The *Rhodomonas salina* mitochondrial genome: bacteria-like operons, compact gene arrangement and complex repeat region

Amy M. Hauth*, Uwe G. Maier¹, B. Franz Lang and Gertraud Burger

Département de Biochimie, Robert Cedergren Research Center for Bioinformatics and Genomics, Canadian Institute for Advanced Research, Université de Montréal, 2900 Boulevard Édouard-Montpetit, Montréal, Québec, Canada H3T 1J4 and ¹Cell Biology, Philipps-University Marburg, Karl-von-Frisch-Strasse, D35032 Marburg, Germany

Received June 29, 2005; Revised and Accepted July 21, 2005 DDBJ/EMBL/GenBank accession no. NC_002572

**ABSTRACT**

To gain insight into the mitochondrial genome structure and gene content of a putatively ancestral group of eukaryotes, the cryptophytes, we sequenced the complete mitochondrial DNA of *Rhodomonas salina*. The 48 063 bp circular-mapping molecule codes for 2 rRNAs, 27 tRNAs and 40 proteins including 23 components of oxidative phosphorylation, 15 ribosomal proteins and two subunits of *tat* translocase. One potential protein (ORF161) is without assigned function. Only two introns occur in the genome; both are present within *cox1* belong to group II and contain RT open reading frames. Primitive genome features include bacteria-like rRNAs and tRNAs, ribosomal protein genes organized in large clusters resembling bacterial operons and the presence of the otherwise rare genes such as *rps1* and *tatA*. The highly compact genome organization contrasts with the presence of a 4.7 kb long, repeat-containing intergenic region. Repeat motifs ~40–700 bp long occur up to 31 times, forming a complex repeat structure. Tandem repeats are the major arrangement but the region also includes a large, ~3 kb, inverted repeat and several potentially stable ~40–80 bp long hairpin structures. We provide evidence that the large repeat region is involved in replication and transcription initiation, predict a promoter motif that occurs in three locations and discuss two likely scenarios of how this highly structured repeat region might have evolved.

**INTRODUCTION**

Mitochondrial DNAs (mtDNAs) of eukaryotes as diverse as mammals, yeasts and ciliates encode essentially the same biological functions. Components specified by mitochondrial genes are involved primarily in oxidative phosphorylation, protein synthesis and in some instances also in transcription, RNA processing, and the import, assembly and maturation of proteins [for a recent review see (1)]. The most conspicuous variation between mtDNA in eukaryotes is significant differences in gene complement, which ranges from as few as five in apicomplexan protists (2) to nearly a hundred in jakobid flagellates (3). Also noticeably, mtDNAs from various taxa have differences in coding density, i.e. the proportion of coding versus non-coding DNA. Highly compact mitochondrial genomes occur in animals ['small is beautiful’ (4)], red algae (5), ciliates (6,7) and some fungi [e.g. *Schizosaccharomyces pombe* (8)], with the highest density found in the bicosoecid flagellate *Cafeteria roenbergensis* (9) where only 3.4% is non-coding. Lowest density occurs in land plants and many fungi as well as the unicellular relatives of animals, the choanozoan and ichthyosporean protists, where extensive intergenic regions account for 50–70% of the total mitochondrial genome [reviewed in (10)].

Recently, we discovered a ‘hybrid’ mtDNA genome in the sense that it has both a highly compact gene arrangement and a single extensive repeat region. This bi-partite structure occurs in the unicellular protist *Rhodomonas salina*, a cryptophyte alga. Cryptophytes (or cryptomonads) owe their name to the fact that their cells harbour remnants of a second eukaryote. The vestigial structures include a chloroplast and a small nucleus [nucleomorph (11,12)]. Molecular phylogenies indicate that cryptophytes are among the earliest-diverging and slowest-evolving eukaryotic lineages (13,14). Therefore, its
mtDNA might have maintained certain ancestral features that have been lost in other, more diverged species. Here, we report the mitochondrial genome organization of *R. salina*, including detailed analysis of its remarkable gene complement, its bacteria-like gene order and the unusual presence of a highly complex intergenic repeat region. Furthermore, we address how the repeat region might have emerged and which biological role it may play.

