# Genome structure and gene content in protist mitochondrial DNAs

**Michael W. Gray[1,*], B. Franz Lang[2], Robert Cedergren[2], G. Brian Golding[5], Claude Lemieux[6], David Sankoff[3], Monique Turmel[6], Nicolas Brossard[4], Eric Delage[2], Tim G. Littlejohn[4,+], Isabelle Plante[4], Pierre Rioux[4], Diane Saint-Louis[4], Yun Zhu[4] and Gertraud Burger[2,4]**

Program in Evolutionary Biology, Canadian Institute for Advanced Research, [1]Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia B3H 4H7, Canada, [2]Département de Biochimie, [3]Centre de Recherche Mathématique and [4]OGMP Sequencing Unit, Université de Montréal, Montréal, Québec H3C 3J7, Canada, [5]Department of Biology, McMaster University, Hamilton, Ontario L8S 4K1, Canada and [6]Département de Biochimie, Université Laval, Québec, Québec G1K 7P4, Canada

Received October 23, 1997; Revised and Accepted November 21, 1997

## ABSTRACT

**Although the collection of completely sequenced mitochondrial genomes is expanding rapidly, only recently has a phylogenetically broad representation of mtDNA sequences from protists (mostly unicellular eukaryotes) become available. This review surveys the 23 complete protist mtDNA sequences that have been determined to date, commenting on such aspects as mitochondrial genome structure, gene content, ribosomal RNA, introns, transfer RNAs and the genetic code and phylogenetic implications. We also illustrate the utility of a comparative genomics approach to gene identification by providing evidence that *orfB* in plant and protist mtDNAs is the homolog of *atp8*, the gene in animal and fungal mtDNA that encodes subunit 8 of the $F_0$ portion of mitochondrial ATP synthase. Although several protist mtDNAs, like those of animals and most fungi, are seen to be highly derived, others appear to be have retained a number of features of the ancestral, proto-mitochondrial genome. Some of these ancestral features are also shared with plant mtDNA, although the latter have evidently expanded considerably in size, if not in gene content, in the course of evolution. Comparative analysis of protist mtDNAs is providing a new perspective on mtDNA evolution: how the original mitochondrial genome was organized, what genes it contained, and in what ways it must have changed in different eukaryotic phyla.**

## INTRODUCTION

Mitochondrial DNA (mtDNA) is extraordinarily diverse in size, gene content and genome organization (1–5) and it is a daunting task to attempt to elucidate the mechanisms and reconstruct the pathways by which this evolutionary diversification has occurred. The preferred approach to answering such evolutionary questions is through comparative analysis of complete mtDNA sequences, which provides a genome-level perspective on such issues as what genes are present, how they are arranged, whether there are introns (and, if so, what types), how spacer sequences are distributed and how large they are, whether segments of the genome are repeated and other relevant information. Currently, 63 complete mtDNA sequences are available through public domain databases; however, the phylogenetic range that these sequences represent is both narrow and biased: 47 (75%) are from animal species (31 vertebrate, 16 invertebrate); five (8%) are from fungi; two (3%) are from plants; only nine (14%) are from protists, in spite of the fact that the latter group of organisms (mostly unicellular) comprises the bulk of the biological diversity of the eukaryotic lineage (6). This limited and highly non-representative data set has made it difficult to draw meaningful conclusions about the ancestral form of the mitochondrial genome, a necessary starting point for inferences about subsequent mitochondrial genome evolution.

To redress this imbalance, the Organelle Genome Megasequencing Program (OGMP) was established in 1992, having as a specific aim the systematic and comprehensive determination of complete protist mtDNA sequences. [Brief descriptions of the OGMP and two allied databases, the Protist Image Database (PID) and the Organelle Genome Database Project (GOBASE), appear at the end of this review]. At that time only three complete protist mitochondrial genome sequences had been published: the 6 kb mtDNA sequences of the apicomplexans *Plasmodium yoelii* (a rodent parasite) (7) and *Plasmodium falciparum* (the human malaria parasite) (8) and the 40 kb mtDNA sequence of the ciliate protozoan *Paramecium aurelia* (9). Partial but extensive mtDNA sequence information was also available for another ciliate protozoan, *Tetrahymena pyriformis*, several trypanosomatid protozoa (in the

*To whom correspondence should be addressed. Tel: +1 902 494 2521; Fax: +1 902 494 1355; Email: m.w.gray@dal.ca

[+]Present address: Australian National Genomic Information Service (ANGIS), University of Sydney, Sydney, New South Wales 2006, Australia

genera *Trypanosoma*, *Leishmania* and *Crithidia*) and the green alga (chlorophyte) *Chlamydomonas reinhardtii*. These limited data suggested that protist mtDNAs might be even more structurally variable than their counterparts in the multicellular eukaryotic lineages (1).

In the ensuing 5 years, a larger selection of complete protist mtDNA sequences has become available through the efforts of the OGMP, a complementary Fungal Mitochondrial Genome Project (FMGP) (5) and other research groups. This review summarizes and comments upon various aspects of protist mitochondrial genome structure, particularly gene content, that have emerged from these new sequences. In recent years comprehensive reviews of animal (10), fungal (5,11) and plant (12,13) mtDNAs have been published, but reviews of protist mtDNAs have been limited to specific groups, e.g. ciliates (14), trypanosomatids (15) and apicomplexans (16). Because protists encompass most of the phylogenetic breadth of the eukaryotic lineage and, by definition, contain a number of clades whose evolutionary depth exceeds that of the traditional animal, plant and fungal kingdoms, it is important to sample widely within this disparate assemblage to obtain a clear perspective on the range of mtDNA structural diversity in protists, in comparison with the more widely studied mitochondrial genomes from other eukaryotes. The data assembled here emphasize that most non-protist mtDNAs, particularly those of animals, are substantially derived relative to most of their protist counterparts, having lost many genes that are commonly still found in protist mitochondrial genomes. The compilation provided here better defines the properties of a typical ancestral (i.e. minimally diverged) protist mtDNA and allows us to suggest with greater confidence what genes were likely contained in the proto-mitochondrial genome (i.e. the last common ancestor of contemporary mitochondrial genomes).

## SCOPE OF THE REVIEW

Table 1 identifies the 23 complete protist mtDNA sequences that to our knowledge have been determined to date. These sequences encompass a reasonably broad selection of protist taxa, although they still represent only a fraction of recognized protist lineages (6). Nine of these sequences are in the public domain; the remainder are unpublished ones determined by the OGMP (eight), the FMGP (two) or other research groups (four). As well, we include complete mtDNA sequences from representative non-protists for purposes of comparison. Figure 1 displays the relative phylogenetic positions (to the extent that these can be inferred or proposed at present) of the protists listed in Table 1, together with other protist species, including future candidates selected by the OGMP for complete mtDNA sequencing.

## METHODOLOGY

### Data collection and analysis

In the case of complete mtDNA sequences published by other groups and deposited in the public domain we have used the standardized and corrected versions available in GOBASE (17; see below). Importantly, annotations accompanying these sequences have been unified with respect to gene and product nomenclature. These particular sequences have also been re-analyzed by us using informatics tools developed in-house and described below.

With the exception of BLAST (used for remote database searches) (18), FASTA (used for detailed sequence comparison) (19) and NIP (the Staden nucleotide sequence analysis package) (20), all of the informatics tools employed for this compilation have been developed by the OGMP Sequencing Unit. Many of the programs make use of the OGMP 'masterfile' (mf) concept, an ASCII-based sequence file format that integrates nucleotide sequence, gene annotations and technical notes.

The sequence retrieval and analysis tools developed by the OGMP have for the most part been written in the Perl programing language. These tools include: BBLAST [batch mode BLAST search of the National Center for Biotechnology Information (NCBI) GenBank database]; BOB (BLAST output browser); FERRET, BADGER and CLEVER, retrieval tools used in conjunction with the NCBI Entrez database; GOBASE2MF [a program for converting from sequence records stored in Sybase tables of GOBASE (17) into mf format]; CLEANMF (used to verify sequence files in mf format as to annotation syntax and logic); PEPPER (for translation of protein coding sequences and extraction of non-coding regions); ONIP (command line interface to the Staden NIP program, used in the creation of codon usage tables of various gene classes); CN (sequence counter and checker). For compiling the body of data presented in Table 2, a number of wrapper scripts were written in the Bourne shell script language; these programs call upon the above tools and produce output files of appropriate layout. Scripts that use genome sequence files in mf format as input include: CODAT (calculation of A+T content of coding and non-coding regions); COTAB [creation of codon usage tables of three types of protein coding regions: genes, intronic open reading frames (ORFs) and unique ORFs]; BFASTA (batch FASTA search, used in comparing the protein sequences of two library files); TRNLIST (which creates a list of tRNA genes present in a genome). Further information about these programs is available at the OGMP website (see below).

## RESULTS AND DISCUSSION

### Mitochondrial genome structure

Complete sequence analysis has provided evidence of both circular mapping and linear mapping protist mtDNAs, with circular mapping genomes predominating (Table 2). Among the protist mitochondrial genomes characterized as linear, no common end structures have been identified (see Table 2 for details).

The protist mtDNAs listed in Table 2 have a median size of ~40 kb, ranging from 6 kb in the three apicomplexan species (the smallest known mtDNAs) to 77 kb in the choanoflagellate *Monosiga brevicollis*. The majority of protist mtDNAs are compact, gene-rich genomes, with few or no large non-coding regions. Intergenic spacers are generally small and sparse, accounting in nine cases for <10% of the mtDNA, with coding regions sometimes overlapping. In *Acanthamoeba castellanii*, *Dictyostelium discoideum*, *M.brevicollis*, *Chlamydomonas eugametos* and *Pedinomonas minor* all genes are transcribed from the same strand of the mtDNA; otherwise, more than one potential transcription unit is present in protist mitochondrial genomes.

