TITLE: Population genetics of *Paramecium* mitochondrial genomes: recombination, mutational spectrum, and efficacy of selection.

AUTHORS: Parul Johri[*,**,1], Georgi K. Marinov[*,‡,**], Thomas G. Doak[*,†], and Michael Lynch[*,§]

[*]Department of Biology, Indiana University, Bloomington, IN 47405, United States

[†]National Center for Genome Analysis Support, Indiana University, Bloomington, IN 47405, United States

[‡]Current address: Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, United States

[§]The Biodesign Center for Mechanisms of Evolution, Arizona State University, Tempe, AZ 85287, United States

[**]These authors contributed equally to this work

[1]Corresponding author:

Parul Johri
Center for Evolution and Medicine
Life Sciences C, Room 360
427 East Tyler Mall
Arizona State University, Tempe, AZ 85287
(812) 320-0687
pjohri@indiana.edu

RUNNING TITLE: Evolution of *Paramecium* mitochondrial genomes.

KEYWORDS: *Paramecium*, mitochondria, telomeres, recombination, mutation spectrum, efficacy of purifying selection.

ABSTRACT

The evolution of mitochondrial genomes and their population-genetic environment among unicellular eukaryotes are understudied. Ciliate mitochondrial genomes exhibit a unique combination of characteristics, including a linear organization and the presence of multiple genes with no known function or detectable homologs in other eukaryotes. Here we study the variation of ciliate mitochondrial genomes both within and across thirteen highly diverged *Paramecium* species, including multiple species from the *P. aurelia* species complex, with four outgroup species: *P. caudatum*, *P. multimicronucleatum,* and two strains that may represent novel related species. We observe extraordinary conservation of gene order and protein-coding content in *Paramecium* mitochondria across species. In contrast, significant differences are observed in tRNA content and copy number, which is highly conserved in species belonging to the *P. aurelia* complex but variable among and even within the other *Paramecium* species. There is an increase in GC content from ~20% to ~40% on the branch leading to the *P. aurelia* complex. Patterns of polymorphism in population-genomic data and mutation-accumulation experiments suggest that the increase in GC content is primarily due to changes in the mutation spectra in the *P. aurelia* species. Finally, we find no evidence of recombination in *Paramecium* mitochondria and find that the mitochondrial genome appears to experience either similar or stronger efficacy of purifying selection than the nucleus.

INTRODUCTION

Mitochondrial genomes have played integral roles in furthering our understanding of relationships among species as well as revealing population structure and demographic history. As a consequence, we have obtained insights into the unique population-genetic properties of mitochondrial genomes. In most species, mitochondria are inherited uniparentally (but see Barr, et al. 2005). Although mitochondrial genomes are known to frequently undergo recombination in plants (Stadler and Delph 2002; Mackenzie 2007) as well as fungi (Fritsch, et al. 2014), no recombination has been detected in animals (Ballard and Whitlock 2004; Piganeau and Eyre-Walker 2004). Because of the unique mode of transmission and lack of recombination in some species, mitochondria have been suggested to have lower effective population sizes than their nuclear counterparts and therefore to accumulate more deleterious mutations (Lynch and Blanchard 1998; Neiman and Taylor 2009). In addition, mitochondrial genomes experience much higher spontaneous rates of mutation than their corresponding nuclear genomes in animals, but exhibit the opposite trend in plants (Lynch, et al. 2006).

Unlike the relatively uniform and conserved properties of metazoan mitochondrial genomes, mitochondrial genomes in unicellular eukaryotes exhibit remarkable variation in genome structure and GC content. Mitochondrial genome structures range from hundreds of short linear segments (0.3-8.3 kb) in the ichtysporean *Amoebidium parasiticum* (Burger, et al. 2003), an opisthokont, to many small (< 10 kb) circular genomes in the diplonemid *Diplonema papillatum* (Vlcek, et al. 2011), an excavate, to a single larger linear or circular chromosome, and a variety of other states (reviewed in Smith and Keeling 2015). In addition to variation in organization,

90 mitochondria from unicellular lineages display widely diverse GC compositions, ranging from as
91 low as 10% in some yeast (Smith 2012) to as high as 60% in *Lobochlamys culleus* (Borza, et al.
92 2009), although most species are AT rich, with an average GC content of 35% (Smith 2012).
93
94 Much has been learned about the structure, evolution and population-genetic environment of
95 mitochondria in the main model systems, especially in plants and metazoans (Lynch 2007; Smith
96 2016). However, we lack such understanding of mitochondria of the majority of unicellular
97 eukaryotes, where the bulk of eukaryotic phylogenetic diversity lies. We address this gap by
98 surveying both within and between-species variation in mitochondrial genomes among multiple
99 ciliate species belonging to the genus *Paramecium*.
100
101 In the large and morphologically and ecologically diverse ciliate lineage, mitochondrial genomes
102 sampled so far are organized into large linear chromosomes, several tens of kilobases in length,
103 with telomeres at the ends (Goddard and Cummings 1975; Morin and Cech 1988; Swart, et al.
104 2012). However, few mitochondrial genomes have until now been fully sequenced among the
105 ciliates, with two in the *Paramecium* genus (*P. tetraurelia* and *P. caudatum*; (Barth and
106 Berendonk 2011)), and only a few others: *Tetrahymena pyriformis* (Burger, et al. 2000),
107 *Euplotes minuta* and *Euplotes crassus* (de Graaf, et al. 2009), *Oxytricha trifallax* (Swart, et al.
108 2012), *Stentor coeruleus* (Slabodnick, et al. 2017), *Ichthyophthirius multifiliis* (Coyne, et al.
109 2011), and the anaerobic ciliate *Nyctotherus ovalis* (de Graaf, et al. 2011).
110
111 While the *Paramecium* genus contains a number of distinct morphospecies, it is especially
112 known for including a species complex consisting of multiple morphologically identical but

113 sexually isolated species – the *P. aurelia* complex (Sonneborn 1975). Species in the *P. aurelia*
114 complex are ancient, with the estimated time of divergence for the complex as a whole being on
115 the order of 300 million years (McGrath, Gout, Johri, et al. 2014), implying that the genus
116 *Paramecium* itself is even more ancient. Interestingly, among *Paramecium* species, there is an
117 increase of GC content in mtDNA in the branch leading to the *P. aurelia* complex (Burger, et al.
118 2000; Barth and Berendonk 2011), allowing us to study the evolution of nucleotide composition
119 across mitochondrial genomes that are structurally very similar.
120

121 The *Paramecium* species offer a particularly interesting system in which to study the
122 evolution of mitochondrial genomes because of the unique population-genetic
123 environment experienced by their cellular organelles. *Paramecium* cells are
124 mitochondria-rich: each individual cell in *P. aurelia* species is estimated to contain about 5000
125 mitochondria, with about 8-10 genomes per mitochondrion (Beale and Tait 1981), which is much
126 larger than in mammalian cells with 1000-2000 mitochondria (Kukat, et al. 2011) and yeast cells
127 with 20-30 mitochondria per cell (Visser, et al. 1995). In addition, *Paramecium* lineages, like
128 other ciliate species, possess two nuclei: the polyploid somatic nucleus (called the
129 macronucleus), which divides amitotically where the bulk of transcriptional activity
130 occurs, and the diploid germline nucleus (known as the micronucleus), which is
131 transcriptionally silent, and which undergoes sexual reproduction. All *Paramecium*
132 species can proliferate asexually for a limited number of generations, after which they
133 senesce unless they undergo sexual reproduction or conjugation. During asexual
134 proliferation, *Paramecium* undergoes binary fission during which mitochondria appear to double
135 in length, replicate their genomes (Perasso and Beisson 1978), and are presumed to be randomly

136 distributed between the two daughter cells. It therefore appears that mitochondria do not
137 experience any bottleneck during mitotic division.
138
139 During conjugation, *Paramecium* cells exhibit cytoplasmic inheritance (Koizumi and
140 Kobayashi 1989), i.e., despite the exchange of micronuclei between the two conjugants
141 there is almost no exchange of cytoplasm and other organelles (reviewed in Meyer and
142 Garnier 2002). Thus, mitochondria are uniparentally inherited. A distinct aspect of
143 *Paramecium* mitochondrial biology is that *Paramecium* mitochondria appear to exist as
144 independent structural units and do not undergo fusion, unlike the constant flux of organelle
145 fusion and fission in other metazoan mitochondrial populations (Kiefel, et al. 2006). Both
146 uniparental inheritance of mitochondria and the absence of fusion in the cytoplasm suggest a lack
147 of recombination among mitochondria genomes.
148
149 In this study, we further the understanding of the biology and the population-genetic
150 environment of ciliate mitochondria by presenting the complete mitochondrial genomes of
151 nine species belonging to the *P. aurelia* complex, four outgroup (relative to *P. aurelia*)
152 species, and additional 5-10 isolates for each of four *Paramecium* species. Using
153 phylogenetic and population-genetic analyses, we investigate variation in protein-coding
154 genes, tRNA content, and the forces governing the evolution of nucleotide composition of
155 mitochondrial genomes across the phylogeny. Finally, we estimate the recombination rate
156 across the genome and address the controversy of whether mitochondrial genomes
157 experience reduced purifying selection in comparison to their nuclear counterparts.
158

159 **RESULTS**

160

161 **Mitochondrial genome sequencing and assembly**

162

163 We assembled complete mitochondrial genomes of seven species belonging to the *P. aurelia*
164 complex: *P. biaurelia*, *P. tetraurelia*, *P. sexaurelia*, *P. octaurelia*, *P. novaurelia*, *P.*
165 *decaurelia*, *P. dodecaurelia*, *P. quadecaurelia*, and *P. jenningsi.* In addition, we analyzed
166 previously reported complete mitochondrial genomes of *P. tetraurelia and P. sexaurelia*
167 belonging to the *P. aurelia complex,* as well as four outgroup species: *P. caudatum*, *P.*
168 *caudatum-C026*, *P. multimicronucleatum*, and *P. multimicronucleatum-Peniche3I* (Figure 1;
169 isolates sequenced in Johri et al., 2017). We note that *P. caudatum-C026* and *P.*
170 *multimicronucleatum-Peniche3I* were initially sampled as individual isolates belonging to
171 the *P. caudatum* and *P. multimicronucleatum* species (based on morphological criteria),
172 respectively. However, the analysis of the mitochondrial sequences revealed that they are
173 highly diverged from the reference strains (see below) and are therefore almost certainly
174 separate species and were treated as such in subsequent analyses. For all 7 new
175 mitochondrial genomes sequenced and assembled in this study, Illumina reads were
176 assembled using SPAdes (Bankevich, et al. 2012), and mitochondrial contigs were identified
177 by BLAST searches against the publicly available *P. caudatum* and *P. tetraurelia*
178 mitochondrial sequences (see the Materials and Methods section for more details).

179

180 In addition, we examined sequenced mitochondrial genomes of 10 isolates of both *P.*
181 *tetraurelia* and *P. sexaurelia*, and 5 isolates of each *P. caudatum* and *P.*

182 *multimicronucleatum* (Supplementary Figure 1) sampled worldwide (Johri, et al. 2017).
183 Illumina paired-end reads from these isolates were mapped to the assembled reference
184 genomes and SNPs were called as in Johri, et al. (2017). In this study, all mitochondrial
185 genomes of individual isolates were also assembled *de novo* (Supplementary Figures 2, 3
186 and 4) in order to examine large-scale genome organization mapping.
187
188 **Genome structure, organization and telomeric repeats**
189
190 All *Paramecium* mitochondrial genomes are linear (Goddard and Cummings 1975; Morin and
191 Cech 1988; Swart, et al. 2012). Our assemblies show that they are ~40 kb in the *P. aurelia*
192 species and ~44 kb in *P. caudatum* and *P. caudatum-C026*. Telomeric repeats and gene
193 content observed at ends of the assembled genomes imply that the full lengths of the linear
194 contigs have been assembled in most species with the exception of the *P.*
195 *multimicronucleatum* and *P. multimicronucleatum* Peniche3I mitocontigs, which appear to
196 be missing small portions of the 3′ end of the chromosome (Figure 2). The length of the
197 assembled contigs in *P. multimicronucleatum* and *P. multimicronucleatum-Peniche3I*
198 suggests an overall size closer to that observed in *P. aurelia* than to the larger mitochondrial
199 genomes in the *P. caudatum* lineage. We also note that the raw assemblies for two of the *P.*
200 *aurelia* species contain extensions (Supplementary Figure 3, 5, and 6). In *P. novaurelia*, an
201 additional ~18 kb is present at the 5′ end of the mitocontig, while in *P. quadecaurelia,* a
202 small, ~1 kb extension is seen at the 3′ end. However, the read coverage over these regions
203 is very different from the rest of the mitocontigs, suggesting either misassembly or
204 heterogeneity within cell populations. We therefore ignored these extensions in subsequent

205 analysis. The organization of the genomes is very similar, with gene order preserved almost
206 perfectly between all species (Figure 2).
207
208 We also identified telomeric repeats based on the sequences at the end of assembled
209 mitocontigs (see the Methods section for more details). In most *P. aurelia* species, we
210 identify almost identical repeats, with a 23-bp consensus sequence
211 GCCCTGGTGGCCCTAGAAGCTCC (Figure 3). However, in *P. jenningsi* and *P.*
212 *sexaurelia* the telomere repeat motif is the same length, but has 1 nucleotide difference
213 from the consensus sequence. Of note, these two species are the earliest diverging ones
214 within the *P. aurelia* complex species included in our analysis. We observed even more
215 divergent telomeric sequences in *P. caudatum*
216 (GCCCTGGTAACGCTGGTCGCCCTTTTAAAATA) and *P. multimicronucleatum*
217 (GCCCTTGTTACACTTGGTGGCTCTTTAAGCTCT). In these species, the core
218 telomeric repeat sequence has been expanded by an additional 10-bp of sequence not
219 present in *P. aurelia*. The *Paramecium* core telomeric repeat is broadly similar to that of
220 *Tetrahymena* (ACCCTCGTGTCTCTTTA; Figure 3), the other Oligohymenophorean
221 genus for which mitochondrial genomes are available, but distinct from what is
222 observed in distantly branching ciliates such as the spirotrichean *Oxytricha trifallax*
223 (Figure 3).
224
225 Examination of existing RNA-seq datasets for several of the species revealed no
226 evidence for RNA editing in *Paramecium* mitochondria (data not shown), in
227 concordance with previous reports (Orr, et al. 1997).

228

229 **Variation in protein coding gene content between and within species**

230 All *Paramecium* mitochondrial genomes contain a core set of 15 genes (Figure 2) involved
231 in electron transport and ATP synthesis (*atp9*, *cob*, *cox1*, *cox2*, *nadh1*, *nadh2*, *nadh3*,
232 *nadh4*, *nadh4L*, *nadh5*, *nadh6*, *nadh7*, nadh9 and *nadh10)*, the heme maturase *YejR*, and 8
233 ribosomal protein genes (*rpL6*, *rpL12*, *rpL14*, *rpL16*, *rpS3*, *rpS12*, *rpS13*, and *rpS14*). One
234 ribosomal protein subunit *rpS19* was found to be absent from *P. tetraurelia*, *P. octaurelia*,
235 *P. novaurelia*, *P. dodecaurelia*, and *P. quadecaurelia* and therefore appears to have been
236 independently lost from mitochondrial genomes at least 3 times along the *P. aurelia*
237 phylogeny (Figure 1). Remarkably, we also find a presence-absence polymorphism of
238 *rpS19* within *P. sexaurelia* isolates, with the ORF not being identifiable in 2 out of 11
239 isolates studied (Supplementary Figure 7). Although this might suggest a mitochondrial
240 gene in the process of transfer to the nucleus, the closest hit in the nuclear genome has an E
241 value of 0.27, implying either a complete loss or perhaps transfer to a mitochondrial
242 plasmid, which has not been captured in existing assemblies.

243

244 *Paramecium* genomes also contain 16 ciliate-specific genes of unknown function, called
245 *Ymf* genes (Burger, et al. 2000), named according to the Commission on Plant Gene
246 Nomenclature (Price and Reardon 2001). Each sequenced strain contains all of these genes,
247 with the exception of *P. multimicronucleatum* strain M13 in which *Ymf80* and *Ymf83*
248 appear to have fused with *nadh4* and *Ymf64*, respectively. In addition, several ORFs
249 originally identified in *P. tetraurelia*, are observed in only one or a few additional *P.*
250 *aurelia* mitochondrial genomes, typically overlapping longer ORFs: *orf189*, *orf285*, and

251 *orf314* restricted to the three *P. tetraurelia* isolates; *orf221* present in *P. tetraurelia*, *P.*
252 *octaurelia* and *P. biaurelia*; *orf265* in *P. biaurelia*; and *orf78* found only in *P. tetraurelia*
253 isolate A.
254
255 In addition, we found a previously unidentified ORF larger than 100 amino acids, which
256 appears to be present in all *Paramecium* mitochondrial genomes and is located at the very 5′
257 end of the chromosome, immediately before the large ribosomal RNA. This ORF has
258 homology to the CRISPR-associated endoribonuclease Cas6; its functional significance is
259 unclear at present. It could potentially represent a case of horizontal gene transfer from
260 endosymbiotic bacteria (Preer 1969; Fokin and Gortz 2009) often associated with
261 *Paramecium*. Sequencing of more ciliate mitochondrial genomes could provide insight into
262 its origin.
263
264 ***Paramecium* species are highly diverged and contain cryptic species complexes**
265 We found most species to be fairly evolutionarily distant from each other, as measured
266 by the average number of nonsynonymous (*dN*) and synonymous (*dS*) site substitutions
267 per site (Supplementary Figure 8) between all pairs of species, using yn00, PAML
268 (Yang 2007; version 4.9a). *P. decaurelia* and *P. dodecaurelia* (average *dS*: 0.85;
269 average *dN*: 0.05, average total divergence: 0.23) are the closest species pair, followed
270 by *P. tetraurelia* and *P. octaurelia* (average *dS*: 0.93; average *dN*: 0.09; average total
271 divergence: 0.27). All other species pairs exhibit on average $dS > 1.0$, i.e., synonymous
272 sites in protein coding genes have on average undergone more than 1 substitution each
273 since the time of divergence. We found that mean *dS* between *P. caudatum* and *P.*

274 *caudatum-C026* is ~2.0 (average *dN*: 0.07; average total divergence: 0.28), implying
275 that these two isolates possibly represent separate species. A similar observation
276 confirms that *P. multimicronucleatum-Peniche3I* and *P. multimicronucleatum* are
277 probably separate species (average *dS* = 1.97; average *dN*: 0.10; average total
278 divergence: 0.26). Our observations suggest the possibility of cryptic species
279 complexes in both *P. caudatum* and *P. multimicronucleatum*, as previously suspected
280 (Hori, et al. 2006; Tarcz, et al. 2012). We also note that several additional *P.*
281 *multimicronucleatum* isolates most closely related to *P. multimicronucleatum-*
282 *Peniche3I* might also be separate species as they are fairly divergent from the reference
283 strain as well as *P. multimicronucleatum-Peniche3I* (Supplementary Figure 1). The
284 phenomenon of species complexes containing numerous morphologically identical
285 species, already known for *P. aurelia* (Sonneborn 1937) and *T. pyriformis* (Gruchy
286 1955) is therefore probably much more widespread among ciliates than previously
287 appreciated.