**MATERIALS AND METHODS**

**Genome sequence techniques**

*R. salina* [formerly *Pyrenomonas salina* (15)] obtained from CCAM (Culture Collection of Algae Marburg, http://staff-www.uni-marburg.de/~cellbio/welcomeframe.html) was grown on F/2 medium (for growth medium recipe, http://ccam.bioinfo.rpi.edu/applications/staff-www.uni-marburg.de/ccam/welcomeframe.html) under permanent cool-white fluorescent light. When reaching the late logarithmic growth phase (after ~2 weeks), cells were harvested by centrifugation, and mtDNA was purified, mechanically sheared, cloned and sequenced essentially as described previously (16). The complete mtDNA sequence is available in GenBank (accession no. NC_002572).

**Bioinformatics analysis**

Individual direct and inverted repeats within the *R. salina* mtDNA genome were identified using TRIPlots (developed by A.H., http://megasun.bch.umontreal.ca/People/ahauth/tools). Complete analysis of the repetitive region required five different similarity measures using various window (w) and maximum nucleotide mismatch (m) combinations: (i) w = 70, m = 20; (ii) w = 50, m = 15; (iii) w = 20, m = 2; (iv) w = 15, m = 1; and (v) w = 10, m = 0. Manual compilation combined individual measures to define the final structure of the repeat region.

For prediction of probable origin and terminus of replication, the cumulative GC skew technique was used that measures the asymmetric strand distribution of G and C for individual fixed-length windows along the sequence, i.e. (G–C)/(G+C), and then sums the scores across the sequence [(17), http://bioinformatics.upmc.edu/GCSkew.html]. Cumulative scores plotted along the sequence indicate probable origin (local minimum) and terminus (local maximum) of replication.

For stem–loop structure prediction and calculation of minimum Gibbs free energies in DNA, we utilized mFold (18,19). All web analyses (http://www.bioinfo.rpi.edu/applications/mfold/) used default parameters that include a temperature of 37°C. Here, we considered only stem–loop structures predicted to have a ΔG° ≤ −20 kcal/mol, although numerous smaller and less stable structures occur in the analyzed repeat region as well.

Promoter search using the TRIPlots software mentioned above identified direct and inverted repeats between the region upstream of the rRNA cluster and within the repeat region centred around an 8 nt sequence, TAAAAAATAT. Subsequent analysis yielded numerous occurrences of this octamer (simple pattern search tool developed by A. Hauth, unpublished). Using the occurrences exclusively upstream of *rns*, the motif was expanded to the decamer TAAAAAATAT. Subsequent observation noted a 13 nt TAAAAAATATGGT sequence in each of the three predicted promoter locations.

**RESULTS**

**Physical properties, gene content and overall organization of *R. salina* mtDNA**

The mtDNA of *R. salina* maps as a circular molecule of 48,063 bp; however, like that of most eukaryotes, it is probably organized in linear, multicentric concatamers (20–23). The genetic map (Figure 1) shows exceptionally dense packing of the 69 genes, 9 of which overlap by 1–39 bp. The two largest non-coding stretches of the genome are a 4.7 kb long region containing repeats (described in more detail below) and a 505 bp long region located between *rps19* and *cox1* without obvious structural features.

*R. salina* mtDNA codes for the small subunit (SSU) and large subunit (LSU) rRNAs, 27 transfer RNAs (tRNAs), and 40 proteins, including 18 components of the respiratory chain, 5 ATP synthase subunits, 15 ribosomal proteins and 2 subunits of the tat translocase. Intriguingly, two separate genes encode the Nad10 protein of *R. salina* (Figure 1); *nad10* a specifies a short N-terminal portion of the protein while the residual 9/10 is encoded by *nad10* b. We assume that the gene modules are transcribed and translated separately because there is no evidence for trans-splicing such as group II intron-like sequences adjacent to the gene modules (24).

Four classes of *R. salina* mitochondrial genes have counterparts in only a few other protist and plant taxa (Figure 1). First, the tatA gene (coding for a subunit of tat translocase) has been recognized so far only in jakobid mtDNAs (25). Second, *rps1* has been found to date only in jakobids and malawaiomnads [for reviews see (26,27)]. Third, genes encoding certain subunits of NADH dehydrogenase subunits and ATPase (*nad8*, *nad10*, *nad11* and *atp1*) are rarely found in eukaryotes. Fourth, subunits of succinate:ubiquinone oxidoreductase (*sdh3* and *sdh4*) are present in only a few green algae and the liverwort *Marchantia polymorpha* (28–31), in several red algae (5,32–34) and in all jakobid flagellates including *Reclinomonas americana* (3,35).