The overall A+T content is high (>70% in 15 cases) in protist mtDNAs and is usually elevated in non-coding intergenic regions compared with coding regions (up to 1.2-fold higher in *M.brevicollis*

**Table 1.** Completely determined mitochondrial genome sequences

| Organism | Abbreviation | Classification | Accession No. | Source[b,c] |
|---|---|---|---|---|
| **Protists[a]** | | | | |
| *Acanthamoeba castellanii* | ACA | amoebid (rhizopod) | U12386 | OGMP (64) |
| *Cafeteria roenbergensis* | CRO | bicosoecid | Unpublished | OGMP |
| *Chlamydomonas eugametos* | CEU | green alga (chlorophyte) | AF008237 | RWL (76) |
| *Chlamydomonas reinhardtii* | CRE | green alga (chlorophyte) | U03843 | (77,78) |
| *Chondrus crispus* | CCR | red alga (rhodophyte) | Z47547 | (34) |
| *Chrysodidymus synuroideus* | CSY | heterokont alga (chrysophyte) | Unpublished | OGMP |
| *Dictyostelium discoideum* | DDI | slime mold | AB000109 | YT |
| *Malawimonas jakobiformis* | MJA | histionid (jakobid flagellate) | Unpublished | OGMP |
| *Monosiga brevicollis* | MBR | choanoflagellate | Unpublished | FMGP |
| *Nephroselmis olivacea* | NOL | green alga (chlorophyte) | Unpublished | CL/MT |
| *Ochromonas danica* | ODA | heterokont alga (chrysophyte) | Unpublished | OGMP |
| *Paramecium aurelia* | PAU | ciliate | X15917 | (9) |
| *Pedinomonas minor* | PMI | green alga (chlorophyte) | Unpublished | OMGP |
| *Phytophthora infestans* | PIN | oomycete | Unpublished | FMGP |
| *Plasmodium falciparum* | PFA | apicomplexan (human malaria parasite) | M76611 | (8) |
| *Plasmodium yoelii* | PYO | apicomplexan (rodent malaria parasite) | M29000 | (7) |
| *Porphyra purpurea* | PPU | red alga (rhodophyte) | Unpublished | OGMP |
| *Prototheca wickerhamii* | PWI | green alga (chlorophyte) | U02970 | OGMP (48) |
| *Reclinomonas americana* | RAM | histionid (jakobid flagellate) | AF007261 | OGMP (24) |
| *Rhodomonas salina* | RSA | cryptophyte alga (cryptomonad) | Unpublished | OGMP |
| *Tetrahymena pyriformis* | TPY | ciliate | Unpublished | OGMP |
| *Theileria parva* | TPA | apicomplexan (bovine parasite) | Z23263 | (79) |
| *Trypanosoma brucei* | TBR | trypanosomatid | Unpublished[d] | PJM |
| **Non-protists** | | | | |
| *Allomyces macrogynus* | AMA | fungus (chytridiomycete) | U41288 | FMGP (22) |
| *Homo sapiens* | HSA | animal (vertebrate) | V00662 | (80) |
| *Metridium senile* | MSE | animal (cnidarian) | AF000023 | DRW (81) |
| *Marchantia polymorpha* | MPO | plant (bryophyte) | M68929 | (23) |
| *Schizosaccharomyces pombe* | SPO | fungus (ascomycete) | X544121 | FMGP |

[a] Descriptions of and detailed information about many of these species may be found at the Protist Image Database (PID; URL http://megasun.bch.umontreal.ca/protists/).

[b] Where the complete sequence is reported in one or two papers, the references are listed here; otherwise, relevant citations can be obtained by consulting the annotation provided in the NCBI entry. Data from unpublished sequences were provided by: OGMP, Organelle Genome Megasequencing Program; FMGP, Fungal Mitochondrial Genome Project (URL http://megasun.bch.umontreal.ca/People/lang/FMGP); RWL, R.W.Lee (Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada); YT, Y.Tanaka (Institute of Biological Sciences, University of Tsukuba, Japan); CL/MT (C.Lemieux and M.Turmel, Département de Biochimie, Université Laval, Québec, Canada); PJM, P.J.Myler (Seattle Biomedical Research Institute, Seattle, WA); DRW, D.R.Wolstenholme (Department of Biology, University of Utah, Salt Lake City, UT).

[c] Data summaries and gene maps for the individual OGMP sequencing projects are available at URL http://megasun.bch.umontreal.ca/ogmp/.

[d] P.J.Myler, personal communication. A different sequence, assembled from a number of separate sources, is available as NCBI accession no. M94286. The sequence of the transcribed region of *Leishmania tarentolae* maxicircle DNA is also available (accession no. M101026).

mtDNA). The numbers in Table 2 suggest that, in general, protist mtDNAs have evolved in the direction of higher A+T content.

In animals, as exemplified by *Homo sapiens* and *Metridium senile* in Table 2, the evolutionary trend has clearly been toward a further compaction of the mitochondrial genome, both by loss of genes and by virtual elimination of intergenic spacers. Conversely, in plants (e.g. *Marchantia polymorpha*) the trend has been in the opposite direction, with the mtDNA tending to increase in size, primarily by acquisition of a large amount of apparently non-coding DNA of currently unknown origin and function (Table 2). In the recently sequenced 366 924 bp mitochondrial genome of the angiosperm *Arabidopsis thaliana* (21), fewer genes are encoded than are found in *M.polymorpha* mtDNA, which is half the size (Table 2); overall <10% of the *A.thaliana* mtDNA has an assigned coding function. A key question is how and why evolution has produced such divergent mitochondrial genome patterns in different eukaryotic lines.

## Gene content

In vertebrate animals, e.g. *H.sapiens* (Hsa), the mitochondrial genome contains genes for 13 inner mitochondrial membrane proteins involved in electron transport and coupled oxidative phosphorylation (*nad1-6* and *4L*, *cob*, *cox1-3* and *atp6* and *8*) (Table 3), as well as genes for large subunit (LSU) and small subunit (SSU) rRNAs (*rnl* and *rns* respectively; Table 4). This 'standard set' of mtDNA-encoded genes (plus *atp9*) is also found in fungal (e.g. *Allomyces macrogynus*, Ama) mtDNAs, except that certain ascomycete fungi (e.g. *Schizosaccharomyces pombe*, Spo) lack all *nad* genes. Animal and fungal mtDNAs do not encode a 5S rRNA (Table 4) nor, with the exception of *rps3* in *A.macrogynus* mtDNA (22), do they carry any ribosomal protein genes (Table 5). In land plant mtDNAs a few extra respiratory chain protein genes are found (e.g. *nad9* and *atp1* in *M.polymorpha*; Table 3); however, the most notable departure from animal and fungal mtDNAs is the presence in plant mtDNA of a set of ribosomal protein genes (Table 5) as well as a gene for 5S rRNA (*rrn5*; Table 4). In the case of *M.polymorpha* mtDNA several homologs of known mitochondrial genes (e.g. *sdh3,4* and *yejR,U,V*; Tables 3 and 6) were initially considered to be unique ORFs (23).

With respect to gene content, protist mtDNAs generally resemble plant rather than animal or fungal mtDNAs. The largest gene repertoire so far identified in any mtDNA is that found in the mitochondrial genome of the heterotrophic flagellate *Reclinomonas americana* (Ram, Tables 3–7; 24). Genes in the other sequenced mtDNAs are all subsets of the *R.americana* set, implying that the *R.americana* pattern is closest to the ancestral pattern of genes carried by the proto-mitochondrial genome (24). The *R.americana*

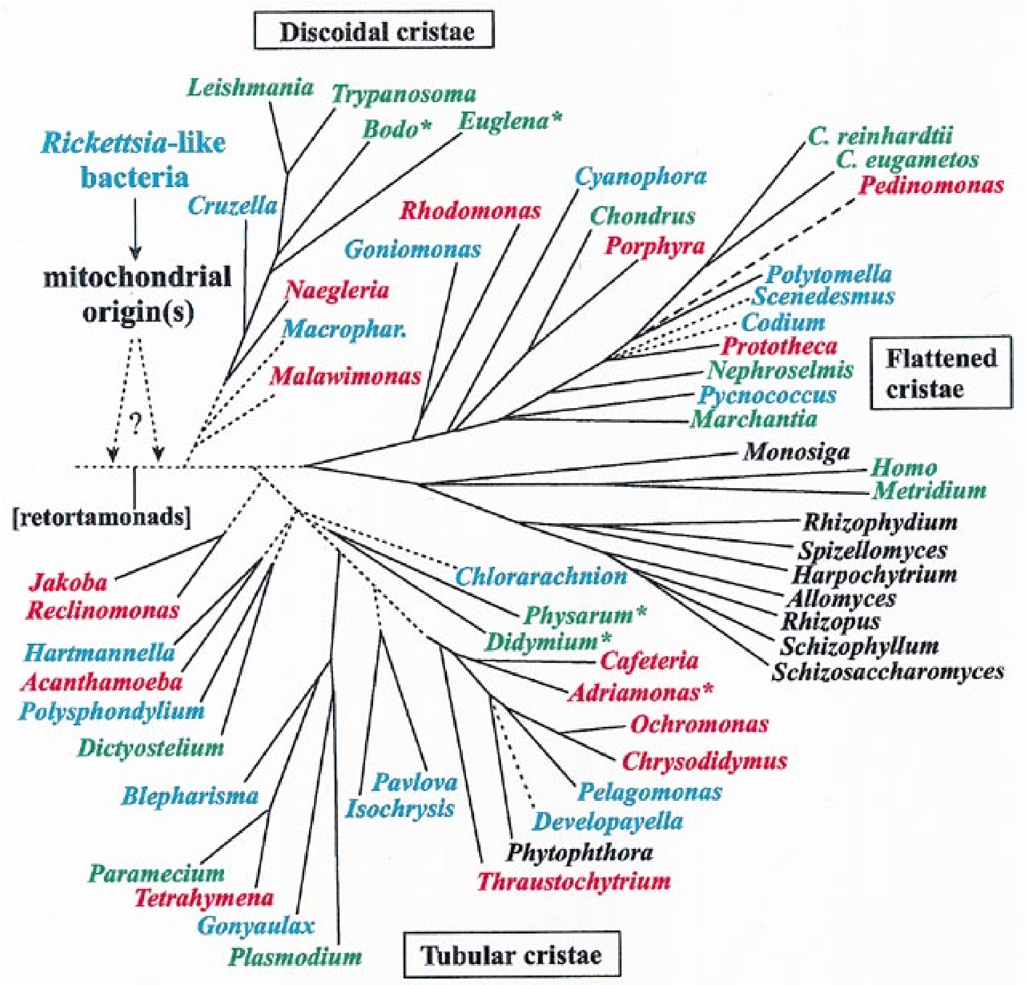


**Figure 1.** Phylogenetic hypothesis of the eukaryotic lineage based on ultrastructural and molecular data. Organisms are divided into three main groups distinguished by mitochondrial cristal shape (either discoidal, flattened or tubular). Unbroken lines indicate phylogenetic relationships that are firmly supported by available data; broken lines indicate uncertainties in phylogenetic placement, resolution of which will require additional data. Color coding of organismal genus names indicates mitochondrial genomes that have been completely (Table 1), almost completely (*Jakoba*, *Naegleria* and *Thraustochytrium*) or partially (*) sequenced by the OGMP (red), the FMGP (black) or other groups (green). Names in blue indicate those species whose mtDNAs are currently being sequenced by the OGMP or are future candidates for complete sequencing. Amitochondriate retortamonads are positioned at the base of the tree, with broken arrows denoting the endosymbiotic origin(s) of mitochondria from a *Rickettsia*-like eubacterium. *Macrophar.*, *Macropharyngomonas*.

results also indicate that gene loss (presumably by transfer to the nucleus) has occurred to different extents in different lineages (25), with many respiratory chain genes and almost all ribosomal protein genes having already been eliminated in the common ancestor of animal and fungal mtDNAs. In support of the view that *R.americana* mtDNA is ancestral (i.e. minimally diverged) is the highly eubacterial character of certain of its genes (e.g. *rnpB*, encoding the RNA component of RNase P) as well as the presence of putative eubacterial translation initiation signals (Shine–Dalgarno motifs; 24). In addition, as in the case of chloroplast genomes (3,26,27), *R.americana* mtDNA encodes subunits of a multi-component, eubacteria-like ($\alpha_2\beta\beta'$) core RNA polymerase. In contrast, in other eukaryotes the core mitochondrial RNA polymerase is a single polypeptide, nuclear DNA-encoded enzyme homologous to bacteriophage T3 and T7 RNA polymerases (28–32). Although *R.americana* mtDNA has a larger number of genes than other sequenced protist mtDNAs, it is notable that these additional genes are all involved in mitochondrial biogenesis and/or function.