288

289 **Selective pressures on ciliate-specific mitochondrial genes**

290 The 16 ciliate-specific mitochondrial genes found in all *Paramecium* species have orthologs
291 in *Tetrahymena* species as well as in *Oxytricha*. Thus, these genes have been preserved for a
292 long evolutionary time, and yet have diverged sufficiently that no known homologs exist in
293 well-studied species in other eukaryotic kingdoms. Across the *Paramecium* species, they
294 appear to be on average faster evolving (Supplementary Figure 9, Supplementary Table 1)
295 with average $dN/dS \cong 0.11$, in comparison to genes that encode enzyme components of the
296 respiratory chains (average $dN/dS \cong 0.04$) and ribosomal proteins (average $dN/dS \cong 0.06$).

297 The *dN*/*dS* values here were obtained for each gene under the assumption that *dN*/*dS*
298 remains constant across the phylogeny using CODEML, PAML (Yang 2007; version
299 4.9a). Ciliate-specific *Ymf* genes therefore exhibit a relatively higher rate of evolution than
300 other genes, with as little as 30% sequence identity between *P. aurelia*, *P. caudatum,* and *P.*
301 *multimicronucleatum*, similar to observations in *Tetrahymena* (Moradian, et al. 2007), as
302 well as with higher values of π_n/π_s in *P. tetraurelia*, *P. sexaurelia*, *P. caudatum*, and *P.*
303 *multimicronucleatum* (Supplementary Figure 10).
304
305 The higher rate of evolution of *Ymf* genes can either be explained by relaxed purifying
306 selection at the sequence level, or recurrent positive selection over long periods of time,
307 or a combination of both. To distinguish between these possibilities, we performed a
308 McDonald-Kreitman test (MK test) in the species *P. tetraurelia*, *P. sexaurelia*, *P.*
309 *caudatum*, and *P. multimicronucleatum*. As most of these species are highly diverged
310 from each other, we used ancestral reconstruction over the set of all 13 taxa to first infer
311 ancestral nucleotides for each internal node. The numbers of synonymous (D_S) and
312 nonsynonymous (D_N) changes were then inferred along each terminal branch leading to
313 each of the 4 species mentioned above. We tested for positive selection by determining
314 whether the proportion of nonsynonymous fixed differences relative to synonymous
315 ones was significantly larger than the proportion of nonsynonymous relative to
316 synonymous polymorphic differences. None of the genes in *P. tetraurelia* were found to
317 have undergone positive selection by this criterion (Holm–Bonferroni corrected $p \geq 0.05$
318 in each case). In *P. sexaurelia*, *cox1*, *Ymf76*, and *Ymf63* are the only genes that showed
319 significant positive selection, whereas in *P. caudatum*, only *Ymf64* exhibits positive

320 selection (Supplementary Table 2). With very few SNPs available in *P.*
321 *multimicronucleatum*, the MK test was not significant for any gene. Nevertheless, we
322 find no statistical enrichment for number of genes evolving under positive selection
323 among the *Ymf* relative to other genes in *P. sexaurelia* ($p = 0.56$; Fisher's exact test) and
324 *P. caudatum* ($p = 0.40$; Fisher's exact test), suggesting that the majority of them are
325 likely evolving fast due to relaxed sequence constraints.
326
327 It is possible that most of the *Ymf* genes are simply highly diverged genes that perform the
328 same functions as genes found in other mitochondrial genomes, but are not identifiable by
329 sequence. In order to gain some insight into their possible functions, we used HHPred
330 (Soding, et al. 2005) to predict their homologs and found that parts of *Ymf* genes display
331 homology to standard mitochondrial genes, encoding ribosomal subunits and proteins
332 involved in electron transport. About 37% of sequence from *Ymf59* was found to be
333 homologous to *rpS10*, ~22% sequence of *Ymf61* to a 30S ribosomal protein S24E, ~28% of
334 *Ymf64* sequence to a 30 S *rpS3*, ~79% of *Ymf65* sequence to the NADH subunit and ~11%
335 of *Ymf67* to cytochrome c oxidase, subunit P. *Ymf59* and *Ymf64* in *Oxytricha* were also
336 previously found to be homologous to *rpsS10* and *rpS3* respectively (Swart, et al. 2012).
337
338 **tRNA content variation across and within species**
339 While the number and identity of protein-coding genes is generally the same across the
340 genus, tRNAs display significant variation. All species in the *P. aurelia* complex have only
341 three tRNA loci, at highly conserved locations (Figure 2), most of the time consisting of a
342 Y, F, and W tRNAs (the latter recognizing the UGA stop codon that has been reassigned to

343 tryptophan in ciliate mitochondria). The one exception is *P. octaurelia*, in which a T
344 tRNA has been substituted for the Y tRNA that other species possess. In stark contrast,
345 large and highly variable sets of tRNA are predicted in mitochondrial genomes in the *P.*
346 *caudatum* and *P. multimicronucleatum* lineages (Figure 3), with as many as 38 tRNA
347 predictions in *P. multimicronucleatum* M05, and with individual *P. caudatum* and *P.*
348 *multimicronucleatum* isolates having different tRNAs at multiple loci in their
349 mitochondrial genomes. These tRNAs often overlap protein-coding genes (in contrast to
350 what is observed in *P. aurelia*), in particular *Ymf* genes. Of note, the three tRNAs found in
351 the *P. aurelia* species are also present in all outgroup species. The absolute numbers of
352 the tRNAs found should be taken with caution, as some of the predicted tRNAs
353 overlap each other as well as *Ymf64* in the *P. multimicronucleatum* strains. However,
354 even if these are excluded, the number of tRNAs in both *P. caudatum* and *P.*
355 *multimicronucleatum* strains is still large and variable and consist of tRNAs not found
356 in the *P. aurelia* species. This is a unique finding among mitochondrial genomes and
357 should be investigated more thoroughly in future studies to better understand the origin
358 of the tRNAs.
359

360 **Mutation spectrum and GC-content variation within the *Paramecium* genus**

361

362 Perhaps the most remarkable discontinuity in *Paramecium* mitochondrial genomes is the
363 difference between the high GC content observed in species of the *P. aurelia* complex (~40%)
364 and the low GC content in the outgroup species (~20%). Very low GC content is also observed
365 in mitochondrial genomes of other ciliates, for instance in the five *Tetrahymena* species

366 (~18-21%), in *Icthyophthirius multifiliis* (~16%) (Coyne, et al. 2011), and in *Oxytricha*
367 *trifallax* (~24%) suggesting a lineage-specific increase in GC content along the branch
368 leading to *P. aurelia* species. In contrast, nuclear genomes have relatively similar levels of GC
369 content across all *Paramecium* species (McGrath, Gout, Doak, et al. 2014). The low GC content
370 of *P. caudatum* and *P. multimicronucleatum* mtDNA is marked by a highly biased codon usage
371 (Barth and Berendonk 2011), with most synonymous positions exhibiting a strong bias for A or
372 T nucleotides (Supplementary Figure 11). Indeed, the difference in GC content is most
373 pronounced at 4-fold degenerate sites, with the GC content at such sites being as low as
374 3.5% in *P. multimicronucleatum* and as high as 54.5% in *P. decaurelia* (Figure 1). The 0-
375 fold redundant sites in the outgroup species have much higher GC content (~26-37%) and
376 the GC content (31-36%) of the rRNAs is similar across the entire genus suggesting strong
377 selection on rRNA nucleotide content.
378
379 At least three forces may be responsible for determining the GC content at 4-fold degenerate
380 sites: (1) mutational processes, (2) codon usage bias due to selection, and/or (3) genome-
381 wide selection for higher/lower GC content. We therefore examined GC composition in other
382 regions of the genomes, such as tRNAs and intergenic regions, which do not experience selection
383 for codon usage. Both tRNAs and intergenic regions have elevated GC content in the *P. aurelia*
384 species (~40-45% and ~34-37%) relative to *P. caudatum* and *P. multimicronucleatum* (~14-22%
385 and ~16-20% respectively; Figure 1), suggesting that either differences in mutational bias or
386 selection for genome-wide GC content is primarily responsible for changes in nucleotide
387 composition across the genome.
388

389 To distinguish between mutation and selection as the possible explanations for these
390 observations, we evaluated the mutational pressures acting on *Paramecium* species, using two
391 approaches. First, we used the mitochondrial mutation spectrum obtained from mutation-
392 accumulation (MA) studies of *P. tetraurelia* (Sung et al, 2012) to calculate the AT mutation bias,
393 $m = v/u$, where v is the mutation rate from G/C to A/T and u is the rate of mutation from A/T to
394 G/C (Table 1). We found m ≅ 1.0, showing little mutation bias and corresponding to expected
395 equilibrium GC content of 51.3% (Table 1). This is very close to the GC content of 4-fold
396 redundant sites (54%) in *P. tetraurelia*. We also analyzed existing sequencing data from
397 MA experiments in *P. biaurelia* and *P. sexaurelia* (Long, Doak, et al. 2018;
398 Supplementary Table 3). These data imply slight mutation bias towards GC in *P. biaurelia*
399 and *P. sexaurelia*, corresponding to expected equilibrium GC percentages (Sueoka 1993) of
400 55.8% and 67.0% respectively. Given the relatively small number of mutations (a few tens)
401 identified in each set of MA experiments, there is considerable uncertainty in these
402 estimates. Nonetheless, these data do not present evidence that 4-fold degenerate sites in *P.*
403 *aurelia* mitochondria are evolving away from mutation equilibrium via selection.
404
405 Next, derived singleton alleles at 4-fold degenerate sites were used to quantify the number of
406 G/C to A/T mutations relative to A/T to G/C mutations and thus infer an estimate of AT
407 mutation bias from population data. Again, we find that m is close to 1.0 (i.e., no mutation bias
408 towards A/T) in both *P. tetraurelia* and *P. sexaurelia*, with the predicted GC content under
409 mutation equilibrium being remarkably close to that of their 4-fold degenerate sites (Table 1).
410 These results suggest that the composition of 4-fold degenerate sites is mostly determined by
411 mutation. The same can be seen by calculating the strength of selection ($S = 4N_e s$) favoring

412 A/T at 4-fold degenerate sites. As *S* usually takes values between 0.1 and 4.0 (Lynch 2007;
413 Long, Sung, et al. 2018) and we obtain much lower magnitudes (0.0-0.5) in the *P. aurelia*
414 species (Table 1), we infer negligible or weak selection at 4-fold degenerate sites, consistent
415 with absence of codon bias at the third position.
416
417 In contrast to the *P. aurelia* species, we infer a significant AT mutation bias in *P. caudatum*
418 (Table 1), yielding a predicted GC content at equilibrium of ~17%, with the observed GC
419 content in intergenic and tRNA regions (16% and 20%) being remarkably close to this value.
420 These calculations suggest that genome-wide GC content in *P. caudatum* mitochondria is mostly
421 governed by mutation bias, and that there has been a major shift in mutation bias along the
422 branch leading to the *P. aurelia* species.
423
424 **Recombination in *Paramecium* mitochondrial genomes**
425
426 Next, we sought to evaluate the evidence for the occurrence of recombination in *Paramecium*
427 mitogenomes by performed multiple tests for its presence. First, we evaluated the relationship
428 between linkage disequilibrium (LD), calculated by r^2 (Hill and Robertson 1968) and distance
429 between sites; r^2 does not decrease with distance in *P. tetraurelia*, *P. sexaurelia*, and *P.*
430 *caudatum* (*P. multimicronucleatum* isolates lacked sufficient number of polymorphic sites;
431 Supplementary Figure 12), consistent with the expectation under no recombination. Next, we
432 conducted the four-gamete test (FGT) (Hudson and Kaplan 1985), which detects recombination
433 by searching for pairs of polymorphic sites with all four segregating haplotypes and assumes that
434 such pairs of sites must have arisen via recombination (under the infinite-sites model). We found
435 all four gametes at 2 pairs of sites (1.45×10^{-5} of all pairs) in *P. tetraurelia*, 312781 pairs (2.14 ×

436 10^{-2} of all pairs) in *P. sexaurelia*, 42647 pairs (1.21×10^{-2} of all pairs) in *P. caudatum*, and 0
437 pairs of sites in *P. multimicronucleatum*. The observed variation between species correlates well
438 with levels of total sequence diversity and indicates the bias in power to detect recombination
439 towards species with more sequence variation. Although results from the FGT suggest the
440 presence of some recombination, the probability of finding all four gametes does not increase
441 with the distance between the pairs of sites (slope= -7.64×10^{-9}, $p = 0.765$ for *P. sexaurelia*;
442 slope = -1.39×10^{-8}, $p = 0.792$ for *P. caudatum*), contrary to the expectation under
443 recombination, and is inconsistent with other distance-based analyses. Violation of the infinite-
444 sites model (i.e., recurrent mutation at the same site) can result in the presence of pairs of sites
445 with all four alleles. The FGT works best for species and genomes whose recombination rate is
446 much larger than the mutation rate, which may not be the case in *Paramecium* mitochondria.
447
448 In order to test if the presence of four gametes was more likely to be caused by recombination or
449 mutation, we used LDhat (McVean, et al. 2002) to test for recombination and estimate $2N_er$,
450 using values of $2N_e\mu$ estimated by nucleotide diversity as calculated in Johri et al., 2017
451 (Supplementary Table 4). Here r is the recombination rate, μ is the mutation rate and N_e is the
452 effective population size of the species under consideration. $2N_er$ estimated under the finite-sites
453 mutation model (Wakeley 1997; Hudson 2001) was found to be 0.0 for all four species, and is
454 consistently 0.0 in all non-overlapping windows (1000 bp) spanning the genome. We also used
455 permutation tests (McVean, et al. 2002) to detect recombination. The idea behind these tests is
456 that in the absence of recombination the relative position of SNPs would not change inferred
457 values of statistics used to measure recombination. Thus, the absence of recombination, statistics
458 like sum-of-distance between sites with all four gametes (G4) (Meunier and Eyre-Walker 2001),

459 and correlation between r^2 or D' and physical distance, would not be significantly different from
460 an expectation obtained by randomly shuffling SNPs and computing these statistics. We find no
461 statistical significance for the presence of recombination using permutation tests (Supplementary
462 Table 4) and thus conclude that there is no recombination in *Paramecium* mitochondria.
463
464 **Efficacy of selection in mitochondria versus the nucleus**
465
466 Finally, we asked if mitochondrial genes experience weaker efficacy of purifying
467 selection compared to nuclear genes, as would be expected due to smaller effective
468 population size and the lack of recombination in the mitochondria. We did so by
469 comparing multiple statistics in each of the two genomes. Because our statistics
470 included divergence at synonymous sites, we conducted these analyses primarily on *P.*
471 *tetraurelia* where values of d*S* were calculated with respect to the closest outgroup
472 species *P. biaurelia*, for which we had available sequences in both the nucleus and
473 mitochondria. Average d*S* values for the set of nuclear and mitochondrial genes were
474 found to be 0.885 and 1.174 respectively (Table 2), a small but significant difference. In
475 order to control for differences in d*S* driving the patterns, we also conducted all analyses
476 restricted to genes with d$S < 1.0$. For the set of genes with d$S < 1.0$, the average value of
477 d*S* is not significantly different between the two genomes, 0.636 among the nuclear and
478 0.658 among the mitochondrial genes. A potential caveat of comparing the efficacy of
479 purifying selection between all genes present in the mitochondria vs. nucleus, is that we
480 might instead be measuring differences in strength of purifying selection. In order to
481 correct for that, we also compared nuclear genes with similar functions to those in the

482 mitochondria. We compared the 14 mitochondrial genes with ~13-27 nuclear genes
483 involved in the oxidative phosphorylation pathway, as well as the 8 ribosomal genes in
484 the mitochondria with ~77-430 nuclear genes that encode for structural components of
485 the ribosomes.

486

487 Average d*N*/d*S* is similar or slightly lower for genes in the nucleus (mt = 0.082, nuc =
488 0.079, p = 0.307; Table 2), with the exception of the well-conserved ribosomal genes,
489 which have lower d*N*/d*S* in the mitochondria (mt = 0.044, nuc = 0.088, p = 0.208). None
490 of these differences in d*N*/d*S* are significant, suggesting that the average amount of
491 purifying selection experienced by genes in the nucleus is not significantly different
492 from that experienced by those in the mitochondria.

