The oxytricha trifallax mitochondrial genome

Estienne C. Swart  
Princeton University

Mariusz Nowacki  
University of Bern

Justine Shum  
Princeton University

Heather Stiles  
Princeton University

Brian P. Higgins  
Princeton University

See next page for additional authors

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs

Please let us know how this document benefits you.

Recommended Citation

Swart, Estienne C.; Nowacki, Mariusz; Shum, Justine; Stiles, Heather; Higgins, Brian P.; Doak, Thomas G.; Schotanus, Klaas; Magrini, Vincent J.; Minx, Patrick; Mardis, Elaine R.; and Landweber, Laura F., "The oxytricha trifallax mitochondrial genome." Genome Biology and Evolution. 4, 2. 136-154. (2012).  
https://digitalcommons.wustl.edu/open_access_pubs/3645

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
The *Oxytricha trifallax* Mitochondrial Genome

Estienne C. Swart1, Mariusz Nowacki1,4, Justine Shum1, Heather Stiles3, Brian P. Higgins1, Thomas G. Doak2, Klaas Schotanus1, Vincent J. Magrini3, Patrick Minx3, Elaine R. Mardis3, and Laura F. Landweber1,*

1Department of Ecology and Evolutionary Biology, Princeton University
2Department of Biology, University of Indiana
3Genome Sequencing Center, Washington University School of Medicine
4Present address: Institute of Cell Biology, University of Bern, Bern, Switzerland

*Corresponding author: E-mail: lfl@princeton.edu.

Accepted: 8 December 2011

Data deposition: Mitochondrial plasmid—JN383842, Mitochondrial genome—JN383843, Macronuclear-encoded mt-DNA polymerase—JN383844, Macronuclear-encoded mt-RNA polymerase—JN383845.

Abstract

The *Oxytricha trifallax* mitochondrial genome contains the largest sequenced ciliate mitochondrial chromosome (~70 kb) plus a ~5-kb linear plasmid bearing mitochondrial telomeres. We identify two new ciliate split genes (rps3 and nad2) as well as four new mitochondrial genes (ribosomal small subunit protein genes: rps-2, 7, 8, 10), previously undetected in ciliates due to their extreme divergence. The increased size of the *Oxytricha* mitochondrial genome relative to other ciliates is primarily a consequence of terminal expansions, rather than the retention of ancestral mitochondrial genes. Successive segmental duplications, visible in one of the two *Oxytricha* mitochondrial subterminal regions, appear to have contributed to the genome expansion. Consistent with pseudogene formation and decay, the subtermini possess shorter, more loosely packed open reading frames than the remainder of the genome. The mitochondrial plasmid shares a 251-bp region with 82% identity to the mitochondrial chromosome, suggesting that it most likely integrated into the chromosome at least once. This region on the chromosome is also close to the end of the most terminal member of a series of duplications, hinting at a possible association between the plasmid and the duplications. The presence of mitochondrial telomeres on the mitochondrial plasmid suggests that such plasmids may be a vehicle for lateral transfer of telomeric sequences between mitochondrial genomes. We conjecture that the extreme divergence observed in ciliate mitochondrial genomes may be due, in part, to repeated invasions by relatively error-prone DNA polymerase-bearing mobile elements.

Key words: split genes, segmental duplication, genome expansion, linear mitochondrial plasmid, mobile elements, extreme mitochondrial divergences.

Introduction

Although ciliates are well known for their dimorphic macronuclear and micronuclear nuclear genomes, they also possess distinctive genomes in their mitochondria. The *Paramecium* and *Tetrahymena* mitochondrial genomes were among the first confirmed to be linear and to have their telomeric sequences identified (Suyama and Miura 1968; Goddard and Cummings 1975; Morin and Cech 1986, 1988). Ciliate mitochondrial genomes are both gene-rich and relatively large (20–60 kb) (Gray et al. 1998), though many mitochondrial genes remain unclassified (Pritchard et al. 1990; Burger et al. 2000; Brunk et al. 2003; Moradian et al. 2007), in part due to their extreme divergences from other eukaryotic mitochondrial genomes (Pritchard et al. 1990; Burger et al. 2000; Moradian et al. 2007). Split ribosomal RNA and *nad1* genes (Seilhamer, Gutell, et al. 1984; Seilhamer, Olsen, et al. 1984; Schnare et al. 1986, 1995; Heinonen et al. 1987; Pritchard et al. 1990; Burger et al. 2000) were also discovered in ciliate mitochondria. Some anaerobic ciliates contain hydrogen-producing organelles, or hydrogenosomes, that may derive from mitochondria, and the ciliate *Nyctotherus* (*Armophorea*) has a partially sequenced hydrogenosome genome (Akhmanova et al. 1998; Boxma et al. 2005). *Nyctotherus* may be more closely related to *Euplotes*...
and *Oxytricha* (*Spirotrichea*) than to *Paramecium* and *Tetrahymena* (*Oligohymenophorea*), though this relationship still lacks convincing phylogenetic support (Ricard et al. 2008; de Graaf et al. 2009). Comparison of the mitochondrial and hydrogenosome genomes will permit examination of these relationships.

Sequencing and assembly of the macronuclear genome of *Oxytricha trifallax* also yielded most of its mitochondrial genome, which we completed by polymerase chain reaction (PCR) and sequencing. With the addition of this genome and the availability of complete mitochondrial genomes from two different ciliate phyla—the spirotrichs *Oxytricha* and *Euplotes* and oligohymenophorans *Paramecium* and *Tetrahymena*—detailed comparative genomic studies of ciliate mitochondria are now possible.

### Materials and Methods

DNA isolation as described in Dawson and Herrick (1982) resulted in partially purified macronuclei, which were then used to produce libraries for Sanger and 454 sequencing from various populations of size-selected DNA. Mitochondrial DNA present in the libraries recovered by sequence information of the majority of the macronuclear genome sequence; assembly with the Newbler (proprietary: www.454.com) produced two large mitochondrial contigs from pooled 454 and Sanger sequence data (currently represented by Contig4281.1 and Contig4553.1 from the 2.1.8 assembly). Additional sequences from PCR products amplified across the missing regions completed the mitochondrial genome sequence. We completed the mitochondrial genome assembly using these additional Sanger sequences, plus smaller contigs not originally merged in the two large contigs, using the Geneious software’s assembler (Drummond et al. 2009).

To investigate the size of the mitochondrial plasmid, DNA was separated on an ethidium bromide-stained agarose gel, depurinated in-gel (0.25% HCl 15 min; washed in 0.4 M NaOH for 15 min) and transferred to Hybond XL membrane (Amersham) in 0.4 M NaOH using a Nytran TurboBlotter (Schleicher & Schuell). Labeled probe was generated by means of random priming (RadPrime, Invitrogen) of a wild-type *Oxytricha* strain JRB310 cloned PCR product. After overnight hybridization at 60 °C (0.5 M NaPO4, pH 7.2, 1% BSA, 1 mM EDTA, 7% SDS), the membrane was washed in 0.2× SSC with 0.1% SDS (30 min, 60 °C) and visualized on a GE Healthcare Storm 840 PhosphorImager.