The emerging data suggest that loss of particular genes from mtDNA happened a number of times, independently, in the course of mitochondrial genome evolution. For example, *sdh* genes have only been found so far (Table 3) in the mtDNA of a cryptophyte [*Rhodomonas salina* (33)], rhodophytes [the red algae *Porphyra purpurea* (33), *Chondrus crispus* (34) and *Cyanidium caldarium* (35)] and land plants [*M.polymorpha* (33,36)], as well as in *R.americana* mtDNA (24,33). These genes are not present in *A.thaliana* mtDNA (21) and so far have not been identified in other, partially sequenced angiosperm mitochondrial genomes. Considering the proposed phylogenetic positions of these lineages (Fig. 1) and the current limited distribution of mtDNA-encoded *sdh* genes, we infer that these genes must have been lost from mtDNA on different occasions (33).

**Table 2.** Characteristics of sequenced mitochondrial genomes

| Organism | Form[a] | Size (bp) | % Coding[b] | % Non-coding | % A+T | | |
|---|---|---|---|---|---|---|---|
| | | | | | Coding[b] | Non-coding | Total |
| **Protists** | | | | | | | |
| *Acanthamoeba castellanii* | C | 41,591 | 93.2 | 6.8 | 70.2 | 76.1 | 70.6 |
| *Cafeteria roenbergensis* | C | 43,159 | 96.5 | 3.5 | 72.4 | 83.2 | 72.7 |
| *Chlamydomonas eugametos* | C | 22,897 | 84.6 | 15.4 | 65.7 | 63.6 | 65.4 |
| *Chlamydomonas reinhardtii* | L | 15,758[c] | 83.1 | 16.9 | 54.9 | 54.5 | 54.8 |
| *Chondrus crispus* | C | 25,836 | 94.8 | 5.2 | 71.6 | 82.1 | 72.1 |
| *Chrysodidymus synuroideus* | C | 34,119 | 94.7 | 5.3 | 75.1 | 88.7 | 75.9 |
| *Dictyostelium discoideum* | C | 55,564 | 90.5 | 9.5 | 72.6 | 72.5 | 72.6 |
| *Malawimonas jakobiformis* | C | 47,325 | 88.5 | 11.5 | 72.4 | 85.1 | 73.8 |
| *Monosiga brevicollis* | C | 76,568 | 47.0 | 53.0 | 77.8 | 93.2 | 86.0 |
| *Nephroselmis olivacea* | C | 45,223 | 78.4 | 21.6 | 65.7 | 72.7 | 67.2 |
| *Ochromonas danica* | L | 41,035[d] | 89.5 | 10.5 | 73.3 | 78.0 | 73.8 |
| *Paramecium aurelia* | L | 40,469[e] | 86.7 | 13.3 | 58.1 | 62.8 | 58.8 |
| *Pedinomonas minor* | C | 25,137 | 60.9 | 39.1 | 76.3 | 80.2 | 77.8 |
| *Phytophthora infestans* | C | 37,957 | 90.1 | 9.9 | 76.5 | 88.6 | 77.7 |
| *Plasmodium falciparum* | L[f] | 5,966[g] | 76.0 | 24.0 | 69.3 | 65.6 | 68.4 |
| *Plasmodium yoelii* | L[f] | 5,952[g] | k | k | k | k | 68.9 |
| *Porphyra purpurea* | C | 36,753 | 90.8 | 9.2 | 66.0 | 72.0 | 66.5 |
| *Prototheca wickerhamii* | C | 55,328 | 70.6 | 29.4 | 69.9 | 84.6 | 74.2 |
| *Reclinomonas americana* | C | 69,034 | 91.3 | 8.7 | 72.8 | 85.1 | 73.9 |
| *Rhodomonas salina* | C | 48,063 | 85.2 | 14.8 | 69.8 | 72.5 | 70.2 |
| *Tetrahymena pyriformis* | L | 47,172[h] | 96.0 | 4.0 | 78.3 | 89.4 | 78.7 |
| *Theileria parva* | L[i] | 5,723[j] | k | k | k | k | 69.8 |
| *Trypanosoma brucei* | C | 22,289 | 63.6 | 36.4 | 73.7 | 81.6 | 76.6 |
| **Non-protists** | | | | | | | |
| *Allomyces macrogynus* | C | 57,473 | 77.4 | 22.6 | 63.0 | 51.8 | 60.5 |
| *Homo sapiens* | C | 16,569 | 92.7 | 7.3 | 55.8 | 52.2 | 55.6 |
| *Metridium senile* | C | 17,443 | 94.6 | 5.4 | 61.6 | 67.2 | 61.9 |
| *Marchantia polymorpha* | C | 186,609 | 56.1 | 33.9 | 56.4 | 59.8 | 57.6 |
| *Schizosaccharomyces pombe* | C | 19,431 | 89.2 | 10.8 | 69.0 | 77.6 | 69.9 |

[a] C, circular mapping; L, linear mapping.

[b] Includes identified genes, unidentified ORFs, introns and intron ORFs.

[c] Includes 492 bp subterminal inverted repeats and terminal 40 nt 3′ single-strand extensions (78).

[d] Includes 2208 bp terminal inverted repeats (OGMP, unpublished results).

[e] Sequence starts at the DNA replication initiation loop, which contains a tandem array of 11 34 bp A+T-rich repeat units. Termination sequence at the other end of the linear DNA (estimated to be ~200 bp) remains unsequenced (14).

[f] Head-to-tail tandem repeats of a 6 kb unit (82).

[g] Length of repeat unit.

[h] Excluding tandemly arrayed telomeric sequences (31 bp repeat unit) of variable length (OGMP, unpublished results).

[i] 7.1 kb DNA element containing incompletely characterized terminal inverted repeats (79).

[j] Excludes terminal inverted repeat sequences (residues 1–59 and 5783–5895 of Z23263).

[k] Identification of fragmented and scrambled rRNA coding modules (see Table 4) is incomplete for these genomes; for that reason the proportion of coding versus non-coding DNA cannot be calculated at present.

As the sorts of comparative data being generated by complete protist mtDNA sequencing continue to accumulate, we should be able to document more precisely the number and timing of individual instances of mitochondrial gene loss, many of which undoubtedly involve mitochondrion to nucleus gene transfer. Even now, the results suggest that gene flux from mitochondrial to nuclear genomes is not only a widespread and on-going phenomenon, but that it has been both more gradual and more frequent than previously appreciated. The *cox2* gene, as one example, appears to have been lost from mtDNA at least three times (see Table 3): in the lineage leading to the Apicomplexa, in the *Pedinomonas*/*Chlamydomonas* lineage of green algae and in certain legumes (dicotyledonous plants) (37,38).

Most protist mtDNAs contain a number of conserved but unidentified ORFs (Table 6). Especially notable in this regard are *ymf16* (which has been shown to code for a membrane protein of unknown function; 39) and *ymf39*, which are present in the mtDNA of many protists and plants (but not in animal or fungal mtDNA). However, most of the unidentified ORFs encountered during mitochondrial genome sequencing are unique: they do not match any sequence in the protein databases. Considering the nature and distribution of identified respiratory chain (Table 3) and ribosomal protein genes (Table 5), we suspect that at least some of these unidentified ORFs may represent highly diverged versions of known mtDNA-encoded genes, no longer recognizable by similarity searches. Additional comparative data should help to address this question and may ultimately permit the functional assignment of conserved ORFs, as in the case of *ymf19* (*orfB*; see below). Assuming that further gene assignments of this type can be made through this comparative approach, differences in protist mtDNA gene content could turn out to be less pronounced than they appear to be at the moment.

## Ribosomal RNA

With only a few exceptions, protist mtDNAs encode LSU and SSU rRNAs whose potential secondary structures deviate minimally from their eubacterial counterparts (OGMP, unpublished results). This corresponds to what has been observed with plant mitochondrial rRNAs, but stands in marked contrast to most fungal but particularly animal mitochondrial rRNAs (40,41). Clearly recognizable in most protist mitochondrial LSU rRNAs are the 5′- and 3′-terminal regions corresponding to the ‘5.8S’ and ‘4.5S’ domains of a eubacterial counterpart such as *Escherichia coli*

**Table 3.** Mitochondrial DNA-encoded genes involved in electron transport and coupled oxidative phosphorylation[a]

| | HSA | MSE | SPO | AMA | MBR | MPO[b] | PWI | NOL | CEU | CRE | PMI | CCR | PPU | RSA | ACA | DDI | ODA | CSY | PIN | CRO | PAU | TPY | PFA[c] | TBR[d] | MJA | RAM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Complex I (*nad*) | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | ■ | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■[e] | ■[e] | ○ | ■ | ■ | ■ |
| 2 | ■ | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ |
| 3 | ■ | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ■[f] | ■ | ■ |
| 4 | ■ | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ■ | ■ | ■ |
| 4L | ■ | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| 5 | ■ | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ■ | ■ | ■ |
| 6 | ■ | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| 7 | ○ | ○ | ○ | ○ | ○ | □[g] | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ■ | ○ | ■ |
| 8 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ○ | ■ |
| 9 | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ■ | ■ | ■ |
| 10 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ○ | ○ | ○ | ■ |
| 11 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ |
| Complex II (*sdh*) | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| 3 | ○ | ○ | ○ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| 4 | ○ | ○ | ○ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| Complex III (*cob*) | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| Complex IV (*cox*) | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■[h] | ■[h] | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| 2 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■[h] | ■[h] | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ■ | ■ | ■ |
| 3 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ | ■ | ■ |
| Complex V (*atp*) | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ |
| 3 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| 6 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ |
| 8[i] | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| 9 | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ |

[a] Full organism names are listed in Table 1. ■, gene present; □ pseudogene; ○ gene absent.
[b] *Arabidopsis thaliana* mtDNA (accession nos Y08501 and Y08502) lacks *sdh* genes but encodes a functional copy of *nad7* (21).
[c] Pyo and Tpa mtDNAs, which have the same gene content as Pfa mtDNA, are not listed in this table.
[d] The same genes are found in the maxicircle DNA of *Leishmania tarentolae* (accession no. M10126). Transcripts of trypanosomatid mitochondrial genes undergo post-transcriptional U addition/deletion RNA editing to generate translatable mRNAs (83).
[e] In both *T.pyriformis* and *P.aurelia* mitochondria the *nad1* gene is split into two pieces and rearranged (OGMP, unpublished results). In *T.pyriformis*, corresponding transcripts have been identified, one (*nad1_a*) encoding the N-terminal portion and the other (*nad1_b*) specifying the C-terminal portion of NADH dehydrogenase subunit 1 (J.Edqvist and M.W.Gray, unpublished results).
[f] Identification of *nad3* in trypanosomatid mtDNA (84) should be regarded as tentative (P.J.Myler, personal communication).
[g] Gene contains six in-frame TGA codons (23); transcript detected but not further processed (85).
[h] A single open reading frame (*cox1_cox2*) encodes both subunits 1 and 2 of cytochrome *c* oxidase in *A.castellanii* (86) and *D.discoideum* (87,88) mtDNAs.
[i] *orf172* (*ymf19*; 89) in *M.polymorpha* mtDNA and *orfB* in angiosperm mtDNA (see text).

23S rRNA. These terminal regions have largely been eliminated from animal mitochondrial LSU rRNAs (41). These observations reinforce the emerging view that the most ancestral (minimally derived) mitochondrial genomes will be found among the protists.