**Vestiges of bacterial operon structures**

Comparisons of mitochondrial gene order reveal vestiges of bacterial operon structure in *R. salina* (Figure 2). The ribosomal protein clusters *rps3*-rpl16-rpl14-rpl5-rps14 and *rps8*-rpl6-rps13-rps11 in *R. salina* mtDNA preserve the same relative gene order as in the adjacent S10, *spc*, and α operons of *Escherichia coli*, which are also maintained with little variation in the gene-rich mtDNAs of several other protists and the land plant *M. polymorpha*. Other indications of operon-like regulation of protein expression are clusters of genes coding for subunits of the same enzyme complex (Figure 1). Note that the mitochondrial ribosomal protein clusters are highly reduced in the closest unicellular relative of animals [e.g. the choanoflagellate *Monosiga brevicollis* (10)] and that only a single gene (*rps3*) was retained in some but not all fungi (36).

**Mitochondrial rRNA and tRNA genes**

The mitochondrial genome of *R. salina* encodes euabacteria-like LSU and SSU rRNAs. Their predicted sizes (2663 and
1482 nt, respectively) are shorter than those of their E.coli counterparts (2904 and 1542 nt, respectively), which reflects size reduction in the variable regions, and reduction and/or lack of a few small helices in more conserved regions of the R.salina mitochondrial rRNAs. Otherwise, the secondary structures of the mitochondrial and eubacterial homologs are essentially superimposable. Secondary structures are provided through GOBASE, the Organelle Genome Database [(37), http://megasun.bch.umontreal.ca/gobase].

R.salina mtDNA encodes 27 tRNA genes, most of which cluster in groups of up to 5 genes. All tRNA sequences have standard cloverleaf secondary structures with only insignificant departures from the bacterial consensus structure. Because four-codon tRNA families are decoded by a single tRNA in mitochondria (i.e. by ‘super-wobble’ of U-34), a total of 25 tRNAs are sufficient to recognize all codons. Therefore, the occurrence of two additional tRNA genes is unexpected. The redundant tRNA^Arg(GCG) might translate CGC codons more efficiently than tRNA Arg(UCG) does.

It is more difficult to rationalize the existence of mitochondrial encoded tRNA^Ile(UAU). One explanation is that the U-34 of the anticodon could be modified in a way to restrict efficient pairing with G residues (i.e. to decode methionine codons), as seen in some animal mtDNAs where the U-34 modification apparently compensates for the lack of separate genes for initiator and elongator methionine tRNAs (38). This rationalization does not apply to R.salina mtDNA because it contains both tRNA^Met genes (Figure 1). Alternatively, the U-34 of the anticodon could be modified in a way to restrict pairing to A residues (39) and thus decode...
AUA-ile codons more efficiently than does tRNA\textsuperscript{ile}(CAU). Irrespective of the precise specificity of the \textit{R. salina} tRNA\textsuperscript{ile}(UAU), similarity searches reveal a high degree of identity with mitochondrial tRNA Phe from both \textit{R. salina} (Figure 3) and several other species. This is an apparent case of gene duplication followed by divergence of one of the copies to a different function, a mechanism, termed ‘authentity theft,’ that was recently identified in sponge mtDNAs (40).

Introns

Two introns of length 4091 and 2482 bp are inserted into the \textit{R. salina} \textit{cox1} gene. These introns belong to group II and harbour typical open reading frames (ORFs) of the so-called RT type (41), which are characterized by three distinctive protein domains: Reverse Transcriptase, maturase and a C-terminal DNA binding domain (42). Intron secondary structures and ORFs are most similar to those of \textit{cox1} introns in the brown alga \textit{Pylaiella littoralis}, a class that is most widespread across mitochondria (43).

