doi:10.1016/j.jmb.2007.06.085 J. Mol. Biol. (2007) 372, 356–368

JMB

Available online at www.sciencedirect.com
ScienceDirect

ELSEVIER

The Highly Reduced and Fragmented Mitochondrial Genome of the Early-branching Dinoflagellate *Oxyrrhis marina* Shares Characteristics with both Apicomplexan and Dinoflagellate Mitochondrial Genomes

Claudio H. Slamovits, Juan F. Saldarriaga, Allen Larocque and Patrick J. Keeling*

Department of Botany University of British Columbia 3529-6270 University Boulevard Vancouver, BC, V6T 1Z4 Canada

The mitochondrial genome and the expression of the genes within it have evolved to be highly unusual in several lineages. Within alveolates, apicomplexans and dinoflagellates share the most reduced mitochondrial gene content on record, but differ from one another in organisation and function. To clarify how these characteristics originated, we examined mitochondrial genome form and expression in a key lineage that arose close to the divergence of apicomplexans and dinoflagellates, *Oxyrrhis marina*. We show that *Oxyrrhis* is a basal member of the dinoflagellate lineage whose mitochondrial genome has some unique characteristics while sharing others with apicomplexans or dinoflagellates. Specifically, *Oxyrrhis* has the smallest gene complement known, with several rRNA fragments and only two protein coding genes, *cox1* and a *cob-cox3* fusion. The genome appears to be highly fragmented, like that of dinoflagellates, but genes are frequently arranged as tandem copies, reminiscent of the repeating nature of the *Plasmodium* genome. In dinoflagellates and *Oxyrrhis*, genes are found in many arrangements, but the *Oxyrrhis* genome appears to be more structured, since neighbouring genes or gene fragments are invariably the same: *cox1* and the *cob-cox3* fusion were never found on the same genomic fragment. Analysing hundreds of cDNAs for both genes and circularized mRNAs from *cob-cox3* showed that neither uses canonical start or stop codons, although a UAA terminator is created in the *cob-cox3* fusion mRNA by post-transcriptional oligoadenylation. mRNAs from both genes also use a novel 5′ oligo(U) cap. Extensive RNA editing is characteristic of dinoflagellates, but we find no editing in *Oxyrrhis*. Overall, the combination of characteristics found in the *Oxyrrhis* genome allows us to plot the sequence of many events that led to the extreme organisation of apicomplexan and dinoflalgellate mitochondrial genomes.

Keywords: mitochondrial genome; fusion protein; RNA editing; translation initiation; translation termination

*Corresponding author

Introduction

Mitochondria are now known to have arisen by endosymbiosis involving an alpha-proteobacterium. The most compelling evidence for this comes from similarities between gene sequences from alpha-proteobacterial genomes and their homologues in the relic genomes of the mitochondria.[1,2] While these genomes may have retained sufficient similarities to betray their origins, they have also changed

Abbreviations used: SSU, small subunit; LSU, large subunit; EST, expressed sequence tags; cob, cytochrome b; cox1, cytochrome oxidase 1; cox3, cytochrome oxidase subunit 3.

E-mail address of the corresponding author: pkeeling@interchange.ubc.ca

a great deal in the course of their evolution. Indeed, many mitochondrial genomes have taken on bizarre forms or adopted peculiar strategies of expression that are either novel to mitochondrial genomes, or taken to unusual extremes. Examples of this are abundant, and include highly fragmented genomes, complex forms of RNA editing, or massive *trans*-splicing, to name a few of those that are better understood.[3–5]

There are occasionally indications of how these unusual genomes came to evolve,[6,7] but in other cases the most highly modified genomes seem to have evolved very rapidly, leaving few traces of the course of events that led to even massive changes. One lineage with a potentially informative history of mitochondrial genome evolution is the alveolates. This is a major assemblage of microbial eukaryotes, primarily comprising three large and diverse subgroups, ciliates, dinoflagellates, and apicomplexans, whose relationships to one another are generally well known.[8,9] Ciliates branched first among the three subgroups, and have fairly conventional mitochondrial genomes. Complete sequences are known from *Paramecium tetraurelia* and five species of *Tetrahymena*,[10–12] and they are all fairly large (40–47 kb), map linearly, and contain a middling number of protein coding genes (44–46). They also employ a non-canonical genetic code where TGA encodes tryptophan and TAG is unassigned, and several genes have been found to use non-canonical start codons (ATA, ATT, TTG and GTG). Both the small subunit (SSU) and large subunit (LSU) rRNAs have been split into two independently expressed fragments, and several tRNAs have been found to be imported.[13] These are odd characteristics, but not outside the range of what one could consider normal for a mitochondrial genome. Apicomplexan mitochondrial genomes, on the other hand, are very reduced and highly derived. *Plasmodium* has the smallest mitochondrial genome known to date: at a mere 6 kb, this linear-mapping repeat harbours only three protein-coding genes (*cox1*, *cox3*, and *cob*) and fragments of SSU and LSU rRNAs.[14,15] Both rRNAs are highly fragmented, with 23 individual segments identified to date that are still insufficient to completely account for either molecule.[15] The *Plasmodium* mitochondrial genome encodes no tRNAs, and they have been shown to be imported in *Toxoplasma*.[16]

The third major alveolate lineage, the dinoflagellates, is the most poorly studied, but emerging evidence suggests that their mitochondrial genomes are by far the most unusual. So far, the same three protein-coding genes found in *Plasmodium* have been identified,[17–20] and several rRNA fragments that are similar to those of apicomplexans have also been characterized.[21] However, there are also two major differences. First, these coding sequences do not seem to map to a single chromosome. Comparing the genomic contexts of four different fragments encoding the *Crypthecodinium cohnii cox1* revealed that all four had unique flanking sequences, and that they did not assemble into a single mitochondrial chromosome.[20,22] Instead, there appear to be many small chromosomes with one or a few genes or gene fragments, perhaps similar to the situation in the ichthyosporean *Amoebidium parasiticum*.[4] The second major difference is that when genomic DNA and cDNA copies of *cox1* and *cob* were compared from several species, extensive evidence for a novel type of RNA editing was found where A-G, U-C and C-U substitutions predominate, along with a few other rarer changes.[18,23,24]

It seems likely that major changes to mitochondrial genome structure and content took place before apicomplexans and dinoflagellates diverged. But the evolutionary history of these changes since then is unclear, as we lack information from any of the deep branching taxa that could reveal the nature of intermediate stages in a stepwise accumulation of novel characteristics. We have therefore examined the mitochondrial genome structure and expression in *Oxyrrhis marina*, a marine predator that represents an ancient branch of the dinoflagellate lineage.[25,26] We have carried out a large scale EST sequencing project on *O. marina* and, taking advantage of the oligoadenylation of mitochondrial transcripts in apicomplexans[27,28] and dinoflagellates (mitochondrial transcripts appear in poly(A) selected libraries from other EST projects[29,30]), identified a large number of transcripts from the mitochondrial genome. We also generated comparable information from genomic DNA and analysed mRNA endpoints in circularized mRNAs, showing that the *O. marina* genome has fewer protein coding units that any other genome known (two *cox1* and a *cob-cox3* fusion), it lacks RNA editing, contains highly fragmented rRNAs, and protein-coding genes are found in multiple but limited genomic contexts. We also show that the *O. marina* mRNAs use non-canonical start codons, are oligo(U) capped at the 5′ end, and lack stop codons altogether, creating one instead by oligoadenylation or by allowing the oligo(A) tail to be translated. This combination of novel characteristics and other characteristics found in either dinoflagellates, apicomplexans, or both, allows an unprecedented view of the progressive evolution of the form and expression of a highly derived genome.

Results and Discussion

Oxyrrhis marina is an early-diverging member of the dinoflagellate lineage

To reconstruct the order of events in the evolution of the mitochondrial genome in the dinoflagellate lineage, an understanding of the nature of this genome in the early branching lineages is important, since they might represent descendents of intermediate stages in this evolution. *O. marina* is likely one such taxon, but we sought first to clarify its relationship to other dinoflagellates.