A minority of protist mtDNAs encode rRNA genes whose structure and/or the structure of their products is very unusual. The 9S (SSU) and 12S (LSU) mitochondrial rRNAs of trypanosomatid protozoa (e.g. *Leishmania tarentolae* and *Trypanosoma brucei*) are among the smallest and structurally most divergent of known rRNAs, having potential secondary structures in which only a few of the expected conserved structural elements are identifiable (40,41). Also unusual are the mitochondrial *rnl* genes of *Paramecium aurelia* (42,43), *Tetrahymena pyriformis* (43) and *Pedinomonas minor* (OGMP, unpublished results), which are split into two pieces that are separated in the genome and interspersed with other genes (Table 4). The *Pedinomonas* situation is particularly intriguing because a more extreme case of *rnl* fragmentation and scrambling is seen in the mtDNA of a phylogenetically later branching green algal genus, *Chlamydomonas* (44–46). Fragmented and dispersed rRNA gene elements, encoded on both strands of the mtDNA, have also been found in the small apicomplexan mtDNAs (8,47). Because most protist mtDNAs encode conventional, 16S-like and 23S-like rRNAs (the ancestral state), these deviant examples must represent derived patterns of mitochondrial rRNA gene structure and organization within the specific lineages in which they occur.

Like animal and fungal mtDNAs, most protist mtDNAs lack a 5S rRNA gene, the current exceptions (Table 4) being the chlorophyte algae *Prototheca wickerhamii* (48) and *Nephroselmis olivacea* (Nol) (M.Turmel, C.Otis and C.Lemieux, unpublished results), the red alga *C.crispus* (see Table 4, footnote g) and the jakobid flagellate *R.americana* (49). As in the case of *sdh* genes noted above, the sporadic phylogenetic distribution of mitochondrial *rrn5* suggests that this gene was lost from mtDNA a number of times.

**Table 4.** RNA-encoding genes in mtDNA[a]

| | HSA | MSE | SPO | AMA | MBR | MPO | PWI | NOL | CEU | CRE | PMI | CCR | PPU | RSA | ACA | DDI | ODA | CSY | PIN | CRO | PAU | TPY | PFA | TBR | MJA | RAM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | b | | | | | | | | | | | | | | | | | c | | | |
| ribosomal RNA | | | | | | | | | | | | | | | | | | | | | | | | | | |
| *rnl* | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■[d] | ■[d] | ■[e] | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■[e] | ■[e] | ■[d] | ■ | ■ | ■ |
| *rns* | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■[d] | ■[d] | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■[f] | ■[f] | ■[d] | ■ | ■ | ■ |
| *rrn5* | ❍ | ❍ | ❍ | ❍ | ❍ | ■ | ■ | ■ | ❍ | ❍ | ❍ | ■[g] | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ■ |
| RNase P RNA | | | | | | | | | | | | | | | | | | | | | | | | | | |
| *rnpB* | ❍ | ❍ | ■[h] | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ■ |
| guide RNAs[i] | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | j | - | - |
| other | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ■[k] | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |

[a] Full organism names are listed in Table 1. ■, gene present; ❍ gene absent.
[b] The same genes are present in *A.thaliana* mtDNA (21).
[c] Pyo and Tpa mtDNAs have the same gene content as Pfa mtDNA.
[d] Multiply split and rearranged *rnl* and *rns* genes → multiply fragmented LSU and SSU rRNAs (44–47).
[e] Split (2 piece) and rearranged *rnl* (42,43; OGMP, unpublished results).
[f] Split (2 piece) *rns* → split (2 piece) SSU rRNA (90,91).
[g] The original claim that *C.crispus* mtDNA encodes a 5S rRNA (34) has since been discounted (49; see also 4) However, re-analysis of the *C.crispus* mtDNA sequence has now revealed a gene for a *bona fide* 5S rRNA, different from the 5S rRNA-like structure originally proposed by Leblanc *et al.* (34). The *C.crispus rrn5* (complement of residues 16043–16152 in Z47547) is located between and in the same transcriptional orientation as *nad3* and *rps11* (G.Burger, unpublished results).
[h] B.F.Lang, unpublished results.
[i] Small RNAs that function in U addition/deletion RNA editing (83).
[j] The number of guide RNAs encoded by the *T.brucei* and *L.tarentolae* maxicircle DNAs is three and 15 respectively. For a compilation of trypanosomatid guide RNAs see http://www.biochem.mpg.de/~goeringe/gRNA/gRNAseqs.html).
[k] Gene encoding a 129 nt RNA of unknown function is located immediately downstream of *rnl* (Y.Tanaka, personal communication).

## Transfer RNAs and the genetic code

Complete sequencing of an organelle genome is the only way to determine unequivocally whether that genome encodes all of the tRNA species necessary to support organellar protein synthesis. Several protist mtDNAs (those of *M.brevicollis*, *P.wickerhamii*, *R.salina* and *Malawimonas jakobiformis* in Table 1) do appear to encode the minimal required tRNA set, if one allows that a single tRNA is able to decode the four-codon family specifying a given amino acid (see Table 7). However, in most cases, tRNAs recognizing one or more codons are evidently absent from the mitochondrial genome, and tRNA import from the cytosol is usually invoked as the mechanism for making up the deficit. Import of nuclear DNA-encoded cytosolic tRNAs into mitochondria is clearly required in the case of *A.castellanii*, *D.discoideum*, *P.aurelia*, *T.pyriformis*, *Chlamydomonas* spp. and *P.minor*, whose mtDNAs encode substantially fewer than the minimal required set (Table 7); in fact, import of tRNA into *Tetrahymena* mitochondria, long inferred on the basis of tRNA population studies (50), has recently been documented experimentally (51). No tRNA genes have been found in the mitochondrial genomes of apicomplexan or trypanosomatid protists, where import of a full set of tRNAs from the cytoplasm is assumed (52,53). The data in Table 7 indicate that mitochondrial tRNA import is not only likely to be widespread among protists [as it is also in plants (54) and several chytridiomycete fungi (5)], but that it emerged early in the evolution of the mitochondrial translation system, probably a number of times independently. Genes for certain tRNAs (e.g. Met and Trp) are encoded by the mitochondrial genomes of virtually all protists, whereas genes for other tRNAs (notably Thr) are found infrequently among protist mtDNAs (Table 7).

Several protist mitochondrial genomes, as well as that of *M.polymorpha*, lack only one or two of the minimal required set of tRNA genes. Again, in these cases it is generally held that import of cytosolic tRNAs makes up the deficit. Indeed, import into *M.polymorpha* mitochondria has recently been documented in the case of nucleus-encoded tRNA$^{\mathrm{Ile}}$(aau) (55) and tRNA$^{\mathrm{Thr}}$(agu) (56), genes for which have not been identified in *M.polymorpha* mtDNA (23). However, an alternative possibility that should be considered is that the anticodon sequence in a single mtDNA-encoded tRNA might be subject to partial editing, such that the unedited and edited versions accept different amino acids and pair with codons corresponding to these amino acids. Partial C→U editing of a tRNA$^{\mathrm{`Gly'}}$(gcc) to generate a tRNA$^{\mathrm{Asp}}$(guc) in opossum mitochondria (57) serves as a precedent for this possibility.

In *A.castellanii*, sequencing of the mtDNA has provided evidence of a novel type of tRNA editing that affects most of the mtDNA-encoded tRNAs (58–62; D.H.Price and M.W.Gray, unpublished results). This editing is confined to one or more of the first three positions at the 5′-end of the tRNA (62). Except for the mismatching in the acceptor stem that is corrected by this editing, the secondary structures of *Acanthamoeba* mitochondrial tRNAs are quite conventional (58–62). What appears to be the same type of mitochondrial tRNA editing has recently been documented in the chytridiomycete fungus *Spizellomyces punctatus* (63) and several other primitive fungi (B.F.Lang, unpublished results); moreover, in the case of tRNAs encoded by *D.discoideum* mtDNA secondary structure modeling strongly suggests that several of these undergo a similar type of editing. Orthodox cloverleaf secondary structures are the rule for mitochondrial tRNAs throughout the protists, one notable variant being an unusual tRNA$^{\mathrm{Met}}$ in *Tetrahymena* mitochondria (64). The structurally aberrant tRNAs characteristic of animal mitochondria (65,66) are therefore exceptional, representing a highly

**Table 5.** Ribosomal protein genes encoded by mtDNA[a]

| | Hsa | Mse | Spo | Ama | Mbr | Mpo[b] | Pwi | Nol | Ceu | Cre | Pmi | Ccr | Ppu | Rsa | Aca | Ddi | Oda | Csy | Pin | Cro | Pau | Tpy | Pfa[c] | Tbr[d] | Mja | Ram |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Small subunit (*rps*) | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | ○ | ○ | ○ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| 2 | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| 3 | ○ | ○ | ■[e] | □[f] | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ |
| 4 | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| 7 | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| 8 | ○ | ○ | ○ | ○ | ■ | ■ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| 10 | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| 11 | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| 12 | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ■ | ■ | ■ |
| 13 | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ |
| 14 | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ |
| 19 | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ■ | ○ | ○ | ■ | ■ |
| Large subunit (*rpl*) | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| 2 | ○ | ○ | ○ | ○ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ |
| 5 | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| 6 | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| 10 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| 11 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| 14 | ○ | ○ | ○ | ○ | ■ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ |
| 16 | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ |
| 18 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| 19 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| 20 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■[g] | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| 27 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| 31 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| 32 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| 34 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |

[a]Full organism names are listed in Table 1. ■ gene present; □ pseudogene; ○ gene absent. Small subunit-associated ribosomal proteins are also encoded by the mtDNAs of yeast (*Saccharomyces cerevisiae*; *var1*) and *Neurospora crassa* (S-5) (see table III in 2); however, these proteins share no obvious sequence similarity with any known eubacterial small subunit ribosomal protein.

[b]Several of these genes have not been identified in the completely sequenced *A.thaliana* mitochondrial genome (accession nos Y08501 and Y08502); these include *rps1*, *rps2*, *rps8*, *rps10*, *rps11*, *rps13* and *rpl6*. Two additional genes (*rps14* and *rps19*) are present as pseudogenes in *A.thaliana* mtDNA (21).

[c]Like the Pfa mitochondrial genome, Pyo and Tpa mtDNAs do not encode any ribosomal protein genes.

[d]Same ribosomal protein gene content in *L.tarentolae* maxicircle DNA (accession no. M10126).

[e]*orf227* (previously named *urfa*; 92); G.Burger and B.F.Lang, unpublished results.

[f]No transcript detected (22).

[g]Not reported in the original publication describing this genome (34).

derived form of mitochondrial tRNA which, nevertheless, is able to assume the required L-shaped tertiary structure (67).

In almost half of the protists listed in Table 7 we infer, on the basis of codon usage and the presence of a tRNA$^{Trp}$ having a CCA anticodon, that the mitochondrial translation system uses the standard genetic code, as is the case in land plants. In the remaining protists UGA appears to be decoded as tryptophan rather than as stop (Table 7), being the preferred Trp codon in all but *P.aurelia*; in fact, UGA is used almost exlusively to encode Trp in *M.brevicollis* and *T.pyriformis* mitochondria. From the phylogenetic distribution of this code variation it is evident that the change in UGA coding must have occurred on more than one occasion.