Large, repeat-containing, intergenic region

The large, 4.7 kb-long, intergenic region in \textit{R. salina} mtDNA consists predominantly of an elaborate, 4.5 kb repeat region. Our analysis using length, conservation and Gibbs free energies parameters (see Materials and Methods) defines numerous direct and inverted repeats that together form a complex, multi-tiered, regular structure (Figure 4A). Distinct repeat units of length 36–693 bp (denoted a–f) occur 2–31 times as complete or partial instances (Figure 4B, Supplementary Table 1) and account for the majority of direct, inverted, palindromic and tandem repeats within the region (Figure 5). Units a–e recur in the same relative order within each of

![Figure 2](image-url). Conservation of ribosomal protein gene organization. Gene order found in mtDNAs is compared with that of the contiguous bacterial \textit{str}, \textit{S10}, \textit{spc} and alpha operons of \textit{E.coli} and \textit{Rickettsia prowazekii}. Solid lines connect adjacent genes and insertion of several additional genes are indicated by triangles. Bacterial data (\textit{E.coli}, \textit{R. prowazekii}) are shaded. \textit{E.coli} (accession no. NC\textunderscore 000913); \textit{R. prowazekii} (accession no. NC\textunderscore 000963); \textit{R. americana} (accession no. NC\textunderscore 001823); \textit{J.libera} (G. Burger, B. F. Lang and M. W. Gray, unpublished data); \textit{M.polymorpha} (accession no. NC\textunderscore 001660); \textit{A.castellanii} (accession no. NC\textunderscore 001637); \textit{P.infestans} (accession no. NC\textunderscore 002387); \textit{R. salina} (this report, accession no. NC\textunderscore 001823); \textit{M.brevicollis} (accession no. NC\textunderscore 004309). Data from \textit{E.coli} were retrieved from NCBI’s complete genome section. Data from \textit{Rickettsia} and protists were retrieved from GOBASE (http://megasun.bch.umontreal.ca/gobase).
five large sections (denoted 1, 2, 3, and 1', 2'); the prime mark indicates reverse complementation; Figure 4C). A large inverted repeat encompasses ~3 kb spanning sections 1 + 2 and 1' + 2'. In fact, the only unique sequence in the entire repeat region is a 112 bp sequence (denoted u) located at the inversion point.

Tandem repeats are the major repeat arrangement within this large intergenic region (aside from the large inverted repeat previously mentioned). Three regions consist of consecutive copies of unit b or d (Figure 4C). Unit b forms regions with 5.3 and 6.3 consecutive copies in sections 1 and 1'. Twenty-eight tandem copies of unit d make up a region in section 3, one having an imperfect superstructure composed of two types of unit d: type I is 36 bp long, while type II extends the same pattern 14 bp further (Supplementary Table 1). In addition, sections exhibit a tandem superstructure with each of sections 1–3 and 1'–2' acting as 'copies' in a higher-order tandem repeat.

Several types of palindromic sequences (in addition to the large inverted repeat mentioned above) occur in this long intergenic region of *Rhodomonas* mtDNA (Figure 4D). The first type (denoted z) has a very low G+C content (25.4%), consists of two inverted copies of unit d and recurs as divergent instances in three different sections at the position where d abuts unit e (z3, z2 and z2' in Figure 4A and D). The other three types of palindromes (denoted w–y) have an unusually high G+C content (53–88%), are perfectly conserved and occur in unit e proximal to the inversion point (u, see above; w1', x1' and y1' in Figure 4A and D). Free energy calculations (18,19) indicate that these four loop structures in *vivo* (for details see Figure 4D). Notably, four hairpins (z2, z2', w1 and w1') have a tertiary-stabilized GTAA (GNRA) tetraloop albeit without a G–C closing base pair (44,45).

Alignment and comparison of recurring sequences within the repetitive region of *R. salina* mtDNA indicates an extraordinarily high degree of sequence similarity (Supplementary Table 2). Further details regarding substitution, insertion and deletion rates are provided in the Supplementary Material.

**Replication and transcription**

Genes are encoded on both DNA strands and their orientation suggests the presence of two transcription units (~12 and 22 kb in length), starting from the repeat region in both directions. Both tandem repeats and stem–loop structures described above might play a role in transcription and/or replication, as suggested for vertebrate animal mtDNAs (46–50). Similarly, in the mtDNA of the red alga *Chondrus crispus*, transcription has been shown to initiate at a bidirectional promoter that is close to a palindromic repeat (51), a feature also present in the red alga *Porphyra purpurea* (5).