To investigate the rRNA split gene structure in *O. trifallax*, RNA was isolated from *O. trifallax* strain JRB 310 using TRIzol according to the manufacturer’s specifications (Invitrogen, Carlsbad, CA) and treated with Ambion DNA-free (Austin, TX) to remove contaminating DNA. Clean RNA was tailed with GTP in a standard reaction (1X NEB Buffer 2, 1 mM GTP, 5 µg RNA, and 2 U Poly (U) Polymerase [New England Biolabs; Ipswich, MA]) at 37 °C for 10 min. A reverse transcriptase reaction was performed using 1.6 µg tailed RNA and the Invitrogen SuperScript III First-Strand Synthesis System (Carlsbad, CA) with a UXR C12D primer (Horton and Landweber 2000). PCR was performed using 1X buffer, 10 mM each dNTP, 1.5 mM MgCl2, 200 nM primer rnsaF (5′-TCCGGATGAACGCCGCGCGA-3′), 200 nM UXR anchor primer (5′-CATCATCATCATCTCGAGAATT-3′), ~80 ng cDNA, and 1.25 U Taq polymerase (Roche Applied Science; Indianapolis, IN). The PCR reaction conditions were: one cycle of 95 °C for 2 min, 35 cycles of 95 °C for 10 s, 58 °C for 10 s, and 72 °C for 30 s, before a final extension at 72 °C for 5 min. The same PCR was performed with an extension time of 60 s rather than 30 s for the primer pair rnsbF (5′-AGTTGCTCTGAAAGGTCGGACAA-3′) and UXR anchor, as well as the pair rnsaF (5′-CCATTAGTGGATGCC-TATATTGAATG-3′) and UXR anchor. Aliquots of each reaction were visualized in 2% agarose gels with SyBr Green using a Typhoon imager (GE Healthcare, Waukesha, WI). PCR products corresponding to expected sizes were cloned into plasmid pSC-Amp/Kan using the StrataClone PCR cloning kit ( Stratagene; Santa Clara, CA) and sequenced.

*Oxytricha* open reading frames (ORFs) were identified using a combination of Blast homology to either the NCBI nrdb or “mitochondrial” proteins from UniProt (UniProt Consortium 2011), when homology could be identified. ORFs were also predicted where no homology was detected by a custom python script, which provides a sliding window score for the probability of being a coding sequence and automatic ORF predictions in Geneious (Drummond et al. 2009). Since we do not know which start codons are employed in the *Oxytricha* mitochondrial genome, we have predicted start codons based on the start codons used in *Tetrahymena* (ATG, ATA, ATT, GTG, TTG). Predicted ORFs were at least 150-bp long.

We were able to detect homologs of most of the ciliate mitochondrial protein coding genes using conventional Blast-based homology searches. However, there are still a number of additional genes that are so divergent that they fall within or beyond the “twilight zone” of protein sequence similarity (Rost 1999), where Blast searches alone are unable to detect homology. An additional complication in these genomes is the presence of split genes, which may reduce sequence search sensitivity by shortening the regions available for local sequence alignment. We therefore used the more sensitive search technique provided by the HHpred web server (Soding et al. 2005), which uses a combination of PSI-Blast (Altschul et al. 1997) and HHsearch (an HMM-profile based search tool; Soding 2005). The latter tool is one of the fastest protein structure prediction tools with reasonable prediction accuracy (Hildebrand et al. 2009) and was recently useful in identifying an additional ciliate mitochondrial gene containing an rps3 C-terminal domain (de Graaf et al. 2009). We also used Quickphyre (Kelley and Sternberg 2009) with default parameters to attempt to find homologs for a limited
number of ORFs. Two additional techniques assisted us in classifying “unknown” ORFs in Oxytricha and other ciliates: transitive homology relationships (“chains of homology”; Brunk et al. 2003) and the inference of orthology based on extensive synteny within spirotrichs and oligohymenophorans (and, to a lesser degree, between these classes).

We used tRNAscan-SE (Lowe and Eddy 1997) with default parameters and the “mitochondrial/chloroplast” source option to identify tRNAs in the Oxytricha mitochondrial genome. Quikfold, from the UNAfold package on the DINAfold server, was used to predict the mitochondrial plasmid DNA hairpins with the temperature set to 20 °C, [Na+] 1 M, and [Mg2+] = 0 M.

Transmembrane helices were predicted using THMMP2 (Krogh et al. 2001) with default parameters. PAML version 3.15 was used to estimate $\theta$ and $\sigma$ with default parameters and the “mitochondrial/chloroplast” source option and the inference of orthology based on extensive synteny within spirotrichs and oligohymenophorans.

**Results**

Structure of the Ciliate Mitochondrial Chromosome

Approximately 70 kb O. trifallax mitochondrial genome shares a number of structural features with the existing ciliate mitochondrial genomes (fig. 1; GenBank accession JN383842). As in the Tetrahymena and Euplotes mitochondrial genomes, the Oxytricha mitochondrial genes are predominantly or exclusively arranged in two transcriptional directions, diverging from an approximately central location, whereas both the Paramecium mitochondrial genome and Nyctotherus hydrogenosome genome have one primary direction of transcription (fig. 1). The Oxytricha mitochondrial DNA has a relatively high AT content (76%, excluding telomeres) as is typical for mitochondria in general (Gray et al. 2004). To date, there seems to be little taxonomic consistency in mitochondrial genome base composition within ciliates: Tetrahymena pfyformis has an AT content similar to O. trifallax at 79% (Burger et al. 2000), whereas P. tetraurelia (Pritchard et al. 1990) and E. minuta (de Graaf et al. 2009) have a considerably lower AT content at 59% and 64%, respectively. The Nyctotherus hydrogenosome DNA (GenBank accession: GU057832.1) is also less AT rich (58.5%).

In all ciliate mitochondrial genomes, including that of Oxytricha, there is either a central (in Tetrahymena [Burger et al. 2000], Euplotes [de Graaf et al. 2009], and the hydrogenosome of Nyctotherus [de Graaf et al. 2011]) or terminal (in Paramecium [Pritchard et al. 1990]) region bearing low sequence complexity repeats. In Paramecium (Goddard and Cummings 1975, 1977; Pritchard and Cummings 1981) and Tetrahymena (Arneberg et al. 1974), the AT-rich region coincides with the origin of DNA replication (in Tetrahymena it is contained within the largest mitochondrial ORF, ym777, encoding translated [Smith et al. 2007] 1,386 aa protein of unknown function. The Tetrahymena paravorax mitochondrial genome contains the longest AT-rich stretch, ~1 kb of 96.5% AT sequence adjacent to the major site of change in transcription direction (Moradian et al. 2007). The central repeats in T. pyriformis, T. pigmentosa, and T. malaccensis are shorter, at a few hundred base pairs each. In Oxytricha, the central region is a ~140-bp long stretch of pure AT, composed of degenerate repeats of the unit (written as a POSX regular expression): ((AAAT) + (AT))+[(4,4)] which contains stretches of potentially self-complementary repeats, typically palindromes such as TATA, TATATA, and TATATATA. The presence of DNA structures that would be refractory to DNA polymerase is indicated by our difficulty in amplifying across this region using conventional PCR. The Euplotes > 1 kb central repeat region is more GC-rich than that of the other ciliate mitochondrial regions (~83.5% AT) and is comprised of semipalindromic 18-bp repeats (TANNATGTACATNTA). Paramecium possesses pure AT repeats in its terminal region (TATTTTTAAAAAAAAAAATATATAATATATAA). Nyctotherus’s hydrogenosome repeat is considerably more GC rich (46.7% AT) than all the other ciliate mitochondrial genome repeats. Since the hydrogenosome genome is still incomplete, it is possible that terminal AT-rich repeats are missing from this genome.