Phylogenies of four protein-coding genes all support the deep-branching position for *O. marina*,[25,26] but the phylogeny of the SSU rRNA has been used to

suggest it is a derived dinoflagellate specifically belonging of the Gonyaulacales.[31] We therefore tested the position of *O. marina* relative to a deep-branching relative of dinoflagellates (*Perkinsus marinus*) and a gonyaulacalean (*Alexandrium tamarense*) by constructing a multigene phylogeny using EST data (see below). From a data set consisting of 30 proteins and 7705 amino acids (Supplementary Data; Table 1), we inferred the phylogeny using several methods, and consistently found *O. marina* to branch between *P. marinus* and the true dinoflagellates (Figure 1). The gonyaulacalean was separated from *O. marina* by a very strongly supported node, and the topology forcing *O. marina* with *A. tamarense* was rejected by AU tests (as were all other positions for *O. marina*; Figure 1). Altogether, the phylogentic data given here and elsewhere[25,26] strongly support the conclusion that *O. marina* is an early branching member of the dinoflagellate lineage.

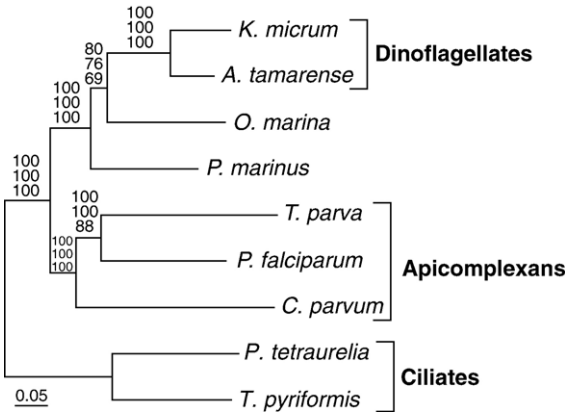

Figure 1. Phylogenetic position of *O. marina* within alveolates based on a concatenation of 30 protein sequences. A maximum likelihood tree is shown, with bootstrap support from ProML (top), PhyML (centre), and weighted neighbour-joining (bottom). All ML, Bayesian and distance trees shared the same topology shown.

Expression of *Oxyrrhis* mitochondrial genes

To determine the nature of the *O. marina* mitochondrial genome, we identified a large number of mitochondrial transcripts from an expressed sequence tag project. Altogether we sequenced 18,102 ESTs from *O. marina*, which assembled into 9876 contigs. From this sample, the two most highly represented clusters corresponded to mitochondrial genes: *cox1* was the most abundant transcript with 339 ESTs assembling into a contig of 1589 bp in length, and a *cob-cox3* fusion was the second most abundant transcript with 130 ESTs assembling into a contig of 2104 bp in length. We manually searched for other potentially mitochondrion-encoded protein-coding genes but found none; this is consistent with the protein–gene complements of both apicomplexans and dinoflagellates.[14,18,20–23,32]

In addition to the major protein-coding genes, we also identified a number of clusters with similarity to alveolate rRNA fragments. Specifically, we identified LSUE and LSUG, which are known from the apicomplexans *Plasmodium* and *Theileria*, and the dinoflagellate *Alexandrium catenella*,[15,21,33] and RNA10, which has been identified in *Plasmodium*[15] and a fragment of the same region also in *Theileria*.[33] We examined the potential for these sequences to form secondary structures compatible with those described for the corresponding regions of LSU, and in all three cases found predicted structures matching the approximate size and location as homologues from *P. falciparum*, *A. catenella* and *C. cohnii*[15,21,22] (Figure 2). This is further supported by the existence of a 9 nt region that is predicted to pair between LSUG and LSUE. RNA10 is also predicted to interact with LSUG, and in our model this interaction occurs between a CGA motif in the RNA10 arm loop and two U residues in LSUG

Table 1. Summary of *O. marina* sequences that are universally edited in known dinoflagellate *cox1* and *cob* genes

Gene/site	Edit	*O. marina* sequence	*O. marina* state	Amino acid Oxy / Dino / Plasmo
cox1-330 190	A-G	U	Unique[a]	L / (I-V) / M
cox1-351 211	U-C	U	DNA	L / (F-L) / I[b]
cox1-495 355	U-C	U	DNA	F / (F-L) / F
cox1-666 523	A-G	G	RNA	A / (I-V) / T
cox1-771 619	A-G	G	RNA	V / (I-V) / V
cox1-855 703	A-G	G	RNA	V / (I-V) / V
cox1-1178 1025	A-G	U	Unique	I / (K-R) / S[c]
cox1-1198 1046	G-C	Deletion	Unique	N/A
cob-161	C-U	A	Unique	Y / (S-F) / F[b]
cob-193	A-G	A	DNA	T / (T-A) / V
cob-230	G-C	C	RNA	A / (G-A) / A[b]
cob-292	A-G	G	RNA	V / (I-V) / L
cob-572	A-G	U	Unique	F / (Y-C) / L[b]
cob-760	U-C	U	DNA	L / (F-L) / N[a]
cob-784	C-U	U	RNA	S / (L-F) / L
cob-844	U-C	U	DNA	V / (V-L) / V[b]

[a] Unique means the *O. marina* sequence is identical to neither the DNA nor RNA sequence conserved in true dinoflagellates.
[b] A variation in the third position of the codon makes this substitution conservative.
[c] Editing at the second position of the codon, all others are edits are at the first position.

(Figure 2), similar to the interaction proposed for *P. falciparum* and *Escherichia coli*.[15,34,35]

LSUG was represented by a single EST of 82 bp with a oligo(A) tail (Figure 3(e)). LSUE and RNA10 were also present as small, oligoadenylated cDNAs, but additional cDNAs with larger flanking regions were found in both cases. LSUE was represented by 22 ESTs that assembled into three different contigs. Each contig shares a 57 bp stretch of similarity to one another and to the *P. falciparum* LSUE and a 34 bp region upstream of this with one another but not *P. falciparum* (this is probably also part of the rRNA), but the remaining 296, 520 and 697 bp at the 5′ end of each contig showed no significant similarity to one another or to any known sequence (Figure 3(c)). At the 3′ end most of the ESTs are oligoadenylated, always at the same position (13 out of the 20 that reach that point) 112 bases downstream of the similarity to LSUE, indicating that this region is part of the functional LSUE fragment. The RNA10 fragment appears in seven ESTs in two different contexts. In two cases short, 26 and 36 bp cDNAs end with an oligo(A) tail at the same position, whereas remaining five cDNAs are longer (37 to almost 600 bp at 5′ and 446 to 740 bp at 3′) (Figure 3(d)). These fall into two different contexts at the 3′ end and no significant similarity can be found between them, whereas the 5′ ends are similar (Figure 3(d)).

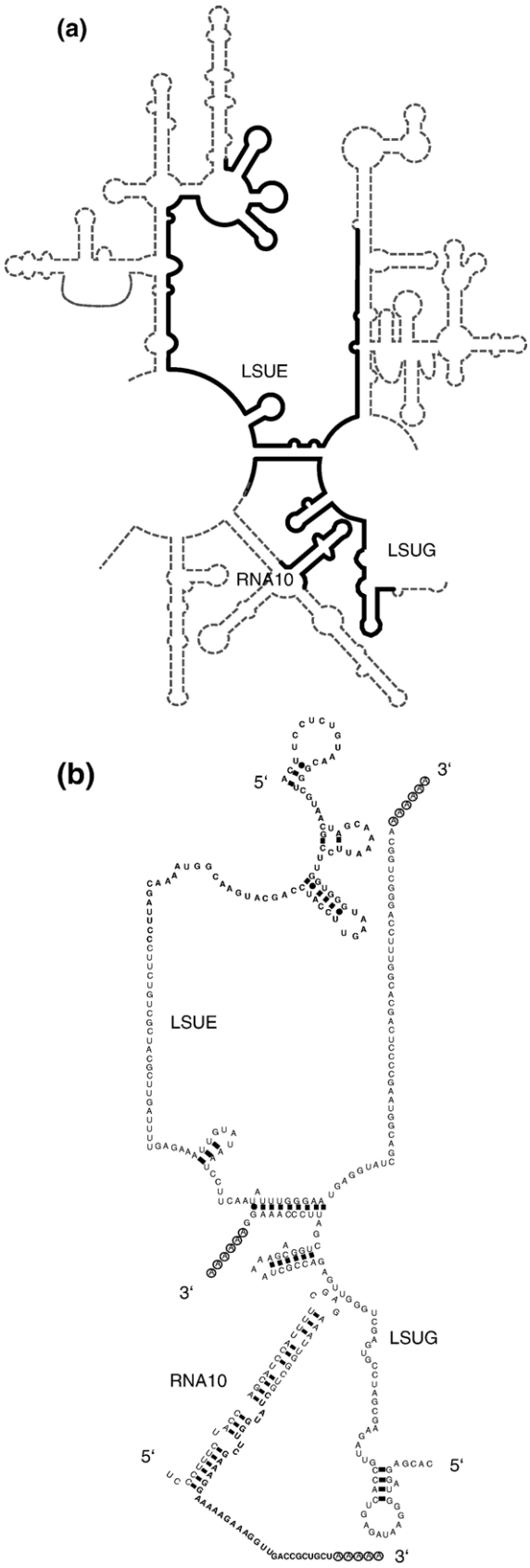

Figure 2. Predicted secondary structure conservation between *O. marina* LSUE and LSUG and *P. falciparum* homologues. (a) *O. marina* LSUE, LSUG, and RNA10 fragments mapped onto predicted secondary structure of *E. coli* LSU rRNA. (b) Primary sequence and predicted structure and interactions between *O. marina* LSUE, LSUG and RNA10 in predicted secondary structure. Structures are based on Kamikawa *et al*.[21]