We searched in *silico* for potential promoters at three positions in *R. salina* mtDNA (see Materials and Methods): upstream of both inferred primary transcription units (within the repeat region) and upstream of the rRNA gene cluster (rns, rnl), a postulated secondary transcription unit to assure high expression levels of tRNAs (arrows in Figure 1). Intriguingly, we found a potential 13 nt promoter sequence TAAAAAATATGTGT that is located exclusively upstream of rns and in the repeat region upstream of both inferred transcription units (within unit d and the z palindromes of the repeat region), but nowhere else in the genome. The proposed promoter for the transcript to the ‘left’ of the repeat region occurs in each copy of the d tandem repeat in section 3. The one to the ‘right’ occurs within the leftmost copy of unit d in section 2'. Relaxing the promoter motif to the decamer TAAAAATAT allows inclusion of all unit d copies throughout the repeat region, including shadow copies in unit e near the u inversion.

![Figure 3](https://example.com/figure3.png)

**Figure 3.** High sequence similarity of tRNA^{Phe}(GAA) and tRNA^{Ile/Met}(UAU) differ by only 7 nt (indicated in boldface), suggesting a recent gene duplication. Both structures are consistent with features of canonical tRNAs that are under evolutionary selection.
Figure 4. Repeat and hairpin structures in the 4.7 kb long intergenic region of the *R. salina* mtDNA. (A) The repeat region spans positions 20210–24672. Five recurring repeat units (a–e, shown as colored arrows/triangles) form five sections (1, 2, 3, 1' and 2') bounded by a small repeat unit (f). The arrows/triangles indicate the relative order of repeat units and their orientation. A major inversion switches the orientation of the last two, relative to the first three sections leaving a small 112 bp unique sequence (u) at the inversion point. Nine potential stem–loop structures (w–z) appear as a hairpin symbol and represent two consecutive, inverted occurrences of the same arrow, a palindrome. Consecutive, direct occurrences of the same repeat unit indicate a tandem repeat. (B) A unique letter and unique arrow/triangle represents each recurring unit in the repetitive region. Five major repeat units (a–e) compose the main recurring structure bounded by a sixth repeat unit (f). Other units include a partial representation of b (o) and four palindromic units (w–z). (C) Magnified representation of all five sections with final two inverted to highlight similar structure. Within sections, repeat units recur in the same relative order albeit as complete, partial, palindromic or tandem occurrences. The tandem repeat in section 3 contains 28 copies of d and the ones in section 1 and 1' has 5.3 and 6.5 copies of b, respectively. Below each d occurrence, specific types (I and II) appear. Nucleotides that differ from the majority are shown above each section with the exception of the z palindromes [for sequence and potential secondary structure of palindromes, see (D)]. Nucleotides at two positions in each of unit a and unit c are shown in all a and c occurrences as they are central to the evolutionary history of the repetitive region. (D) Several palindromic sequences that exhibit a strong potential to form hairpin secondary structures are shown. Three divergent z palindromes are shown (ΔG37 of −14.1, −15.5 and −24.5 kcal/mol for z3, z2 and z2', respectively); arrows indicate nucleotide divergence from z2'. Palindromes w–y recur with complete identity in each unit e (ΔG37 of −26.6, −50.8 and −38.2 kcal/mol, respectively).
point although the decamer also occurs once within a coxl intron albeit in the ‘wrong’ orientation. Notably, the z palindrome in section 3 and z1 contain a copy of either the 10 or 13 nt promoter motif within the stem of the predicted hairpin (z3 and z1 in Figure 4D).

In some mitochondrial systems, origins of replication have been associated with both transcription initiation and repeat regions. Therefore, we asked the question whether the large intergenic region in Rhodomonas mtDNA might coincide with the replication start site of this genome. One approach to predict replication origin and terminus is to measure the asymmetric strand distribution of G and C. We determined this distribution using cumulative GC skew [(17), http://bioinformatics.upmc.edu/GCSkew.html] that can indicate both a putative replication origin (local minimum) and terminus (local maximum). Indeed, in Rhodomonas mtDNA, a local minimum is proximal to the large, intergenic, repeat-containing region and a local maximum occurs where the two opposing directions of transcription meet, thus corroborating the predicted location of the terminus of replication (Figure 6). The precise origin and terminus remain to be mapped by biochemical methods, particularly whether the replication origin is to the ‘left’, to the ‘right’ or within the repeat region.