The O. trifallax mitochondrial genome is capped by telomeres consisting of 35-bp repeats of CGACTCCTCTATCCTCATCCTAGACTCCGCTTACT, with an unknown repeat number (the longest assembled mitochondrial telomeric repeat consists of approximately 15 repeat units) and appears to be linear, like the mitochondrial genome of Tetrahymena. As in Tetrahymena and Paramecium, we have found no macronuclear genome–encoded telomerase RNA with this repeat. The telomeric repeat units are in the same size range as those for a variety of Tetrahymena species (Morin and Cech 1988) (31–53 bp) but more GC-rich (51.4% vs. 26.0–40.0%). No sequence data for similar telomeric repeats has been published for Paramecium, for which a different end-replication model, based on cross-links between the two DNA strands, has been proposed (Pritchard and Cummings 1981; Nosek et al. 1998) nor for Euplotes or Nyctotherus.

Like Tetrahymena, the Oxytricha mitochondrial genome also has a terminal inverted repeat (TIR) just inside the telomeric repeats, comprised of a somewhat smaller region (~1,800 bp; 87.8% identical, including a 96 bp indel) (fig. 1) than Tetrahymena’s (~2,680 bp). This region is roughly bounded by a trnC and a putative trnC pseudogene (trnC-ψ). The Tetrahymena inverted repeat is largely comprised of the large subunit ribosomal RNAs and also contains tRNAs, including trnL paralogs, whereas Oxytricha’s appears to be largely comprised of protein-coding ORFs of unknown function. The presence of potentially unrelated terminal inverted duplicated genes in both Oxytricha and Tetrahymena suggests that this region may be an important source for gene duplications in these genomes. Aside from ciliates, TIRs...
are characteristic of many linear mitochondrial genomes from diverse eukaryotes, including yeasts, such as *Pichia piperi* (~1.8 kb) and *Williopsis saturnus* (~1.9 kb) (Dinouel et al. 1993); chytridomycete fungi, such as *Hyaloraphidium curvatum* (~1.4 kb) (Forget et al. 2002); cnidarians, such as *Hydra oligactis* (Kayal and Lavrov 2008); slime molds, such as *Physarum polycephalum* (~0.6 kb) (Takano et al. 1994); and unicellular green algae, such as *Chlamydomonas*...
reinhardtii (580 bp) and Polytomella parva, which have a well-conserved TIR of \(\sim 1.5\) kb shared by the four ends of its bipartite mitochondrial genome (Fan and Lee 2002). TIRs appear to be a common characteristic not only of mitochondrial genomes but of many linear eukaryotic and bacterial plasmids as well (Meinhardt et al. 1990, 1997) and have been proposed to be a solution to the end-replication problem for linear mitochondrial molecules (Dinouel et al. 1993; Vahrenholz et al. 1993).

Mitochondrial Genome Gene Content

Protein-Coding Genes

As can be seen in table 1, with the exception of a few gene losses, ciliate mitochondrial genomes share largely the same complement of known protein-coding genes. The partially sequenced hydrogenosome genome from Nyctotherus contains a subset of Oxytricha mitochondrial protein-coding genes: Nyctotherus has lost genes required for aerobic metabolism, in particular the cox genes (Boxma et al. 2005; de Graaf et al. 2011). Details of the identification and annotation of previously undiscovered or unannotated proteins in ciliate mitochondrial genomes and the Nyctotherus hydrogenosomal genome are provided in supplementary table 1 (Supplementary Material online). In total, we have been able to annotate seven previously unidentified genes in Euplotes, six in Tetrahymena and Paramecium, and three in Nyctotherus.

Oxytricha’s complement of small ribosomal proteins, in particular, is fairly complete, compared with other protist repertoires (Gray et al. 2004): all ribosomal proteins except for rpl1 and rps11 have been identified in all four sequenced ciliate mitochondrial genomes. We found homologs for all but one (ymf61) of the Tetrahymena putative ribosomal proteins, for which no homologs were found using conventional Blast searches (Brunk et al. 2003). The fact that three of the newly classified ribosomal proteins (rps4, 7, 10) are commonly encoded in protist mitochondrial genomes (Gray et al. 2004) but were missing from the Tetrahymena mitochondrial proteome survey (Smith et al. 2007) (which would have detected nuclear versions of these proteins, had they been transferred to the nucleus), instills confidence in these gene predictions. With the addition of these small subunit ribosomal proteins, most of the common mitochondrial encoded protist ribosomal proteins (Gray et al. 2004) appear to have been discovered, in the mitochondrial or nuclear genomes in ciliates.

Our annotations of a number of previously unannotated Tetrahymena mitochondrial-encoded genes, plus the availability of proteomic and bioinformatic identification of nuclear-encoded Tetrahymena mitochondrial genes (Smith et al. 2007), indicates that the remainder of the unknown Tetrahymena ORFs are largely nonribosomal, in agreement with a previous study (Brunk et al. 2003). These unknown ORFs could be novel mitochondrial proteins or proteins that...
have diverged beyond our ability to detect homology to known proteins. The sequencing of additional ciliate mitochondrial genomes (particularly ciliate classes other than spirotrichs and oligohymenophorans) may be beneficial in identifying the remaining unidentified genes in ciliate mitochondrial genomes, especially if HMM-HMM profile comparison tools (such as HHpred) are used to improve the information content and quality of the alignments underlying the query HMM profiles.