RNA-circularization and analysis of cDNA ends shows *cob* and *cox3* are expressed as a fusion

From the complete set of ESTs, no clear instance of a stand-alone transcript of either *cob* or *cox3* was found, only large numbers of the *cob-cox3* fusion and truncated cDNAs. We examined the region between *cob* and *cox3* to determine if they truly represent a fusion as opposed to a polycistronic message. The two coding units are in-frame, and regions of detectable homology with other *cob* and *cox3* homologues are separated by 82 bp. Within this region there is no ATG codon or canonical stop codon, although it is not clear whether this is of significance in this genome (see below). More importantly, there is no indication of partial transcripts that only correspond to *cob* or *cox3*: there are many 5′-truncated cDNAs and eight that are truncated at various positions within the linker, but many more are truncated within *cob* (75) or within *cox3* (39). Overall truncation occurs randomly across the coding region, suggesting these are aberrant transcripts or truncated products commonly seen in cDNA synthesis. Likewise, most ESTs are polyadenylated at a specific point at the end of the *cox3* segment (see below). There is a minority of transcripts that are polyadenylayed at

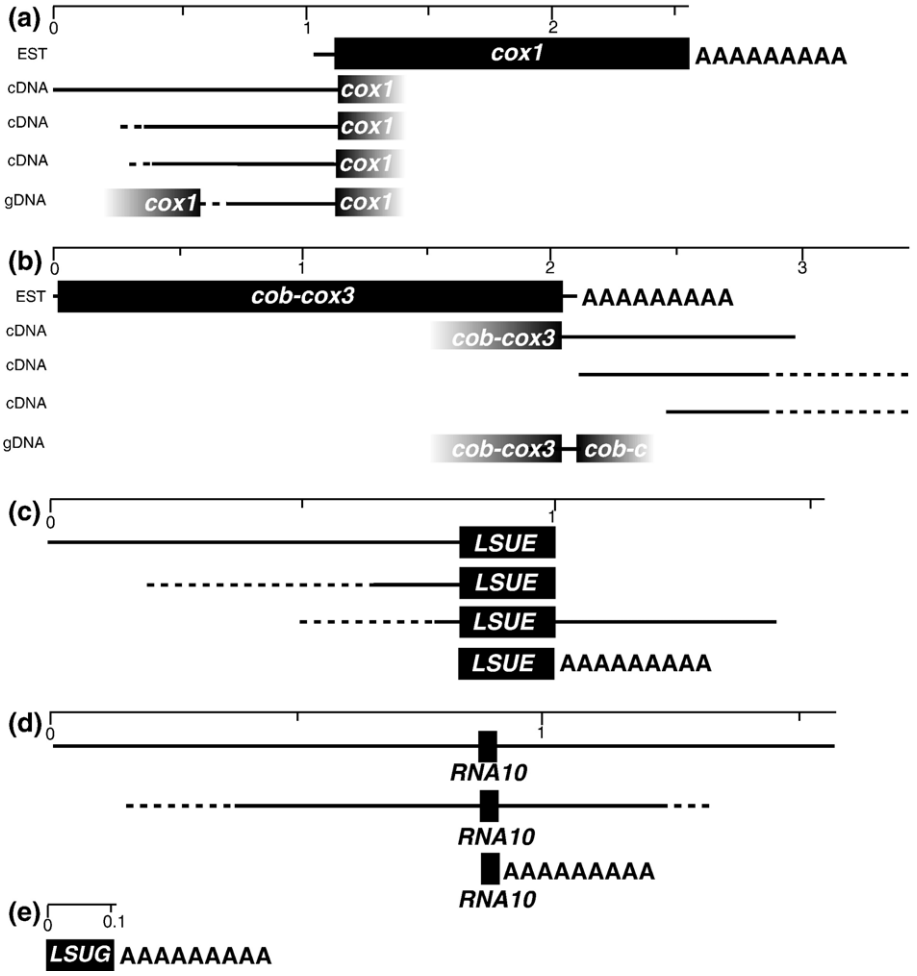

Figure 3. Genomic contexts of *O. marina* mitochondrial genes. (a)–(e) Examples of different contexts found in EST data or by amplification between genes. Scale bars are shown in kb, and are different for protein coding genes and rRNAs (the latter being smaller are shown with a 2× scale). Genes are in black boxes and non-coding sequence a line. Continuous lines represent sequence homologous to the topmost example, whereas dotted lines represent the point at which the homology breaks off and the genomic context is different. (a) and (b) For protein-coding genes, the majority of ESTs began or ended close to the predicted ends of the protein-coding sequence, and these are labelled EST. A small number of ESTs were found to extend upstream of *cox1* or downstream of *cob-cox3*, and these are labelled cDNA to distinguish them. Amplification from genomic DNA yielded several clones linking two copies of *cox1* or *cob-cox3*, and only one representative of each is shown, labelled gDNA. (c)–(e) rRNA fragments LSUE, RNA10, and LSUG were each found as a short, oligoadenylated cDNAs encoding only the region of similarity with rRNA. Longer cDNAs encoding identical LUSE and RNA10 fragments were also found, and the sequence flanking each were found to vary.

earlier positions, suggesting that the *cob* portion could be transcribed individually, but these shorter transcripts are not consistent and end randomly, in some cases well before the end of conserved *cob* regions. We therefore examined polyadenylation patterns by performing 3′-RACE with a primer located at the terminal portion of the *cob* region. Consistent with the pattern of EST data, the vast majority of product extended through *cob* to the end of *cox3* and was oligoadenylated at the same position as the ESTs. A minor fraction ended at various positions throughout the gene, and of these no two clones shared the same oligoadenylation site (not shown). In the analysis of genomic context for *cob-cox3* (see below), genomic fragments corresponding to *cob* alone were found, but many partial gene fragments were found for both genes, suggesting all are non-functional recombination products rather than functional genes.

To more definitively determine if stand-alone transcripts of either *cob* or *cox3* exist in *O. marina*, we performed RT-PCR on circularized RNA with divergent primers (cRT-PCR). This technique had been previously employed to characterise mitochondrial transcripts in plants.[36,37] cRT-PCR using outwardly oriented primers against the 5′ end of *cob* and the 3′ end of *cox3* produced fragments of variable size. Sequencing 14 of these revealed a basic structure where the 3′ end of the *cox3* coding region is followed by 10 to 16 A residues at the exact position of the oligo(A) tail in the ESTs, all followed by the 5′ end of *cob* (Figure 4(a)). This is consistent with the presence of *cob-cox3* fusion mRNA. Several clones were apparently truncated at either the 5′ end