**DISCUSSION**

**Primitive features of the Rhodomonas mitochondrial genome**

With 40 protein genes, the mtDNA of R. salina is more gene-rich than that of animals, fungi and plants. Notably, mitochondrion-encoded rpsl and tatA are otherwise found exclusively in jakobids and/or malawimonads, believed to be the most primitive mitochondrial eukaryotes known. Other ancestral features of R. salina mtDNA include bacteria-like tRNA and rRNA structures and ribosomal protein gene clusters that resemble bacterial operons. Molecular phylogenies based on mitochondrion-encoded protein sequences place Rhodomonas basally in the eukaryotic tree, but without significant support (results not shown). More sequence data and broader taxon sampling will be required to substantiate the view that cryptophytes are an early diverging clade.

**Structure and extent of repeat regions in mitochondrial genomes**

The R. salina repetitive region has a complex, regular structure containing numerous well conserved direct and inverted repeats (Figure 4A–C). Most conspicuous is a large, near perfect, 3 kb palindrome that contains several smaller, yet strong hairpins (Figure 4D). For mtDNA, this region is extremely unusual as it not only is large in size but also has a complex repetitive structure and a high conservation of both sequence and secondary structure.

Densely packed genes and a single, long, repeat-containing intergenic region is also found in mtDNA of a few other taxa. The repeat region of the green alga Pedinomonas minor (52) is very different from the one in Rhodomonas. It is twice as large (9 kb) and is made up of about twice as many (i.e. 13) distinct families of repeat motifs, with an even broader size range (6–389 bp). Greater sequence variation (up to 25%) occurs
among the copies of a given repeat family, and portions form a highly irregular structure of first-, second- and third-order patterns.

The single intergenic region in animals contains control signals for transcription and replication (‘control region’) and is typically 1 kb long [e.g. in humans (53)] but expands to as large as 10 kb [e.g. weevils (54)] due to variable number tandem repeats [reviewed in (49)]. Within a species, these repeats contain well conserved motif copies, but have variable copy numbers both within a population and between generations: an indication of ongoing expansion/contraction of this region. Closely related species often have a repeat region in the same location but with diverging sequence pattern and a pattern size quite similar in some clades [e.g. 151 ± 27 bp in lagomorphs (rabbits and hares) (55)] or widely varying in others [e.g. 9–287 bp in Crocodylidae (56)].

Finally, a distinct type of long, non-coding repeat regions is the inverted telomeric repeats present at both ends of many linear mitochondrial chromosomes. For example, telomeres of ~200 bp to ~11 kb long in the ciliate Tetrahymena consist of a tandemly arranged short motif {8–40 copies of a 31 bp in Tetrahymena pyriformis (57), ~190 copies of an ~52 bp in Tetrahymena malaccensis (58). In addition, those in the yeast Candida parapsilosis are composed of a much longer recurring motif [e.g. 738 bp (59)].

Formation, expansion and contraction of massive tandem repeats probably take place by slipped mispairing during replication, while repeat inversions seem to involve strand-switching. Experimental evidence for the underlying molecular mechanism is discussed in more detail in the Supplementary Material.

Here, we focused on massed repeat regions in mtDNA. It should be noted that numerous instances of dispersed repeats including recombinationally active repeats and gene duplications have been reported in mtDNAs of plants, fungi, chlorophyte algae, as well as others that are not addressed in the context of this study.