Split Protein-Coding Genes in Ciliates

\( \text{rps}3 \). The \textit{Euplotes} \text{rps}3 is unusually long (767 and 768 amino acids for \textit{E. minuta} and \textit{Euplotes crassus}, respectively) in comparison to the \text{rps}3 orthologues found in the \textit{Oxytricha} (349 aa), \textit{Tetrahymena} (330 aa), and \textit{Paramecium} (234 aa) and was shown to contain the C-terminal domain of \text{rps}3 in the 5’-terminal half of this gene (de Graaf et al. 2009) (fig. 2). The 3’ half of the \textit{Euplotes} gene has no detectable similarity to \text{rps}3. In \textit{Oxytricha}, this same gene is divided into a shorter, 5’-terminal portion containing the C-terminal \text{rps}3 domain, followed by a longer portion of unknown function. We identified an \textit{Oxytricha} homolog to a gene previously classified as \text{rps}3 (Burger et al. 2000) in \textit{Tetrahymena} and \textit{Paramecium} but disputed as such (Brunk et al. 2003). As for \textit{Tetrahymena} and \textit{Paramecium} and unlike \textit{Euplotes} (de Graaf et al. 2009), HHpred predicts with high probability (5.1 \times 10^{-06} for \textit{Oxytricha}) that an N-terminal \text{rps}3 domain is present in the mitochondrial genome, in an ORF that we label \text{rps}3\_a. It is possible that the \text{rps}3 N-terminal domain is encoded in a missing portion of the \textit{Euplotes} mitochondrial genome. It therefore appears that this is another split gene present in most, if not all, sequenced ciliate mitochondrial genomes. Accordingly, the previously disputed \text{rps}3 (N-terminal) can now be called \text{rps}3\_a, and the recently classified C-terminal \text{rps}3 portion, \text{rps}3\_b, consistent with the split gene nomenclature in Burger et al. (2000).

The long gene annotated as \text{rps}3 in \textit{Euplotes} may represent either a novel gene fusion or an incorrect annotation.
due to sequencing errors. The sum of the lengths of the Oxy-
tricha, Tetrahymena, and Paramecium rps3 domains is
roughly consistent with typical UniProt rps3 entries (e.g.,
~480 aa for Tetrahymena), although there is considerable
length variation in rps3 among species—even within fungi
alone (Sethuraman et al. 2009). Some of the rps3 genes we
inspected appear to be missing a domain (e.g., we found no
N-terminal rps3 domain in Schizosaccharomyces pombe). It
appears that there is some flexibility in the intervening rps3
domain spacer: in humans the intervening spacer between
the N- and C-terminal domains contains a single-stranded
nucleic acid binding domain (KH domain) required for stable
NF-κB regulatory complex binding, an extra-ribosomal func-
tion (Wan et al. 2007). In plants, the N- and C-terminal
rps3 domains are separated by a domain of unknown function
(Smits et al. 2007). One other case of a split arrangement
of the N- and C-terminal rps3 domains has been docu-
tmented in the slime mold Dictyostelium, where the domains

![Fig. 2.](http://gbe.oxfordjournals.org/)

*Fig. 2.* rps3 genes in ciliate mitochondrial genomes (Euplotes minuta—GQ903130; Tetrahymena pyriformis—AF160864; Paramecium aurelia—NC_001324). The rps3_a and rps3_b multiple sequence alignments are indicated above and below a schematic representation of the split rps3 genes. Regions with substantial sequence similarity are indicated in dark purple, whereas those that are poorly conserved are indicated in pink; the rps3_a and rps3_b parts are on distant loci. The rps3 extension annotated as a part of this gene in Euplotes does not align to any of the other rps3 sequences. Multiple sequence alignments were generated with Muscle (Edgar 2004) with default parameters.
are separated by long peptide sequences of unknown function (Iwamoto et al. 1998; Smits et al. 2007). Unlike Dictyostelium, the ciliate split rps3 ORFs are located some distance from one another, with multiple genes separating them, and in both Oxytricha and Euplotes, they are encoded on opposite strands.

nad2. Part of this gene was previously identified in Paramecium, Tetrahymena, and Euplotes, but its length varies greatly: the Tetrahymena and Paramecium "nad2" genes are unusually short, just 166 and 178 aa long, respectively. By contrast, the shortest curated nad2 gene in UniProt is 346 aa in Branchiostoma, whereas the longest is 538 aa in Usilago. Euplotes crassus (391 aa) and E. minuta (722 aa) are thought to possess N-terminal extensions with no substantial sequence similarity to other nad2 genes (de Graaf et al. 2009). HHpred searches using the original Tetrahymena nad2 ORF revealed that this ORF contains only a C-terminal portion of nad2.

We found an ORF (372 aa) in the Oxytricha mitochondrial genome that appears to be weakly similar to a "putative nonribosomal" Tetrahymena protein, ymf65 (Blastp e value 0.028 to T. malaccensis nad2) and to the Hydra and Phytophthora nad2 genes (e values of 1.2 and 4.8 to NCBI's nrdb, respectively). In Oxytricha, this region is separated from an ORF (167 aa) with BlastP hits to the existing annotated ciliate nad2 entries in GenBank (e values of 10^-6 to 10^-8), by a 60 aa unknown ORF. HHpred predictions for either the 372 aa Oxytricha ORF or Tetrahymena ymf65 correspond to the N-terminal half of nad2. ymf65 was previously predicted to have ten transmembrane helices (Brunk et al. 2003), the same number we obtained for the 372 aa Oxytricha ORF using THMM2 (Krogh et al. 2001) (fig. 3).

In Paramecium, the nad2 N-terminal region appears to be further split into two ORFs: ymf65_a and ymf65_b. Like the transmembrane helix sums, the length sums of the ORFs corresponding to nad2 for Tetrahymena, Paramecium, and Oxytricha (526, 518, and ~543 aa) are within the range of eukaryotic nad2 lengths in UniProt protein sequences.

The annotated nad2 gene from the Nyctotherus hydorgenosome genome (GenBank accession: AI871267.1) encodes the C-terminal nad2 portion and is a similar length (166 aa) to its Oxytricha, Tetrahymena, and Paramecium orthologs. An ORF preceding the Nyctotherus nad2 C-terminal region does not share substantial sequence similarity with—and is approximately 100 amino acids shorter than—the Oxytricha, Tetrahymena, and Paramecium nad2 N-terminal regions. The N-terminal half of the Nyctotherus nad2 appears to be the ORF currently annotated as orf371.

The sums of the transmembrane helices from the N- and C-terminal nad2 ciliate gene portions, for all but Euplotes (14 for Tetrahymena and Oxytricha, 13 for Paramecium, and 15 for Nyctotherus), are in accord with the number predicted for most nonmetazoan eukaryotes (13–14, e.g., Arabidopsis thaliana [UniProt: O05000], Dictyostelium discoideum [UniProt: O21048]). The overall spacing of the helices in the N-to-C terminal concatenated sequences of nad2 from Oxytricha, Nyctotherus, and Tetrahymena and Paramecium THMM2 profiles are also in good agreement.

In E. minuta, the entire ORF annotated as nad2 is predicted by THMM2 (Krogh et al. 2001) to contain 17 transmembrane helices, whereas E. crassus appears to have 12. The annotated nad2 from E. crassus is just half the length (391 aa) of E. minuta (774 aa) and appears to encode only the C-terminal portion of the E. minuta nad2 (57.8% pairwise identity). The pairwise alignments of the concatenated translations of the ORFs upstream of the shorter E. crassus nad2 (orf175 and orf147) to the remaining N-terminal portion of E. minuta nad2 are only 18.7% identical, suggesting that these ORFs are either highly divergent or unrelated. Judging from the lengths and transmembrane numbers of nad2 from Oxytricha, Tetrahymena, and Paramecium, the E. crassus nad2 is also a split gene, though we have not identified with certainty the location of the missing part. Barring sequencing and annotation errors, these Euplotes nad2 genes may indicate that nad2 can be split or fused in multiple ways. It therefore appears that nad2 may be split to different extents in different ciliate species (possibly independent evolutionary events), though a common nad2 N/C-terminal split appears to be shared by Tetrahymena, Paramecium, Oxytricha, and Nyctotherus.

