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

Matus Valach1,*, Gertraud Burger1, Michael W. Gray2 and B. Franz Lang1,*

1Department of Biochemistry and Robert-Cedergren Centre of Bioinformatics and Genomics, Université de Montréal, Montréal, QC, H3C 3J7, Canada and 2Department 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 archaean, 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 archaean) 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 archaean 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

<sup>1</sup>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

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
of the yeast mitoribosome LSU is composed of the LSU mt-rRNA expansion segments, which (together with additional proteins) spatially replace SS rRNA without adopting a SS-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 SS 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 SS 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 mitoribosomal ribosome in these cases, are not entirely clear.

Although replacement of SS rRNA by protein in some mitochondrial (and perhaps plastid) systems can account for the absence of an organelle SS 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 SS 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 SS 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 SS 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 SS rRNAs with permuted secondary structure encoded in brown algal mitochondria, and the identification of two apicoplast SS 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-genre 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 Cnbuild and Cncalibrate programs of Infernal v1.1, after masking -wgsc’ (21) increases the chance of detecting sequences in and plastid-specific CMs (referred to as mt-5S and pt-5S sequence positions are used for building mitochondrion- and the false discovery rate (FDR) by:

\[ FDR = \frac{No. \ of \ wrongly \ assigned \ mt\-genes}{No. \ of \ all \ mt\-genes \ in \ the \ mt\-genome \ test \ set}. \]

Strains and cultures

Andalucia godoyi PRA-185, Malavimonas jakobiformis (ATCC 50310), M. californiana (ATCC 50740), Malavimonas 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 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 Resources website (as of 22 July 2014). Homologs with the MFannot tool (http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl), essentially as described (27), except for Klebsormidium flaccidum, which grows on synthetic media.

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-CHE (25) and secondary structures were drawn with R2R v1.0.3 (26).
Quantitative analyses of small RNA expression in Toxoplasma gondii using RNA-Seq.

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*, *Petalo尼亚 fascia*, *Punctaria latifolia*, *Sargassum hemiphyllum*, *Sa. henslowianum*, *Sa. integerrimum*, *Sa. thunbergii*, *Sa. vachellianum*, *Scytosiphon dotyi* and *Sc. lomentaria*; http://onekp.com).

Quantitative analyses of small RNA expression in Toxoplasma gondii using RNA-Seq.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Test dataset</th>
<th>mt-genomes</th>
<th>pt-genomes</th>
<th>mt-genomes</th>
<th>pt-genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RF-5S</td>
<td>mt-5S</td>
<td>pt-5S</td>
<td></td>
</tr>
<tr>
<td>True positives</td>
<td></td>
<td>76</td>
<td>497</td>
<td>108</td>
<td>500</td>
</tr>
<tr>
<td>False positives</td>
<td></td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>False negatives</td>
<td></td>
<td>32</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TPR</td>
<td></td>
<td>70.4%</td>
<td>99.4%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>FDR</td>
<td></td>
<td>2.8%</td>
<td>0%</td>
<td>3.7%</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

4Hits within the inclusion threshold as reported by Cmsearch used with default settings. The mt-genome test data set includes 4884 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 $B$ (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.

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 $10^{-22}$ 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
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., Paracercamornas marina (ATCC 50344, earlier misidentified as Cercomonas longicauda (45)), Stachyamoeba lipophora and Thecamonas troehs.

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 (46,47). 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 Ulania (Syedera) acus and the eustigmatophyte Nanochloropsis 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 these 12, the RF-5S model detects eight. Interestingly, pt-5S (but not RF-5S) revealed rrn5 in the apicoplast (plastid-derivred organelle) genome of Toxoplasma gondii and Eimeria tenella. Further, in both cases, the corresponding E-values (<10^-3) 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 Trebouxiphyccae 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 syntetic rns-rrn5-rrnM 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, transcriptome 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 \( \beta \) or \( \gamma \) 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, syriniids, 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, glauco-
phytes, 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 bahamensis* due to a U-rich \( \gamma \) 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 \( \alpha \) 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 oomyete, 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 bahamensis*, *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. bahamensis* 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 (*Phyropia*), two diatoms (*Phaeodactylum* and *Thalassiosira*), an
oomycete (*Phytophthora sojae*), a brown alga (*Ectocarpus*), and a coccidian (*Toxoplasma*) confirm that organelle SS 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 SS rRNA (Figure 6). Still, the analysis of overrepresented read starts and ends corroborates the predicted organelle SS 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 SS 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 10⁻⁶⁻⁷).