Evolutionary history of repeat regions in Rhodomonas mtDNA

The regular structure of the five sections in the repetitive region of Rhodomonas mtDNA suggests two expansion events starting from a hypothetical, ancestral sequence (Figure 7): duplication/inversion (e.g. strand-switching) to create $1 + 2$ and $2' + 1'$, and tandem duplication (e.g. by slipped strand mispairing) to create sections $2, 2'$ and $3$ (for similarity see Figure 4C and Supplementary Table 1). The order of the two events is unclear with competing arguments supporting alternative histories. The most parsimonist is a two-step history that first duplicates and inverts the seed sequence to form sections $1'$ and $2'$, and then duplicates in tandem section $2$ to yield $2$ and $3$ (Figure 7, i–ii). This scenario requires that selective pressure, e.g. a functional constraint, maintains the high degree of similarity within the inversion. Alternatively, a three-step history first duplicates in tandem section $2$ to yield $2$ and $3$, then duplicates and inverts the entire repeat region to add sections $1', 2'$ and $3'$ and last excises section $3'$ (Figure 7, i–III). In this scenario, the excision (e.g. a deletion caused by slipped strand mispairing) acts as a correction event. Although more steps are necessary, this latter history implies that the high sequence similarity between $1 + 2$ and $1' + 2'$ is not due to selective pressure, but rather to a recent duplication/inversion event.

Intriguingly, a tandem repeat region is present in a similar position in the mtDNA Guillardia theta, a cryptophyte that is distantly related to Rhodomonas (S. Douglas, T. Cavalier-Smith and U. Maier, unpublished data). This corroborates the notion that the Rhodomonas repeat region is ancient, thus pointing to a biological role of this large intergenic genome portion.

Biological role of the Rhodomonas repeat region

In animal mtDNAs, control region tandem repeats contain H-strand and L-strand transcription promoters in each copy of the repeats (55,60). The repeats found in lagomorph mtDNAs contain not only promoters but also a 20 bp sequence conserved across the group, which could represent an element involved in replication initiation (55). Similarly, terminal inverted repeats of linear mitochondrial chromosomes have been implicated in replication initiation (61,62) and in circularization of linear molecules prior to replication (63).

Therefore, we propose that the repeat region in Rhodomonas mtDNA is involved in the control of transcription and replication initiation, with several findings supporting this view. First, asymmetric strand distribution of G and C indicates proximity of the intergenic repeat-containing region to a replication origin (Figure 6). As shown in animal mtDNAs, control region repeats occur proximal to the replication origin and may even bind regulatory proteins (55).

Second, two of the three predicted promoters (those upstream of the primary transcription units) fall within the
repeat region. The third promoter lies upstream of the rRNA cluster in order to produce the high expression levels necessary for these rRNAs. The putative promoter motif occurs exclusively in these three locations and is included in a 28-copy tandem repeat. In several mitochondrial genomes, promoters have been predicted to reside in tandem repeats (49,50).

Third, promoters have been demonstrated experimentally to be associated with hairpin structures (46–48,51). Several potentially stable secondary structures are evident within the repeat region, but not elsewhere in *Rhodomonas* mtDNA (Figure 4D). The z palindromes have the potential to form a hairpin structure with a putative transcription promoter in the stem.

**OUTLOOK**

The repeat region of the *R. salina* mtDNA occupies a total of 10% of the genome: an expensive addition to an otherwise highly compact genome. If selective pressure minimized non-coding sequence in 90% of the genome, the massive repeat region is unlikely to be parasitic but rather, plays an important biological role. Comparative analysis of further cryptophyte mitochondrial genomes will be instrumental in testing the predictions presented here, inferred from *Rhodomonas* mtDNA. Ultimate testing of these predictions involves biochemical characterization of mitochondrial replication and transcription initiation. Such experiments will require optimization of culture conditions and subcellular separation methods in *Rhodomonas*, work ongoing in our laboratories.

**SUPPLEMENTARY MATERIAL**

Supplementary Material is available at NAR Online.

**ACKNOWLEDGEMENTS**

The authors would like to thank Isabelle Plante for excellent technical assistance. This work was supported by the Canadian Institutes of Health Research (CIHR; SP-34 and MOP-15331) and the Deutsche Forschungsgemeinschaft, Germany. We acknowledge salary and interaction support from the Canadian Institute for Advanced Research (CIAR; B.F.L. and G.B.) and a strategic training fellowship in bioinformatics from the CIHR-Genetics Institute (A.H.). Laboratory equipment and informatics infrastructure was financed in part by Genome Quebec/Canada. Funding to pay the Open Access publication charges for this article was provided by the CIHR.

Conflict of interest statement. None declared.

**REFERENCES**