493

494 The efficacy of purifying selection can also be measured by estimating the fraction of
495 segregating nonsynonymous polymorphisms that undergo fixation. Such a measure can
496 be calculated using the neutrality index (NI = $(P_n / P_s) / (D_n / D_s)$, where P_n and P_s are
497 the number of nonsynonymous and synonymous polymorphisms respectively; D_n and
498 D_s are the number of nonsynonymous and synonymous fixed differences respectively).
499 The neutrality index is usually found to be larger than 1.0 for genes experiencing
500 purifying selection because mildly deleterious variants are allowed to segregate among
501 individuals, but rarely fix in populations. Higher values of the NI suggest that smaller
502 proportion of segregating nonsynonymous variants are allowed to fix in the population,
503 indicating stronger efficacy of purifying selection. The number of branch-specific
504 substitutions at synonymous (D_s) and nonsynonymous (D_n) sites was calculated by

505 ancestral reconstruction, using *P. tetraurelia*, *P. biaurelia*, *P. sexaurelia*, *P. caudatum*,
506 and *P. multimicronucleatum*, for both the nucleus and the mitochondria, and restricted
507 to sites for which changes could be confidently inferred along specific branches. The
508 advantage of using ancestral reconstruction to infer D_n and D_s is that particular sites that
509 are very fast evolving can be excluded from the analysis, thus minimizing problems
510 arising due to saturation of divergence at synonymous sites. Comparing the distributions
511 of values of NI in the mitochondrion and nucleus shows that overall there is no
512 significant difference between the two sets when all genes are included (whether we
513 include or exclude *Ymf* genes) in *P. tetraurelia* (Table 2). However, for genes involved
514 in oxidative phosphorylation, nuclear genes appear to be experiencing significantly
515 stronger efficacy of purifying selection than mitochondrial genes (mt NI = 1.696, nuc
516 NI = 6.087, $p = 0.02$).
517
518 We also calculated the neutrality index as $(\pi_n/\pi_s)/\ (dN/dS)$ (Betancourt, et al. 2012;
519 denoted by NIπ below) for each gene with dN/dS obtained from pairwise comparison
520 with respect to *P. biaurelia*. In this case all sites contribute to the analysis, but the
521 maximum likelihood estimate of pairwise d*N*/d*S* can minimize biases due to saturation
522 of d*S*. Again, we observe that neutrality index of nuclear genes is either similar or
523 higher than those of mitochondrial genes. Genes involved in oxidative phosphorylation
524 in the nucleus have higher but not significantly different NIπ compared to those in the
525 mitochondria (mt = 2.384, nuc = 6.712, $p = 0.422$).
526
527 In order to minimize statistical biases that can arise due to sampling, we also estimated a

528 variation of the neutrality index $NI_{DoS} = D_n/(D_n+D_s) - P_n/(P_n+P_s)$ (Stoletzki and Eyre-Walker
529 2011), where more negative values represent stronger efficacy of purifying selection. As
530 previously observed, we find no significant difference between the neutrality index of genes in
531 the nucleus vs. mitochondria except for Oxphos genes, where nuclear genes have significantly
532 lower values of NI_{DoS} than mitochondrial genes (mt $NI_{DoS} = 0.001$, nuc $NI_{DoS} = -0.227$, $p = 0.01$)
533 This suggests that there are more deleterious variants segregating in the nucleus than in
534 the mitochondrion in *P. tetraurelia*. Overall, mitochondria appear to experience either
535 similar or stronger efficacy of purifying selection than the nucleus.
536
537 Lastly, we used π_n/π_s as a proxy for the efficacy of recent purifying selection. This
538 comparison can be extended to all four species (Supplementary Table 5). We find that
539 average π_n/π_s of mitochondrial (mt) and nuclear genes (nuc) is not significantly different
540 in *P. tetraurelia* (mean mt = 0.163; mean nuc = 0.235; $p = 0.18$) and *P. caudatum* (mean
541 mt = 0.135; mean nuc = 0.170; $p = 0.22$) respectively. However, mean π_n/π_s in the
542 mitochondrial genes is significantly lower than that of nuclear genes in *P. sexaurelia*
543 (mean mt = 0.051; mean nuc = 0.268; $p < 2.2 \times 10^{-16}$) and *P. multimicronucleatum* (mean
544 mt = 0.099; mean nuc = 0.178; $p = 1.87 \times 10^{-3}$) respectively, suggesting that
545 mitochondrial genes might be under stronger purifying selection than nuclear genes.
546
547 As side note, values of NI for *P. sexaurelia*, *P. caudatum* and *P. multimicronucleatum*
548 were consistently found to be much less than 1.0, likely due to under-estimation of
549 changes at synonymous sites. We aimed to reduce the bias caused by saturation of D_s
550 by re-calculating NI only in the mitochondrion using all 13 taxa (in order to break up

551 longer branches). We continue to recover extremely low values of NIs (Supplementary
552 Table 6) and fail to obtain values close to those obtained using NIπ, suggesting that the
553 absolute values of NI can be misleading and have to be interpreted with caution.
554
555
556
557
558 **DISCUSSION**
559 In this study, we greatly expand the set of sequenced ciliate mitochondrial genomes by
560 presenting a wider sampling of the mitogenome diversity within the *Paramecium* genus.
561 Using this wealth of sequence data, we characterize the diversity and conservation of
562 genome organization and gene content, and we study in depth the population genetic
563 characteristics such as mutational and selection pressures acting on mtDNA within the
564 genus.
565
566 ***Ciliate-specific mitochondrial genes***
567 *Paramecium* mitochondrial genomes possess 16 lineage-specific open-reading frames (referred
568 to as *Ymf* genes) that have no known homologs in non-ciliate species and lack assigned
569 functions, but are nonetheless conserved across *Paramecium* and *Tetrahymena* species. Other
570 ciliates like *Oxytricha* have also been found to have unidentified open-reading frames (Swart, et
571 al. 2012), but not all *Ymf* genes have homologs identified in ciliates outside of
572 *Oligohymenophora* (which contains both *Tetrahymena* and *Paramecium*). We find that *Ymf*
573 genes are evolving faster, mostly due to relaxed purifying selection, as Ymf genes are not

574 significantly more likely to undergo positive selection than other genes. Shedding light on the
575 identity of the *Ymf* genes could possibly indicate entirely new sets of genes present in the
576 ancestor of *Oligohymenophorean* mitochondrial genome.
577
578 We found that parts of five of the 16 *Ymf* genes show homology to standard mitochondrial
579 proteins, especially ribosomal proteins. As ciliates are evolutionarily distant from most model
580 organisms in the eukaryotic tree, it is possible that some *Ymf* genes are ribosomal subunits or
581 genes belonging to the NADH complex and are simply not identifiable because of extreme
582 divergence. Similar findings have been reported for other protozoan mitochondrial genomes in
583 the past (e.g. de Graaf, et al. 2009; Pombert, et al. 2013; Burger, et al. 2016; Skippington, et al.
584 2017) that have diverged substantially from the most well-studied mammalian mitochondrial
585 genomes. In some species, RNA editing (including insertions and deletions) can be substantial
586 (Gray 2003), which can mask the proteins encoded from identification from genomic sequence
587 data. We however found no evidence of RNA editing, consistent with a previous report in *P.*
588 *tetraurelia* mitochondria (Orr, et al. 1997), thus it is unlikely to account for the observed
589 divergence. An interesting possibility is that most of the *Ymf* genes are either part of or interact
590 with the ATP synthase complex. In *Tetrahymena thermophila*, the ATP synthase has been
591 reported to form an unusual structure possessing completely novel subunits whose orthologs are
592 not identifiable in other organisms, some of which are *Ymf* genes (Balabaskaran Nina, et al.
593 2010). It is therefore possible that ciliates possess structurally unique proteins that perform
594 relatively conventional functions in the mitochondria, but are difficult to identify based on other
595 known sequences. Although the origin of these genes is unclear, it is intriguing that a set of such
596 fast evolving genes is preserved across highly diverged species.

597

598 ***Change in mutation spectrum and nucleotide composition***

599 We used a combination of previous mutation-accumulation studies and our population-genomic
600 data to determine that the change in nucleotide composition of *P. aurelia* mitochondrial genomes
601 toward higher GC is most likely the result of changes in mutational biases. Unfortunately, due to
602 our modest sample sizes, some of the SNPs observed as singletons in our data may in fact be
603 fairly common in the population at large. Our estimates of mutation spectra might thus be biased
604 by selection. Mutation-accumulation lines in *P. caudatum* would help further disentangle these
605 forces. Because ciliates have very low mutation rates, their mutation-accumulation study requires
606 a large number of generations to produce only a handful of mutations, making it very difficult to
607 obtain precise estimates of the mutation spectrum. The most feasible strategy to more precisely
608 estimate these spectra in *Paramecium* would thus be to acquire larger population samples in
609 order to observe lower-frequency variants. We also note that although an alternative explanation
610 for higher GC content in the *P. aurelia* species could be biased gene conversion, we found no
611 evidence of mitochondrial recombination (including non-crossovers) in the *P. aurelia* species.

612

613 An interesting question raised by our results is how fast the mutation spectrum in mitochondria
614 evolves across species (Lynch, et al. 2008; Montooth and Rand 2008). Re-analyzing data from
615 previous work on mutation-accumulation lines in mitochondria of other model organisms
616 presents a number of interesting observations (Table 1). First, within opisthokonts, there is a
617 huge variation in mutation bias (m) in the mitochondria, ranging from nearly 0 (strongly biased
618 towards GC) to values much larger than 1 (biased towards AT; Table 1). However, most species
619 except *S. cerevisiae* have a mutation bias towards A/T, consistent with most mitochondrial

genomes being AT-rich. Second, across opisthokonts, we find that selection can have a significant impact on mitochondrial GC content. The effect of selection (*S*) can be observed by comparing observed genome-wide nucleotide composition (or specifically at 4-fold degenerate sites) with the expected composition under mutation equilibrium (Table 1). Again, in most species selection favors higher G/C, but in *S. cerevisiae* there appears to be strong selection favoring A/T genome-wide. Thus, there might not be a universal direction for mutation bias or selection for nucleotide composition in mitochondria. Finally, closely related species in other lineages have been shown to have significant differences in mutation bias in the mitochondria as observed for example between *C. elegans* and *C. briggsae* (mutation bias re-calculated by combining observations from Howe, et al. 2010; Konrad, et al. 2017). All of these observations suggest that it may not be that such an unusual event for the mutational biases to shift away from AT along the long branch leading to the *P. aurelia* species. However, the exact nature of the biochemical mechanisms responsible for remains an intriguing open question for future research.

Similar efficacy of purifying selection experienced by the mitochondria and nucleus

Mitochondrial genomes are often non-recombining and are usually passed on via uniparental inheritance. Mitochondria are therefore expected to have lower effective population sizes than that of the nucleus within the same organism (Lynch and Blanchard 1998; Neiman and Taylor 2009). One consequence of reduced effective population sizes is an increased probability of segregation and fixation of slightly deleterious mutations. There have been multiple contradictory reports on whether mitochondrial genes experience stronger or weaker purifying selection than the nuclear genes. Studies

643 examining a small number of protein-coding loci across a large number of species have
644 concluded that mitochondria experience less effective purifying selection than the
645 nucleus (Weinreich and Rand 2000; Betancourt, et al. 2012; Popadin, et al. 2013).
646 However a recent study conducted on multiple individuals and whole genomes in
647 *Drosophila melanogaster* and humans found no significant difference between the
648 efficacy of purifying selection in the mitochondrial vs. nuclear protein-coding genes
649 (Cooper, et al. 2015).
650
651 We confirm the absence of recombination in mitochondria of three species of
652 *Paramecium*. Previous studies in *P. tetraurelia* (Adoutte, et al. 1979; Barth, et al. 2008)
653 and *P. primaurelia* (Beale, et al. 1972) had reached similar conclusions using a small
654 number of markers. We also evaluated the extent of purifying selection in ciliate
655 mitochondria relative to the nucleus primarily for *P. tetraurelia*. We find that despite
656 the lack of recombination in the mitochondria, the efficacy of selection in the
657 mitochondria is similar to if not stronger than that of the nucleus.
658
659 Our results seem to be in discordance with theoretical predictions according to which
660 one would expect the efficacy of selection in the mitochondria to be lower than in the
661 nucleus. One possibility is that we observe stronger efficacy of selection in the
662 mitochondria because the magnitude of negative selection itself is stronger in the
663 mitochondrial genes (Popadin, et al. 2013), which differ from the average nuclear gene
664 in several key ways (Adrion, et al. 2016). Most mitochondrial genes are expressed at
665 higher levels compared to nuclear genes (Havird and Sloan 2016). Mitochondrial

666 encoded proteins (*cox1* and *cox2*) that are part of the oxidative phosphorylation
667 (OXPHOS) pathway are core enzyme catalytic subunits (Tsukihara, et al. 1996; Zhang
668 and Broughton 2013; Havird and Sloan 2016). Finally, most genes retained in the
669 mitochondria encode for highly hydrophobic proteins and have a high GC content
670 (Johnston and Williams 2016). Thus, finding comparable sets of genes between these
671 two genomes is admittedly difficult (but see Lynch 1996; Lynch 1997).
672
673 Lastly, our results need to be evaluated in the light of *Paramecium*-specific life cycle. In
674 *Paramecium*, there is almost no exchange of cytoplasm during conjugation (Koizumi
675 and Kobayashi 1989; Meyer and Garnier 2002). Thus it appears that both parents pass
676 on their mitochondria to their respective offspring, without exchange or degradation of
677 mitochondrial genomes. In addition, mitochondria double before binary fission and are
678 distributed randomly across the daughter cells. Thus, mitochondria do not appear to
679 undergo bottlenecks during any stage of the life cycle in *Paramecium,* and experience
680 no associated reduction in effective population size. On the other hand, *Paramecium*
681 species frequently undergo asexual reproduction, thereby reducing the nuclear effective
682 population size. Therefore, effective population sizes of mitochondrial and nuclear
683 genomes may be more similar in *Paramecium* than in many other organisms. Due to the
684 combination of sexual and asexual reproduction, under equilibrium conditions, expected
685 heterozygosity in the nucleus would take values between $4N_e\mu_n$ - $2N_e\mu_n$. Similarly, in
686 the mitochondria, expected hetrozygosity would be ~$2N_e\mu_m$. Assuming that the ratio of
687 divergence at silent sites can be used as a proxy for the ratio of mutation rate between
688 the two compartments, it is possible to approximately estimate the ratio of effective

689 population sizes of the two compartments as $N_{e(m)}/N_{e(n)} = y \times (\pi_m/\pi_n)/(d_m/d_n)$, where y
690 would be some number between 1 and 2. For *P. tetraurelia*, for which we have
691 relatively closer outgroup species and thus more reliable divergence estimates, our
692 estimated range of $N_{e(m)}/N_{e(n)}$ is 0.94 – 1.88. Although underestimation of neutral
693 divergence in the mitochondria relative to the nucleus could skew our inference slightly,
694 the above calculation suggests that effective population sizes of mitochondria may be
695 similar or, larger than that of the nucleus in *Paramecium*.
696
697 We therefore conclude that our finding of similar or stronger efficacy of selection in the
698 mitochondria relative to the nucleus in *Paramecium* may lie within theoretical expectations given
699 *Paramecium*'s unique life cycle and mode of mitochondrial transmission. A better understanding
700 of the *Paramecium* life cycle in the wild might help build more appropriate null expectations in
701 the future. Our results suggest the possibility that unicellular eukaryotes in general may have
702 larger mitochondrial than nuclear effective population sizes and more efficacious purifying
703 selection in the mitochondria might be more common than believed.
704
705 **METHODS:**
706
707 **Genome sequencing and assembly**
708 Single isolates of *P. jenningsi* (strain: M), *P. octaurelia*, *P. decaurelia* (strain: 223), *P.*
709 *dodecaurelia* (strain: 274), *P. novaurelia* (strain: TE) and *P. quadecaurelia* (strain: 328) were
710 used to extract macronuclear DNA. DNA extraction, sequencing library preparation and genome
711 sequencing were previously described (Johri et al. 2017). Sequencing reads were assembled

712 using SPAdes (Bankevich, et al. 2012; version 3.5.0) after removing potential adapter sequence
713 with Trimmomatic (Bolger, et al. 2014; version 0.33). Mitochondrial contigs were identified
714 from the resulting assemblies by BLAST (Altschul, et al. 1997) searches against the published *P.*
715 *caudatum* and *P. tetraurelia* mitochondrial genomes. Mitochondrial re-sequenced data from 5-
716 10 isolates of *P. tetraurelia*, *P. sexaurelia*, *P. caudatum* and *P. multimicronucleatum* were
717 downloaded from SRA (PRJNA490059) and SNPs were called as described in Johri et al. 2017,
718 using reference genomes of strain 99 for *P. tetraurelia*, strain 130 for *P. sexaurelia*, C104 for *P.*
719 *caudatum*, and M04 for *P. multimicronucleatum*.
720

721 **Mitochondrial genome annotation**

722 Genome annotation was carried out as follows. Protein coding genes were identified by
723 generating all ORFs longer than 60aa in all six frames, using the Mold, Protozoan, and
724 Coelenterate Mitochondrial Code (i.e., UGA codes for W instead of being a stop codon) and all
725 alternative start codons specific to *Paramecium* (AUU, AUA, AUG, AUC, GUG, and GUA),
726 and retaining the longest ORFs associated with each stop codon. BLASTP was then used to
727 identify homologs of annotated mitochondrial proteins in *P. tetraurelia* and *P. caudatum*.
728 Additional ORFs were identified by imposing the requirement that their length exceeds 100aa,
729 and subsequently annotated using BLASTP against the Non-redundant protein sequences (nr)
730 database and HMMER3.0 (Eddy 2011) scans against the PFAM 27.0 database (Finn, et al.
731 2014). tRNA genes were annotated with tRNAscan-SE (Schattner, et al. 2005; version 1.21),
732 using the "Mito/Chloroplast" source. rRNA genes were identified using Infernal (Nawrocki,
733 et al. 2009; version 1.1.1).
734

735 **Identification of telomeres**

736 Telomeric repeats were identified as follows. The first and the last 200bp of each *de novo*
737 assembled mitochondrial genome were used as input to the MEME *de novo* motif finding
738 program (Bailey, et al. 2009; version 4.6.1), which was run with the following parameters: "-
739 *maxw 25 -dna -nmotifs 5 -mod anr*". The repetitive units defined that way were then manually
740 aligned to each other and refined to arrive at final telomeric repeats comparable across all
741 species.

742

743 **RNA-seq analysis**

744 For each species, sequencing reads were aligned against a combined Bowtie (Langmead, et al.
745 2009) index containing both the nuclear and mitochondrial genomes using TopHat2 (Kim, et al.
746 2013; version 2.0.8) with the following settings: "*--bowtie1 --no-discordant --no-mixed --*
747 *microexon-search --read-realign-edit-dist 0 --read-edit-dist 4 --read-mismatches 4 --min-intron-*
748 *length 10 --max-intron-length 1000000 --min-segment-intron 10 --min-coverage-intron 10*".
749 Custom python scripts were then used to identify sequence variants relative to the mitogenomes
750 assemblies.