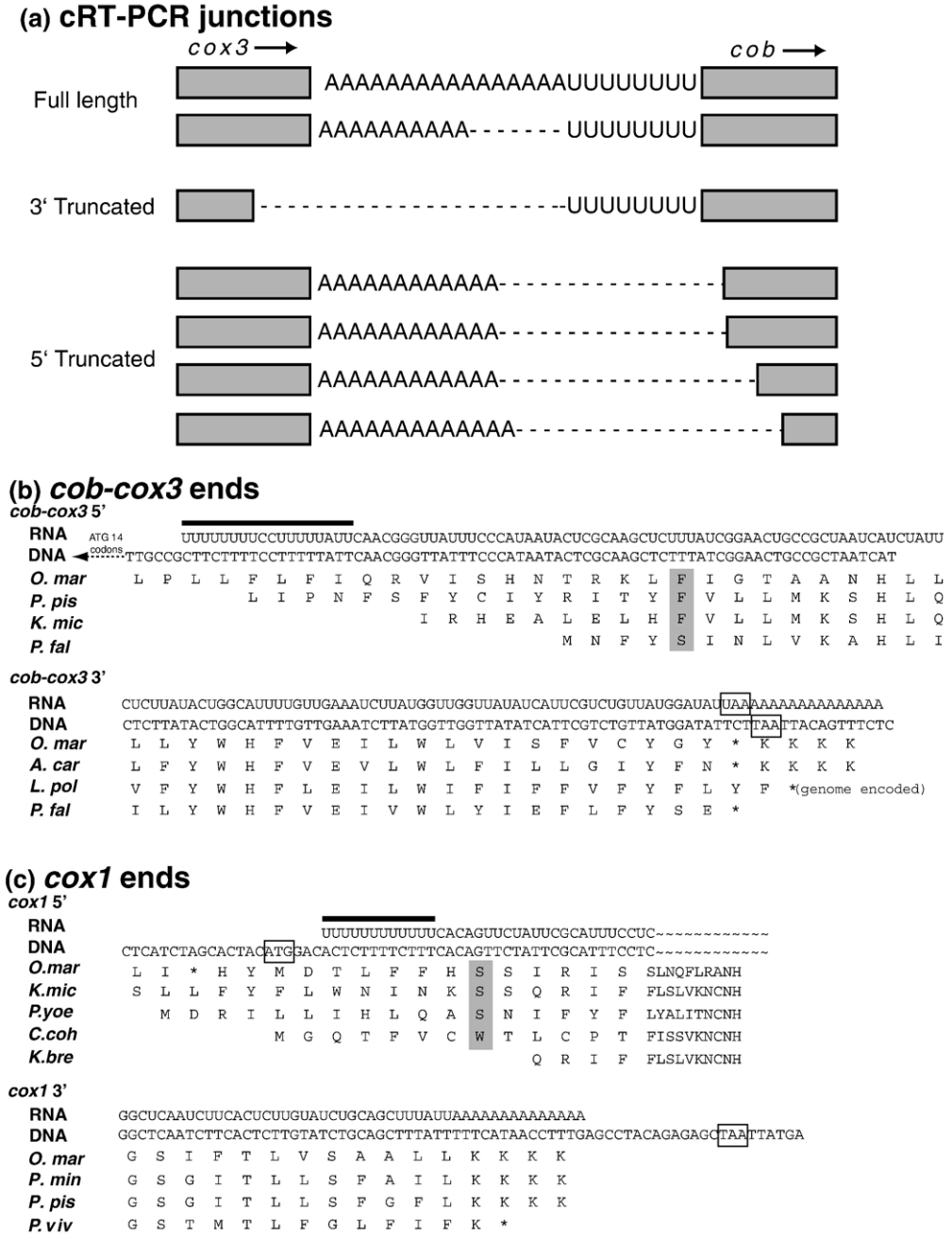

Figure 4. Initiation and termination points protein-coding mRNAs. (a) Analysis of *cob-cox3* mRNA ends by cRT-PCR. Full-length mRNAs generate products with the 3′ end of cox3, an oligo(A) tail, an oligo(U) track, all followed by the 5′ end of *cob*. In 5′ or 3′ truncated products, the presence of oligo(A) strictly corresponded with the presence of the 3′ end of *cox3*, whereas the presence of oligo(U) strictly corresponded with the presence of the 5′ end of *cob*. This confirms the oligo(A) is a 3′ tail and the oligo(U) is a 5′ cap, which was also supported by EST sequencing. (b) mRNA ends of *cob-cox3* transcripts. (c) mRNA ends of *cox1* transcripts. For each gene-end the *O. marina* mRNA and DNA sequences are aligned, and below both are translations from *O. marina*, several dinoflagellates and *Plasmodium*. The oligo(U) is indicated by a line over the RNA sequence. ATG initiation codons and TAA termination codons encoded by the genomic sequence are indicated by boxes in the DNA sequence. Note that all mRNAs initiate downstream of these ATG codons and terminate upstream of these TAA codons. The 5′ position where amino acid sequence similarity between *O. marina* and dinoflagellates is first detected is indicated by a shaded box. We propose the protein probably initates at a AUU codon near this position. The 3′ position where oligoadenylation occurs is indicated by the oligo(A) sequence. In *cob-cox3* this creates a new UAA codon 3 bp upstream of the genome encoded one (indicated by a box in the RNA sequence), and in *cox1* there is no termination codon. Consistent with *O. marina*, dinoflagellate *cox1* and *cob* mRNAs lack termination codons, and *cox3* mRNAs create one by oligoadenylation (except for *L. polyedrum*, where the UAA is encoded in the genome). Abbreviations (and accession numbers) are: *O. mar, Oxyrrhis marina; K. mic., Karlodinium micrum* (DQ082989)*; K. brev., Karenia brevis* (CO060303)*; C. con., Crypthecodinium cohnii* (AF182643)*; P. min., Prorocentrum minimum* (AF463415)*; A. car. Amphidinium carterae* (CF064811)*; P. pis., Pfiesteria piscicida* (AF463413)*; L. pol., Lingulodinium polyedrum* (CD809384)*; P. yoe., Plasmodium yoelii* (M29000)*; P. fal., Plasmodim falciparum* (AY910014)*; P. viv., Plasmodium vivax* (NC007243).

of *cob* or the 3′ end of *cox3* (never the opposite), suggesting some mRNA degradation at each end. In contrast, in all amplifications where primers corresponding to the 5′ and 3′ ends of one moiety (i.e. either *cob* or *cox3* alone) were used together, no products were observed. This suggests that no stand-alone mRNAs for either cob or *cox3* exist, or at least are not sufficiently abundant to be detected by RT-PCR.

The possible existence of stand-alone transcripts of cob and cox3 is not exhaustively eliminated by our data, but the fusion mRNA certainly must be the major mRNA, and we predict that Northern analyses will generate a similar result to the many ESTs characterized here. Given the absence of canonical start and stop codons from *O. marina* mRNAs (see below), we also cannot rule out the possibility that the fusion mRNA is a polycistronic message from which Cob and Cox3 are translated individually. However, since the *cob* and *cox3* moieties are in the same frame and there is no evidence for non-UAA based termination, it is more likely they are translated as a fusion protein. Cob and Cox3 are members of different electron transport complexes, so it is difficult to imagine how they might both function as a fusion protein. In an analogous situation in *Acanthamoeba castellanii*, *cox1* and *cox2* are fused and co-transcribed, but Western blotting with a Cox2 antibody revealed only a Cox2-sized protein and no evidence for a fusion-protein.[38] Whether the *O. marina* Cob and Cox3 proteins are post-translationally processed in an analogous fashion to what appears to happen with *A. castellanii* Cox1 and Cox2 could be determined by Western blotting. Another interesting possibility is that Cox3 is no longer functional in *O. marina*. Indeed, *cox3* is absent from the mitochondrial genome of several ciliates,[10–12] and there is no evidence of a mitochondrion-targeted homologue in the *T. thermophila* nuclear genome.[39] Moreover, the *O. marina cox3* moiety is highly divergent, with less than half its length sharing even detectable sequence similarity to *cox3* homologues.