## Introns

Compared with plant mtDNA, protist mtDNAs seem to have remarkably few introns (Table 8). At least half of these genomes entirely lack group I and group II introns. So far, among the 23 completely sequenced protist mtDNAs listed in Table 1, group I introns have only been found (and then only in small numbers) in the amoeboid protozoa *A.castellanii* and *D.discoideum*, the green algae *P.wickerhamii*, *N.olivacea* and *C.eugametos* and the choanoflagellate *M.brevicollis*. *Prototheca wickerhamii* and *M.polymorpha* mtDNAs share with one another (and with fungal mtDNA) positionally equivalent and structurally homologous *cox1* introns, suggesting that these introns have been inherited vertically from a mitochondrial ancestor of fungi, green algae and plants (68). On the other hand, horizontal transfer of other group I introns is suggested by the fact that in the *rnl* gene of *A.castellanii* mtDNA and in the chloroplast DNA of certain *Chlamydomonas* species, several mobile group I introns are not only positionally identical, but have homologous intron core structures and intron ORFs (69).

Very few group II introns have been found in protist mtDNAs (a total of seven such introns in five of 23 completely sequenced protist mtDNAs). Again, we have some evidence suggesting acquisition of certain of these introns by horizontal transfer (OGMP, unpublished results), as appears also to be the case for certain group II introns found in the *rnl* gene of the brown alga *Pylaiella littoralis* (70). In our view the paucity of group II introns in protist mtDNAs coupled with their sporadic distribution and evidence of horizontal transfer makes it quite unlikely that there was a wholesale acquisition of group II introns by the eukaryotic cell via the α-proteobacteria-like proto-mitochondrial endosymbiont.

**Table 6.** Additional protein genes encoded by mtDNA[a]

| | HSA | MSE | SPO | AMA | MBR | MPO | PWI | NOL | CEU | CRE | PMI | CCR | PPU | RSA | ACA | DDI | ODA | CSY | PIN | CRO | PAU | TPY | PFA | TBR[b] | MJA | RAM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| transcription | | | | | | | | | | | | | | | | | | | | | | | | | | |
| *rpoA* | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| *rpoB* | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| *rpoC* | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| *rpoD* | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| translation | | | | | | | | | | | | | | | | | | | | | | | | | | |
| *tufA* | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| cyt. *c* biosynthesis | | | | | | | | | | | | | | | | | | | | | | | | | | |
| *yejR* (*ccl1*) | ○ | ○ | ○ | ○ | ○ | ■[c] | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ○ | ○ | ■ | ■ |
| *yejU* | ○ | ○ | ○ | ○ | ○ | ■[d] | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| *yejV* | ○ | ○ | ○ | ○ | ○ | ■[e] | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| *yejW* | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| cyt. oxidase assembly | | | | | | | | | | | | | | | | | | | | | | | | | | |
| *cox11* | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| protein transport | | | | | | | | | | | | | | | | | | | | | | | | | | |
| *secY* | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| conserved ORFs | | | | | | | | | | | | | | | | | | | | | | | | | | |
| *ymf16*[f] | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| *ymf39*[g] | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| other[h] | | | | | | | | | | | | | | | | | | | | | | | | | | |
| *dpo*[i] | ○ | ○ | ○ | ○ | ○ | □[j] | ○ | ○ | ○ | ○ | ○ | ○ | ■[k] | ○ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| *rtl*[l] | ○ | ○ | ○ | ○ | ○ | ■[m] | ○ | ○ | ○ | ■[n] | ○ | ○ | ■[o] | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| *end*[p] | ○ | ○ | ○ | ■[q] | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| unique ORFs[r] | 0 | 0 | 0 | 1 | 2 | 73[s] | 2 | 1 | 0 | 0 | 0 | 1[t] | 4 | 2 | 3 | 8 | 13 | 6 | 5 | 4 | 11[u] | 9[u] | 0 | 5 | 4 | 3 |

[a]Full organism names are listed in Table 1. ■ gene present; □ pseudogene; ○ gene absent. Intron ORFs not included (see Table 8).
[b]Same in *L.tarentolae* maxicircle DNA.
[c]Gene is split into three separate ORFs in both *M.polymorpha* (*orf509* = *ymf4*; *orf169* = *ymf3*; *orf322* = *ymf2*) and *A.thaliana* (*ccb382*, *ccb203* and *ccb452*). *M.polymorpha orf509* is equivalent to *A.thaliana ccb382* + *ccb203*, whereas *A.thaliana ccb452* is homologous to *M.polymorpha orf169* + *orf322* (21).
[d]*orf228* = *ymf5* (*ccb256* in *A.thaliana* mtDNA; 21).
[e]*orf277* = *ymf6* (*ccb206* in *A.thaliana* mtDNA; 21).
[f]*orf244* in Mpo mtDNA.
[g]*orf183* in Mpo mtDNA (*orf25* in angiosperms).
[h]A putative *mutS* homolog, identified in a coral mtDNA (93), has not been found in any of the sequenced protist mtDNAs listed in Table 1.
[i]ORF showing similarity to mitochondrial plasmid-encoded DNA polymerase.
[j]Remnants of *dpo* gene (94).
[k]Coding sequence distributed over three separate ORFs (OGMP, unpublished results).
[l]ORF showing similarity to reverse transcriptase.
[m]Oda *et al.* (23).
[n]Boer and Gray (95).
[o]Coding sequence distributed between two separate ORFs (OGMP, unpublished results).
[p]ORF showing similarity to DNA endonuclease of type GIF-YIG (96).
[q]Three ORFs of this type have been found in Ama mtDNA (22).
[r]Comprising >60 codons and not overlapping one another or other identified genes.
[s]Only 29 ORFs >60 codons were predicted as possible genes using a defined index of G+C content in the first, second and third positions of codons (23).
[t]In the course of re-analyzing the Ccr mtDNA sequence one of the two previously annotated (34) unique ORFs, *orf94*, has been identified as *rpl20* (G.Burger, unpublished results).
[u]An additional 13 ORFs in Tpy (equivalent to 14 Pau ORFs) are defined as 'ciliate-specific' (shared between Tpy and Pau but not other mtDNAs). Of the 25 ORFs (unique + ciliate-specific) in Pau mtDNA 12 were previously annotated (9), whereas an additional 13 have been found in the course of re-analyzing the Pau mtDNA sequence (G.Burger, unpublished results).

## A comparative genomics approach to gene identification: the case of *orfB* and *atp8*

Accumulating sequence data are aiding in the identification of some of the unassigned ORFs that have been uncovered in the course of sequencing mitochondrial genomes. As an example we provide evidence here that *orfB*, a conserved gene of unknown function originally identified in plant mtDNA (see Table 3, footnote i), is the homolog of *atp8*, which encodes subunit 8 of the $F_0$ portion of the ATP synthase. The latter gene has been found in a number of animal and fungal mtDNAs, but up to now has not been identified in plant or protist mitochondrial genomes. Conversely, *orfB* is found in almost all plant and protist mtDNAs, but not in those of animals or fungi. Both Atp8 and OrfB proteins are characterized by the same block of three identical amino acids at the N-terminus, followed by an otherwise quite variable sequence (Fig. 2). The known OrfB proteins of plants differ from

**Table 7.** Transfer RNA genes encoded by mtDNA[a]

| A.A. | Anticodon | Codons | Hsa | Mse | Spo | Ama | Mbr | Mpo | Pwi | Nol | Ceu | Cre | Pmi | Ccr | Ppu | Rsa | Aca | Ddi | Oda | Csy | Pin | Cro | Pau | Tpy | Pfa | Tbr | Mja | Ram |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | ugc | GCN | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| C | gca | UGR | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ■ | ■[b] | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| D | guc | GAY | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| E | uuc | GAR | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ■ | ○ | ○ | ■[b] | ■ |
| F | gaa | UUY | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■[b] | ■ |
| G | gcc | GGY | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ■ |
| G | ucc | GGN | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| H | gug | CAY | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ■ | ○ | ○ | ■ | ■ |
| I | cau[f] | AUA | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ■[d] | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■[b] | ■ |
| I | gau | AUY | ■ | ○ | ■ | ■ | ■ | ○ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■[c] | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■[b] | ■ |
| K | uuu | AAR | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■[h] | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■[b] | ■ |
| L | caa | UUG | ○ | ○ | ○ | ○ | ○ | ■[b] | ○ | ○ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| L | uaa | UUR | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ■[b] | ○ | ○ | ■[b] | ■ |
| L | uag | CUN | ■ | ○ | ■ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| Me | cau | AUG | e | ○ | ■ | ■ | ■ | ■ | ■ | ■ | c, e | e | ○ | ■ | ■ | ■ | e | e | ■ | ■ | ■ | ■ | e | e | ○ | ○ | ■ | ■ |
| Mf | cau | AUG | e | ■ | ■ | ■ | ■ | ■[b] | ■ | ■ | c, e | e | ○ | ■ | ■ | ■ | e | e | ■[b] | ■ | ■ | ■ | e | e | ○ | ○ | ■ | ■ |
| N | guu | AAY | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■[b] | ■ |
| P | ugg | CCN | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| Q | uug | CAA | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■[b] | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| R | acg[g] | CGN | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| R | gcg | CGY | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| R | ucg | CGN | ■ | ○ | ■ | ■ | ■ | ■ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| R | ucu | AGR | ○ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| S | gcu | AGY | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■ | ■ |
| S | uga | UCN | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■[b] | ■ |
| T | ggu | ACY | ○ | ○ | ○ | ○ | ○ | ■ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| T | ugu | ACN | ■ | ○ | ■ | ■ | ■ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ■[b] | ○ |
| V | uac | GUN | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ■[b] | ■ |
| W | cca | UGG | ○ | ○ | ■[i] | ■ | ○ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ■ | ■[j] | ■ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ |
| W | uca | UGR | ■ | ■ | ○ | ○ | ■ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ○ | ○ | ○ | ■ | ■ | ■ | ○ | ○ | ○ | ○ |
| Y | gua | UAY | ■ | ○ | ■ | ■ | ■ | ■[b] | ■ | ■ | ○ | ○ | ■[c] | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ○ | ○ | ■[b] | ■ |
| Distinct *trn* genes | | | 22 | 2 | 25 | 25[k] | 25 | 27 | 26 | 26 | 3 | 3 | 8 | 23 | 24 | 27[m] | 15[n] | 17[n] | 24[q] | 24[r] | 25 | 22 | 4 | 7 | 0 | 0 | 25 | 26[s] |
| Total (incl. duplicates) | | | | | | | | 30 | | | 4 | | 10[l] | | | | 16[o] | 19[p] | | 29 | | | | 8 | | | 36 | |

[a]See Table 1 for complete organism names. ■ gene present; ○ gene absent. Aminoacylation specificity (a.a.) is indicated by the standard one letter symbols for amino acids (Me, elongator methionine; Mf, initiator methionine). The predicted anticodon of each tRNA is shown in lower case letters, with the predicted codon(s) that would be recognized shown in upper case letters (N = any nucleotide; R = A or G; Y = C or U). Expanded wobble base pairing is assumed, such that anticodons beginning with uridine are considered to recognize all codons in a four-codon family.

[b]Duplicate identical genes.

[c]Duplicate non-identical genes.

[d]Triplicate genes, two of which are identical, the third differing by a single T→C transition.

[e]Genome specifies a single *trnM*(cau).