751

752 **Building phylogenetic trees**

753 Nucleotide sequences were extracted, aligned using MUSCLE, and concatenated.
754 Missing data was encoded as "N". RAxML was used to build the tree with
755 GTRGAMMA as the substitution model. Boostrap values for 1000 replicates were
756 obtained via the fast method recommended by RAxML with the following command
757 line:

758 *raxmlHPC -f a -s sequences.fasta –n sequences_boot -m GTRGAMMA -T 50 -p 31 -x 7777 -N*
759 *1000*
760
761 **Estimation of *dN*/*dS*, D_n, D_s, P_n, P_s, and π_n/π_s**
762 *dN*/*dS* was estimated either across the phylogeny using CODEML, PAML (Yang 2007;
763 version 4.9a) and for the *P. aurelia* species it was also estimated pairwise w.r.t the
764 closest outgroup species using yn00, PAML. π_n/π_s was obtained for all protein coding
765 genes in the 4 species- *P. tetraurelia*, *P. sexaurelia*, *P. caudatum*, and *P.*
766 *multimicronucleatum*, where total number of changes in synonymous sites was greater
767 than 1. This filter was executed in order to reduce errors in π_n/π_s due to very low values
768 of synonymous polymorphisms in a gene.
769
770 D_n and D_s, the number of nonsynonymous and synonymous changes, were inferred by
771 performing ancestral reconstruction at each site, and then counting branching-specific
772 substitutions. Ancestral reconstruction (PAML, baseml, GTR model) was conducted
773 over the phylogeny of all available taxa. For comparing statistics between mitochondrial
774 and nuclear genes, the ancestral reconstruction was performed over the same set of taxa
775 for both, i.e. over *P. tetraurelia*, *P. sexaurelia*, *P. caudatum*, and *P.*
776 *multimicronucleatum*. For all analyses that involved ancestral reconstruction, only sites
777 whose posterior probability of the inferred ancestral state ≥ 0.85 were used in the
778 analyses—this filter was imposed for counting synonymous and nonsynonymous
779 polymorphisms (P_s and P_n, respectively) as well as divergent sites (D_n and D_s). There
780 may be some concern that ancestral reconstruction could end up biasing the ratio of

781 D_n/D_s as many more changes at synonymous sites might result in lower confidence in
782 inferring ancestral states at synonymous but not nonsynonymous sites. Such a bias
783 would increase D_n relative to D_s and thus decrease values of NI. Therefore, all analyses
784 involving D_n and D_s were also performed including all sites, with no filter, and results
785 remained unchanged.
786
787 **Calculation of multiple estimators of Neutrality Index (NI) and statistical tests**
788 Several estimators of neutrality index have been proposed in order to counter different biases.
789 The estimators we used were calculated as follows:
790
791 Simple neutrality index, NI = $P_n/P_s/D_n/D_s$ (Rand and Kann 1996).
792 $NI_\pi = \pi_n/\pi_s/d_N/d_S$, where d_N/d_S was calculated pairwise, w.r.t closest outgroup species
793 (Betancourt, et al. 2012).
794 $NI_{TG} = \Sigma_i[D_{si}.P_{ni}/(P_{si} + D_{si})] / \Sigma_i [P_{si}.D_{ni}/(P_{si} + D_{si})]$, where i is the ith gene (Tarone 1981;
795 Greenland 1982).
796 $NI_{DoS} = D_n/(D_n+D_s) - P_n/(P_n+P_s)$ (Stoletzki and Eyre-Walker 2011).
797
798 For McDonald Kreitman test (McDonald and Kreitman 1991), Fisher's exact test in R (R-
799 Core-Team 2014) was used to test significance. In all cases p value was corrected by
800 Bonferonni-Holm method (Holm 1979) for multiple tests, using R.
801
802 **Mutation spectrum from mutation accumulation (MA) lines**
803 MA line experiments for *P. tetraurelia* had previously been published (Sung et al. 2012) and

804 results from their analysis were used directly in this study.

805

806 MA experiments carried out in order to obtain nuclear mutation rates in *P. biaurelia* and *P.*
807 *sexaurelia* (Long, Doak, et al. 2018) were reanalyzed as follows. Sequencing reads were
808 assembled for each MA line individually and mitochondrial contigs identified as described
809 above. A composite consensus mitochondrial genome sequence was then created from the
810 individual assemblies by creating multiple sequence alignments of all mitochondrial contigs
811 using MAFFT (Katoh and Standley 2013; version 7.221) and retaining the most frequent base
812 for each alignment column (with the exception of telomeres, which were manually curated).
813 Adapter-trimmed reads were then aligned in a 2×100bp format against a combined Bowtie
814 index, containing a combination of the nuclear and consensus mitochondrial genomes, allowing
815 for up to 3 mismatches and retaining only unique reads. Putative mutations were identified by
816 requiring that any variant is supported by at least 3 non-redundant read pairs on each strand, is
817 supported by not more than 4 times more reads on one strand than on the other, and is also
818 observed in ≥5% of reads covering a given position. Telomeric sequences were excluded due to
819 an excessively high number of sequence variants observed in those regions.

820

821 **Mutation spectrum from population genomics data**

822 For each SNP, the ancestral allele was inferred by performing ancestral reconstruction on the
823 13-taxa phylogeny to predict the nucleotides on internal nodes (see above). The ancestral allele
824 was used to infer the derived allele segregating in *P. tetraurelia*, *P. sexaurelia*, *P. caudatum* and
825 *P. multimicronucleatum*. This analysis was restricted to sites where the ancestral nucleotide was
826 inferred with confidence score >= 0.90, where the derived allele was present in only a single

individual, i.e. was a singleton, and was at 4-fold degenerate site. Of these, we counted all mutations that were from G/C to A/T or from A/T to G/C, with respect to the total number of utilizable sites that were counted according to the same criteria as above. Care was also taken to remove all sites that were part of overlapping ORFs.

Calculation of bias in mutation spectrum

Mutation rate of A/T → G/C (u) and G/C → A/T (v) mutations was calculated as follows:

u = (number of A/T → G/C mutations) / (total number of utilizable A/T sites)

v = (number of G/C → A/T mutations) / (total number of utilizable G/C sites)

Mutation bias towards A/T (m) was calculated as, $m = v/u$ and the expected equilibrium G/C content was calculated as $1/(1+m)$ following Lynch 2007.

$S = 4N_e s$ (or $2N_e s$), is the population-scaled strength of selection towards A/T nucleotides and can be calculated using the equation, $P_{AT} = 1/(1 + m^{-1}e^{-S})$, where P_{AT} is the observed fraction of A/T sites and s is the average selective advantage of A/T over G/C nucleotides (Bulmer 1991). Thus $S = -\ln[(m.(1-P_{AT}))/P_{AT}] = \ln(P_{AT}/P_{GC}/v/u)$.

Recombination analyses

All analyses to detect recombination were restricted to SNPs that were biallelic, homozygous, and had a known ancestral state. The statistic (r^2) to measure linkage disequilibrium was calculated as $r^2 = (f_{Aa} - f_A \times f_a)^2 / [f_A \times f_a \times (1-f_A) \times (1-f_a)]$. The program "pairwise" in LDhat 2.2 (McVean, et al. 2002) was used to infer recombination rates using the permuted composite likelihood test as well other permutation tests. These tests were performed under both the gene

850 conversion (average tract length: 500) and cross-over models with 2 values of θ (= $4N_e\mu$) for
851 each species: the closest allowed θ value lower than that estimated from nucleotide diversity
852 values, and the closest higher value.
853
854 **Identifying nuclear genes belonging to Oxidative Phosphorylation pathway and Ribosomal**
855 **complex**
856 KEGG (Kanehisa, et al. 2017) was used to obtain the full list of genes that are part of complexes
857 involved in Oxidative Phosphorylation (Complexes I-V) for *P. tetraurelia*. A total number of 87
858 genes in *P. tetraurelia* were obtained, and their corresponding orthologs were identified in other
859 species. For genes encoding proteins that are part of the ribosomal complex, we used the
860 PANTHER (Mi, et al. 2017) annotation obtained in a previous study (McGrath, Gout, Johri, et
861 al. 2014) for all species and selected all genes that were structural constituents of ribosome. This
862 allowed us to start with a set of 585 genes in *P. tetraurelia*, 564 in *P. sexaurelia*, and 213 genes
863 in *P. caudatum*. However, analyses requiring orthologs from all 3 species were conducted with a
864 smaller subset of genes.
865
866
867 **ACKNOWLEDGEMENTS:**
868 We thank Jean-Francois Gout for technical help, as well as for a critical reading of the
869 manuscript. We thank Hongan Long for providing us with sequencing data for *P. biaurelia* and
870 *P. sexaurelia* nuclear mutation-accumulation experiments. We also thank Jeffrey Palmer for
871 critical reading of the manuscript and discussions about the project. This work was financially
872 supported by the National Science Foundation (MCB-1050161 and DEB-1257806).

873

874 **DATA AVAILABILITY:**

875 Assembled mitogenomes and protein coding gene annotations can be accessed at

876 https://github.com/georgimarinov/Paramecium_mitochondrial_genomes.

877

878

879

880 **REFERENCES**

881 (Xu, et al. 2012), (Haag-Liautard, et al. 2008),

882 Adoutte A, Knowles JK, Sainsard-Chanet A. 1979. Absence of detectable mitochondrial
883 recombination in *Paramecium*. Genetics 93:797-831.
884 Adrion JR, White PS, Montooth KL. 2016. The Roles of Compensatory Evolution and
885 Constraint in Aminoacyl tRNA Synthetase Evolution. Mol Biol Evol 33:152-161.
886 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped
887 BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids
888 Res 25:3389-3402.
889 Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS.
890 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202-208.
891 Balabaskaran Nina P, Dudkina NV, Kane LA, van Eyk JE, Boekema EJ, Mather MW, Vaidya
892 AB. 2010. Highly divergent mitochondrial ATP synthase complexes in *Tetrahymena*
893 *thermophila*. PLoS Biol 8:e1000418.
894 Ballard JW, Whitlock MC. 2004. The incomplete natural history of mitochondria. Mol Ecol
895 13:729-744.
896 Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko
897 SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: a new genome assembly algorithm and its
898 applications to single-cell sequencing. J Comput Biol 19:455-477.
899 Barr CM, Neiman M, Taylor DR. 2005. Inheritance and recombination of mitochondrial
900 genomes in plants, fungi and animals. New Phytol 168:39-50.
901 Barth D, Berendonk TU. 2011. The mitochondrial genome sequence of the ciliate *Paramecium*
902 *caudatum* reveals a shift in nucleotide composition and codon usage within the genus
903 *Paramecium*. BMC Genomics 12:272.
904 Barth D, Przybos E, Fokin SI, Schlegel M, Berendonk TU. 2008. Cytochrome b sequence data
905 suggest rapid speciation within the *Paramecium aurelia* species complex. Mol Phylogenet Evol
906 49:669-673.
907 Beale GH, Knowles JK, Tait A. 1972. Mitochondrial genetics in Paramecium. Nature 235:396-
908 397.
909 Beale GH, Tait A. 1981. Mitochondrial genetics of Paramecium aurelia. In: G.H. Bourne JFD,
910 K.W. Jeon, editor. International review of cytology: Academic Press. p. 19-40.
911 Betancourt AJ, Blanco-Martin B, Charlesworth B. 2012. The relation between the neutrality
912 index for mitochondrial genes and the distribution of mutational effects on fitness. Evolution
913 66:2427-2438.

914 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence
915 data. Bioinformatics 30:2114-2120.
916 Borza T, Redmond EK, Laflamme M, Lee RW. 2009. Mitochondrial DNA in the
917 *Oogamochlamys* clade (*Chlorophyceae*): high Gc content and unique genome architecture for
918 Green algae(1). J Phycol 45:1323-1334.
919 Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics
920 129:897-907.
921 Burger G, Forget L, Zhu Y, Gray MW, Lang BF. 2003. Unique mitochondrial genome
922 architecture in unicellular relatives of animals. Proc Natl Acad Sci U S A 100:892-897.
923 Burger G, Moreira S, Valach M. 2016. Genes in Hiding. Trends Genet 32:553-565.
924 Burger G, Zhu Y, Littlejohn TG, Greenwood SJ, Schnare MN, Lang BF, Gray MW. 2000.
925 Complete sequence of the mitochondrial genome of *Tetrahymena pyriformis* and comparison
926 with *Paramecium aurelia* mitochondrial DNA. J Mol Biol 297:365-380.
927 Cooper BS, Burrus CR, Ji C, Hahn MW, Montooth KL. 2015. Similar efficacies of selection
928 shape mitochondrial and nuclear genes in both *Drosophila melanogaster* and *Homo sapiens*. G3
929 5:2165-2176.
930 Coyne RS, Hannick L, Shanmugam D, Hostetler JB, Brami D, Joardar VS, Johnson J, Radune D,
931 Singh I, Badger JH, et al. 2011. Comparative genomics of the pathogenic ciliate *Ichthyophthirius*
932 *multifiliis*, its free-living relatives and a host species provide insights into adoption of a parasitic
933 lifestyle and prospects for disease control. Genome Biol 12:R100.
934 de Graaf RM, Ricard G, van Alen TA, Duarte I, Dutilh BE, Burgtorf C, Kuiper JW, van der
935 Staay GW, Tielens AG, Huynen MA, et al. 2011. The organellar genome and metabolic potential
936 of the hydrogen-producing mitochondrion of *Nyctotherus ovalis*. Mol Biol Evol 28:2379-2391.
937 de Graaf RM, van Alen TA, Dutilh BE, Kuiper JW, van Zoggel HJ, Huynh MB, Gortz HD,
938 Huynen MA, Hackstein JH. 2009. The mitochondrial genomes of the ciliates *Euplotes minuta*
939 and *Euplotes crassus*. BMC Genomics 10:514.
940 Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195.
941 Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K,
942 Holm L, Mistry J, et al. 2014. Pfam: the protein families database. Nucleic Acids Res 42:D222-
943 230.
944 Fokin SI, Gortz HD. 2009. Diversity of *Holospora* bacteria in *Paramecium* and their
945 characterization. Microbiol Monogr 12:161-199.
946 Fritsch ES, Chabbert CD, Klaus B, Steinmetz LM. 2014. A genome-wide map of mitochondrial
947 DNA recombination in yeast. Genetics 198:755-771.
948 Goddard JM, Cummings DJ. 1975. Structure and replication of mitochondrial DNA from
949 *Paramecium aurelia*. J Mol Biol 97:593-609.
950 Gray MW. 2003. Diversity and evolution of mitochondrial RNA editing systems. IUBMB Life
951 55:227-233.
952 Greenland S. 1982. Interpretation of summary measures when interaction is present. Am J
953 Epidemiol 116:587-587.
954 Gruchy DF. 1955. The breeding system and distribution of *Tetrahymena pyriformis*. J Protozool
955 2:178-185.
956 Haag-Liautard C, Coffey N, Houle D, Lynch M, Charlesworth B, Keightley PD. 2008. Direct
957 estimation of the mitochondrial DNA mutation rate in Drosophila melanogaster. PLoS Biol
958 6:e204.
959 Havird JC, Sloan DB. 2016. The roles of mutation, selection, and expression in determining

960 relative rates of evolution in mitochondrial versus nuclear genomes. Mol Biol Evol 33:3042-
961 3053.
962 Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. Theor Appl Genet
963 38:226-231.
964 Holm S. 1979. A simple sequentially rejective multiple test procedure. Scand J Stat 6:65-70.
965 Hori M, Tomikawa I, Przybos E, Fujishima M. 2006. Comparison of the evolutionary distances
966 among syngens and sibling species of *Paramecium*. Mol Phylogenet Evol 38:697-704.
967 Howe DK, Baer CF, Denver DR. 2010. High rate of large deletions in *Caenorhabditis briggsae*
968 mitochondrial genome mutation processes. Genome Biol Evol 2:29-38.
969 Hudson RR. 2001. Two-locus sampling distributions and their application. Genetics 159:1805-
970 1817.
971 Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in
972 the history of a sample of DNA sequences. Genetics 111:147-164.
973 Johnston IG, Williams BP. 2016. Evolutionary inference across eukaryotes identifies specific
974 pressures favoring mitochondrial gene retention. Cell Syst 2:101-111.
975 Johri P, Krenek S, Marinov GK, Doak TG, Berendonk TU, Lynch M. 2017. Population
976 genomics of *Paramecium* species. Mol Biol Evol 34:1194-1216.
977 Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on
978 genomes, pathways, diseases and drugs. Nucleic Acids Res 45:D353-D361.
979 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
980 improvements in performance and usability. Mol Biol Evol 30:772-780.
981 Kiefel BR, Gilson PR, Beech PL. 2006. Cell biology of mitochondrial dynamics. International
982 Review of Cytology - a Survey of Cell Biology 254:151-213.
983 Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate
984 alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome
985 Biol 14:R36.
986 Koizumi S, Kobayashi S. 1989. Microinjection of plasmid DNA encoding the A surface antigen
987 of *Paramecium tetraurelia* restores the ability to regenerate a wild-type macronucleus. Mol Cell
988 Biol 9:4398-4401.
989 Konrad A, Thompson O, Waterston RH, Moerman DG, Keightley PD, Bergthorsson U, Katju V.
990 2017. Mitochondrial mutation rate, spectrum and heteroplasmy in *Caenorhabditis elegans*
991 spontaneous mutation accumulation lines of differing population size. Mol Biol Evol 34:1319-
992 1334.
993 Kukat C, Wurm CA, Spahr H, Falkenberg M, Larsson NG, Jakobs S. 2011. Super-resolution
994 microscopy reveals that mammalian mitochondrial nucleoids have a uniform size and frequently
995 contain a single copy of mtDNA. Proc Natl Acad Sci U S A 108:13534-13539.
996 Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment
997 of short DNA sequences to the human genome. Genome Biol 10:R25.
998 Long H, Doak TG, Lynch M. 2018. Consistently low base-substitution mutation rates within the
999 *Paramecium aurelia* species complex.
1000 Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, Patterson C, Gregory C,
1001 Strauss C, Stone C, et al. 2018. Evolutionary determinants of genome-wide nucleotide
1002 composition. Nature ecology & evolution.
1003 Lynch M. 1997. Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA
1004 genes. Mol Biol Evol 14:914-925.
1005 Lynch M. 1996. Mutation accumulation in transfer RNAs: molecular evidence for Muller's

1006 ratchet in mitochondrial genomes. Mol Biol Evol 13:209-220.
1007 Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.
1008 Lynch M, Blanchard JL. 1998. Deleterious mutation accumulation in organelle genomes.
1009 Genetica 102-103:29-39.
1010 Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle
1011 genomic architecture. Science 311:1727-1730.
1012 Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K,
1013 Kulkarni S, Hartl DL, et al. 2008. A genome-wide view of the spectrum of spontaneous
1014 mutations in yeast. Proc Natl Acad Sci U S A 105:9272-9277.
1015 Mackenzie SA. 2007. The unique biology of mitochondrial genome instability in plants. In:
1016 Logan D, editor. Plant Mitochondria: Blackwell Publishing, Oxford. p. 36–46.
1017 McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*.
1018 Nature 351:652-654.
1019 McGrath CL, Gout JF, Doak TG, Yanagi A, Lynch M. 2014. Insights into three whole-genome
1020 duplications gleaned from the *Paramecium caudatum* genome sequence. Genetics 197:1417-
1021 1428.
1022 McGrath CL, Gout JF, Johri P, Doak TG, Lynch M. 2014. Differential retention and divergent
1023 resolution of duplicate genes following whole-genome duplication. Genome Res 24:1665-1675.
1024 McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and
1025 estimating recombination from gene sequences. Genetics 160:1231-1241.
1026 Meunier J, Eyre-Walker A. 2001. The correlation between linkage disequilibrium and distance:
1027 implications for recombination in hominid mitochondria. Mol Biol Evol 18:2132-2135.
1028 Meyer E, Garnier O. 2002. Non-Mendelian inheritance and homology-dependent effects in
1029 ciliates. Advances in genetics 46:305-337.
1030 Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. 2017. PANTHER
1031 version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data
1032 analysis tool enhancements. Nucleic Acids Res 45:D183-D189.
1033 Montooth KL, Rand DM. 2008. The spectrum of mitochondrial mutation differs across species.
1034 PLoS Biol 6:e213.
1035 Moradian MM, Beglaryan D, Skozylas JM, Kerikorian V. 2007. Complete mitochondrial
1036 genome sequence of three *Tetrahymena* species reveals mutation hot spots and accelerated
1037 nonsynonymous substitutions in Ymf genes. PLoS ONE 2:e650.
1038 Morin GB, Cech TR. 1988. Mitochondrial telomeres: surprising diversity of repeated telomeric
1039 DNA sequences among six species of *Tetrahymena*. Cell 52:367-374.
1040 Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments.
1041 Bioinformatics 25:1335-1337.
1042 Neiman M, Taylor DR. 2009. The causes of mutation accumulation in mitochondrial genomes.
1043 Proc. R. Soc. B 276:1201-1209.
1044 Orr AT, Rabets JC, Horton TL, Landweber LF. 1997. RNA editing missing in mitochondria. Rna
1045 3:335-336.
1046 Perasso R, Beisson J. 1978. Temporal Pattern of Mitochondrial Multiplication during Cell-Cycle
1047 of Paramecium. Biol Cellulaire 32:275-290.
1048 Piganeau G, Eyre-Walker A. 2004. A reanalysis of the indirect evidence for recombination in
1049 human mitochondrial DNA. Heredity 92:282-288.
1050 Pombert J-F, Smirnov A, James ER, Janouškovec J, Gray MW, Keeling PJ. 2013. The complete
1051 mitochondrial genome from an unidentified *Phalansterium* species. Protist Genomics 1:25-32.