A nad2 split gene is also present in angiosperm mitochondrial genomes (Malek et al. 1997). In these plants, nad2 is joined by the transsplicing of a group II intron (Binder et al. 1992). We have not detected any group II introns in the ciliate mitochondrial genomes by scans of the RFAM (Griffiths-Jones et al. 2005) group II intron model with Infernal (Nawrocki et al. 2009). In Oxytricha, we think it is unlikely that RNA editing removes all the stops necessary to join the nad2 ORFs (at least two stop codons would need to be eliminated or read through). Instead, it appears that, like nad1 (Seilhamer, Gutell, et al. 1984; Seilhamer, Olsen, et al. 1984; Heinenen et al. 1987; Pritchard et al. 1990; Schnare et al. 1995; Burger et al. 2000), this gene is not transspliced (supported by cDNA PCR results, data not shown), and therefore, the gene pieces are translated as separate subunits that require co-assembly to form the functional protein structure.

Split rRNA Genes

In both Tetrahymena and Paramecium, the large and small subunit rRNA genes are further split (Seilhamer, Gutell, et al. 1984; Seilhamer, Olsen, et al. 1984; Heinenen et al. 1987; Pritchard et al. 1990; Schnare et al. 1995; Burger et al. 2000) into large and small portions. The current T. pyriformis and P. tetraurelia GenBank annotations present two different structures for the rns split: Tetrahymena has a short rns_a followed by a long rns_b, whereas Paramecium has a long...
FIG. 3.—THMM2 transmembrane profile predictions for the concatenated nad2 split ORFs from *Tetrahymena pyriformis* (AF160864), *Paramecium tetraurelia* (NC_001324), *Oxytricha trifallax* (JN383842), and *Nyctotherus ovalis* (GU057832.1). THMM2 posterior probabilities are given on the y axis; the x axis length is in amino acids. Concatenation points are indicated by arrows.
**Fig. 4.**—Comparison of ciliate split rRNA genes. Solid red and blue bars represent small and large subunit rRNA coding sequences, respectively, drawn approximately to scale; discontinuous red and blue bars represent longer sequences that have been compressed due to figure space constraints; black lines represent intervening sequences which are approximately halved relative to the coding sequences, except the central, large discontinuous intervening region in ciliates, indicated by the dashed line, which represents an extensive, primarily protein-coding, region; tRNA genes are purple. The duplicated large subunit region of *Tetrahymena pyriformis* is represented here by a single locus: the primary difference between the two loci is a different rRNA succeeding *rnl_a* (*trnY* and *trnM_i*). Homology between the different segments is indicated by pastel-colored parallelograms. Question marks indicate the lack of experimental evidence in *Oxytricha* or *Euplotes* supporting or rejecting gene splits. GenBank accession numbers for the rRNA loci are: *Mus musculus* nuclear genome rRNA (BK000964.1); *Escherichia coli* genome (NC_000913); *Chlamydomonas reinhardtii* chloroplast genome (NC_005353.1); *Paramecium tetraurelia* (NC_001324), *T. pyriformis* (AF160864.1), *Euplotes minuta* (GQ903130), and *Oxytricha trifallax* (JN383842) mitochondrial genomes.

**Oxytricha trifallax** Mitochondrial Genome

*Oxytricha* and *Euplotes* were misinterpreted instead as *rns* having the same structure in both ciliates (Schnare et al. 1995). For *Euplotes*, no split rRNAs were detected using local sequence alignments of portions of the identified rRNA sequence (de Graaf et al. 2009), but no experimental support was provided for this conclusion. Northern analysis in *Nyctotherus* suggests that its SSU rRNA may be fragmented into three pieces of 1.7 kb, 750 bp, and 600 bp (Akhmanova et al. 1998).

Our inspection of alignments to the *Tetrahymena* and *Paramecium* split rRNA genes, as well as unpublished expressed sequence tag (EST) data, indicate that *Oxytricha* has the same splits identified in the LSU and SSU genes for *Paramecium* and *Tetrahymena* (fig. 4). EST data suggest that a long AT-rich DNA spacer (~141 bp; 96% AT) divides *Oxytricha* rns into two subunits, as in *Tetrahymena* and *Paramecium*, because no EST reads span this region. We also verified this split by obtaining RACE products corresponding to the 3′ end of *rns_a* (supplementary fig. 1).

**Material online.** Sequence alignments suggest that the ribosomal RNA fragment following *rns_b* is orthologous to the *rnl_a* fragment of *Tetrahymena's* SSU RNA, which is physically separated from the remaining *Oxytricha rnl_b* fragment. A short AT-rich tract (~32 bp; 91.6% AT) between the *rns_b* and *rnl_a* in *Oxytricha* is also poorly covered by expression data (Swart EC, Landweber LF, unpublished data) and hence a likely splitting point. Alignments of rns from the different ciliate species, including the *Nyctotherus* rns gene (see NCBI AJ871267.1) suggest that an rnl_b portion is located at the end of the gene annotated as rns in *Euplotes*. Furthermore, there are discrepancies in length of the annotated *Euplotes rnl* (2,230 bp) in comparison to that of other ciliate rns (~2,550–2,600 bp) (see table 3). These lines of evidence suggest that the *Euplotes* ribosomal LSU is split like that of *Tetrahymena*, *Paramecium*, and *Oxytricha*. Alignments of the ciliate SSU regions suggest that *Euplotes* could also possess the split LSU. At least one split (between *rnl_a* and *rnl_b*) is shared by all the ciliates and likely arose in their common ancestor. Additional
Table 2
Ciliate Mitochondrially Encoded tRNAs

<table>
<thead>
<tr>
<th></th>
<th>Oxytricha trifallax</th>
<th>Euploides minuta</th>
<th>Nyctotherus ovalis</th>
<th>Tetrahymena pyriformis</th>
<th>Paramecium tetraurelia</th>
</tr>
</thead>
<tbody>
<tr>
<td>trnC</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>trnE</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>trnF</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>trnH</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>trnK</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>trnL</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>trnM_i</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>trnM_ii</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>trnQ</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>trnW</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>trnY</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

Table 3
Ciliate Split Ribosomal RNA Segment Lengths (bp)