Protein-coding mRNAs have a 5′ oligo(U) cap

Of the cRT-PCR clones characterized, only three began at the same 5′ position as any other clone, and all three of these began with exactly eight U residues (Figure 4(a)). In the genomic sequence this corresponds to a T-rich region, but not a stretch of uninterrupted T residues, suggesting the oligo(U) cap is not simply the result of transcription. These U residues are inferred to be at the 5′ end of the mRNA because they do not appear in any of the cRT-PCR products that start downstream of this position (i.e. clones that are 5′ truncated), whereas they are present in the one clone that starts at this position but lacks a oligo(A) tail (i.e. one that is 3′ truncated) (Figure 4(a)). Two sense-strand *cob-cox3* ESTs start 126 and 75 bp before the putative 5′ end suggested by cRT-PCR, but they do not contain additional U residues. A third EST, however, starts at exactly the same point as the cRT-PCR clones and has also one non-genome encoded U residue at the 5′end (Figure 4(b)). This is consistent with the likely start point of the mRNA (see below), and suggests that the longer transcripts are likely spurious. Examined *cox1* ESTs were consistent with this, since the longest at the 5′ end also begins with a run of nine U residues that are not encoded by the DNA sequence, but that do map to a T-rich region (Figure 4(c)). We examined all ESTs corresponding to rRNA fragments and none were observed to encode additional 5′ U residues (not shown), altogether suggesting that the *O. marina* mitochondrion uses a unique 5′ oligo(U) cap on the mRNAs of its two protein-coding genes that are not encoded in the genome. Whether true dinoflagellates share such a characteristic has not been tested, but in apicomplexa oligo(C)-mediated 5′-RACE shows no sign of such a 5′ modification.[27]

How the mRNAs come to have a 5′ oligo(U) cap is not clear. It is possible that a short stretch of U residues is simply added to nascent mRNAs. However, the fact that the oligo(U) caps of both transcripts correspond to T-rich regions of the genome suggests an imperfect template-directed process may be at work. It is possible that transcription initiates with a U-bias resulting in the C residues encoded in the genome being transcribed as U, but this is not supported by the lack of such transitions at 5′ ends of rRNA transcripts or the mRNAs that extend beyond this position for *cob-cox3*. It is also possible these positions are post-transcriptionally edited. C-U transitions are one of the common edits in dinoflagellate mitochondria. As discussed below, we find no evidence for editing in the protein-coding regions of *O. marina* mitochondrial genes, and no evidence that editing has ever operated in this genome. If these sites are edited, they are quite extraordinary because they could represent a restricted form editing system, where edits were constrained to the non-coding regions of the mRNA, presumably for some reason relating to mRNA structure or stability. Altogether we feel it is unlikely that an editing system like that of dinoflagellates could originate but remain so restricted in one lineage while proliferating in another, so until more evidence is brought to bear on the oligo(U) cap of *O. marina* mRNAs we do not conclude they are edited by a homologous system to that operating in true dinoflagellates.

Oxyrrhis mitochondrial transcripts use alternate start codons

In the *Plasmodium* mitochondrion, *cox1* and *cox3* lack canonical AUG initiation codons.[14,27] In true dinoflagellates there is no compelling case for the use of AUG initiators in any of the three mitochondrial genes (unpublished data). We therefore compared the 5′ ends of the two *O. marina* protein-coding genes with their transcripts and found a similar situation to that of dinoflagellates. The *cox1* gene encodes an ATG, but no sense transcripts extend that far, whereas over 200 cDNAs start

approximately 45 bp downstream of this codon. Based on sequence similarity between *O. marina* and dinoflagellates, we propose the protein begins at an isoleucine codon (AUU) nine bases downstream of the oligouridylylated *cox1* EST (Figure 4(c)). The *cob-cox3* fusion gene also encodes an ATG, but once again neither cRT-PCR nor cDNA sequences contained this region (Figure 4(a) and (b)). All of the oligouridylylated clones begin at the same position and there are no AUG codons between the oligouridylylation site and the region of sequence similarity with other *cob* genes. There is an AUU codon 23 bp downstream of the oligouridylylation site (Figure 4(b)). Altogether we conclude AUG is no longer an initiation codon in the *O. marina* mitochondrion, and AUU is more likely used in both genes.

Stop codons are absent from *cox1* and created by oligoadenylation in *cob-cox3*

An in-frame TAA codon appears several codons downstream of the expected endpoint of the *cox1* coding sequence, but in all cDNAs oligoadenylation occurs prior to this position (Figure 4(a)). The position of oligoadenylation is extremely consistent: in all 23 cDNAs where the extreme 3′ end was sequenced, processing occurs at exactly the same nucleotide. Moreover, the processing point corresponds to the expected end of *cox1* based on sequences from true dinoflagellates, so it appears that *cox1* lacks a termination codon, and that the protein is therefore oligolysinated at the carboxy terminus.

There is also an in-frame TAA codon only slightly downstream of the expected end point of the *cob-cox3* protein, but once again the mRNA is oligoadenylated prior to this position (Figure 4(b)). As with *cox1*, the position of this processing point is very consistent, being identical in all 26 cDNAs examined, and in all cRT-PCR products including a oligo (A) tail. In this case, however, oligoadenylation takes place downstream of a T residue, creating another UAA codon a mere 3 bp upstream of the one encoded by the gene (Figure 4(b)). The creation of termination codons by oligoadenylation is also known from human mitochondria.[40,41]

In ciliates and *Plasmodium* UAA is the sole termination codon: UGA encodes tryptophan and UAG is not used.[10–12,27,42] The situation we observe in *O. marina* is similar to that of true dinoflagellates, where oligoadenylation takes place upstream of termination codons, either eliminating them or making a new UAA[20] (unpublished data). Indeed, we have observed a strict pattern where *cox1* and *cob* transcripts invariably lack a stop codon, but *cox3* transcripts invariably use UAA, in all but one instance created by oligoadenylation. *O. marina* and dinoflagellates are therefore striking because the translation system is able to recognize UAA in *cox3* or *cob-cox3* mRNAs, but also cope with *cox1* and *cob* mRNAs that lack stop codons. If the termination system is at least partially maintained to service *cox3* transcripts, why is the absence of termination codons so prevalent in *cox1* and *cob*? Cases of individual genes lacking termination codons have been reported from plant mitochondria[37] and functional proteins can be produced from human mitochondrial genes lacking a termination codon.[40] *O. marina* and dinoflagellates are therefore not unique, although their systematic absence of termination codons is not seen elsewhere.

Oxyrrhis mitochondrial genes are found in two distinct genomic contexts

The multiple genomic contexts of LSUE and RNA10 rRNA fragments are reminiscent of the highly fragmented and complex ones that appear to be common to dinoflagellate mitochondrial genomes.[21,22] Indeed, the ends of both LSUE and RNA10 contigs overlap with many additional EST sequences that are further differentiated, suggesting additional complexity (not shown). Similarly, three individual ESTs overlapping the 5′ end of *cox1* were found, and although they are identical to one another, further upstream they are non-homologous and show no similarity to any known sequence (not shown). Three ESTs extend downstream of *cob-cox3* several hundreds of base-pairs, and they too lack similarity to any known sequence. They are also further differentiated and exhibit fragmented similarity to other, non-coding ESTs. All this suggests that the mitochondrial genome of *O. marina* is fragmented and possibly organized similarly to that of dinoflagellates, so to directly investigate the contexts of both protein-coding genes we characterized the genes and their flanking regions from genomic DNA.

First, we confirmed the existence in the genome of the three distinct non-coding regions upstream of *cox1*, as well as the coding regions themselves, and by amplification using gene-specific primers based on cDNA sequences found all to be present in the genome. To examine the flanking regions of both *cox1* and *cob-cox3* we used two approaches. Initially, we carried out PCR using different combinations of outward-facing, gene-specific primers to attempt to amplify fragments where all possible orientations of both genes are physically linked in the genome. Following this, we carried out PCR using a single outward-facing, gene-specific primer by itself, to generate potentially random fragments upstream and downstream of both genes.

In all the products that were characterized using either approach, we could only find a gene linked to itself or a fragment of itself, but never observed both genes on the same DNA fragment. For example, all genomic sequences found upstream of *cox1* corresponded to another *cox1* gene. In all cases where both 5′ and 3′ gene-specific primers were used for one gene, the two genes were separated from one another by stretches of non-coding DNA of varying length and otherwise appeared to be intact genes (e.g. Figure 3(a) and (b); Supplementary Data, Figure 2). In all cases where a single gene-specific

primer was used, the resulting genomic fragment consisted of the end of one apparently intact gene, a spacer, and then a truncated fragment of the same gene in the opposite strand (Supplementary Data, Figure 2). In no case was *cox1* found on the same genomic fragment as *cob* or *cox3*, and in no case was an rRNA segment found on the same genomic fragment as a protein-coding gene.

This contrasts with the EST sequences that overlap with the 5′ end of *cox1* and the 3′ end of *cob-cox3*, which extend for over 1 kb and almost 1.5 kb, respectively, without hitting another copy of the same gene (Figure 3). We examined the GC content these sequences to assess if they are nucleus-encoded pseudo-genes. The GC content of three non-coding regions upstream of *cox1* were 40, 38 and 39%, and the three downstream of *cob-cox3* were 39, 36 and 39%. Mitochondrial coding sequences were similar at 35%, and sequences between copies of the same gene were 40%. All these contrast with nuclear genes (56%, based on 15,418 bp of coding sequence) and intergenic regions (56% based on 1200 bp of 3′ UTR). This is most consistent with the non-coding regions upstream of *cox1* and downstream of *cob-cox3* being genuine mitochondrial sequences. This means that the protein-coding genes are most likely found in at least two different kinds of context, closely spaced tandem repeats of a particular gene, or within larger tracts of non-coding sequence.