1052 Popadin KY, Nikolaev SI, Junier T, Baranova M, Antonarakis SE. 2013. Purifying selection in
1053 mammalian mitochondrial protein-coding genes is highly effective and congruent with evolution
1054 of nuclear genes. Mol Biol Evol 30:347-355.
1055 Preer LB. 1969. Alpha, an infectious macronuclear symbiont of *Paramecium aurelia*. J Protozool
1056 16:570-578.
1057 Price CA, Reardon EM. 2001. Mendel, a database of nomenclature for sequenced plant genes.
1058 Nucleic Acids Res 29:118-119.
1059 R-Core-Team. 2014. R: A language and environment for statistical computing. R Foundation for
1060 Statistical Computing, Vienna, Austria.
1061 Rand DM, Kann LM. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts
1062 among genes from *Drosophila*, mice, and humans. Mol Biol Evol 13:735-748.
1063 Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers
1064 for the detection of tRNAs and snoRNAs. Nucleic Acids Res 33:W686-689.
1065 Skippington E, Barkman TJ, Rice DW, Palmer JD. 2017. Comparative mitogenomics indicates
1066 respiratory competence in parasitic Viscum despite loss of complex I and extreme sequence
1067 divergence, and reveals horizontal gene transfer and remarkable variation in genome size. BMC
1068 Plant Biol 17:49.
1069 Slabodnick MM, Ruby JG, Reiff SB, Swart EC, Gosai S, Prabakaran S, Witkowska E, Larue GE,
1070 Fisher S, Freeman RM, Jr., et al. 2017. The macronuclear genome of *Stentor coeruleus* reveals
1071 tiny introns in a giant cell. Current biology : CB 27:569-575.
1072 Smith DR. 2016. The past, present and future of mitochondrial genomics: have we sequenced
1073 enough mtDNAs? Brief Funct Genomics 15:47-54.
1074 Smith DR. 2012. Updating our view of organelle genome nucleotide landscape. Front Genet
1075 3:175.
1076 Smith DR, Keeling PJ. 2015. Mitochondrial and plastid genome architecture: Reoccurring
1077 themes, but significant differences at the extremes. Proc Natl Acad Sci U S A 112:10177-10184.
1078 Soding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology
1079 detection and structure prediction. Nucleic Acids Res 33:W244-248.
1080 Sonneborn TM. 1975. *Paramecium aurelia* complex of 14 sibling species. Trans Am Microsc
1081 Soc 94:155-178.
1082 Sonneborn TM. 1937. Sex, sex inheritance and sex determination in *Paramecium aurelia*. Proc
1083 Natl Acad Sci U S A 23:378-385.
1084 Stadler T, Delph LF. 2002. Ancient mitochondrial haplotypes and evidence for intragenic
1085 recombination in a gynodioecious plant. Proc Natl Acad Sci U S A 99:11730-11735.
1086 Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. Mol Biol Evol 28:63-70.
1087 Sueoka N. 1993. Directional mutation pressure, mutator mutations, and dynamics of molecular
1088 evolution. J Mol Evol 37:137-153.
1089 Swart EC, Nowacki M, Shum J, Stiles H, Higgins BP, Doak TG, Schotanus K, Magrini VJ, Minx
1090 P, Mardis ER, et al. 2012. The *Oxytricha trifallax* mitochondrial genome. Genome Biol Evol
1091 4:136-154.
1092 Tarcz S, Potekhin A, Rautian M, Przybos E. 2012. Variation in ribosomal and mitochondrial
1093 DNA sequences demonstrates the existence of intraspecific groups in *Paramecium*
1094 *multimicronucleatum* (Ciliophora, Oligohymenophorea). Mol Phylogenet Evol 63:500-509.
1095 Tarone RE. 1981. On summary estimators of relative risk. J Chronic Dis 34:463-468.
1096 Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima
1097 R, Yaono R, Yoshikawa S. 1996. The whole structure of the 13-subunit oxidized cytochrome c

1098 oxidase at 2.8 A. Science 272:1136-1144.
1099 Visser W, van Spronsen EA, Nanninga N, Pronk JT, Gijs Kuenen J, van Dijken JP. 1995. Effects
1100 of growth conditions on mitochondrial morphology in Saccharomyces cerevisiae. Antonie van
1101 Leeuwenhoek 67:243-253.
1102 Vlcek C, Marande W, Teijeiro S, Lukes J, Burger G. 2011. Systematically fragmented genes in a
1103 multipartite mitochondrial genome. Nucleic Acids Res 39:979-988.
1104 Wakeley J. 1997. Using the variance of pairwise differences to estimate the recombination rate.
1105 Genet Res 69:45-48.
1106 Weinreich DM, Rand DM. 2000. Contrasting patterns of nonneutral evolution in proteins
1107 encoded in nuclear and mitochondrial genomes. Genetics 156:385-399.
1108 Xu S, Schaack S, Seyfert A, Choi E, Lynch M, Cristescu ME. 2012. High mutation rates in the
1109 mitochondrial genomes of Daphnia pulex. Mol Biol Evol 29:763-769.
1110 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-
1111 1591.
1112 Zhang F, Broughton RE. 2013. Mitochondrial-nuclear interactions: compensatory evolution or
1113 variable functional constraint among vertebrate oxidative phosphorylation genes? Genome Biol
1114 Evol 5:1781-1791.
1115

FIGURES and TABLES

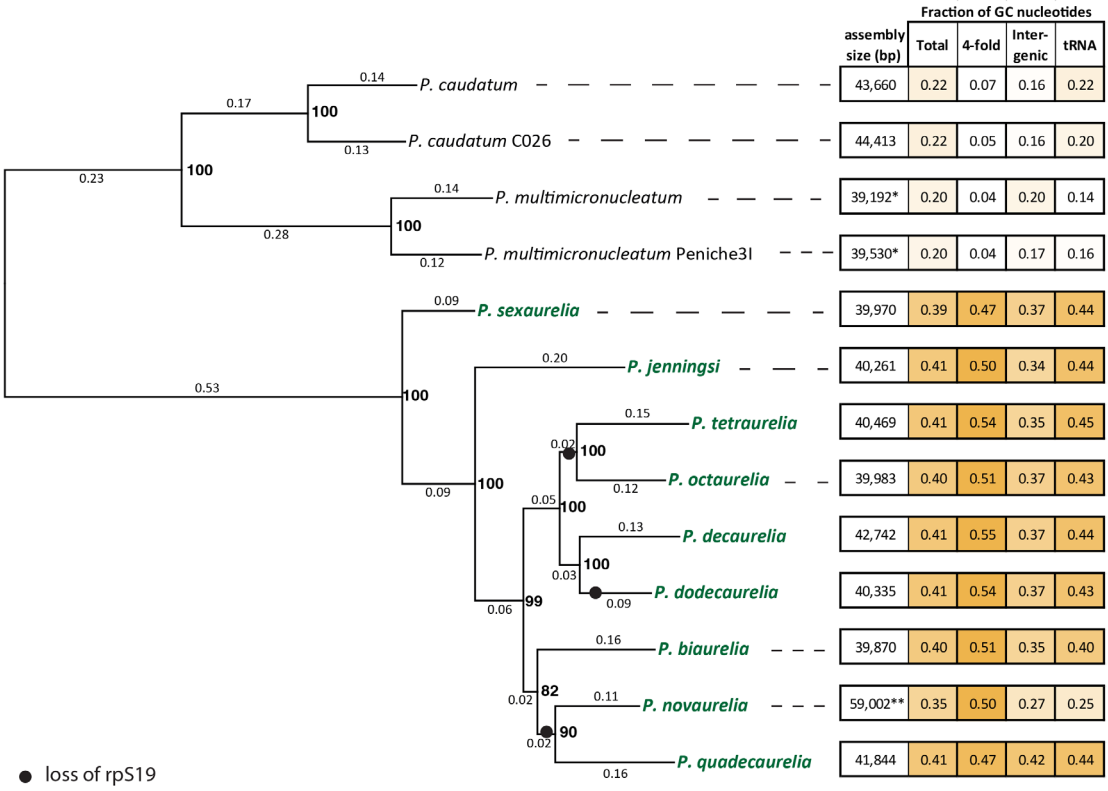


Figure 1: Mitochondrial phylogeny and change in nucleotide composition across *Paramecium* species. The species in green belong to the *P. aurelia* complex. Numbers on branches indicate total number of substitutions per site and bold numbers on nodes show bootstrap values. Black solid circles show the inferred event of loss of the ribosomal protein *rpS19*. The numbers on the right show the total assembly size (in bp) and mean GC content of the whole mitochondrial genome (Total), at 4-fold degenerate sites, at intergenic regions, and of tRNAs respectively. Starred (*) numbers indicate incomplete assembly of mitochondrial genomes.

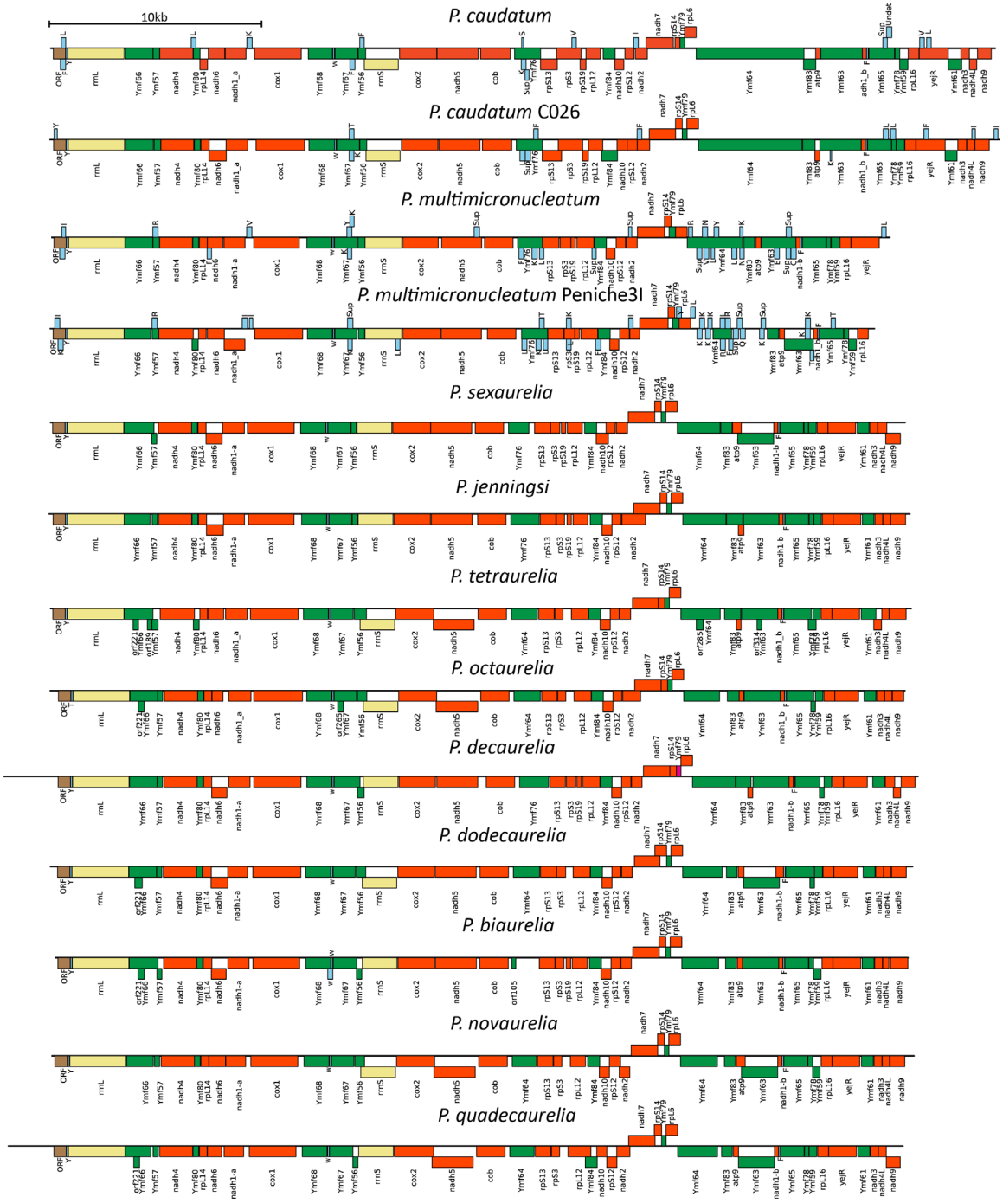


Figure 2: Structure of mitochondrial genomes in *Paramecium*. All *Paramecium* mitochondrial genomes are linear with 24 protein-coding genes (shown in reddish orange), 2 rRNA genes (yellow), lineage-specific genes referred to as *Ymf* genes (in green) and varying numbers and content of tRNA genes (shown in light blue).

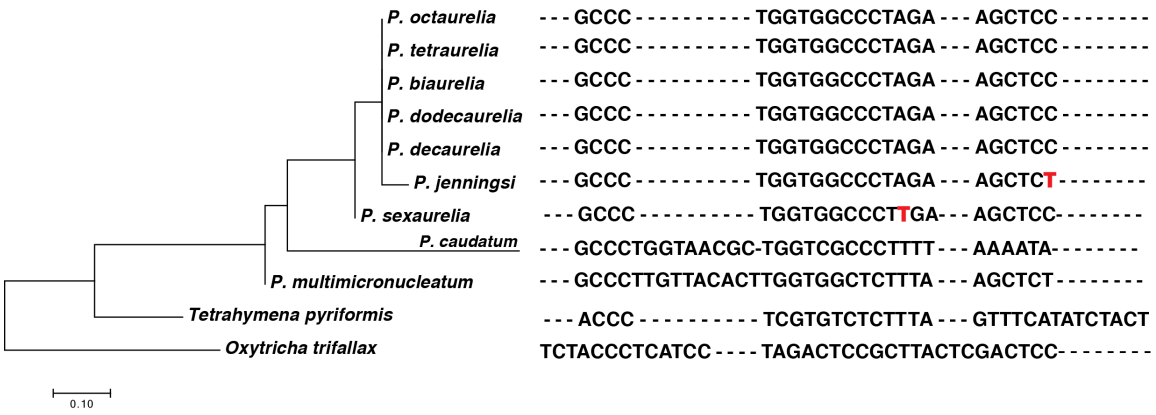


Figure 3: Telomeric repeat sequences in the *Paramecium* genus. Nucleotides in red show the single base pair differences among the *P. aurelia* species. The phylogenetic tree on the left is built using the telomeric repeat sequences.

AA	Anticodon	*Tetrahymena* species	*P. multimicronucleatum* (reference)	*P. multimicronucleatum* (strain: M05)	*P. multimicronucleatum* (strain: M13)	*P. multimicronucleatum-Penlche3I*	*P. caudatum* (reference)	*P. caudatum* (strain: C065)	*P. caudatum* (strain: C083)	*P. caudatum-C026*	*Paramecium aurelia* species
C	GCA				1	1					
E	TTC	1	1			1	1				
F	AAA		3	2	3	2	1	1	3	3	
F	GAA	1	1	1	1	1	1	1	1	1	1
I	AAT				1					1	
I	GAT		2	3							
I	TAT		2	1		1	1		1	1	
K	TTT		4	11	6	5	4	4	2	2	
H	GTG	1									
L	CAA		2	2	1						
L	TAA	2	5	4	4	3	3	3	3	3	
L	TAG				1	1					
Met	CAT	1									
N	ATT					1					
N	GTT				1	1					
Q	TTC		1	1				1			
R	TCT		2	3	1	2					
S	AGA								1		
S	TGA						1	1			
W (SeC)	TCA	1	1	1	2	2	1	1	1	1	1
Sup	CTA		1	1	1		1	1			
Sup	TTA		4	3		6	2	2	2	1	
T	AGT		1	1	1						
T	GTA										
T	TGT		1	2	1					1	
Undet	Undet				1				1		
V	AAC				1	1	2	1	1		
V	TAC		1			1			1		
Y	ATA		1	1		1				1	
Y	GTA	1	1	1	1	1	2	1	1	1	1

Figure 4: tRNA content variation in *P. caudatum* and *P. multimicronucleatum* strains. tRNAs shaded in grey are present in all *Paramecium* and *Tetrahymena* species. tRNAs whose anticodons were ambiguous are displayed as "Undet".

Table 1: Mutation bias towards AT (m) in the mitochondrial genomes of *Paramecium* and other species, where v and u are mutation rates from G/C to A/T, and A/T to G/C respectively. Values of m less than 1 indicate that mutation spectrum is biased towards G/C. $S = 4N_e s$ (or $2N_e s$) for diploids (or haploids) represents the strength of selection favoring A/T composition, where negative values of S represent selection favoring G/C. S is calculated using the equation, $P_{AT} = 1/(1 + m^{-1}e^{-S})$, where P_{AT} is the fraction of AT nucleotides.