<table>
<thead>
<tr>
<th></th>
<th>Oxytricha trifallax</th>
<th>Euploides minuta</th>
<th>Tetrahymena pyriformis</th>
<th>Paramecium tetraurelia</th>
</tr>
</thead>
<tbody>
<tr>
<td>ml a</td>
<td>301</td>
<td>281*</td>
<td>289</td>
<td>280</td>
</tr>
<tr>
<td>ml b</td>
<td>2,289</td>
<td>2,230</td>
<td>2,279</td>
<td>2,315</td>
</tr>
<tr>
<td>ml</td>
<td>2,590</td>
<td>2,511* (2,230)</td>
<td>2,568</td>
<td>2,595</td>
</tr>
<tr>
<td>ms a</td>
<td>200</td>
<td>206*</td>
<td>208</td>
<td>210 (1,477)</td>
</tr>
<tr>
<td>ms b</td>
<td>1,426</td>
<td>&gt;1,431*</td>
<td>1,407</td>
<td>1,415 (204)</td>
</tr>
<tr>
<td>ms</td>
<td>1,626</td>
<td>1,637* (2,257)</td>
<td>1,615</td>
<td>1,627</td>
</tr>
</tbody>
</table>

* Length estimates from our sequence alignments.

b Estimates based on an experimental reassessment (Schnare et al. 1986) of the original P. tetraurelia tRNA split gene structure (Seilhamer, Olsen et al. 1984). Bracketed values indicate the lengths of the tRNA segments as annotated in the mitochondrial genomes deposited in GenBank (E. minuta—GQ903130 and P. tetraurelia—NC_001324).
66.7%) for the remaining pairs of tRNA orthologs (trnE, trnF, trnH, trnQ, trnW, trnY, locARNAs [Will et al. 2007]), suggesting that this gene has either been evolving under relatively relaxed constraints or positive selection since divergence from the common ancestor of these two ciliates. The divergence between the two *Oxytricha* trnM's (55.9%) is greater than that between typical eukaryotic initiator and elongator trnMs; for instance, in *Reclinomonas*, these genes are 67.1% identical, suggesting that there may be substantial functional divergence between the two classes of ciliate mitochondrial trnMs. Given the substantial divergence of ciliate trnMs, we are hesitant at this point to ascribe the role of initiator or elongator tRNA to any of the ciliate trnMs.

### A Linear Mitochondrial Plasmid

During inspection of the *Oxytricha* mitochondrial genome assembly, we discovered an additional large contig (~4.9 kb) possessing an internal region with substantial sequence similarity—251 bp at 82% identity—to one of the *Oxytricha* mitochondrial contigs (position indicated on fig. 1). We subsequently confirmed that this contig is a ~5.28-kb linear plasmid (fig. 5a), which we call m0 (GenBank accession: JN383842). The 251-bp region appears to be a “footprint” of a past recombination event between the plasmid and mitochondrial genome. The region of similarity is reminiscent of a 473-bp sequence shared by the *Physarum* mitochondrial genome and its linear plasmid mf, which has been shown to permit integration of the linear plasmid into the mitochondrial genome via homologous recombination (Takano et al. 1992).

Consistent with a mitochondrial origin, the plasmid genes appear to use the same genetic code as the ~70 kb *Oxytricha* mitochondrial chromosome. Translation with either the standard or ciliate genetic codes would produce much shorter ORFs, due to in-frame UGA codons. However, we do note that the tryptophan codon bias (the telltale signature of the mitochondrial genetic code) of these ORFs (12 UGA vs. 9 UGG) is much weaker than that of known or “ciliate-specific” mitochondrial ORFs (i.e., those that appear to have orthologs in other ciliates; 229 UGA vs. 20 UGG). The deviation from the standard *Oxytricha* mitochondrial tryptophan codon usage suggests that either this plasmid may be a relatively recent acquisition that has not yet acquired the standard mitochondrial codon usage or that selection on codon usage is weaker on this plasmid.

The plasmid contains two large ORFs. The 3’ ORF encodes a linear mitochondrial plasmid/phage-like DNA polymerase that is easily identifiable by Blast (BlastP best-hit e value 8 × 10⁻¹⁵ to a *Fusarium proliferatum* linear plasmid DNA polymerase [GenBank accession: YP_001713860]). The DNA polymerase has the characteristic “DTDS” residues of the DNA Pol B family active site (Hopfner et al. 1999) and appears to be a member of phage and linear plasmid (including mitochondrial plasmids) DNA polymerases (Kempken et al. 1992) (see next section).

The 5’ ORF has no convincing Blast or HHpred hits but appears to be an RNA polymerase based on a QuickPhyre (Kelley and Sternberg 2009) prediction: the best QuickPhyre prediction (e value 0.41; estimated precision 80%) is to an X-ray crystal structure of a phage N4 virion RNA polymerase, related to the T7-like RNA polymerases (Kazmierczak et al. 2002). Linear plasmids, including mitochondrial ones, bearing both an RNA and DNA polymerase are a typical form (Meinhardt et al. 1990; Handa 2008). We also discovered that the longest ORF in the TIR of the *Oxytricha* mitochondrial genome, encoding a 411 aa protein, is predicted by Phyre (e value 0.67; 80% precision) to be related to the same phage N4 RNA polymerase as the mO RNA polymerase (65% precision, e value 1.4). The sequence similarity shared between the mO RNA polymerase and the TIR protein is a meager 17% (global pairwise alignment, gap opening, and extensions penalties of 12 and 3, using the BLOSUM62 matrix).

We were unable to find any evidence of protein homologs of the mitochondrial plasmid ORFs in either *Paramecium* or *Tetrahymena* mitochondrial genome data. Linear mitochondrial plasmids have been identified in *P. caudatum*, *P. jenningsi* and *P. micromultinucleatum* (Endoh et al. 1994; Tsukii et al. 1994) but none of their sequences has been published. In a low coverage *Stylonychia lemmnae* genome assembly, we were able to find a telomere-lacking contig with a substantial TBlatN hit to the DNA polymerase ORF (contig12708 [http://lamella.princeton.edu/blat/getseq.cgi?454AllContigs.fna&contig12708]; e value 7 × 10⁻¹⁸). The entire contig appears to be translated with the same mitochondrial genetic code as the one used in *Oxytricha* and not the *Oxytricha/Stylonychia* macronuclear genetic code or standard genetic code (which would introduce two premature stop codons). This *Stylonychia* Blast hit suggests that mitochondrial plasmids may be present in other spirotrichous ciliates.

The m0 5’ terminal ~250 bp contains three types of short semiallomorphic repeats (fig. 5b), which are capable
of producing stem-loop structures of similar size to those of the *Physarum* linear plasmid mF 205-bp TIR region. The 3' end of mO is capped by the same type of telomeric repeats as the main mitochondrial genome. At least one example of short (5 bp), telomere-like repeats on a linear plasmid has been reported for the fungus *Fusarium oxysporum* (Walther and Kennell 1999) (the *F. oxysporum* genome does not contain telomeres since it is circular [Marriott et al. 1984]). The *P. polycephalum* mitochondrial plasmid (mF) has longer—144 bp—subterminal repeats following the plasmid TIRs and is capable of in vivo linearization of the circular mitochondrial genome by recombining with it (Sakurai et al. 2000). Unlike *Physarum*, the *Oxytricha* mitochondrial genome’s telomeric repeats appear to be established, rather than new extensions from the plasmid. This suggests that we have found the first possible case of a stable transfer of a telomere between a mitochondrial genome and a plasmid.