Overall, therefore, there are two different kinds of context for protein-coding genes in the *O. marina* mitochondrion: there are closely spaced repeats of a single homologue or fragments of it, some of which are circular-mapping, and there are less dense fragments where a single gene has been identified to date. The two protein-coding genes appear to be found on distinct molecules where they are arranged as tandem linear repeats of the same gene (some closely spaced, others distantly spaced perhaps), circles with one or more copies of the same gene, or a mixture of both. This is a unique combination of characteristics found in both apicomplexa and the true dinoflagellates. The apicomplexans share with this model the presence of tandem repeats,[14,15] while the true dinoflagellates appear to have multiple chromosomes, but containing fragments of more than one gene, all found in a variety of flanking contexts.[20–22] In *O. marina* we find no evidence for DNA molecules containing two different mitochondrial genes or gene fragments, something that suggests that the genome is more orderly than that of dinoflagellates, although still more complex in structure than the apicomplexan one.

RNA editing is absent in *Oxyrrhis* mitochondria

The best-studied characteristic of dinoflagellate mitochondrial genomes is RNA editing. This unique form of substitutional editing has been found in *cob* and *cox1* from several diverse dinoflagellate genera and affects about 2% of the sequence, nearly all at first and second positions, resulting in a change to the predicted amino acid sequence of the protein.[18] The most common kinds of edits found to date are A-G, U-C and C-U substitutions, but G-U, G-A, G-C, U-A, and U-G edits have also been observed.[18,23] Overall, the pattern for changes and the diverse types of substitutions suggest that the mechanism or combination of mechanisms responsible for these changes are unlike other known forms of editing.[18,24]

To determine if editing is present in *O. marina*, we compared the genomic and cDNA copies of both *cox1* and the *cob-cox3* fusion. All transcripts from both genes were 100% identical to all corresponding gene fragments, demonstrating that editing is not present. This does not take into account the oligo(U) caps, which could result from several C-U edits, but given the lack of editing at any other sight, we think it premature to conclude this is editing until further evidence supports this possibility. We also examined the distribution of character states in the *O. marina* sequences to determine if they are predominantly pre-edited *versus* post-edited compared to dinoflagellates. This is significant for two reasons. If *O. marina* had lost the ability to edit its genome, it might contain many pre-edited states. More importantly, if the genome is ancestrally free of editing, we can use it to determine the effects that editing had on the protein products of dinoflagellate genes. Although editing almost certainly affects dinoflagellate *cox3*, abundant data are only available for *cox1* and cob, and since the *O. marina cox3* is also relatively divergent, only *cox1* and *cob* were analysed.

We identified all positions in both genes where all known dinoflagellate sequences are uniformly edited, and compared these pre and post-edited sequence with that of *O. marina*. Altogether, eight *cox1* sites and eight *cob* sites were uniformly edited in a variety of dinoflagellates (Table 1), and the ratio of post-edited to pre-edited states in *O. marina* is 7:4 (in the other five the *O. marina* state matches neither). When the amino acid residues encoded at these sites are examined, an interesting trend emerges. For nearly all of the sites where the *O. marina* gene encodes the post-edited state, the predicted amino acid sequence of *O. marina* and the dinoflagellates is also identical (with the exceptions of *cox1*-523 and *cob*-784, where the *O. marina* amino acid matches neither). This shows that editing of these positions in dinoflagellates restores the ancestral state in the protein, as expected of editing.[18] The situation with the five positions where the *O. marina* state matches the pre-edited dinoflagellate sequence is somewhat more complex. In two of these (*cox1*-211 and *cob*-760), the *O. marina* amino acid sequence matches the translation of the post-edited dinoflagellate sequence, despite the fact *O. marina* has the pre-edited gene sequence. This is possible because a second substitution within the codon results in the use of two different leucine codon-sets (TTR and CTN). In the other three cases (*cox1*-355, cob-193, and cob-844), the *O. marina*

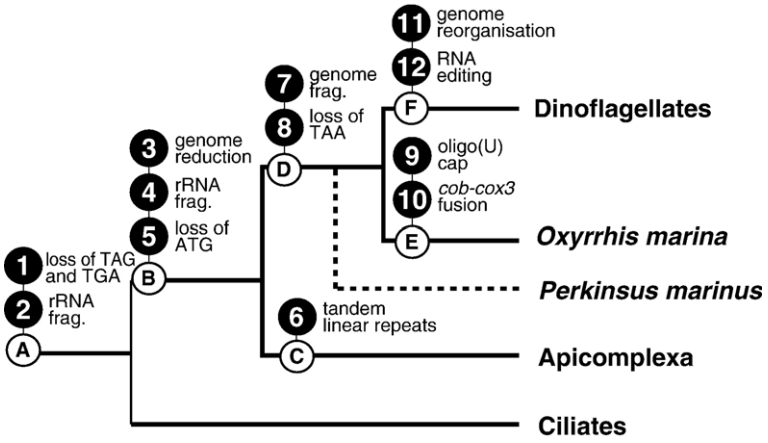

Figure 5. Evolution of alveolate mitochondrial genomes. A hypothetical phylogeny of alveolates based on Figure 1 and other data,[8,9,25,26] plotting the distribution of mitochondrial genome characters (black circles with short descriptions). The state of characters on branch D may potentially be ordered by characterisation of the *P. marinus* genome. More detailed descriptions of characters are as follows. (1) Loss of TAG and conversion of TGA to tryptophan, leaving only TAA as stop. (2) Partial fragmentation of rRNA genes. (3) Massive movement of genes to nucleus leaving only *cox1*, *cox3*, and cob. (4) Massive fragmentation of rRNA genes (may also continue on branches C–F). (5) Loss of canonical ATG start codons from most genes. (6) Tandem repetition of linear genome (could also take place on branch B). (7) Fragmentation of the genome. (8) Loss of termination codons (or creation by oligoadenylation). (9) 5′ Oligo(U) caps on transcripts (could be on branch D or E). (10) Fusion of *cob-cox3*. (11) Large-scale reorganisation/recombination of genome. (12) RNA editing.

protein sequence matches the translation of the pre-edited dinoflagellate sequence, and these are most interesting because in two of them the *O. marina* and pre-edited dinoflagellate sequence both match that of *Plasmodium* (Table 1). This suggests that the *O. marina* gene corresponds to the ancestral sequence, and that the dinoflagellate editing machinery is actively altering the protein sequence to be more divergent that the genes at some positions.[18]

Conclusions: reconstructing the evolution of alveolate mitochondrial genomes

The mitochondrial genome of *O. marina* is an interesting mixture of characteristics shared with dinoflagellates or apicomplexans with some found nowhere else. As such, it allows a relatively detailed reconstruction of the order of events that led to the very unusual conditions we see in alveolate mitochondrial genomes today. Plotting the various innovations onto a tree of major alveolate groups (Figure 5) reveals a stepwise accumulation of unusual structure and function traits. Some characteristics (e.g. fragmentation of rRNAs) began early, continued throughout the evolution of these genomes, and are potentially ongoing. Others arose relatively recently (e.g. RNA editing in dinoflagellates) and are therefore highly restricted in their distribution. *O. marina* and other species that are not strictly associated with one of the three major alveolate groups are particularly important in the fine-scale resolution of this evolutionary pathway, since they are descendents of early ancestors of these major lineages, and therefore may have diverged before the patterns that define these lineages were entirely established (e.g. the absence of RNA editing in *O. marina*). By the same token, examining the structure and gene expression of a number of other such species, the so-called "protalveolates", will be of interest.[43] *Perkinsus marinus* is shown on Figure 5 because its evolutionary position is well known, but other taxa such as *Colpodella*, *Colponema*, and gregarine apicomplexans could also prove to be very informative.