[f]C in the first position of the anticodon presumed to be modified to lysidine, which converts the tRNA to an AUA-decoding isoleucine acceptor (97).

[g]A in first the position of the anticodon presumed to be modified to inosine, with the resulting tRNA able to pair with codons ending in C, U and A, and perhaps also G (see 98).

[h]*trnK*(cuu), the corresponding tRNA of which would be expected to recognize AAG but not AAA (61).

[i]Only UGG Trp codons appear in conserved protein coding genes in *S.pombe* mtDNA, however, several UGA codons occur in *rps3* and intron ORFs (92).

[j]Both UGG and UGA are decoded as Trp in *A.castellanii* mitochondria (61), whereas the tRNA specified by *trnW*(cca) would be expected to recognize only UGG.

[k]Includes a *trnL*(aag) not listed in the table.

[l]Includes a presumptive *trnE* pseudogene, unrelated in sequence to authentic *trnE*.

[m]Includes a *trnI*(uau) not listed in the table.

[n]Transcripts of most Aca mitochondrial tRNA genes (12 of 15) undergo substitutional RNA editing at one or more of the first three positions of the acceptor stem (61,64; D.H.Price and M.W.Gray, unpublished results). Transcripts of at least half of the Ddi mitochondrial tRNA genes are predicted to undergo a similar type of editing.

[o]Includes a *trnX*(uuua) pseudogene (D.H.Price and M.W.Gray, unpublished results), the transcript of which is predicted to have an 8 nt anticodon loop (61).

[p]Includes an unusual tRNA-like element whose anticodon sequence would pair with UAA and UAG (99), which are normally termination codons.

[q]Includes a *trnI*(aau) not listed in the table.

[r]Includes a *trnX*(cua), the corresponding tRNA of which would be expected to recognize UAG (normally a termination codon).

[s]Includes a *trnL*(gag) not listed.

**Table 8.** Introns and intron ORFs in mtDNA[a]

| | HSA | MSE | SPO | AMA | MBR | MPO | PWI | NOL | CEU | CRE | PMI | CCR | PPU | RSA | ACA | DDI | ODA | PIN | CSY | CRO | PAU | TPY | PFA | TBR | MJA | RAM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Group I** | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Introns** | 0 | 2 | 2 | 25 | 4 | 7 | 2 | 4 | 9 | 0 | 0 | 0 | 0 | 0 | 3 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **ORFs** | - | 3[b] | 2 | 10 | 4 | 2 | 2 | 4 | 7 | - | - | - | - | - | 3 | 4 | - | - | - | - | - | - | - | - | - | - |
| **Group II** | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Introns** | 0 | 0 | 1 | 0 | 0 | 25 | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **ORFs** | - | - | 1 | - | - | 8 | - | - | - | - | 0 | - | 2 | 3 | - | - | - | - | - | - | - | - | - | - | - | 0 |

[a] Full organism names are listed in Table 1.

[b] A group I intron in *nad5* contains *nad1* and *nad3* genes (100).

Atp8 essentially in their increased length. Because there is also much length variation among OrfB homologs in some protist mtDNAs, we were prompted to assess the possibility that *atp8* and *orfB* are homologous genes.

The N-terminal functional domain (71) of ATP synthase subunit 8 is well conserved in different fungi compared with the central hydrophobic domain (72) and the C-terminal domain (73). The latter domain contains a region enriched in positively charged amino acid residues (73), which are thought to play an important role in assembly of the $F_0$ complex (see below). If OrfB is indeed homologous to Atp8, we should find similar amino acid signatures in a multiple alignment of a phylogenetically diverse collection of both types of sequences. Such a collection has recently become available through the sequencing efforts of the OGMP and FMGP.

As shown in Figure 2, the highly conserved N-terminal domain provides the best evidence for homology between *orfB* and *atp8*. Further evidence supporting this inference is the presence of perfectly aligned central hydrophobic and positively charged domains. Based on the alignment of the first 57 amino acids shown in Figure 2, we suggest that there is little basis for a distinction between the 'Atp8' and 'OrfB' classes of protein. With two notable exceptions, this sequence compilation further demonstrates that a long C-terminal extension (position 78 and beyond in Fig. 2) is only found among plants and protists. In the stramenopiles *Cafeteria roenbergensis* and *Ochromonas danica* the mtDNA codes for a shorter protein, about as long as the longest fungal sequences. This feature is not clade specific because in another stramenopile, *Phytophthora infestans*, the mitochondrial genome specifies an Atp8 protein that is rather typical in size for protists. The C-terminal extension is not only quite variable in size, but indeed is so divergent in sequence that it can only be reasonably well aligned among very closely related species (e.g. land plants). Thus the presence or absence of a C-terminal extension also does not distinguish between 'Atp8' and 'OrfB' classes.

Conserved sequence motifs within the hydrophobic and C-terminal domains of the Atp8/OrfB protein are restricted to the boundaries between these domains, the 'LP motif' (71), which is immediately followed by a region with one or several positively charged amino acids. Previous studies in fungi have shown that these positively charged amino acids play an important role in assembly of subunits 6, 8 and 9 (73).

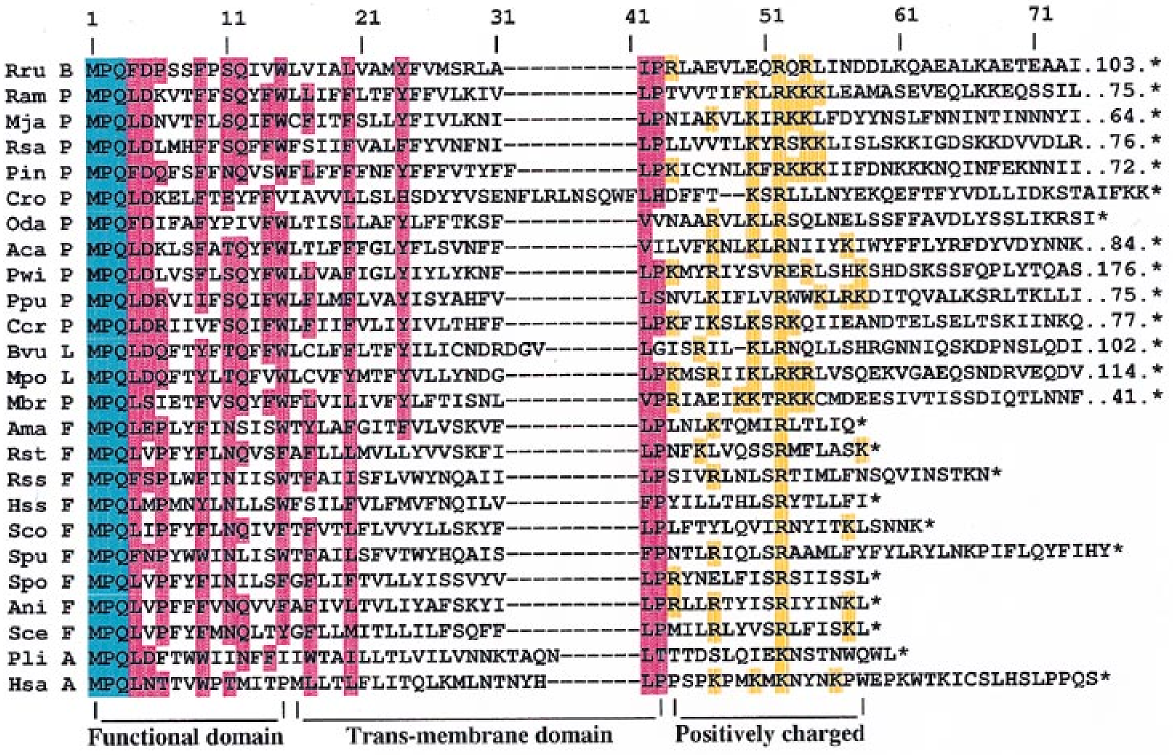
In summary, plant and protist mitochondrial OrfB proteins contain all of the conserved sequence elements characteristic of animal and fungal Atp8 proteins. Thus the *orfB* gene represents the best candidate for the previously 'missing' *atp8* homolog in plant and protist mtDNAs.

## Phylogenetic implications

The mitochondrial gene content and genome organization data being generated by the OGMP and other groups are serving to further clarify our views about the origin and evolution of the mitochondrial genome. One example involves the relationship between land plant and *Chlamydomonas* mtDNAs, which are so different in structure, organization and mode of expression that they show little evidence of having a common evolutionary origin (1,2,74). In the absence of a phylogenetically broad database of comparative information we at one time entertained the possibility that the plant mitochondrial genome might have had a different, more recent evolutionary ancestry than *Chlamydomonas* and other mitochondrial genomes (75). However, sequencing of *P.wickerhamii* (48) and other (24,34,61) protist mtDNAs has clearly demonstrated that plant mtDNA has retained an ancestral pattern that has evidently been lost in the more rapidly evolving and highly derived *Chlamydomonas* mtDNA (74). It is worth emphasizing that the majority of the protist mtDNAs sequenced to date by the OGMP, particularly those from more obscure protists selected from the wild on the basis of ultrastructural or other phylogenetic considerations, retain a more or less ancestral pattern of gene content and organization. In contrast, most of the mtDNAs that had been sequenced prior to the inception of the OGMP (those from animals, most fungi, chlamydomonadalean green algae, ciliates and trypanosomatid protozoa) are highly derived. It is curious that the majority of the protists that have been selected as models for biochemical, genetic and molecular biological research happen to have mtDNAs that are the least representative of the ancestral form.