Species	**Type of study**	**Total number of base-substitution mutations or SNPs**	**AT mutation bias ($m = v/u$)**	**Expected equilibrium GC content [1/(1+m)]**	**Whole-genome GC content**	**4-fold site GC content**	***S* (avg)**	***S* (4-fold)**
P. tetraurelia[1]	MA	Not known	0.88	0.53	0.42	0.54	0.197	-0.013
P. tetraurelia	population	37	0.95	0.51	0.42	0.54	0.163	-0.047
P. biaurelia[2]	MA	55	0.79	0.56	0.40	0.51	0.278	0.085
P. sexaurelia[2]	MA	87	0.49	0.67	0.39	0.47	0.504	0.368
P. sexaurelia	population	112	1.56	0.39	0.39	0.47	0.000	-0.137
P. caudatum	population	96	5.02	0.17	0.22	0.07	-0.151	0.397
Other species from previous studies:								
S. cerevisiae[3]	MA	13	0.00	1.00	0.17	0.08	>> 0	>> 0
Caenorhabditis Elegans[4,5]	MA	25	3.79	0.21	0.24	0.14	-0.073	0.224
Daphnia pulex[6]	MA	6	2.48	0.29	0.38	0.30	-0.182	-0.026
Caenorhabditis Briggsae[7]	MA	15	18.10	0.05	0.25	0.14	-0.780	-0.451
Drosophila melanogaster[8]	MA	28	381.25	0.00	0.18	0.06	-1.923	-1.386

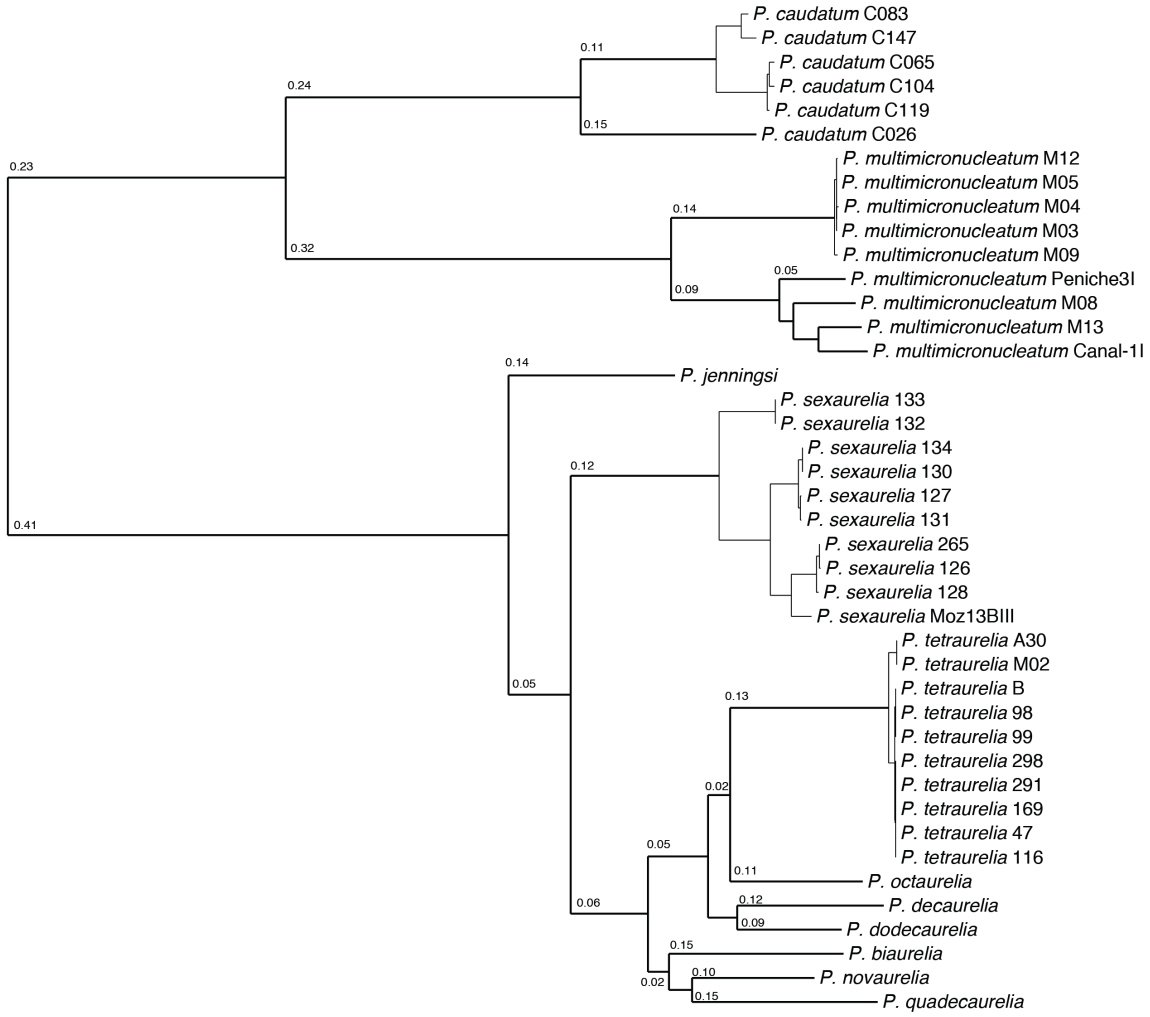
1: Sung et al., 2012; **2** : Long et al., 2018 ; **3**: Lynch et al., 2008; **4:** Denver et al., 2000; **5**: Konrad et al., 2017; **6:** Xu et al., 2012; **7**: Howe et al., 2010; **8:** Haag-Liautard et al., 2008

Table 2: Neutrality index (NI) of mitochondrial and nuclear genes in *P. tetraurelia*, calculated using multiple statistics. All *p* values (corrected for multiple testing by Holm-Bonferroni method) represent comparisons between nuclear and mitochondrial statistics. Statistically significant scores ($p < 0.05$) are underlined.

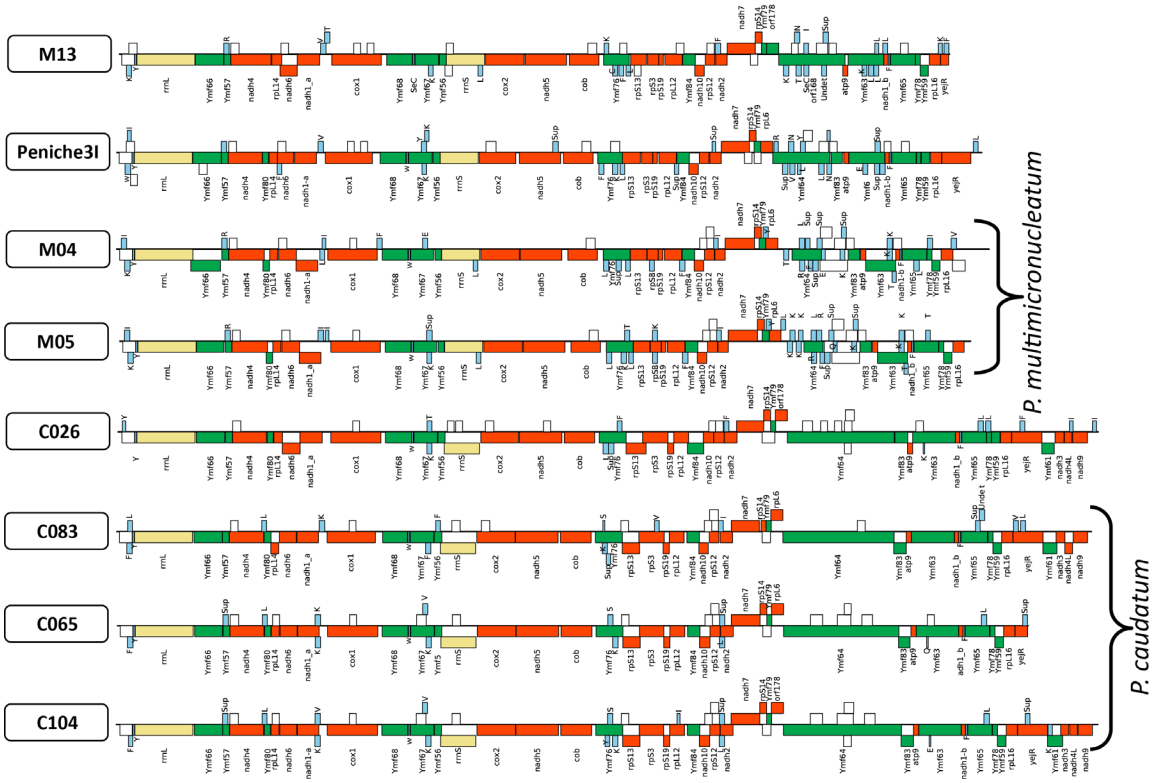
	All (excluding Ymf)	**d*S* < *1.0***	***Oxphos***	***Ribo***
d*S*				
nuclear	0.885	0.636	0.774	0.404
mitochondrial	1.174	0.658	1.324	1.581
p	$\underline{2.74 \times 10^{-6}}$	0.807	<u>0.035</u>	$\underline{4.22 \times 10^{-5}}$
d*N*/d*S*				
nuclear	0.079	0.086	0.029 [18]	0.088
mitochondrial	0.082	0.142	0.052 [15]	0.044
p	0.307	0.210	0.208	0.208
π_N/π_S				
nuclear	0.235	0.235	0.200	0.092
mitochondrial	0.163	0.180	0.127	0.062
p	0.183	0.618	0.223	0.387
NIπ [(π_N/π_S)/(d*N*/d*S*)]				
nuclear	4.018	3.436	6.712 [14]	1.653 [88]
mitochondrial	2.294 (2.495)	1.218	2.384	2.248
p	0.124 (0.537)	<u>0.024</u>	0.422	0.262
NI [(P_N/P_S)/(D_N/D_S)]				
nuclear	2.145	2.271	6.087 [18]	2.873 [80]
mitochondrial	1.528 (1.855)	1.294	1.696 [15]	2.195 [9]
p	0.272 (0.817)	0.208	<u>0.018</u>	0.072
NI_{DoS} [($D_N/(D_N + D_S)$) − ($P_N/(P_N + P_S)$)]				
nuclear	-0.041	-0.053	-0.227	-0.136
mitochondrial	0.026 (0.024)	0.046	0.001	0.073
p	0.272 (0.738)	0.208	<u>0.010</u>	0.072

SUPPLEMENTARY MATERIALS

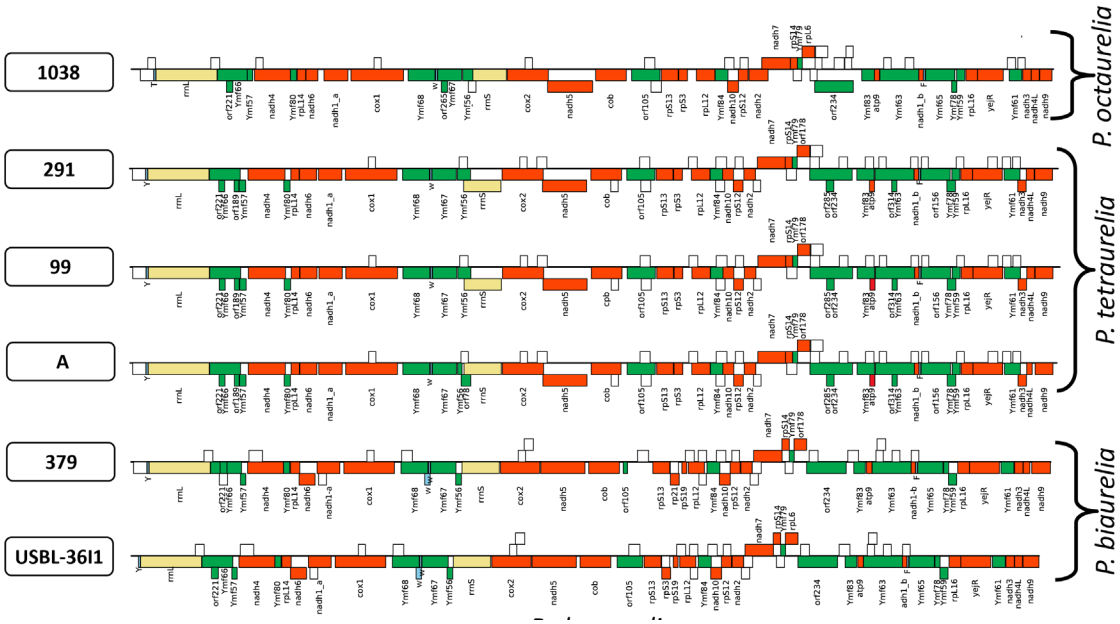
SUPPLEMENTAL FIGURES



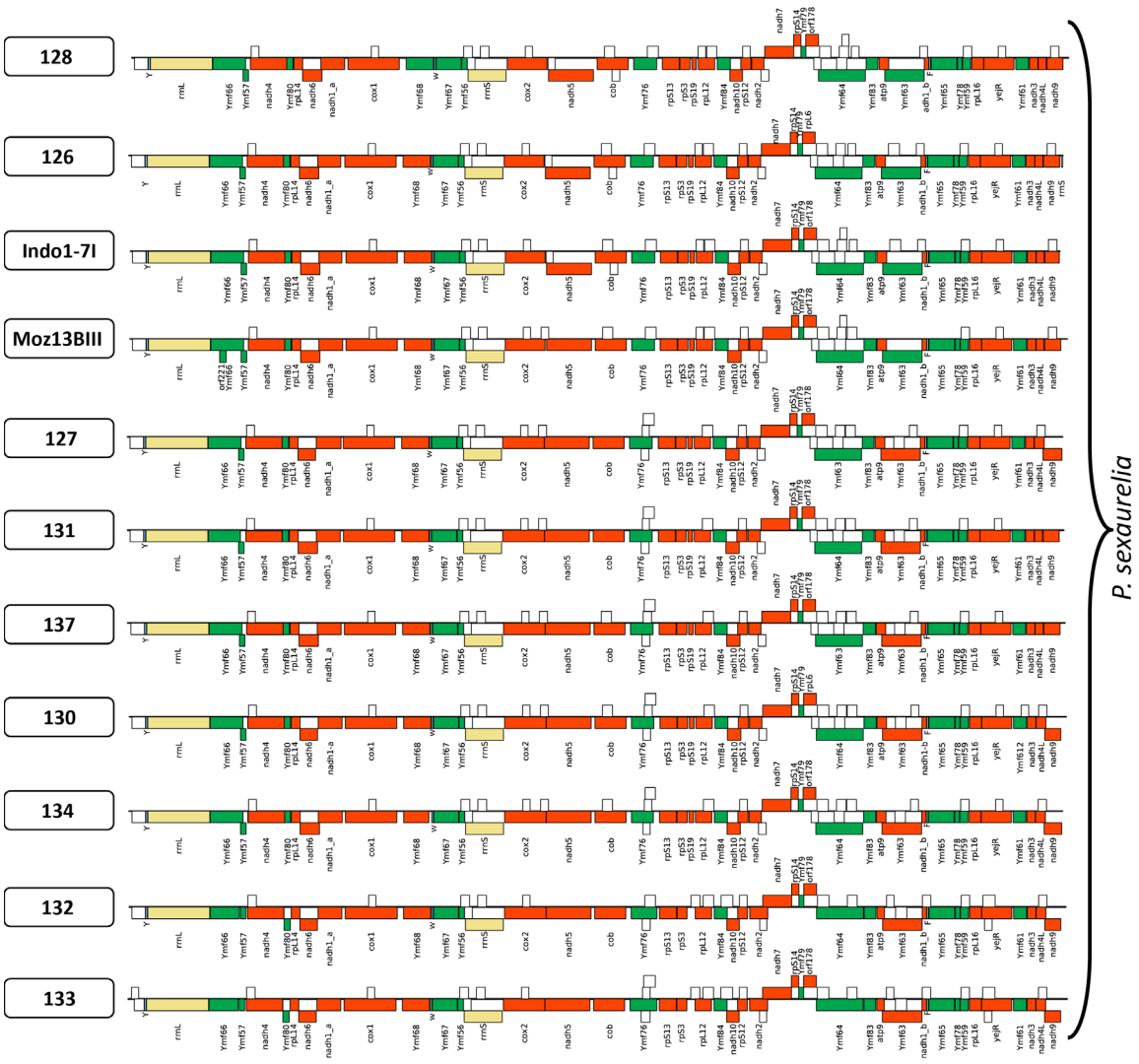
Supplementary Figure 1: Mitochondrial phylogeny of the *Paramecium* genus including individual isolates within each species. The phylogeny was built using RAxML (version 8.2.11), under the substitution model GTRGAMMA. Numbers on branches indicate total number of substitutions per site.



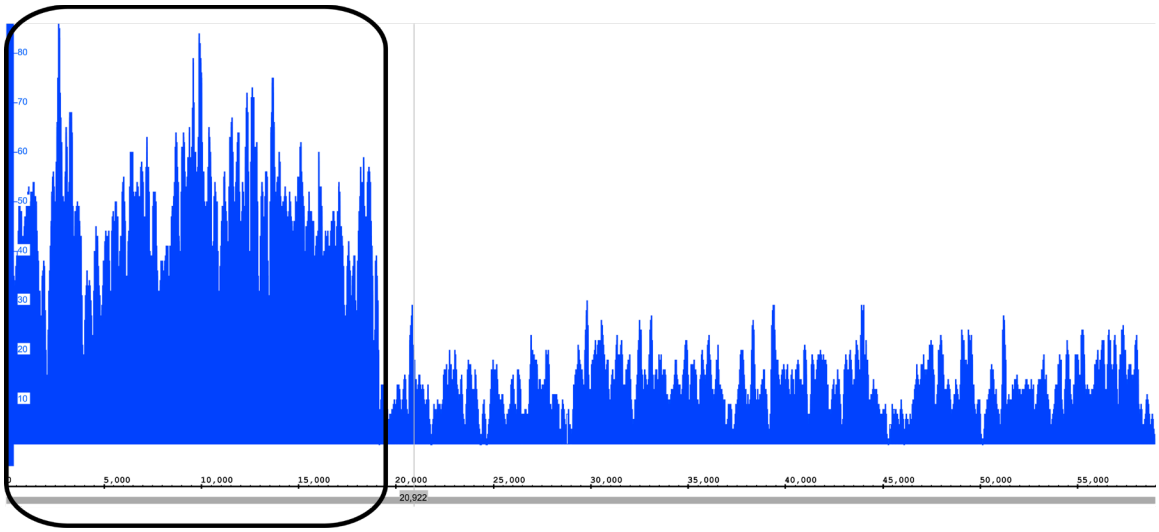
Supplementary Figure 2: Mitochondrial genomes of isolates belonging to the *P. caudatum* and *P. multimicronucleatum* outgroup lineages. Red: protein coding genes; white: novel predicted ORFs longer than 100aa; yellow: ribosomal RNAs; blue: tRNAs. Top layers: forward-strand genes; bottom layers: reverse-strand genes.