Materials and Methods

Strains, cultivation, and library construction

Oxyrrhis marina strain CCMP1788 was cultivated axenically in Droop's Ox-7 medium. 20 l of culture was harvested in a continuous-flow centrifuge and stored in Trizol (Invitrogen). Total RNA was prepared in 20 ml batches according to the manufacturer's directions, resulting in 2 mg of RNA. A directional cDNA library was constructed in pBluescript II SK using EcoR1 and XhoI sites (Amplicon Express), and shown to contain 5.3×10^5 colony-forming units. Total genomic DNA was purified from the Trizol homogenate after the RNA purification step according to the manufacturer's protocol.

EST and genomic sequencing

The 5′ end-sequence from a total of 23,702 clones were assembled using TBestDB†,[44] yielding 9876 discrete clusters based on 18,102 ESTs passing quality control. Clusters were compared to public databases using a variety of search strategies.[44] Clones corresponding to the two major clusters of identifiable mitochondrial protein-coding genes were identified (*cox1* and *cob-cox3* fusion), as well as several clusters with short stretches of similarity to mitochondrial rRNAs. Twenty-three clones from the *cox1* cluster and 35 clones from the *cob-cox3* cluster were completely sequenced to verify the end points of apparently full-length and apparently truncated cDNAs.

The genomic context of each cDNA cluster was examined by amplification from genomic DNA. Gene-specific primers based on the 5′ and 3′ of *cox1*, the 5′ of *cob* and the 3′ of *cox3* were used in all possible combinations to amplify from genomic DNA. Major products of any size

† http://tbestdb.bcm.umontreal.ca

were cloned by TOPO TA-cloning (Invitrogen) and multiple copies sequenced. Approximately 80 genomic clones were characterized, and the ends of each were aligned with cDNA clusters to determine the sequence flanking each and their relationship to one another in genomic DNA. New sequences have been deposited in GenBank under accession numbers EF680822-EF680839.

Sequence analysis

Putative identification of ESTs was carried out by AutoFact automated annotation software,[45] as implemented in TBestDB. Results were analysed manually and all potentially mitochondrial genes were identified based on similarity to other genes known to be mitochondrion-encoded in other organisms. Of all the protein-coding genes, only *cox1* and the *cob-cox3* fusion protein could unequivocally be assigned to the mitochondrial genome. Other mitochondrial proteins had evidence of being nucleus-encoded, or had previously been demonstrated to be so.[46] Several smaller clusters were also identified as encoding short fragments with high similarity to dinoflagellate and apicomplexan mitochondrial rRNAs.[15,18,21,28] RNA sequences were aligned to genomic sequences and the flanking regions examined manually and by Blast to determine variations in the context of a given gene.

RNA editing was examined by aligning cDNA sequences with genomic sequences from *O. marina* and all other dinoflagellates for which both DNA and RNA sequences for either *cox1* and *cob* sequences were available. For *cox1* this consisted of *Crypthecodinium cohnii, Pfiesteria piscicida,* and *Prorocentrum minimum*. For *cob* this consisted of *C. cohnii*, *Alexandrium tamarense*, *Pfisteria shumwayae*, *P. piscicida*, *Karlodinium micrum*, *Prorocentrum micans* and *P. minimum*.

The phylogenetic position of *O. marina* was analysed independently using non-mitochondrial genes from the EST project. Thirty protein-coding genes with sampling from the gonyaulacalean *A. tamarense* were individually aligned with those from another dinoflagellate (*Karlodinium micrum*), ciliates (*Tetrahymena thermophila* and *Paramecium tetraurelia*), apicomplexans (*Plasmodium falciparum*, *Theileria parva* and *Cryptosporidium parvum*), and *Perkinsus marinus*. Alignments were concatenated and phylogenetic analysis performed using Bayesian, maximum likelihood, and distance methods. Maximum likelihood trees and 100 bootstrap replicates were inferred using ProML 3.6[47] with the JTT substitution matrix, MSRV modelled on a gamma distribution with four rate categories and invariable sites (shape parameter estimated to be 0.814) plus invariable sites (estimated to be 0). ML trees and 1000 bootstrap replicates were also inferred using PhyML 2.4.4[48] with the WAG substitution matrix and four gamma categories. Bayesian trees were inferred using MrBayes 3[49] with a WAG substitution matrix. A total of 800,000 generations were calculated with trees sampled every 100 generations and with a prior burn-in of 16,000 generations. Sampled trees were imported into PAUP*, and a majority rule consensus tree was constructed from the post-burn-in trees. Distances were also calculated using Tree-Puzzle 5.2[50] with the same settings (and parameters estimated from the data), and trees constructed using WEIGHBOUR 1.0.1a.[51] One-hundred bootstrap replicates were carried out using puzzleboot (shell script by A. Roger & M. Holder‡).

‡ http://www.tree-puzzle.de

Circular RNA RT-PCR analysis of cDNA ends

The fusion of *cob* and *cox3* was examined following the method described.[37] Briefly, 3 μg of RNA was digested with DNase I (Fermentas) and incubated with 10 units of T4 RNA ligase (Fermentas) as instructed. Self-ligated RNA was used as template for RT-PCR experiments using RT-PCR Superscript II (Invitrogen) with outwardly oriented primers at 255 bp before the end and about 400 bp after the start of the *cob-cox3* fusion gene. PCR products were cloned using TOPO TA-cloning (Invitrogen).

Acknowledgements

This work was supported by a grant from the Natural Sciences and Engineering Council of Canada (227301), and by a grant to the Centre for Microbial Diversity and Evolution from the Tula Foundation. EST sequencing from *O. marina* was supported by the PEP project of Genome Canada/Genome Atlantic. We thank TBestDB for supporting analyses, R. Anderson for assistance with large-scale culturing, M. Gray for useful discussion and Kate Thompson for assistance with cloning and sequencing. P.J.K. is a Fellow of the Canadian Institute for Advanced Research, and a Senior Scholar of the Michael Smith Foundation for Health Research.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.06.085

References

1. Gray, M. W., Burger, G. & Lang, B. F. (2001). The origin and early evolution of mitochondria. *Genome Biol.* **2**, REVIEWS1018.
2. Gray, M. W., Lang, B. F. & Burger, G. (2004). Mitochondria of protists. *Annu. Rev. Genet.* **38**, 477–524.
3. Simpson, L. & Emeson, R. B. (1996). RNA editing. *Annu. Rev. Neurosci.* **19**, 27–52.
4. Burger, G., Forget, L., Zhu, Y., Gray, M. W. & Lang, B. F. (2003). Unique mitochondrial genome architecture in unicellular relatives of animals. *Proc. Natl Acad. Sci. USA*, **100**, 892–897.
5. Marande, W., Lukes, J. & Burger, G. (2005). Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. *Eukaryot. Cell*, **4**, 1137–1146.
6. Simpson, A. G., Lukes, J. & Roger, A. J. (2002). The evolutionary history of kinetoplastids and their kinetoplasts. *Mol. Biol. Evol.* **19**, 2071–2083.
7. Lukes, J., Hashimi, H. & Zikova, A. (2005). Unexplained complexity of the mitochondrial genome and transcriptome in kinetoplastid flagellates. *Curr. Genet.* **48**, 277–299.
8. Fast, N. M., Xue, L., Bingham, S. & Keeling, P. J. (2002). Re-examining alveolate evolution using multiple