## Descriptions

*Organelle Genome Megasequencing Program (OGMP)* (http://megasun.bch.umontreal.ca/ogmp/ ). The OGMP was initiated as a multi-disciplinary and inter-university consortium of Canadian investigators interested in organelle genome evolution and eukaryotic phylogeny. As currently constituted it consists of a Team (B.F.Lang, administrative coordinator; M.W.Gray, scientific coordinator; G.Burger, C.Lemieux and M.Turmel) and an Advisory Board (R.Cedergren, G.B.Golding, D.Sankoff, T.G.Littlejohn and C.J.O'Kelly), with external collaborators on some individual projects. The experimental arm of the OGMP, the Sequencing Unit



```
      1         11        21        31        41        51        61        71
      |         |         |         |         |         |         |         |
Rru B MPQFDPSSFPSQIVWLVIALVAMYFVMSRLA---------IPRLAEVLEQRQRLINDDLKQAEALKAETEAAI.103.*
Ram P MPQLDKVTFFSQYFWLLIFFLTFYFFVLKIV---------LPTVVTIFKLRKKKLEAMASEVEQLKKEQSSIL..75.*
Mja P MPQLDNVTFLSQIFWCFITFSLLYFIVLKNI---------LPNIAKVLKIRKKLFDYYNSLFNNINTINNNYI..64.*
Rsa P MPQLDLMHFFSQFFWFSIIFVALFFYVNFNI---------LPLLVVTLKYRSKKLISLSKKIGDSKKDVVDLR..76.*
Pin P MPQFDQFSFFNQVSWFLFFFFNFYFFFVTYFF--------LPKICYNLKFRKKKIIFDNKKKNQINFEKNNII..72.*
Cro P MPQLDKELFTEYFFVIAVVLLSLHSDYYVSENFLRLNSQWFLHDFFT--KSRLLLNYEKQEFTFYVDLLIDKSTAIFKK*
Oda P MPQFDIFAFYPIVFWLTISLLAFYLFFTKSF---------VVNAARVLKLRSQLNELSSFFAVDLYSSLIKRSI*
Aca P MPQLDKLSFATQYFWLTLFFFGLYFLSVNFF---------VILVFKNLKLRNIIYKIWYFFLYRFDYVDYNNK..84.*
Pwi P MPQLDLVSFLSQYFWLLVAFIGLYIYLYKNF---------LPKMYRIYSVRERLSHKSHDSKSSFQPLYTQAS.176.*
Ppu P MPQLDRVIIFSQIFWLFLMFLVAYISYAHFV---------LSNVLKIFLVRWWKLRKDITQVALKSRLTKLLI..75.*
Ccr P MPQLDRIIVFSQIFWLFIIFVLIYIVLTHFF---------LPKFIKSLKSRKQIIEANDTELSELTSKIINKQ..77.*
Bvu L MPQLDQFTYFTQFFWLCLFFLTFYILICNDRDGV------LGISRIL-KLRNQLLSHRGNNIQSKDPNSLQDI.102.*
Mpo L MPQLDQFTYLTQFVWLCVFYMTFYVLLYNDG---------LPKMSRIIKLRKRLVSQEKVGAEQSNDRVEQDV.114.*
Mbr P MPQLSIETFVSQYFWFLVILIVFYLFTISNL---------VPRIAEIKKTRKKCMDEESIVTISSDIQTLNNF..41.*
Ama F MPQLEPLYFINSISWTYLAFGITFVLVSKVF---------LPLNLKTQMIRLTLIQ*
Rst F MPQLVPFYFLNQVSFAFLLLMVLLYVVSKFI---------LPNFKLVQSSRMFLASK*
Rss F MPQFSPLWFINIISWTFAIISFLVWYNQAII---------LPSIVRLNLSRTIMLFNSQVINSTKN*
Hss F MPQLMPMNYLNLLSWFSILFVLFMVFNQILV---------FPYILLTHLSRYTLLFI*
Sco F MPQLIPFYFLNQIVFTFVTLFLVVYLLSKYF---------LPLFTYLQVIRNYITKLSNNK*
Spu F MPQFNPYWWINLISWTFAILSFVTWYHQAIS---------FPNTLRIQLSRAAMLFYFYLRYLNKPIFLQYFIHY*
Spo F MPQLVPFYFINILSFGFLIFTVLLYISSVYV---------LPRYNELFISRSIISSL*
Ani F MPQLVPFFFVNQVVFAFIVLTVLIYAFSKYI---------LPRLLRTYISRIYINKL*
Sce F MPQLVPFYFMNQLTYGFLLMITLLILFSQFF---------LPMILRLYVSRLFISKL*
Pli A MPQLDFTWWIINFFIIWTAILLTLVILVNNKTAQN-----LTTTDSLQIEKNSTNWQWL*
Hsa A MPQLNTTVWPTMITPMLLTLFLITQLKMLNTNYH------LPPSPKPMKMKNYNKPWEPKWTKICSLHSLPPQS*
      |_______________||_________________________||______________|
      Functional domain   Trans-membrane domain    Positively charged
```

**Figure 2.** Alignment of Atp8 and OrfB amino acid sequences. Sequences from bacteria (B), protists (P), land plants (L), fungi (F) and animals (A) are compared. Three letter abbreviations of organism names are listed in Table 1. Additional abbreviations: Rru, *Rhodospirillum rubrum*; Bvu, *Beta vulgaris*; Rst, *Rhizopus stolonifer*; Rss, *Rhizophydium* ssp.; Hss, *Harpochytrium* ssp.; Sco, *Schizophyllum commune*; Spu, *Spizellomyces punctatus*; Ani, *Aspergillus nidulans*; Sce, *Saccharomyces cerevisiae*; Pli, *Paracentrotus lividus*. Sequences were obtained from the NCBI databases except for Pin, Mbr, Rst, Rru, Hss, Sco and Spu, which are unpublished FMGP sequences, and Mja, Rsa, Cro, Oda and Ppu, which are unpublished OGMP sequences. Color highlighting is as follows: blue, invariant amino acids; magenta, identical residues comprising at least 10 (40% or more) of the total number of residues in a given column (also colored in magenta are those residues that according to the PAM matrix are positive or neutral exchanges with reference to the most abundant residue in the column); yellow, positively charged amino acids. Dashes (–) denote a missing residue at this position in comparison with other sequence(s). Asterisks (*) mark translation termination codons; numbers preceding an asterisk indicate the remaining length of sequence that is not shown.

(directed by G.Burger), is located in the Département de Biochimie, Université de Montréal. The Sequencing Unit comprises two divisions: Molecular Biology (I.Plante, D.Saint-Louis and Y.Zhu), which constructs clone libraries, performs the actual sequencing and works out improved cloning and sequencing methods; Informatics (N.Brossard and P.Rioux), which develops and implements tools required for project management, data handling, sequence analysis and annotation. As the data production arm of the OGMP, the Sequencing Unit delivers analyzed and fully annotated mitochondrial genome sequences for submission to public domain databases. The OGMP website (URL given above) contains additional information about the program, as well as data summaries and gene maps for the individual OGMP sequencing projects completed to date (Table 1).

*Protist Image Database (PID)* (http://megasun.bch.umontreal.ca/protists/ ). The PID (T.G.Littlejohn and C.J.O'Kelly) is a compilation of images and short descriptions of selected protist genera, especially those whose species are frequently used as experimental organisms or are important in studies of organismal evolution. The intent of the PID is to provide integrated on-line information about the morphology, taxonomy and phylogenetic relationships of these organisms. The PID, which was initiated within the OGMP, contains descriptions of most of the species whose mtDNAs have been sequenced by the OGMP. The PID is being continued independently from but in close collaboration with the OGMP, with its web pages maintained on the OGMP web server.

*Organelle Genome Database Project (GOBASE)* (http://megasun.bch.umontreal.ca/gobase/ ). Shortly after the OGMP was established it became apparent that there were serious limitations in accessing all of the relevant information associated with organelles. Data are dispersed among a number of sources (World Wide Web, public data repositories, scientific journals and books) and in many cases are difficult even to locate. Usually only limited links exist among data sources (e.g. there is no easy way to connect from a GenBank record containing an rRNA sequence to the corresponding secondary structure contained in another database). It is even more difficult to perform the sort of cross-genome comparisons that were essential for the present review. Further, the data sets are often incomplete and/or contain errors, which are sometimes hard to identify and to rectify in the underlying data source. In such a disorganized state organelle genomic data constitute a major underexploited information resource. The GOBASE project (17) was initiated by a subset of OGMP members (B.F.Lang, M.W.Gray, G.Burger and

T.G.Littlejohn) to rectify this situation. GOBASE, which is a taxonomically broad database that organizes and integrates diverse data related to organelles, has been constructed as a relational database with a web-based user interface. The current version focuses on the mitochondrial subset of data.

## ACKNOWLEDGEMENTS

We thank J.Chesnick, R.W.Lee, P.J.Myler, K.Stuart, Y.Tanaka and D.R.Wolstenholme for providing unpublished sequence data and other information in advance of publication. We are also grateful to C.J.O'Kelly and T.Nerad for provision of organisms and for invaluable advice on phylogeny, taxonomy and culture conditions. The OGMP has been supported by a Special Project grant (SP-34) from the Medical Research Council of Canada and by a grant (G0-12323) from the Canadian Genome Analysis and Technology Program (CGAT), which has also provided funding to GOBASE (GO-12984). Generous grants of equipment from Sun Microsystems Inc. and LI-COR Inc. are gratefully acknowledged, as is salary and interaction support from the Canadian Institute for Advanced Research.