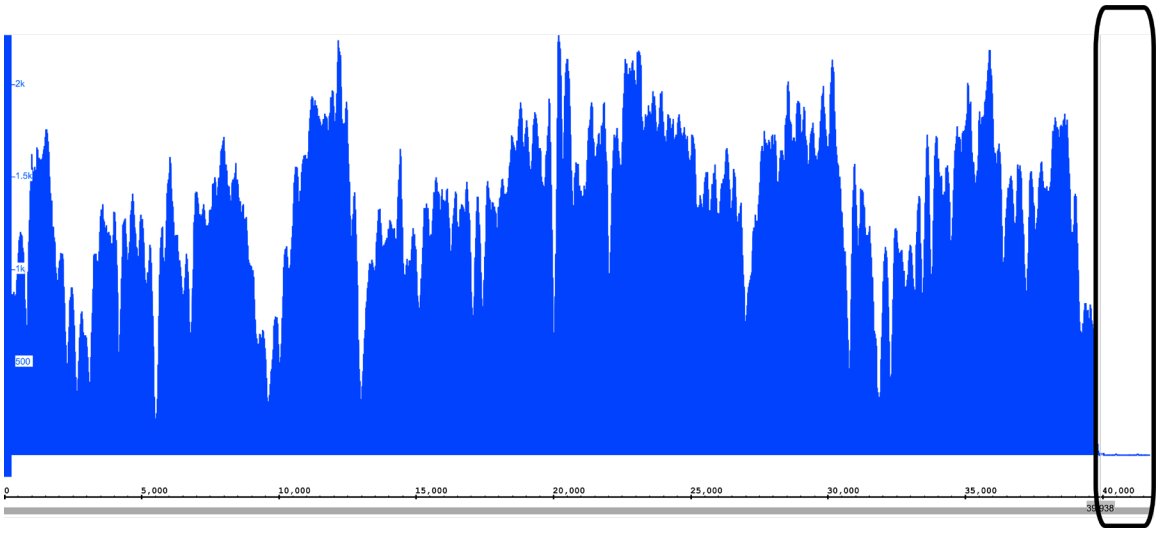
Supplementary Figure 3: Mitochondrial genomes of isolates belonging to the *P. aurelia* species complexes (except for *P. sexaurelia* , shown in the next figure). Red: protein coding genes; white: novel predicted ORFs longer than 100aa; yellow: ribosomal RNAs; blue: tRNAs. Top layers: forward-strand genes; bottom layers: reverse-strand genes.



Supplementary Figure 4: Mitochondrial genomes of isolates belonging to *P. sexaurelia*. Red: protein coding genes; white: novel predicted ORFs longer than 100aa; yellow: ribosomal RNAs; blue: tRNAs. Top layers: forward-strand genes; bottom layers: reverse-strand genes.



Supplementary Figure 5: **Raw mitogenome assembly and sequencing read coverage for *P. novaurelia*.** An ~19kb 3′ extension is present in the raw assembly. However, read coverage over that region is significantly higher than what is observed over the rest of the mitochondrial genome, thus the two regions most likely exist in different copy numbers and the extension might be the result of missassembly.



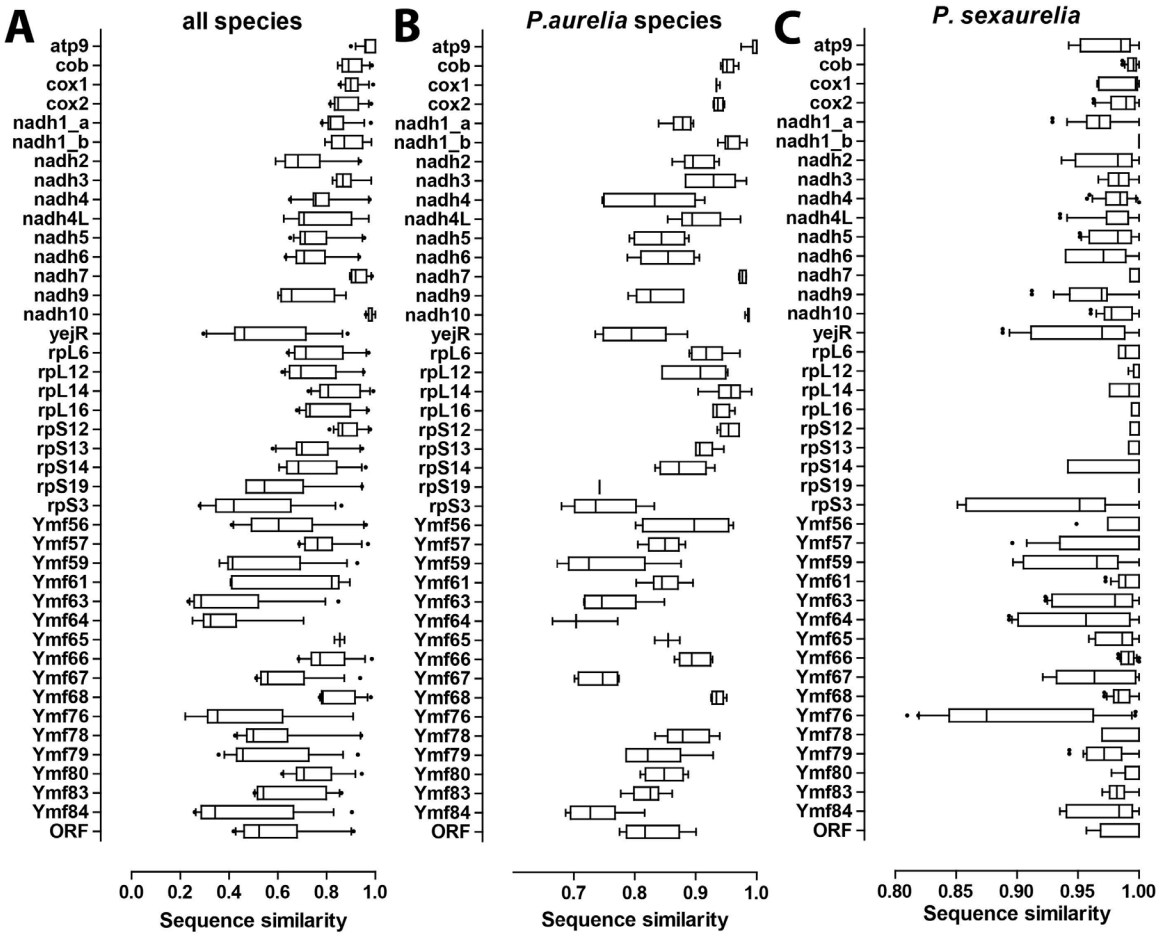
Supplementary Figure 6: Raw mitogenome assembly and sequencing read coverage for *P. quadecaurelia*. A small 5′ extension is present in the raw assembly. However, read coverage over that region is very low relative to what is observed over the rest of the mitochondrial genome, thus the extension is probably be the result of missassembly.

Protein			*P. multimicronucleatum*			*P. caudatum*			*P. biaurelia*		*P. tetraurelia*			*P. sexaurelia*																
	M13	Peniche3I	M04	M05	C026	C065	C083	C104	379	USBL-36I1	291	99	A	126	127	128	130	131	132	133	134	137	Indo1-7I	Moz13BIII	*P. octaurelia*	*P. novaurelia*	*P. decaurelia*	*P. dodecaurelia*	*P. quadecaurelia*	*P. jenningsi*
atp9	75	75	80	80	78	77	77	77	78	78	72	72	72	136	136	136	136	136	104	104	136	136	136	136	72	136	75	86	80	85
cob	444	444	444	444	457	466	463	466	442	442	442	442	430	442	442	442	442	442	442	442	442	442	442	442	442	442	442	442	442	442
cox1	747	709	746	747	757	757	728	757	729	729	746	746	746	729	729	729	729	729	739	739	729	729	729	729	813	729	733	748	728	711
cox2	579	550	582	582	567	584	580	584	567	567	594	594	594	567	591	567	591	591	591	591	591	591	567	567	592	591	567	567	613	567
nadh10	141	141	141	141	157	157	141	157	156	156	156	156	156	178	189	178	189	189	177	177	189	189	178	189	156	156	156	156	156	156
nadh1-a	338	327	325	325	336	323	323	323	328	328	336	336	336	327	352	339	339	339	338	338	339	339	327	339	345	307	330	310	303	326
nadh1-b	63	63	61	61	64	63	63	63	64	64	64	64	64	63	63	63	63	63	63	63	63	63	63	63	63	64	64	64	63	63
nadh2	170	170	154	154	191	186	186	191	162	162	166	166	166	173	173	173	173	173	257	257	173	173	173	173	276	162	162	162	205	367
nadh3	nd	nd	nd	nd	134	nd	132	120	120	120	120	120	120	120	120	120	120	120	120	120	120	120	120	120	120	120	145	120	120	120
nadh4	607	512	536	536	512	512	518	512	517	514	534	534	534	514	517	517	517	517	517	517	517	517	517	517	517	517	517	517	512	512
nadh4L	nd	nd	nd	nd	115	nd	115	115	96	113	113	113	78	113	139	113	139	113	139	139	139	113	113	113	113	113	113	113	122	113
nadh5	622	622	627	627	641	641	641	641	638	638	637	637	637	638	638	638	638	638	638	638	638	638	638	638	648	638	638	638	638	638
nadh6	259	243	266	266	267	318	256	256	230	230	245	245	245	275	249	275	249	275	249	249	275	275	275	275	170	254	238	263	221	262
nadh7	423	423	442	442	411	433	423	433	411	411	400	400	400	411	411	411	411	411	411	411	411	411	411	411	411	411	411	400	411	411
nadh9	nd	nd	nd	nd	223	nd	215	215	270	229	256	256	256	229	269	229	229	229	229	229	249	229	205	229	175	229	218	229	218	229
rpL12	262	262	262	262	223	223	223	223	232	232	265	265	265	232	232	232	232	232	232	232	232	232	232	232	265	232	265	265	356	232
rpL14	150	125	119	119	128	119	119	119	134	134	119	119	119	125	125	125	125	125	125	125	125	125	125	125	119	119	125	128	119	125
rpL16	168	168	178	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168
orf178/rpL6	181	181	182	182	181	181	181	181	181	181	181	181	181	181	181	181	181	181	181	181	181	181	181	181	181	181	181	181	181	181
rpS12	144	144	149	149	156	153	143	153	143	143	139	139	139	143	143	143	143	143	139	139	143	143	143	143	139	174	143	139	157	143
rpS13	237	271	249	249	302	261	261	261	244	244	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
rpS14	103	103	103	103	103	103	76	103	105	105	102	102	102	103	103	103	103	103	103	103	103	103	103	103	102	105	102	183	115	105
rpS19	68	68	68	68	92	92	92	92	69	69	-	-	-	66	66	66	66	66	-	-	66	66	66	66	-	-	66	-	-	66
rpS3	166	166	165	165	381	368	382	368	127	127	125	125	125	144	144	144	144	144	141	141	144	144	144	144	125	127	146	139	127	127
yejR	120	439	nd	nd	453	190	452	452	432	432	429	429	429	429	429	429	429	429	429	429	429	429	429	429	370	429	370	370	433	429
Ymf56	105	107	94	94	106	106	114	106	82	82	189	189	63	78	78	78	78	78	78	78	78	78	78	78	146	150	103	146	78	82
Ymf57	100	100	100	100	100	100	100	100	76	76	93	93	93	77	76	77	77	76	77	77	77	77	77	80	77	80	80	80	80	80
Ymf59	122	122	122	122	128	128	128	128	116	116	113	113	113	116	116	116	116	116	116	116	116	116	116	116	116	116	116	123	116	116
Ymf61	nd	nd	nd	nd	186	nd	176	227	185	185	228	228	228	182	182	182	182	182	185	185	182	182	182	182	182	182	182	185	185	179
Ymf63	538	540	450	450	642	641	647	641	561	561	563	563	563	561	561	561	561	561	561	561	561	561	561	561	561	561	561	559	559	551
Ymf64/orf234	1,000	863	442	288	1,614	1,742	1,679	1,783	572	572	613	613	613	666	666	666	666	666	666	666	666	666	666	666	557	572	668	613	619	671
Ymf65/orf156	374	365	370	370	371	366	397	367	365	365	423	423	423	365	365	365	365	365	365	365	365	365	365	365	422	365	385	434	365	365
Ymf66	428	428	447	409	427	419	427	427	445	445	445	445	445	466	466	466	466	456	413	413	466	466	466	466	429	413	428	428	398	398
Ymf67	354	354	344	344	354	354	354	354	366	378	348	348	348	355	355	355	355	355	355	355	355	355	355	355	345	345	402	345	362	355
Ymf68	387	387	387	387	395	389	389	389	387	387	387	387	387	387	387	387	387	387	387	387	357	387	387	387	387	387	390	387	387	387
Ymf76/orf105	385	385	405	405	405	405	405	405	61	369	405	405	405	324	321	327	317	321	366	380	310	320	326	324	410	372	453	409	379	441
Ymf78	72	63	73	73	69	73	73	73	72	72	122	122	122	66	66	66	66	66	66	66	66	66	66	66	72	72	72	70	66	66
Ymf79	78	99	104	104	75	75	75	75	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70	70
Ymf80	-	89	96	96	90	92	104	90	89	89	89	89	89	89	89	89	89	89	89	89	89	89	89	89	89	89	89	89	89	89
Ymf83	-	167	172	172	192	172	188	187	166	166	250	250	250	166	166	166	175	166	166	166	175	175	166	175	183	166	233	166	166	250
Ymf84	188	188	188	188	245	186	186	186	185	185	185	185	185	185	185	185	185	185	185	185	185	185	185	185	185	185	185	185	186	188
orf189	-	-	-	-	-	-	-	-	-	-	67	67	67	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
orf221	-	-	-	-	-	-	-	-	97	97	84	84	84	-	-	-	-	-	-	-	-	-	-	92	90	-	129	116	90	-
orf265	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	88	-	-	-	-	-
orf285	-	-	-	-	-	-	-	-	-	-	108	108	108	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
orf314	-	-	-	-	-	-	-	-	-	-	73	73	73	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
orf78	-	-	-	-	-	-	-	-	-	-	-	-	138	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
newORF	191	191	192	192	191	191	191	186	191	-	191	191	191	191	191	191	191	191	191	132	191	191	191	191	191	191	191	191	191	-

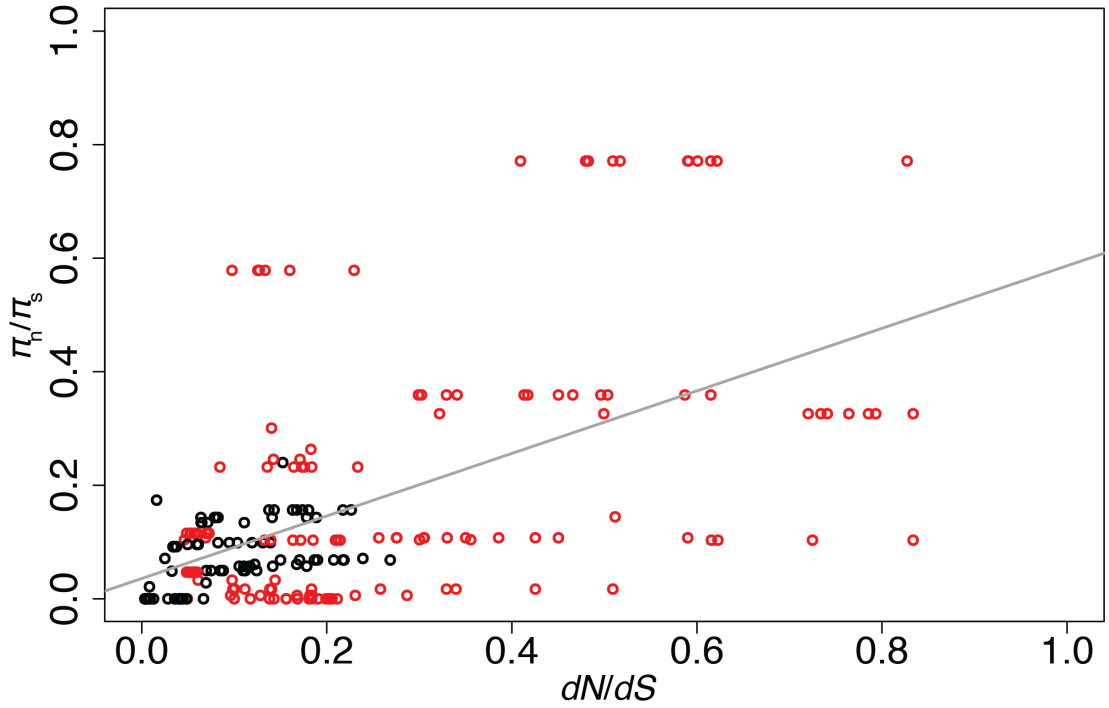
Supplementary Figure 7: Presence/absence of proteins encoded in mitochondrial genomes in species and individual isolates of species of the *Paramecium* genus. The numbers indicate the length of each protein in amino acids. Most other ORFs exhibit relatively constant length across all *Paramecium* mitochondrial genomes, with the exceptions of *atp9*, which is elongated by ~30-60 AA in all *P. sexaurelia* isolates, *rpS3* and *rps19*, which are specifically elongated in *P. caudatum* and *P. caudatum-C026*, and *Ymf76*, which is very short in one of the *P. biaurelia* isolates. *Ymf64* accounts for a significant part of the difference between the length of *P. aurelia* and *P. caudatum* mitochondrial genomes, being much longer in the latter (~1700 AA vs. ~600 AA). In two isolates of *P. multimicronucleatum* (M04 and M05), the locus appears to have split into multiple ORFs, with the one most clearly homologous to *Ymf64* being shortened even relative to *P. aurelia*. However, in the *P. multimicronucleatum* isolate M13, as well as *P. multimicronucleatum-Peniche3I*, *Ymf64* is relatively long (1000 and 863 AA respectively).

mean_dS	*bi*	*tet*	*sex*	*oct*	*nov*	*dec*	*dodec*	*quadec*	*jenningsi*	*caudatum*	*caudatum* C026	*multimicronucleatum*
bi	-	-	-	-	-	-	-	-	-	-	-	-
tet	1.22	-	-	-	-	-	-	-	-	-	-	-
sex	1.31	1.39	-	-	-	-	-	-	-	-	-	-
oct	1.17	0.93	1.42	-	-	-	-	-	-	-	-	-
nov	1.07	1.18	1.41	1.20	-	-	-	-	-	-	-	-
dec	1.32	0.99	1.27	1.12	1.11	-	-	-	-	-	-	-
dodec	1.20	0.94	1.30	1.04	1.23	0.85	-	-	-	-	-	-
quadec	1.19	1.27	1.29	1.20	1.06	1.24	1.21	-	-	-	-	-
jenningsi	1.42	1.38	1.37	1.41	1.26	1.31	1.27	1.28	-	-	-	-
caudatum	2.30	2.56	2.45	2.39	2.77	2.56	2.44	2.62	2.61	-	-	-
caudatum C026	4.86	2.48	2.37	2.53	2.48	2.46	5.05	2.48	2.45	1.36	-	-
multimicronucleatum	2.75	2.69	2.80	2.71	2.85	2.77	2.73	2.71	2.76	7.69	5.12	-
multimicronucleatum Peniche3I	2.67	2.88	2.79	2.57	2.84	2.86	2.69	2.76	2.79	4.95	2.45	1.97

Supplementary Figure 8: Mean pairwise *dS* between all *Paramecium* species. Species pairs with *dS* < 1.0 are shown in shaded boxes.