protein molecular phylogenies. *J. Eukaryot. Microbiol.* **49**, 30–37.
9. Gajadhar, A. A., Marquardt, W. C., Hall, R., Gunderson, J., Ariztia-Carmona, E. V. & Sogin, M. L. (1991). Ribosomal RNA sequences of *Sarcocystis muris*, *Theileria annulata* and *Crypthecodinium cohnii* reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates. *Mol. Biochem. Parasitol.* **45**, 147–154.
10. Brunk, C. F., Lee, L. C., Tran, A. B. & Li, J. (2003). Complete sequence of the mitochondrial genome of *Tetrahymena thermophila* and comparative methods for identifying highly divergent genes. *Nucl. Acids Res.* **31**, 1673–1682.
11. Burger, G., Zhu, Y., Littlejohn, T. G., Greenwood, S. J., Schnare, M. N., Lang, B. F. & Gray, M. W. (2000). Complete sequence of the mitochondrial genome of *Tetrahymena pyriformis* and comparison with *Paramecium aurelia* mitochondrial DNA. *J. Mol. Biol.* **297**, 365–380.
12. Pritchard, A. E., Seilhamer, J. J., Mahalingam, R., Sable, C. L., Venuti, S. E. & Cummings, D. J. (1990). Nucleotide sequence of the mitochondrial genome of *Paramecium*. *Nucl. Acids Res.* **18**, 173–180.
13. Rusconi, C. P. & Cech, T. R. (1996). Mitochondrial import of only one of three nuclear-encoded glutamine tRNAs in *Tetrahymena thermophila*. *EMBO J.* **15**, 3286–3295.
14. Feagin, J. E. (1992). The 6-kb element of *Plasmodium falciparum* encodes mitochondrial cytochrome genes. *Mol. Biochem. Parasitol.* **52**, 145–148.
15. Feagin, J. E., Mericle, B. L., Werner, E. & Morris, M. (1997). Identification of additional rRNA fragments encoded by the *Plasmodium falciparum* 6 kb element. *Nucl. Acids Res.* **25**, 438–446.
16. Esseiva, A. C., Naguleswaran, A., Hemphill, A. & Schneider, A. (2004). Mitochondrial tRNA import in *Toxoplasma gondii*. *J. Biol. Chem.* **279**, 42363–42368.
17. Norman, J. E. & Gray, M. W. (1997). The cytochrome oxidase subunit 1 gene (*cox1*) from the dinoflagellate, *Crypthecodinium cohnii*. *FEBS Letters*, **413**, 333–338.
18. Lin, S., Zhang, H., Spencer, D. F., Norman, J. E. & Gray, M. W. (2002). Widespread and extensive editing of mitochondrial mRNAS in dinoflagellates. *J. Mol. Biol.* **320**, 727–739.
19. Zhang, H. & Lin, S. (2005). Development of a cob-18S rRNA gene real-time PCR assay for quantifying *Pfiesteria shumwayae* in the natural environment. *Appl. Environ. Microbiol.* **71**, 7053–7063.
20. Nash, E. A., Barbrook, A. C., Edwards-Stewart, R. K., Bernhardt, K., Howe, C. J. & Nisbet, R. E. (2007). Organisation of the mitochondrial genome in the dinoflagellate *Amphidinium carterae*. *Mol. Biol. Evol.* **24**, 1528–1536.
21. Kamikawa, R., Inagaki, Y. & Sako, Y. (2007). Fragmentation of mitochondrial large subunit rRNA in the dinoflagellate *Alexandrium catenella* and the evolution of rRNA structure in alveolate mitochondria. *Protist*, **158**, 239–245.
22. Norman, J. E. & Gray, M. W. (2001). A complex organization of the gene encoding cytochrome oxidase subunit 1 in the mitochondrial genome of the dinoflagellate, *Crypthecodinium cohnii*: homologous recombination generates two different *cox1* open reading frames. *J. Mol. Evol.* **53**, 351–363.
23. Zhang, H. & Lin, S. (2005). Mitochondrial cytochrome b mRNA editing in dinoflagellates: possible ecological and evolutionary associations? *J. Eukaryot. Microbiol.* **52**, 538–545.
24. Gray, M. W. (2003). Diversity and evolution of mitochondrial RNA editing systems. *IUBMB Life*, **55**, 227–233.
25. Leander, B. S. & Keeling, P. J. (2007). Early evolutionary history of dinoflagellates and apicomplexans (Alveolata) as inferred from HSP90 and actin phylogenies. *J. Phycol.* **40**, 341–350.
26. Saldarriaga, J. F., McEwan, M. L., Fast, N. M., Taylor, F. J. R. & Keeling, P. J. (2003). Multiple protein phylogenies show that *Oxyhrris marina* and *Perkinsus marinus* are early branches of the dinoflagellate lineage. *Int. J. Sys. Evol. Microbiol.* **53**, 355–365.
27. Rehkopf, D. H., Gillespie, D. E., Harrell, M. I. & Feagin, J. E. (2000). Transcriptional mapping and RNA processing of the *Plasmodium falciparum* mitochondrial mRNAs. *Mol. Biochem. Parasitol.* **105**, 91–103.
28. Gillespie, D. E., Salazar, N. A., Rehkopf, D. H. & Feagin, J. E. (1999). The fragmented mitochondrial ribosomal RNAs of *Plasmodium falciparum* have short A tails. *Nucl. Acids Res.* **27**, 2416–2422.
29. Patron, N. J., Waller, R. F., Archibald, J. M. & Keeling, P. J. (2005). Complex protein targeting to dinoflagellate plastids. *J. Mol. Biol.* **348**, 1015–1024.
30. Patron, N. J., Waller, R. F. & Keeling, P. J. (2006). A tertiary plastid uses genes from two endosymbionts. *J. Mol. Biol.* **357**, 1373–1382.
31. Cavalier-Smith, T. & Chao, E. E. (2004). Protalveolate phylogeny and systematics and the origins of Sporozoa and dinoflagellates (phylum Myzozoa nom. nov.). *Eur. J. Protistol.* **40**, 185–212.
32. Chaput, H., Wang, Y. & Morse, D. (2002). Polyadenylated transcripts containing random gene fragments are expressed in dinoflagellate mitochondria. *Protist*, **153**, 111–122.
33. Kairo, A., Fairlamb, A. H., Gobright, E. & Nene, V. (1994). A 7.1 kb linear DNA molecule of *Theileria parva* has scrambled rDNA sequences and open reading frames for mitochondrially encoded proteins. *EMBO J.* **13**, 898–905.
34. Van de Peer, Y., Chapelle, S. & De Wachter, R. (1996). A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucl. Acids Res.* **24**, 3381–3391.
35. Gutell, R. R., Gray, M. W. S. & chnare, M. N. (1993). A compilation of large subunit (23S and 23S-like) ribosomal RNA structures. *Nucl. Acids Res.* **21**, 3055–3074.
36. Kuhn, J. & Binder, S. (2002). RT-PCR analysis of 5′ to 3′-end-ligated mRNAs identifies the extremities of cox2 transcripts in pea mitochondria. *Nucl. Acids Res.* **30**, 439–446.
37. Raczynska, K. D., Le Ret, M., Rurek, M., Bonnard, G., Augustyniak, H. & Gualberto, J. M. (2006). Plant mitochondrial genes can be expressed from mRNAs lacking stop codons. *FEBS Letters*, **580**, 5641–5646.
38. Lonergan, K. M. & Gray, M. W. (1996). Expression of a continuous open reading frame encoding subunits 1 and 2 of cytochrome c oxidase in the mitochondrial DNA of *Acanthamoeba castellanii*. *J. Mol. Biol.* **257**, 1019–1030.
39. Eisen, J. A., Coyne, R. S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J. R. *et al.* (2006). Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.* **4**, e286.
40. Chrzanowska-Lightowlers, Z. M., Temperley, R. J., Smith, P. M., Seneca, S. H. & Lightowlers, R. N. (2004). Functional polypeptides can be synthesized from human mitochondrial transcripts lacking termination codons. *Biochem. J.* **377**, 725–731.

41. Ojala, D., Montoya, J. & Attardi, G. (1981). tRNA punctuation model of RNA processing in human mitochondria. *Nature*, **290**, 470–474.
42. Cummings, D. J. (1992). Mitochondrial genomes of the ciliates. *Int. Rev. Cytol.* **141**, 1–64.
43. Leander, B. S. & Keeling, P. J. (2003). Morphostasis in alveolate evolution. *Trends Ecol. Evol.* **18**, 395–402.
44. O'Brien, E. A., Koski, L. B., Zhang, Y., Yang, L., Wang, E., Gray, M. W. *et al.* (2007). TBestDB: a taxonomically broad database of expressed sequence tags (ESTs). *Nucl. Acids Res.* **35**, D445–D451.
45. Koski, L. B., Gray, M. W., Lang, B. F. & Burger, G. (2005). AutoFACT: an automatic functional annotation and classification tool. *BMC Bioinformat.* **6**, 151.
46. Waller, R. F. & Keeling, P. J. (2006). Alveolate and chlorophycean mitochondrial cox2 genes split twice independently. *Gene*, **383**, 33–37.
47. Felsenstein, J. (1993). In (Felsenstein, J., ed.), *PHYLIP. Phylogeny Inference Package*, 3.5 edit. University of Washington, Seattle.
48. Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704.
49. Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.
50. Strimmer, K. & von Haeseler, A. (1996). Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964–969.
51. Bruno, W. J., Socci, N. D. & Halpern, A. L. (2000). Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.* **17**, 189–197.

Edited by J. Karn

(Received 24 April 2007; received in revised form 18 June 2007; accepted 26 June 2007)
Available online 3 July 2007