondary structure deviates most notably in phaeophytes, pelagophytes and several other photosynthetic stramenopiles. In these molecules, the α domain is likely permuted (Figure 4B), with the usual helix I replaced by a 5′ (or 3′) hairpin so that the three 5S rRNA domains are linked together by an open three-way junction (Figure 2C). The permuted structure is experimentally supported by RNA-Seq data from one brown alga (Figure 6C). Even more divergent molecules appear to exist in other stramenopiles (Table 2), with all three structural domains deviating considerably in length and sequence (Figure 4); however, experimental support is lacking. In sum, stramenopiles possibly encompass the largest structural diversity of 5S rRNA among eukaryotes.

The domain shuffling of brown algal mt-5S rRNA reported here is not the first instance of circular permutation in structural RNAs: it has been demonstrated previously for nucleus-encoded tRNA genes of the red alga *C. merolae*, with the 5′ and 3′ portions of the tRNA specified in inverted succession. The precursor tRNAs are cleaved and pieces are ligated in the correct order post-transcriptionally (58,59). Another case is *ssrA* (specifying transfer-messenger RNA, or tmRNA), which is circularly permuted in certain bacteria and in mitochondria of jakobids and oomycetes (60–62). In contrast, the permuted mt-5S rRNAs of stramenopiles discovered here are not only permuted at the level of the gene but also in the final product. Precedents for continuous RNAs with shuffled domains are hammerhead (63,64) and twister ribozyme RNAs (65), as well as RNA aptamers (66).

An intriguing question bears on the consequences for the global ribosome structure when 5S rRNA domains are rearranged or carry indels as described above, given that domains physically interact with ribosomal proteins. For example, the β domain (in particular, helix III and loop C) engages in contacts with the (bacteria-type) L18 and L5, and the γ domain (helix IV and loop E) with L25 (67). It will be interesting to gain insight into the tertiary interactions within organelle ribosomes having an unorthodox 5S rRNA.

Additional hidden organelle 5S rRNAs?

Blastocystis seems to be the second group whose 5S-like mt-rRNA genes are too A+T-rich to be detected by covariance analysis. As in Amoebozoa (49), biochemical methods will be required for a confident gene assignment. Equally undetectable with current computational tools would be split *rrn5* genes—although there is currently no evidence for the existence of such a gene configuration—even if well conserved at the sequence level. In the absence of a predicted organelle *rrn5* gene, advanced biochemical studies will be needed to determine whether a 5S rRNA is indeed part of the organelle ribosome and is organelle-encoded. Alternatively, in the course of evolution, this RNA might have been functionally substituted by the product of a genuine nuclear gene (with import of the corresponding nucleus-encoded 5S rRNA into mitochondria), or replaced entirely by proteins in the mitochondrial ribosome in question.

A recent proteomic analysis (68) identified a nucleus-encoded mitochondrial L25 in *A. castellanii*, and BLAST searches detected mt-L25 homologs in the nuclear genomes of other organisms known to contain mtDNA-encoded 5S or 5S-like rRNAs (e.g. other amoebozoans, red algae, land plants, some green algae), as well as in *Phytophtora* and *Blastocystis*, whose deviant mtDNA-encoded 5S rRNAs are reported here. Thus, mitochondrial homologs of proteins that bind to bacterial 5S rRNA might in certain cases serve as proxies for the possible existence of highly divergent mitochondrial 5S rRNAs that remain refractory to discovery by the comparative modeling approach described here, which has otherwise proven so successful.

CMs deposited in RFAM

Organelle CM models and the corresponding seed alignments will be made available through RFAM.

ACCESSION NUMBERS

GenBank: KP165385 (*Paracercomonas marina*), KP165386 (*Klebsormidium flaccidum*), KP165387 (*Malawimonas cali-

forniana), KP165388 (*Stachyamoeba lipophora*), KP165389 (*Thecamonas trahens*), KP165390 and KP165391 (*Malawimonas* sp.).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We acknowledge the excellent assistance of Lise Forget (Université de Montréal) in genomics and transcriptomics data generation.

FUNDING

Genome Quebec/Canada [to B.F.L. and G.B.]; Canadian Institute of Health Research [CIHR (MCR), MSP-14226 and SP-34 to B.F.L., G.B., and M.W.G.; MOP-79309 to G.B.]; Natural Sciences and Engineering Research Council of Canada (NSERC; to B.F.L.]. Funding for open access charge: Genome Quebec/Canada, CIHR, NSERC.
Conflict of interest statement. None declared.