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

---

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

<table>
<thead>
<tr>
<th>Lineage</th>
<th>Number of species</th>
<th>α domain</th>
<th>β domain</th>
<th>γ domain</th>
<th>mt-5S score (E-value) range</th>
<th>mtAT-5S score (E-value) range</th>
<th>mtPerm-5S score (E-value) range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacillariophyta (rapliph, pennate diatoms)</td>
<td>1</td>
<td>conventional</td>
<td>conserved</td>
<td>conserved</td>
<td>25.8 (1.3 × 10⁻⁳)</td>
<td>36.7 (2.5 × 10⁻⁴)</td>
<td>n.a.</td>
</tr>
<tr>
<td>Bacillariophyta (rapliph, pennate diatoms)</td>
<td>5</td>
<td>5′ permuted or conventional</td>
<td>divergent</td>
<td>moderately conserved</td>
<td>11–18.1 (3.4–5.5 × 10⁻²)</td>
<td>34.4 (1.1 × 10⁻³)</td>
<td>21.7–59.7 (2.5 × 10⁻⁵–2.5 × 10⁻⁶)</td>
</tr>
<tr>
<td>Bacillariophyta (centric diatoms)</td>
<td>1</td>
<td>conventional</td>
<td>moderately conserved</td>
<td>divergent</td>
<td>not predicted</td>
<td>23.1 (1 × 10⁻¹)</td>
<td>n.a.</td>
</tr>
<tr>
<td>Blastocystis</td>
<td>5</td>
<td>conventional</td>
<td>moderately conserved</td>
<td>divergent</td>
<td>11.3 (1.4)</td>
<td>18.7–42.9</td>
<td>n.a.</td>
</tr>
<tr>
<td>Chrysophyceae</td>
<td>1</td>
<td>5′ permuted or short conventional</td>
<td>divergent</td>
<td>moderately conserved</td>
<td>8.9 (7)</td>
<td>16.9 (1.5)</td>
<td>n.a.</td>
</tr>
<tr>
<td>Eustigmatophytes</td>
<td>5</td>
<td>short conventional</td>
<td>conserved</td>
<td>conserved</td>
<td>41–48.5 (4.6 × 10⁻⁷–1 × 10⁻⁸)</td>
<td>45.5–63.1 (3.6 × 10⁻⁹–1.4 × 10⁻⁹)</td>
<td>n.a.</td>
</tr>
<tr>
<td>Labyrinthulomycetes</td>
<td>1</td>
<td>short</td>
<td>moderately divergent</td>
<td>14.6</td>
<td>20.9</td>
<td>n.a.</td>
<td></td>
</tr>
<tr>
<td>Oomycetes</td>
<td>14</td>
<td>conventional</td>
<td>moderately conserved</td>
<td>moderately conserved</td>
<td>26.7 (1 × 10⁻⁴)</td>
<td>30–30.1 (4.3 × 10⁻⁸–6.4 × 10⁻⁸)</td>
<td>n.a.</td>
</tr>
<tr>
<td>Pelagophytes</td>
<td>1</td>
<td>5′ permuted</td>
<td>conserved</td>
<td>conserved</td>
<td>59</td>
<td>58.2</td>
<td>56.9</td>
</tr>
<tr>
<td>Phaeophytes</td>
<td>31</td>
<td>5′ permuted</td>
<td>conserved</td>
<td>conserved</td>
<td>49.4–64.3 (4.5 × 10⁻¹¹)</td>
<td>41–65.1 (1.2 × 10⁻⁸)</td>
<td>(1 × 10⁻⁹)</td>
</tr>
<tr>
<td>Raphidiophytes</td>
<td>2</td>
<td>extended conventional</td>
<td>conserved</td>
<td>62.9–69.1 (9.1 × 10⁻⁹–2.3 × 10⁻¹²)</td>
<td>(3.2 × 10⁻⁵–4.6 × 10⁻⁶)</td>
<td>(2.6 × 10⁻¹⁰–2.1 × 10⁻¹¹)</td>
<td>n.a.</td>
</tr>
<tr>
<td>Synurophyceae</td>
<td>1</td>
<td>5′ permuted</td>
<td>divergent</td>
<td>moderately conserved</td>
<td>20.3 (1.7 × 10⁻²)</td>
<td>23.9 (5.5 × 10⁻⁷)</td>
<td>63.1</td>
</tr>
</tbody>
</table>

---

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

In contrast to current perception, our study has uncovered mtDNA-encoded \textit{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 [apurzoans, amoebozoans; Figure 1; for taxonomy, see (50)]. While it seems as if the bulk of mitochondrial \textit{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-\textit{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-\textit{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-\textit{rrn5} sequences are much more highly conserved than their mitochondrial homologs, with rare secondary structure variations in the otherwise ultra-conserved elements of the \(\beta\) and \(\gamma\) domains (Figure 2B). Less drastic deviations include a supernumerary \(~25\text{-nt-long}\) stem-loop at the base of loop C that characterizes pt-5S rRNAs of gnetophytes [the\textit{enfants terribles} in seed plant phylogeny due to fast-evolving plastid gene sequences (51)] (Supplementary Figure S4A). Other deviations involve a shortened \(\gamma\) domain and absent loop E and helix IV as in the chlorarachniophyte \textit{Bigelowiella}, or an A+T-rich \(\gamma\) domain with extended helices IV and V as in \textit{Euglena longa} and \textit{Chromera velia} (Supplementary Figure S4B–D).

**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, \textit{C. synuroideus} and \textit{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 \textit{rrn5} reported above the threshold.
Figure 5. Transcriptome data from jakobid and malawimond mt-5S rRNAs. (A and B) Andalucia godoyi; (C and D) Jakoba bahamiensis; (E and F) Malavimonas californiana; (G and H) Malavimonas 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 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 rnl 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 domain (Table S1). Other diatom sequences are even more divergent and at least two appear to be expressed (Supplementary Table S1). 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 mAT-5S model, a situation resembling that of 5S rRNA-like sequences in Amoebozoa (49). Similarly, mt-5S and mAT-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-
formiana), 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 Institute of Health Research [CIHR (MCR), MSP-14226 to G.B.]; NSERC; to B.F.L.]. Funding for open access charge: Genome Quebec/Canada, CIHR, NSERC.

Conflict of interest statement.

REFERENCES