**Putative Macronucleus-Encoded Mitochondrial DNA and RNA Polymerases**

While no DNA or RNA polymerase genes have been documented in the mitochondrial genomes of *Tetrahymena* and *Paramecium*, there are nuclear-encoded candidates for these genes, which appear to have orthologs in the *Oxytricha* macronuclear genome as well. We sought to clarify the relationship between these putative mitochondrial polymerases and the plasmid-encoded polymerases found in *Oxytricha*.
Mitochondrial DNA polymerases are largely unknown or uncharacterized in most eukaryotes (Shutt and Gray 2006) with the exception of humans and yeast (Kaguni 2004). The opisthokont (metazoa/fungi) mitochondrial DNA polymerase (Pol gamma) is a Pol A family DNA polymerase, like bacterial DNA Pol I but a distinct (Lecrenier and Foury 2000) and divergent member of this family (20–25% identity relative to the Escherichia coli Klenow fragment [Lecrenier et al. 1997]). However, Pol gamma does not appear to exist in many eukaryotic lineages (Burgers et al. 2001) and so a different mitochondrial DNA polymerase must take its place in these organisms. In Arabidopsis and the red alga Cyanidioschyzon merolae, a single putative mitochondrial DNA polymerase, which is not orthologous to Pol gamma, is targeted to both mitochondria and plastids (Eli et al. 2003; Moriyama et al. 2008). These polymerases are more similar to bacterial DNA Pol I polymerases than to Pol gamma (Mori et al. 2005) and form a distinct clade—“plant organellar polymerases” (POPs; Moriyama et al. 2008)—comprised of diverse eukaryotic members, including the amoebobozoan, D. discoideum (Shutt and Gray 2006) the heterokont Phytophthora ramorum; diatoms; plants; red alga; and the ciliates T. pyriformis (GenBank accession: JN383844), appear to possess characteristic mitochondrial targeting signal peptides (table 4) and, therefore, are putative ciliate mitochondrial DNA polymerases.

A T-odd phage RNA polymerase homolog was identified for T. pyriformis during searches for homologues to the yeast mitochondrial and T3/T7 RNA polymerases (Cermakian et al. 1996). A complete homolog of the T. pyriformis sequence was predicted in the T. thermophila macronuclear genome assembly (Eisen et al. 2006; GenBank accession: XP_001014571) and P. tetraurelia (GenBank accession: XP_001431083) (Moriyama et al. 2008). The ciliate POPs, including that of Oxytricha (GenBank accession: JN383844), appear to possess characteristic mitochondrial targeting signal peptides (table 4) and, therefore, are putative ciliate mitochondrial DNA polymerases.

An Abundance of Subterminal Unknown Open Reading Frames

Both ends of the Oxytricha mitochondrial genome—corresponding to ~5 kb and ~14.5 kb (or in total just over 1/4 of the Oxytricha mitochondrial genome length)—contain almost exclusively ORFs without obvious homologues in any known organism or in the Oxytricha macronuclear genome. These regions constitute over half (54%) of the total unknown ORF length in the Oxytricha mitochondrial genome. The sum of these end regions accounts for the majority of the size difference between the smaller Tetrahymena and Euplotes mitochondrial genomes and Oxytricha’s larger one. The structure of the Oxytricha mitochondrial genome resembles the core Euplotes mitochondrial genome, with the exception of one large translocated region—the cob-to-nad5 gene block—located at opposite ends of these mitochondrial genomes (fig. 1), with two large blocks of unknown ORFs appended to either end.

Segmental duplications are evident from a dot plot of the ~14.5 kb telomeric end (fig. 6). The largest of these duplications is closest to the telomeric end and is ~1,450 bp long with ~91% pairwise identity. An ~170-bp region represents the sequence that has been duplicated most often. Pairwise identities relative to the first repeat (from the telomeric end) from this region decrease with increasing distance: 93.4%, 88.9%, and 74.7%. If we assume that these regions are evolving approximately neutrally, then the duplications closest to the telomeric end are younger than the distal ones. This suggests that the ~14.5-kb region arose, in part, through successive expansions resulting in up to three successive terminal duplication events. Curiously, the 251-bp segment shared by the plasmid and mitochondrial genome is located near (~120 bp from) the end of the most recent duplication of these repeats.

The largest duplication within the 14.5-kb end contains at least one long ORF (~600 bp), which appears to be evolving under similar levels of evolutionary constraint (dS/dN = 0.288) to that of the pair of ORFs from the TIRs (dS/dN = 0.302). In Tetrahymena, both nonterminal duplications of nad9 and the TIRs appear to be maintained by concerted evolution (Bruk et al. 2003), which is also likely to be the case for the Oxytricha TIR regions. These levels of constraint are somewhat lower than those we recently reported for the micronuclear-encoded Oxytricha TBE transposase paralogs (Nowacki et al. 2009). The synonymous substitution rate is also lower in these genes—0.095 for ~600 bp duplicated ORF and 0.181 for the TIR ORF pair—than the TBE transposase paralogs (0.287 average), indicating that these genes have either duplicated more recently than the TBEs (since ciliate mitochondrial genes evolve more rapidly than their nuclear genes) and/or that substitution has been suppressed by concerted evolution. Assuming no translocations and that the strength of concerted evolution either declines with
distance from the telomeres or remains approximately the same throughout the genome, the lower overall substitution rates in the 600 bp duplicated ORF pair, relative to the TIR ORF pair, suggest that these duplications arose both internally to, and after, the TIR ORF pair.

The pattern of sequence conservation in the mitochondrial terminal regions suggests both that purifying selection acting upon a duplicated ORF permitted the detection of duplications and that selective constraints have been lost in many of the surrounding ORFs leading to pseudogene formation. Two lines of evidence suggest pseudogenization: 1) the intervening regions between ORFs are longer in the terminal unknown ORF regions (~179 bp mean; ~178 bp standard deviation [SD]; excluding zero length regions) than the central region (~53 bp mean; 63 bp SD); 2) inter-ORF regions constitute ~19.5% of the terminal unknown ORF regions, whereas these spacers constitute ~5% of the *Oxytricha* mitochondrial genome excluding the terminal unknown ORF regions, a figure close to that of typical tightly packed ciliate mitochondrial genomes, such as those of *Tetrahymena* and *Euplotes* (Burger et al. 2000; de Graaf et al. 2009); 3) ORFs in the unknown ORF regions are shorter (419 bp mean; 310 bp SD) compared with those from the central region (501 bp mean; 412 bp SD).

In the 14.5-kb subterminal region, we also noticed an overabundance of tryptophan UGG (15) versus UGA (43) anticodons, in comparison to 20 UGG versus 229 UGA anticodons in conserved or ciliate-specific ORFs in the central region (the ~5-kb subtelomeric region has 0 UGG and 17 UGA tryptophan anticodons). This deviation from the standard tryptophan codon usage in the larger subterminal region could be an indication either of a relatively recent incorporation of foreign genetic material derived from the mO plasmid and/or relaxed constraint associated with possible pseudogene formation.