REFERENCES

1. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
2. Zhang,Z., Green,B.R. and Cavalier-Smith,T. (1999) Single gene circles in dinoflagellate chloroplast genomes. *Nature*, **400**, 155–159.
3. Bullerwell,C.E., Schnare,M.N. and Gray,M.W. (2003) Discovery and characterization of *Acanthamoeba castellanii* mitochondrial 5S rRNA. *RNA*, **9**, 287–292.
4. Sharma,M.R., Booth,T.M., Simpson,L., Maslov,D.A. and Agrawal,R.K. (2009) Structure of a mitochondrial ribosome with minimal RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9637–9642.
5. Greber,B.J., Boehringer,D., Leitner,A., Bieri,P., Voigts-Hoffmann,F., Erzberger,J.P., Leibundgut,M., Aebersold,R. and Ban,N. (2014) Architecture of the large subunit of the mammalian mitochondrial ribosome. *Nature*, **505**, 515–519.
6. Greber,B.J., Boehringer,D., Leibundgut,M., Bieri,P., Leitner,A., Schmitz,N., Aebersold,R. and Ban,N. (2014) The complete structure of the large subunit of the mammalian mitochondrial ribosome. *Nature*, doi:10.1038/nature13895.
7. Amunts,A., Brown,A., Bai,X.-c., Llácer,J.L., Hussain,T., Emsley,P., Long,F., Murshudov,G., Scheres,S.H.W. and Ramakrishnan,V. (2014) Structure of the yeast mitochondrial large ribosomal subunit. *Science*, **343**, 1485–1489.
8. Brown,A., Amunts,A., Bai,X.-C., Sugimoto,Y., Edwards,P.C., Murshudov,G., Scheres,S.H.W. and Ramakrishnan,V. (2014) Structure of the large ribosomal subunit from human mitochondria. *Science*, **346**, 718–722.
9. Sharma,M.R., Koc,E.C., Datta,P.P., Booth,T.M., Spremulli,L.L. and Agrawal,R.K. (2003) Structure of the mammalian mitochondrial ribosome reveals an expanded functional role for its component proteins. *Cell*, **115**, 97–108.
10. Yoshionari,S., Koike,T., Yokogawa,T., Nishikawa,K., Ueda,T., Miura,K. and Watanabe,K. (1994) Existence of nuclear-encoded 5S-rRNA in bovine mitochondria. *FEBS Lett.*, **338**, 137–142.
11. Magalhães,P.J., Andreu,A.L. and Schon,E.A. (1998) Evidence for the presence of 5S rRNA in mammalian mitochondria. *Mol. Biol. Cell*, **9**, 2375–2382.
12. Smirnov,A., Entelis,N., Martin,R.P. and Tarassov,I. (2011) Biological significance of 5S rRNA import into human mitochondria: role of ribosomal protein MRP-L18. *Genes Dev.*, **25**, 1289–1305.
13. Burger,G., Plante,I., Lonergan,K.M. and Gray,M.W. (1995) The mitochondrial DNA of the amoeboid protozoon, *Acanthamoeba castellanii*: complete sequence, gene content and genome organization. *J. Mol. Biol.*, **245**, 522–537.
14. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
15. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
16. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
17. Lang,B.F., Goff,L.J. and Gray,M.W. (1996) A 5 S rRNA gene is present in the mitochondrial genome of the protist *Reclinomonas americana* but is absent from red algal mitochondrial DNA. *J. Mol. Biol.*, **261**, 607–613.
18. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
19. Smith,S.W., Overbeek,R., Woese,C.R., Gilbert,W. and Gillevet,P.M. (1994) The genetic data environment an expandable GUI for multiple sequence analysis. *Comput. Appl. Biosci.*, **10**, 671–675.
20. Bernhart,S.H., Hofacker,I.L., Will,S., Gruber,A.R. and Stadler,P.F. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.
21. Gerstein,M., Sonnhammer,E.L.L. and Chothia,C. (1994) Volume changes in protein evolution. *J. Mol. Biol.*, **236**, 1067–1078.
22. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
23. Grigoriev,I.V., Nordberg,H., Shabalov,I., Aerts,A., Cantor,M., Goodstein,D., Kuo,A., Minovitsky,S., Nikitin,R., Ohm,R.A. *et al.* (2012) The genome portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res.*, **40**, D26–D32.
24. Lorenz,R., Bernhart,S.H., Höner Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
25. Lai,D., Proctor,J.R., Zhu,J.Y.A. and Meyer,I.M. (2012) R-CHIE: a web server and R package for visualizing RNA secondary structures. *Nucleic Acids Res.*, **40**, e95.
26. Weinberg,Z. and Breaker,R.R. (2011) R2R - software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics*, **12**, 3.
27. Burger,G., Gray,M.W., Forget,L. and Lang,B.F. (2013) Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biol. Evol.*, **5**, 418–438.
28. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using *Phred*. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
29. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. *Genome Res.*, **8**, 186–194.
30. Lang,B.F. and Burger,G. (2007) Purification of mitochondrial and plastid DNA. *Nat. Protoc.*, **2**, 652–660.
31. Burger,G., Lavrov,D.V., Forget,L. and Lang,B.F. (2007) Sequencing complete mitochondrial and plastid genomes. *Nat. Protoc.*, **2**, 603–614.
32. Liang,C., Zhang,X., Zou,J., Xu,D., Su,F. and Ye,N. (2010) Identification of miRNA from *Porphyra yezoensis* by high-throughput sequencing and bioinformatics analysis. *PLoS ONE*, **5**, e10698.
33. Huang,A., He,L. and Wang,G. (2011) Identification and characterization of microRNAs from *Phaeodactylum tricornutum* by high-throughput sequencing and bioinformatics analysis. *BMC Genom.*, **12**, 337.
34. Norden-Krichmar,T.M., Allen,A.E., Gaasterland,T. and Hildebrand,M. (2011) Characterization of the small RNA transcriptome of the diatom, *Thalassiosira pseudonana*. *PLoS ONE*, **6**, e22870.
35. Cock,J.M., Sterck,L., Rouzé,P., Scornet,D., Allen,A.E., Amoutzias,G., Anthouard,V., Artiguenave,F., Aury,J.-M., Badger,J.H. *et al.* (2010) The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature*, **465**, 617–621.
36. Xu,M.J., Zhou,D.H., Huang,S.Y., Zhao,F.R., Nisbet,A.J., Lin,R.Q., Song,H.Q. and Zhu,X.Q. (2013) Comparative characterization of microRNA profiles of different genotypes of *Toxoplasma gondii*. *Parasitology*, **140**, 1111–1118.
37. Qutob,D., Chapman,B.P. and Gijzen,M. (2013) Transgenerational gene silencing causes gain of virulence in a plant pathogen. *Nat. Commun.*, **4**, 1349.

38. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.*, **17**, 10–12.
39. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
40. Carver,T., Harris,S.R., Berriman,M., Parkhill,J. and McQuillan,J.A. (2012) Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, **28**, 464–469.
41. Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.-A. and Barrell,B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.
42. Gajria,B., Bahl,A., Brestelli,J., Dommer,J., Fischer,S., Gao,X., Heiges,M., Iodice,J., Kissinger,J.C., Mackey,A.J. *et al.* (2008) ToxoDB: an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Res.*, **36**, D553–D556.
43. Richardson,A.O. and Palmer,J.D. (2007) Horizontal gene transfer in plants. *J. Exp. Bot.*, **58**, 1–9.
44. Smith,D.R. (2014) Mitochondrion-to-plastid DNA transfer: it happens. *New Phytol.*, **202**, 736–738.
45. Karpov,S.A., Bass,D., Mylnikov,A.P. and Cavalier-Smith,T. (2006) Molecular phylogeny of Cercomonadidae and kinetid patterns of *Cercomonas* and *Eocercomonas* gen. nov. (Cercomonadida, Cercozoa). *Protist*, **157**, 125–158.
46. Gray,M.W., Lang,B.F. and Burger,G. (2004) Mitochondria of protists. *Annu. Rev. Genet.*, **38**, 477–524.
47. Burger,G., Gray,M.W. and Lang,B.F. (2003) Mitochondrial genomes: anything goes. *Trends Genet.*, **19**, 709–716.
48. Silberfeld,T., Leigh,J.W., Verbruggen,H., Cruaud,C., de Reviers,B. and Rousseau,F. (2010) A multi-locus time-calibrated phylogeny of the brown algae (Heterokonta, Ochrophyta, Phaeophyceae): investigating the evolutionary nature of the 'brown algal crown radiation'. *Mol. Phylogenet. Evol.*, **56**, 659–674.
49. Bullerwell,C.E., Burger,G., Gott,J.M., Kourennaia,O., Schnare,M.N. and Gray,M.W. (2010) Abundant 5S rRNA-like transcripts encoded by the mitochondrial genome in Amoebozoa. *Eukaryot. Cell*, **9**, 762–773.
50. Adl,S.M., Simpson,A.G.B., Lane,C.E., Lukeš,J., Bass,D., Bowser,S.S., Brown,M.W., Burki,F., Dunthorn,M., Hampl,V. *et al.* (2012) The revised classification of eukaryotes. *J. Eukaryot. Microbiol.*, **59**, 429–493.
51. Zhong,B., Deusch,O., Goremykin,V.V., Penny,D., Biggs,P.J., Atherton,R.A., Nikiforova,S.V. and Lockhart,P.J. (2011) Systematic error in seed plant phylogenomics. *Genome Biol. Evol.*, **3**, 1340–1348.
52. Lang,B.F., Burger,G., O'Kelly,C.J., Cedergren,R., Golding,G.B., Lemieux,C., Sankoff,D., Turmel,M. and Gray,M.W. (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature*, **387**, 493–497.
53. Condon,C., Brechemier-Baey,D., Beltchev,B., Grunberg-Manago,M. and Putzer,H. (2001) Identification of the gene encoding the 5S ribosomal RNA maturase in *Bacillus subtilis*: mature 5S rRNA is dispensable for ribosome function. *RNA*, **7**, 242–253.
54. Verbruggen,H., Maggs,C.A., Saunders,G.W., Le Gall,L., Yoon,H.S. and De Clerck,O. (2010) Data mining approach identifies research priorities and data requirements for resolving the red algal tree of life. *BMC Evol. Biol.*, **10**, 16.