## REFERENCES

1 Gray,M.W. (1989) *Annu. Rev. Cell Biol*., **5**, 25–50.
2 Gray,M.W. (1992) *Int. Rev. Cytol*., **141**, 233–357.
3 Gillham,N.W. (1994) *Organelle Genes and Genomes*. Oxford University Press, New York, NY.
4 Leblanc,C., Richard,O., Kloareg,B., Viehmann,S., Zetsche,K. and Boyen,C. (1997) *Curr. Genet*., **31**, 193–207.
5 Paquin,B., Laforest,M.-J., Forget,L., Roewer,I., Wang,Z., Longcore,J. and Lang,B.F. (1997) *Curr. Genet*., **31**, 380–395.
6 Patterson,D.J. and Sogin,M.L. (1992) In Hartman,H. and Matsuno,K. (eds), *The Origin and Evolution of the Cell*. World Scientific, Singapore, Singapore, pp. 13–46.
7 Vaidya,A.B., Akella,R. and Suplick,K. (1989) *Mol. Biochem. Parasitol*., **35**, 97–108.
8 Feagin,J.E., Werner,E., Gardner,M.J., Williamson,D.H. and Wilson,R.J.M. (1992) *Nucleic Acids Res*., **20**, 879–887.
9 Pritchard,A.E., Seilhamer,J.J., Mahalingam,R., Sable,C.L., Venuti,S.E. and Cummings,D.J. (1990) *Nucleic Acids Res*., **18**, 173–180.
10 Wolstenholme,D.R. (1992) *Int. Rev. Cytol*., **141**, 173–216.
11 Clark-Walker,G.D. (1992) *Int. Rev. Cytol*., **141**, 89–127.
12 Hanson,M.R. and Folkerts,O. (1992) *Int. Rev. Cytol*., **141**, 129–172.
13 Wolstenholme,D.R. and Fauron,C.M.-R. (1995) In Levings,C.S. and Vasil,I.K. (eds), *The Molecular Biology of Plant Mitochondria*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 1–59.
14 Cummings,D.J. (1992) *Int. Rev. Cytol*., **141**, 1–64.
15 Stuart,K. and Feagin,J.E. (1992) *Int. Rev. Cytol*., **141**, 65–88.
16 Feagin,J.E. (1994) *Annu. Rev. Microbiol*., **48**, 81–104.
17 Korab-Laskowska,M., Rioux,P., Brossard,N., Littlejohn,T.G., Gray, M.W., Lang,B.F. and Burger,G. (1998) *Nucleic Acids Res*., **26**, 139–146.
18 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol*., **215**, 403–410.
19 Pearson,W.R. (1990) *Methods Enzymol*., **183**, 63–98.
20 Staden,R. (1990) *Methods Enzymol*., **183**, 193–211.
21 Unseld,M., Marienfeld,J.R., Brandt,P. and Brennicke,A. (1997) *Nature Genet*., **15**, 57–61.
22 Paquin,B. and Lang,B.F. (1996) *J. Mol. Biol*., **255**, 688–701.
23 Oda,K., Yamato,K., Ohta,E., Nakamura,Y., Takemura,M., Nozato,N., Akashi,K., Kanegae,T., Ogura,Y., Kohchi,T. and Ohyama,K. (1992) *J. Mol. Biol*., **223**, 1–7.
24 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) *Nature*, **387**, 493–497.
25 Palmer,J.D. (1997) *Nature*, **387**, 454–455.
26 Bogorad,L. (1991) In Bogorad,L. and Vasil,I.K. (eds), *The Molecular Biology of Plastids*. Academic Press Inc., San Diego, CA, pp. 93–124.
27 Reith,M. (1995) *Annu. Rev. Plant Physiol. Plant Mol. Biol*., **46**, 549–575.
28 Masters,B.S., Stohl,L.L. and Clayton,D.A. (1987) *Cell*, **51**, 89–99.
29 Chen,B., Kubelik,A.R., Mohr,S. and Breitenberger,C.A. (1996) *J. Biol. Chem*., **271**, 6537–6544.
30 Cermakian,N., Ikeda,T.M., Cedergren,R. and Gray,M.W. (1996) *Nucleic Acids Res*., **24**, 648–654.
31 Weihe,A., Hedtke,B. and Börner,T. (1997) *Nucleic Acids Res*., **25**, 2319–2325.
32 Hedtke,B, Börner,T. and Weihe,A. (1997) *Science*, **277,** 809–811.
33 Burger,G., Lang,B.F., Reith,M. and Gray,M.W. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 2328–2332.
34 Leblanc,C., Boyen,C., Richard,O., Bonnard,G., Grienenberger,J.-M. and Kloareg,B. (1995) *J. Mol. Biol*., **250**, 484–495.
35 Viehmann,S., Richard,O., Boyen,C. and Zetsche,K. (1996) *Curr. Genet*., **29**, 199–201.
36 Daignan-Fornier,B., Valens,M., Lemire,B.D. and Bolotin-Fukuhara,M. (1994) *J. Biol. Chem*., **269**, 15469–15472.
37 Nugent,J.M. and Palmer,J.D. (1991) *Cell*, **66**, 473–481.
38 Covello,P.S. and Gray,M.W. (1982) *EMBO J*., **11**, 3815–3820.
39 Prioli,L.M., Huang,J. and Levings,C.S. (1993) *Plant Mol. Biol*., **23**, 287–295.
40 Gutell,R.R. (1994) *Nucleic Acids Res*., **22**, 3502–3507.
41 Gutell,R.R., Gray,M.W. and Schnare,M.N. (1993) *Nucleic Acids Res*., **21**, 3055–3074.
42 Seilhamer,J.J., Gutell,R.R. and Cummings,D.J. (1984) *J. Biol. Chem*., **259**, 5173–5172.
43 Heinonen,T.Y.K., Schnare,M.N., Young,P.G. and Gray,M.W. (1987) *J. Biol. Chem*., **262**, 2879–2887.
44 Boer,P.H. and Gray,M.W. (1988) *Cell*, **55**, 399–411.
45 Denovan-Wright,E.M. and Lee,R.W. (1994) *J. Mol. Biol*., **241**, 298–311.
46 Nedelcu,A.M. (1997) *Mol. Biol. Evol*., **14**, 506–517.
47 Feagin,J.E., Mericle,B.L., Werner,E. and Morris,M. (1997) *Nucleic Acids Res*., **25**, 438–446.
48 Wolff,G., Plante,I., Lang,B.F., Kück,U. and Burger,G. (1994) *J. Mol. Biol*., **237**, 75–86.
49 Lang,B.F., Goff,L.J. and Gray,M.W. (1996) *J. Mol. Biol*., **261**, 607–613.
50 Suyama,Y. (1986) *Curr. Genet*., **10**, 411–420.
51 Rusconi,C.P. and Cech,T.R. (1996) *Genes Dev*., **10**, 2870–2880.
52 Simpson,A.M., Suyama,Y., Dewes,H., Campbell,D.A. and Simpson,L. (1989) *Nucleic Acids Res*., **17**, 5427–5445.
53 Hancock,K. and Hajduk,S.L. (1990) *J. Biol. Chem*., **265**, 19208–19215.
54 Dietrich,A., Weil,J.H. and Maréchal-Drouard,L. (1992) *Annu. Rev. Cell Biol*., **8**, 115–131.
55 Akashi,K., Sakurai,K., Hirayama,J., Fukuzawa,H. and Ohyama,K. (1996) *Curr. Genet*., **30**, 181–185.
56 Akashi,K., Hirayama,J., Takenaka,M., Yamaoka,S., Suyama,Y., Fukuzawa,H. and Ohyama,K. (1997) *Biochim. Biophys. Acta*, **1350**, 262–266.
57 Börner,G.V., Mörl,M., Janke,A. and Pääbo,S. (1996) *EMBO J*., **15**, 5949–5957.
58 Lonergan,K.M. and Gray, M.W. (1993) *Science*, **259**, 812–816.
59 Lonergan,K.M. and Gray,M.W. (1993) *Nucleic Acids Res*., **21**, 4402.
60 Gray,M.W. and Lonergan,K.M. (1993) In Brennicke,A. and Kück,U. (eds), *Plant Mitochondria: With Emphasis on RNA Editing and Cytoplasmic Male Sterility*. VCH, Weinheim, Germany, pp. 15–22.
61 Burger,G., Plante,I., Lonergan,K.M. and Gray,M.W. (1995) *J. Mol. Biol*., **245**, 522–537.
62 Price,D.H. and Gray,M.W. (1998) In Grosjean,H. and Benne,R. (eds), *Modification and Editing of RNA: The Alteration of RNA Structure and Function*. American Society for Microbiology, Washington, DC, in press.
63 Laforest,M.-J., Roewer,I. and Lang,B.F. (1997) *Nucleic Acids Res*., **25**, 626–632.
64 Schnare,M.N., Greenwood,S.J. and Gray,M.W. (1995) *FEBS Lett*., **362**, 24–28.
65 Anderson,S., Bankier,A.T., Barrell,B.G., de Bruijn,M.H.L., Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A., Sanger,F., Schreier,P.H., Smith,A.J.H., Staden,R. and Young,I.G. (1982) In Slonimski,P., Borst, P. and Attardi,G. (eds), *Mitochondrial Genes*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 5–43.
66 Okimoto,R. and Wolstenholme,D.R. (1990) *EMBO J*., **9**, 3405–3411.
67 Steinberg,S. and Cedergren,R. (1994) *Nature Struct. Biol*., **1**, 507–510.
68 Wolff,G., Burger,G., Lang,B.F. and Kück,U. (1993) *Nucleic Acids Res*., **21**, 719–726.
69 Turmel,M., Côté,V., Otis,C., Mercier,J.-P., Gray,M.W., Lonergan,K. and Lemieux,C. (1995) *Mol. Biol. Evol*., **12**, 533–545.
70 Fontaine,J.M., Rousvoal,S., Leblanc,C., Kloareg,B. and Loiseaux-de Goër,S. (1995) *J. Mol. Biol*., **251**, 378–389.

71 Devenish,R.J., Papakonstantinou,T., Galanis,M., Law,R.H., Linnane,A.W. and Nagley,P. (1992) *Annls NY Acad. Sci*., **671**, 403–414.

72 Papakonstantinou,T., Law,R.H., Nesbitt,W.S., Nagley,P. and Devenish,R.J. (1996) *Curr. Genet*., **30**, 12–18.

73 Papakonstantinou,T., Galanis,M., Nagley,P. and Devenish,R.J. (1993) *Biochim. Biophys. Acta*, **1144**, 22–32.

74 Gray,M.W. (1995) In Levings,C.S. and Vasil,I.K. (eds), *The Molecular Biology of Plant Mitochondria*. Kluwer Academic, Dordrecht, The Netherlands, pp. 635–659.

75 Gray,M.W., Cedergren,R., Abel,Y. and Sankoff,D. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 2267–2271.

76 Denovan-Wright,E.M., Nedelcu,A.M. and Lee,R.W. (1998) *Plant Mol. Biol*., **36**, 285–295.

77 Boer,P.H. and Gray,M.W. (1991) *Curr. Genet*., **19**, 309–312.

78 Vahrenholz,C., Rieman,G., Pratje,E., Dujon,B. and Michaelis,G. (1993) *Curr. Genet*., **24**, 241–247.

79 Kairo,A., Fairlamb,A.H., Gobright,E. and Nene,V. (1994) *EMBO J*., **13**, 898–905.

80 Anderson,S., Bankier,A.T., Barrell,B.G., de Bruijn,M.H.L., Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A., Sanger,F., Schreier,P.H., Smith,A.J., Staden,R. and Young,I.G. (1981) *Nature*, **290**, 457–465.

81 Beagley,C.T., Okimoto,R. and Wolstenholme,D.R. (1998) *Genetics*, in press.

82 Vaidya,A.B. and Arasu,P. (1987) *Mol. Biochem. Parasitol*., **22**, 249–257.

83 Hajduk,S.L., Harris,M.E. and Pollard,V.W. (1993) *FASEB J*., **7**, 54–63.

84 Read,L.K., Wilson,K.D., Myler,P.J. and Stuart,K. (1994) *Nucleic Acids Res*., **22**, 1489–1495.

85 Takemura,M., Nozato,N., Oda,K., Kobayashi,Y., Fukuzawa,H. and Ohyama,K. (1995) *Mol. Gen. Genet*., **247**, 565–570.

86 Lonergan,K.M. and Gray,M.W. (1996) *J. Mol. Biol*., **257**, 1019–1030.

87 Ogawa,S., Matsuo,K., Angata,K., Yanagisawa,K. and Tanaka,Y. (1997) *Curr. Genet*., **31**, 80–88.

88 Pellizzari,R., Anjard,C. and Bisson,R. (1997) *Biochim. Biophys. Acta*, **1320**, 1–7.

89 Commission on Plant Gene Nomenclature (1994) *Plant Mol. Biol. Rep*., **12** (CPGN suppl.), S1–S109.

90 Seilhamer,J.J., Olsen,G.J. and Cummings,D.J. (1984) *J. Biol. Chem*., **259**, 5167–5172.

91 Schnare,M.N., Heinonen,T.Y.K., Young,P.G. and Gray,M.W. (1986) *J. Biol. Chem*., **261**, 5187–5193.

92 Lang,B.F., Ahne,F., Distler,S., Trinkl,H., Kaudewitz,F. and Wolf,K. (1983) In Schweyen,R.J., Wolf,K. and Kaudewitz,F. (eds), *Mitochondria 1983, Nucleo-Mitochondrial Interactions*. Walter de Gruyter, Berlin, Germany, pp. 313–329.

93 Pont-Kingdom,G.A., Okada,N.A., Macfarlane,J.L., Beagley,C.T., Wolstenholme,D.R., Cavalier-Smith,T. and Clark-Walker,G.D. (1995) *Nature*, **375**, 109–111.

94 Weber,B., Börner,T. and Weihe,A. (1995) *Curr. Genet*., **27**, 488–490.

95 Boer,P.H. and Gray,M.W. (1988) *EMBO J*., **7**, 3501–3508.

96 Burger,G. and Werner,S. (1985) *J. Mol. Biol*., **186**, 231–242.

97 Muramatsu,T., Nishikawa,K., Nemoto,F., Kuchino,Y., Nishimura,S., Miyazawa,T. and Yokoyama,S. (1988) *Nature*, **336**, 179–181.

98 Pfitzinger,H., Weil,J.H., Pillay,D.T.N. and Guillemaut,P. (1990) *Plant Mol. Biol*., **14**, 805–814.

99 Pi,M., Angata,K., Ikemura,T., Yanagisawa,K. and Tanaka,Y. (1996) *J. Plant Res*., **109**, 1–6.

100 Beagley,C.T., Okada,N.A. and Wolstenholme,D.R. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 5619–5623.