**Discussion**

The *Oxytricha* mitochondrial genome, at ~70 kb, is the largest ciliate mitochondrial genome sequenced to date. It is approximately 22–30 kb larger than the other completely sequenced ciliate mitochondrial genomes; mitochondrial genomes of the distantly related oligohymenophorans *P. tetraurelia* and *T. pyriformis* are ~40 kb (Pritchard et al. 1990).
2 rRNAs in because they contain a relatively large number of mitochon-
ddensities and are considered to be “large” (Gray et al. 2004) of the plasmid might have only mildly deleterious effects.

element in this region of the genome. Such an integration
tions or possibly a higher probability of incorporation of the
either an association between this plasmid and the duplica-
drial plasmid that contains a region that matches segmental
 duplications on the primary mitochondrial genome indicates
d of the ciliate mitochondrial proteins; supplementary
sequence data from the broader diversity of ciliates.
It was previously shown that the oligohymenophoran
cox1, cox2, and cob genes are extremely divergent relative
to other eukaryotic mitochondrial genes, and this is partly
responsible for the difficulty in classifying a large number
of ciliate mitochondrial ORFs (Burger et al. 2000). Extreme
divergence appears to be a general property of ciliate mito-
chondrial protein-coding genes, even for the highly con-
served iron-sulfur proteins nad7 and nad10 (the least
divergent of the ciliate mitochondrial proteins; supplementary
fig. 2, Supplementary Material online). There appears to
be no functional association with such extreme divergence
because genes with unrelated functions (ribosomal, electron
transport, and protein transport/maturation genes) all exhibit
extreme divergences. Though ciliate mitochondrial rRNA
genese do not appear to be evolving at exceptional rates (Gray
and Spencer 1996) in relation to other eukaryotic mitochon-
drial rRNAs, their distances and divergence rates may be
underestimated, due to saturation (supplementary fig. 3,
Supplementary Material online). Therefore, gene substitution
rates appear to be generally elevated in ciliate mitochondria,
irrespective of whether the gene encodes a protein or RNA
product. A neutral evolutionary process due to low fidelity
replication or error-prone repair would be consistent with the
elevated ciliate mitochondrial substitution rates.

There is also evidence to suggest that ciliate nuclear genes
have elevated substitution rates relative to that of other eu-
karyotes, with ciliates such as Oxytricha and Euplotes—that
possess a highly fragmented macronuclear genome structure
—evolving the most rapidly (Zufall et al. 2006). Unlike the
case of mitochondrially encoded genes, we have not ob-
served evidence of extreme divergences in any of the
nuclear-encoded, putatively mitochondrially targeted, genes
that we examined in this study (RF1, mtDNA polymerase,
mtRNA polymerase [Moriyama et al. 2008], nad8, nad11).
Furthermore, since different DNA polymerase complexes
are responsible for nuclear versus mitochondrial replication,
we do not expect a correlation between elevated mitochondrial
substitution rates and nuclear substitution rates.

Based on comparisons of the oligohymenophorean and
spirotrich mitochondrial genomes, we propose that their
common ancestor possessed: 1) a linear mitochondrial ge-
nome; 2) a replication origin within- or in close proximity to
an AT-rich region of low complexity; and 3) TIRs capped by
telomeric repeats.

In ciliate mitochondrial genomes, both the putative rep-
ication origin (Arneg et al. 1974; Goddard and Cummings
1975, 1977; Pritchard and Cummings 1981) and primary
region of transcription initiation appear to lie in close prox-
imity to, or coincide with, a low-complexity/repeat region.
TATA-like elements in multiple Paramecium species have
been proposed as a motif for transcription recognition
(Pritchard et al. 1983) but this now seems unlikely given that
a different T-odd phage-like eukaryotic mitochondrial RNA
polymerase is most likely the primary mitochondrial RNA
polymerase, and such phage RNA polymerases are TATA in-
de pendent (Cermakian et al. 1996; Shutt and Gray 2006). In
Tetrahymena species, a highly conserved, GC box-like re-
gion in the central region of divergent transcription has been
proposed as a motif that may be responsible for initiating
transcription and possibly also DNA replication (Moradian
et al. 2007). Experimental evidence is necessary to pinpoint
the precise location of transcription initiation in these
genomes.

Both mitochondrial TIRs and telomeric sequences, such as
those of Tetrahymena and Oxytricha, were proposed to be
of foreign origin (Nosek and Tomáška 2003). Nosek and
Tomáška (2003) also proposed that linear mitochondrial ge-
nomes owe their linearity to mobile elements, which would
provide both the need and means to replicate linear ge-
nomes by providing DNA sequences/structures and a poly-
merase necessary for replicating linear DNA. The Oxytricha
linear mitochondrial plasmid appears to lack the TIRs char-
acteristic of most known linear plasmids (Meinhardt et al.
1990; Handa 2008). Instead, it had a 5′ end with complex
repeats and a 3′ end with the same telomeric repeats as the
primary mitochondrial genome. The latter feature demon-
strates that it is possible to transfer telomeric sequence re-
peats between mitochondrial genomes and linear plasmids.
One possible scenario for such a transfer is that the original Oxytricha linear plasmid may have possessed a terminal inverted structure, which was lost during mitochondrial genome integration, followed by capture of a telomere-bearing end from the primary genome. Alternatively, the plasmid may have possessed a similar structure to its current form, with a telomeric repeat sequence that was transferred to the Oxytricha mitochondrial genome during an integration event. We propose that, as for horizontal gene transfer, the phagotrophic lifestyles of ciliates may predispose them to periodic mitochondrial invasions by mobile elements bearing error-prone DNA polymerases, such as the Oxytricha mO plasmid. These foreign polymerases may in turn interfere with or partially substitute for the primary, higher fidelity mitochondrial DNA polymerase, contributing to the extreme evolutionary divergences observed in ciliate mitochondria.

Supplementary Material
Supplementary table 1 and figures, 1–3 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

Acknowledgments
We would like to thank the Hans Lipps laboratory (Universität Witten/Herdecke), in particular Franziska Jönsson, for providing us with DNA from Stylonychia and Jingmei Wang for general laboratory assistance. This research was supported by the National Institutes of Health grant GM59708 to L.F.L.

Literature Cited
Barth D, Berendonk TU. 2011. The mitochondrial genome sequence of the ciliate Paramecium caudatum reveals a shift in nucleotide composition and codon usage within the genus Paramecium. BMC Genomics 12:272.
Iwamoto M, et al. 1998. A ribosomal protein gene cluster is encoded in
Moradian MM, et al. 2007. Complete mitochondrial genome sequence
Meinhardt F, et al. 1990. Linear plasmids among eukaryotes: funda-
Massey SE, Garey JR. 2007. A comparative genomics analysis of codon
Marriott AC, Archer SA, Buck KW. 1984. Mitochondrial DNA in
Lowe T, Eddy S. 1997. tRNAscan-SE: a program for improved detection
Lecrenier N, Foury F. 2000. New features of mitochondrial DNA


**Associate editor:** Shu-Miaw Chaw