55. Ohta,N., Sato,N. and Kuroiwa,T. (1998) Structure and organization of the mitochondrial genome of the unicellular red alga *Cyanidioschyzon merolae* deduced from the complete nucleotide sequence. *Nucleic Acids Res.*, **26**, 5190–5198.
56. Gray,M.W., Lang,B.F., Cedergren,R., Golding,G.B., Lemieux,C., Sankoff,D., Turmel,M., Brossard,N., Delage,E., Littlejohn,T.G. *et al.* (1998) Genome structure and gene content in protist mitochondrial DNAs. *Nucleic Acids Res.*, **26**, 865–878.
57. Burger,G., Saint-Louis,D., Gray,M.W. and Lang,B.F. (1999) Complete sequence of the mitochondrial DNA of the red alga *Porphyra purpurea*. Cyanobacterial introns and shared ancestry of red and green algae. *Plant Cell*, **11**, 1675–1694.
58. Soma,A. (2014) Circularly permuted tRNA genes: their expression and implications for their physiological relevance and development. *Front Genet.*, **5**, 63.
59. Soma,A., Onodera,A., Sugahara,J., Kanai,A., Yachie,N., Tomita,M., Kawamura,F. and Sekine,Y. (2007) Permuted tRNA genes expressed via a circular RNA intermediate in *Cyanidioschyzon merolae*. *Science*, **318**, 450–453.
60. Keiler,K.C., Shapiro,L. and Williams,K.P. (2000) tmRNAs that encode proteolysis-inducing tags are found in all known bacterial genomes: A two-piece tmRNA functions in *Caulobacter*. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 7778–7783.
61. Jacob,Y., Seif,E., Paquet,P.-O. and Lang,B.F. (2004) Loss of the mRNA-like region in mitochondrial tmRNAs of jakobids. *RNA*, **10**, 605–614.
62. Hafez,M., Burger,G., Steinberg,S.V. and Lang,B.F. (2013) A second eukaryotic group with mitochondrion-encoded tmRNA. In silico identification and experimental confirmation. *RNA Biol.*, **10**, 1117–1124.
63. de la Peña,M. and García-Robles,I. (2010) Ubiquitous presence of the hammerhead ribozyme motif along the tree of life. *RNA*, **16**, 1943–1950.
64. Perreault,J., Weinberg,Z., Roth,A., Popescu,O., Chartrand,P., Ferbeyre,G. and Breaker,R.R. (2011) Identification of hammerhead ribozymes in all domains of life reveals novel structural variations. *PLoS Comput. Biol.*, **7**, e1002031.
65. Roth,A., Weinberg,Z., Chen,A.G.Y., Kim,P.B., Ames,T.D. and Breaker,R.R. (2014) A widespread self-cleaving ribozyme class is revealed by bioinformatics. *Nat. Chem. Biol.*, **10**, 56–60.
66. Vu,M.M.K., Jameson,N.E., Masuda,S.J., Lin,D., Larralde-Ridaura,R. and Lupták,A. (2012) Convergent evolution of adenosine aptamers spanning bacterial, human, and random sequences revealed by structure-based bioinformatics and genomic SELEX. *Chem. Biol.*, **19**, 1247–1254.
67. Yusupov,M.M., Yusupova,G.Z., Baucom,A., Lieberman,K., Earnest,T.N., Cate,J.H.D. and Noller,H.F. (2001) Crystal structure of the ribosome at 5.5 Å resolution. *Science*, **292**, 883–896.
68. Gawryluk,R.M.R., Chisholm,K.A., Pinto,D.M. and Gray,M.W. (2014) Compositional complexity of the mitochondrial proteome of a unicellular eukaryote (*Acanthamoeba castellanii*, supergroup Amoebozoa) rivals that of animals, fungi, and plants. *J. Proteomics*, **109**, 400–416.