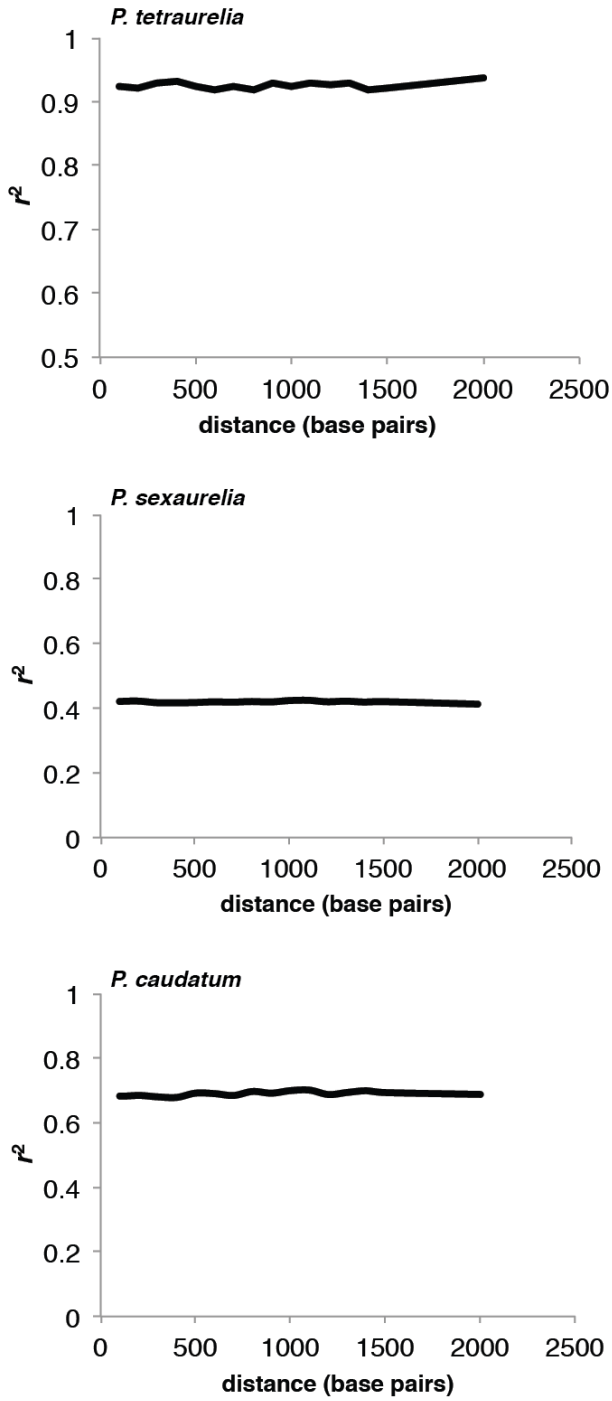
Supplementary Figure 9: **Conservation and divergence of mitochondrial protein sequences in the *Paramecium* genus.** Shown is the percent sequence identity for each protein within all species included in this study (A), within the *P. aurelia* species complex (B), and between the 11 individual *P. sexaurelia* isolates (C). Only one isolate per species was included in (A) and (B). The whiskers of the box plot correspond to the 5-95 percentile range.



Supplementary Figure 10: **Relationship between observed π_n/π_s and pairwise *dN*/*dS* with respect to the closest outgroup species.** This analysis was performed for all protein-coding genes, with the conserved genes in black and the *Ymf* genes indicated in red, for *P. tetraurelia*, *P. sexaurelia*, *P. caudatum* and *P. multimicronucleatum*. Only genes with $dS < 1.0$ and $\theta_s > 1$, where θ_s is the total number of polymorphic sites at synonymous positions, were used in this analysis, resulting in 208 total pairs of observations. This filter was executed in order to reduce errors in π_n/π_s due to very low values of synonymous polymorphisms in a gene.

		P. multimicronucleatum					*P. caudatum*			*P. octaurelia*	*P. tetraurelia*			*P. biaurelia*		*P. sexaurelia*											*P. decaurelia*	*P. dodecaurelia*	*P. jenningsi*	*P. novaurelia*	*P. quadecaurelia*
AA	**codon**	M13	Peniche3I	M04	M05	C026	C083	C065	C104	1038	291	99	A	379	USBL-36I1	126	128	Indo1-7I	Moz13BIII	127	131	137	130	134	132	133	223	274	M	TE	N1A
A	GCA	0.45	0.45	0.44	0.44	0.41	0.41	0.41	0.39	0.21	0.21	0.21	0.21	0.20	0.21	0.21	0.21	0.22	0.21	0.21	0.21	0.21	0.20	0.21	0.22	0.22	0.20	0.26	0.21	0.22	0.19
	GCC	0.02	0.02	0.01	0.01	0.03	0.03	0.03	0.04	0.38	0.38	0.38	0.38	0.38	0.38	0.36	0.35	0.36	0.35	0.37	0.38	0.38	0.37	0.37	0.37	0.37	0.37	0.36	0.36	0.34	0.34
	GCG	0.01	0.01	0.01	0.01	0.02	0.03	0.03	0.02	0.10	0.10	0.10	0.10	0.09	0.09	0.04	0.04	0.04	0.05	0.06	0.05	0.06	0.06	0.06	0.06	0.06	0.12	0.10	0.07	0.07	0.15
	GCU	0.52	0.52	0.53	0.53	0.55	0.54	0.53	0.54	0.31	0.31	0.31	0.32	0.32	0.32	0.39	0.39	0.39	0.39	0.36	0.36	0.36	0.37	0.36	0.35	0.35	0.31	0.29	0.35	0.36	0.32
C	UGC	0.18	0.21	0.20	0.18	0.20	0.19	0.19	0.17	0.67	0.68	0.68	0.68	0.60	0.60	0.52	0.52	0.52	0.53	0.56	0.55	0.57	0.57	0.57	0.63	0.64	0.59	0.72	0.65	0.59	0.57
	UGU	0.82	0.79	0.80	0.82	0.80	0.81	0.81	0.83	0.33	0.32	0.32	0.32	0.40	0.40	0.48	0.48	0.48	0.47	0.44	0.45	0.43	0.43	0.43	0.37	0.36	0.41	0.28	0.35	0.41	0.43
D	GAC	0.13	0.13	0.12	0.11	0.17	0.15	0.20	0.17	0.62	0.63	0.63	0.63	0.56	0.55	0.55	0.56	0.55	0.53	0.60	0.59	0.58	0.58	0.58	0.63	0.64	0.55	0.61	0.58	0.61	0.54
	GAU	0.87	0.87	0.88	0.89	0.83	0.85	0.80	0.83	0.38	0.37	0.37	0.37	0.44	0.45	0.45	0.44	0.45	0.47	0.40	0.41	0.42	0.42	0.42	0.37	0.36	0.45	0.39	0.42	0.39	0.46
E	GAA	0.80	0.86	0.93	0.93	0.79	0.85	0.83	0.82	0.39	0.36	0.37	0.37	0.36	0.35	0.39	0.39	0.39	0.40	0.37	0.37	0.38	0.39	0.38	0.37	0.36	0.35	0.30	0.37	0.31	0.37
	GAG	0.20	0.14	0.07	0.07	0.21	0.15	0.17	0.18	0.61	0.64	0.63	0.63	0.64	0.65	0.61	0.61	0.61	0.60	0.63	0.63	0.62	0.61	0.62	0.63	0.64	0.65	0.70	0.63	0.69	0.63
F	UUC	0.05	0.04	0.03	0.03	0.06	0.05	0.06	0.05	0.34	0.36	0.36	0.36	0.34	0.35	0.33	0.33	0.33	0.33	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.32	0.32	0.31
	UUU	0.95	0.96	0.97	0.97	0.94	0.95	0.94	0.95	0.66	0.64	0.64	0.64	0.66	0.65	0.67	0.67	0.67	0.67	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.68	0.68	0.69
G	GGA	0.32	0.33	0.38	0.38	0.21	0.23	0.20	0.21	0.22	0.20	0.20	0.20	0.20	0.21	0.21	0.22	0.21	0.21	0.20	0.20	0.20	0.20	0.20	0.21	0.21	0.20	0.21	0.22	0.19	0.20
	GGC	0.02	0.04	0.02	0.02	0.03	0.03	0.05	0.07	0.36	0.36	0.36	0.36	0.32	0.33	0.31	0.30	0.31	0.30	0.32	0.32	0.32	0.32	0.32	0.30	0.31	0.30	0.33	0.34	0.32	0.33
	GGG	0.02	0.02	0.01	0.00	0.03	0.02	0.05	0.04	0.20	0.27	0.27	0.27	0.26	0.25	0.20	0.19	0.20	0.20	0.21	0.21	0.21	0.21	0.21	0.23	0.23	0.29	0.24	0.20	0.27	0.25
	GGU	0.64	0.61	0.59	0.59	0.72	0.72	0.71	0.69	0.21	0.17	0.17	0.17	0.22	0.21	0.28	0.29	0.28	0.29	0.28	0.28	0.27	0.27	0.27	0.26	0.26	0.21	0.21	0.24	0.23	0.23
H	CAC	0.24	0.17	0.12	0.12	0.33	0.25	0.28	0.28	0.60	0.61	0.61	0.61	0.57	0.57	0.59	0.60	0.59	0.63	0.66	0.66	0.65	0.64	0.65	0.64	0.63	0.63	0.63	0.68	0.61	0.60
	CAU	0.76	0.83	0.88	0.88	0.67	0.75	0.72	0.72	0.40	0.39	0.39	0.39	0.43	0.43	0.41	0.40	0.41	0.37	0.34	0.34	0.35	0.36	0.35	0.36	0.37	0.37	0.38	0.32	0.39	0.40
I	AUA	0.39	0.39	0.40	0.39	0.41	0.42	0.43	0.43	0.25	0.25	0.25	0.24	0.26	0.25	0.24	0.24	0.24	0.24	0.23	0.23	0.24	0.24	0.24	0.25	0.26	0.23	0.23	0.26	0.29	0.25
	AUC	0.03	0.03	0.02	0.02	0.06	0.04	0.05	0.07	0.39	0.41	0.41	0.41	0.39	0.39	0.38	0.37	0.38	0.38	0.40	0.40	0.39	0.39	0.39	0.41	0.41	0.40	0.40	0.42	0.37	0.36
	AUU	0.58	0.59	0.58	0.59	0.52	0.54	0.52	0.50	0.36	0.34	0.34	0.35	0.35	0.35	0.39	0.39	0.39	0.38	0.37	0.37	0.37	0.37	0.37	0.34	0.34	0.38	0.37	0.33	0.34	0.39
K	AAA	0.90	0.92	0.96	0.96	0.88	0.91	0.88	0.89	0.47	0.45	0.45	0.45	0.43	0.43	0.49	0.49	0.48	0.50	0.47	0.48	0.48	0.48	0.48	0.44	0.45	0.45	0.43	0.42	0.40	0.43
	AAG	0.10	0.08	0.04	0.04	0.12	0.09	0.12	0.11	0.53	0.55	0.55	0.55	0.57	0.57	0.51	0.51	0.52	0.50	0.53	0.52	0.52	0.52	0.52	0.56	0.55	0.55	0.57	0.58	0.60	0.57
L	CUA	0.08	0.06	0.05	0.05	0.06	0.05	0.08	0.07	0.13	0.13	0.13	0.13	0.14	0.14	0.11	0.11	0.11	0.10	0.11	0.11	0.11	0.10	0.11	0.12	0.11	0.15	0.15	0.11	0.13	0.10
	CUC	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.26	0.27	0.27	0.27	0.25	0.25	0.24	0.23	0.24	0.24	0.26	0.26	0.26	0.26	0.26	0.27	0.28	0.26	0.27	0.25	0.26	0.24
	CUG	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.07	0.09	0.09	0.08	0.07	0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.09	0.08	0.07	0.07	0.10
	CUU	0.14	0.14	0.13	0.13	0.13	0.12	0.12	0.13	0.21	0.21	0.21	0.22	0.22	0.22	0.27	0.27	0.28	0.25	0.25	0.25	0.25	0.25	0.25	0.23	0.23	0.20	0.20	0.26	0.23	0.25
	UUA	0.69	0.70	0.75	0.75	0.69	0.73	0.68	0.68	0.18	0.16	0.16	0.16	0.20	0.19	0.18	0.18	0.18	0.19	0.18	0.18	0.18	0.18	0.18	0.18	0.17	0.17	0.17	0.15	0.18	0.15
	UUG	0.07	0.08	0.05	0.05	0.09	0.08	0.09	0.10	0.14	0.14	0.14	0.14	0.13	0.12	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.15	0.15	0.14	0.13	0.16	0.13	0.16
M	AUG	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
N	AAC	0.20	0.21	0.14	0.14	0.25	0.20	0.23	0.24	0.56	0.63	0.63	0.63	0.53	0.54	0.57	0.57	0.57	0.57	0.57	0.57	0.55	0.55	0.56	0.63	0.63	0.55	0.54	0.67	0.59	0.62
	AAU	0.80	0.79	0.86	0.86	0.75	0.80	0.77	0.76	0.44	0.37	0.37	0.37	0.47	0.46	0.43	0.43	0.43	0.43	0.43	0.43	0.45	0.45	0.44	0.37	0.37	0.45	0.46	0.33	0.41	0.38
P	CCA	0.49	0.50	0.48	0.48	0.44	0.48	0.43	0.42	0.19	0.15	0.15	0.15	0.17	0.17	0.16	0.16	0.16	0.16	0.15	0.16	0.16	0.16	0.17	0.15	0.15	0.15	0.18	0.17	0.22	0.18
	CCC	0.01	0.01	0.01	0.01	0.03	0.02	0.05	0.05	0.45	0.45	0.45	0.45	0.42	0.42	0.43	0.41	0.43	0.41	0.45	0.44	0.44	0.43	0.43	0.40	0.40	0.43	0.50	0.45	0.43	0.36
	CCG	0.02	0.02	0.01	0.01	0.04	0.03	0.03	0.03	0.04	0.09	0.09	0.09	0.07	0.07	0.07	0.08	0.07	0.09	0.08	0.08	0.08	0.08	0.08	0.08	0.09	0.12	0.06	0.06	0.05	0.17
	CCU	0.48	0.48	0.50	0.50	0.50	0.48	0.48	0.50	0.32	0.30	0.31	0.31	0.33	0.33	0.34	0.35	0.34	0.34	0.32	0.33	0.32	0.33	0.32	0.37	0.36	0.30	0.26	0.32	0.29	0.29
Q	CAA	0.91	0.91	0.93	0.93	0.84	0.88	0.85	0.85	0.57	0.46	0.46	0.46	0.42	0.42	0.54	0.53	0.54	0.53	0.49	0.49	0.50	0.50	0.50	0.56	0.56	0.47	0.47	0.48	0.45	0.40
	CAG	0.09	0.09	0.07	0.07	0.16	0.12	0.16	0.15	0.43	0.54	0.54	0.54	0.58	0.58	0.46	0.47	0.46	0.47	0.51	0.51	0.50	0.50	0.50	0.44	0.44	0.53	0.53	0.52	0.55	0.60
R	AGA	0.88	0.86	0.87	0.87	0.73	0.79	0.73	0.76	0.19	0.18	0.18	0.18	0.22	0.22	0.24	0.24	0.24	0.21	0.20	0.20	0.21	0.21	0.21	0.22	0.22	0.20	0.18	0.22	0.20	0.14
	AGG	0.06	0.08	0.04	0.04	0.16	0.10	0.15	0.14	0.36	0.38	0.38	0.38	0.43	0.43	0.46	0.46	0.46	0.50	0.53	0.53	0.52	0.52	0.52	0.49	0.50	0.41	0.40	0.47	0.51	0.37
	CGA	0.02	0.02	0.03	0.04	0.02	0.03	0.03	0.02	0.13	0.13	0.13	0.14	0.08	0.08	0.09	0.10	0.10	0.09	0.09	0.09	0.09	0.09	0.09	0.08	0.07	0.12	0.12	0.09	0.08	0.09
	CGC	0.01	0.01	0.00	0.00	0.02	0.02	0.01	0.01	0.16	0.15	0.15	0.15	0.14	0.14	0.11	0.11	0.11	0.11	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.13	0.14	0.10	0.12	0.18
	CGG	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.08	0.07	0.07	0.07	0.05	0.05	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.07	0.08	0.03	0.04	0.12
	CGU	0.03	0.03	0.06	0.06	0.07	0.05	0.07	0.06	0.08	0.08	0.08	0.08	0.08	0.07	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.06	0.06	0.07	0.07	0.07	0.07	0.09	0.06	0.10
S	AGC	0.07	0.07	0.04	0.04	0.07	0.05	0.08	0.08	0.24	0.24	0.24	0.24	0.26	0.26	0.24	0.25	0.24	0.24	0.24	0.24	0.25	0.25	0.25	0.25	0.25	0.23	0.25	0.24	0.27	0.24
	AGU	0.25	0.28	0.24	0.24	0.24	0.27	0.21	0.21	0.08	0.08	0.08	0.09	0.08	0.08	0.10	0.09	0.10	0.09	0.10	0.10	0.10	0.10	0.10	0.08	0.08	0.09	0.07	0.09	0.09	0.09
	UCA	0.23	0.23	0.28	0.28	0.21	0.22	0.24	0.24	0.12	0.11	0.11	0.11	0.12	0.11	0.13	0.14	0.13	0.13	0.12	0.12	0.12	0.12	0.12	0.11	0.11	0.09	0.09	0.10	0.11	0.11
	UCC	0.01	0.01	0.01	0.01	0.02	0.01	0.03	0.03	0.18	0.19	0.19	0.19	0.18	0.18	0.16	0.16	0.16	0.17	0.17	0.17	0.17	0.17	0.17	0.19	0.19	0.21	0.21	0.18	0.18	0.18
	UCG	0.01	0.01	0.00	0.00	0.01	0.02	0.03	0.03	0.12	0.14	0.14	0.14	0.12	0.13	0.12	0.12	0.12	0.11	0.13	0.12	0.12	0.12	0.12	0.13	0.14	0.15	0.14	0.13	0.10	0.12
	UCU	0.42	0.41	0.42	0.42	0.44	0.43	0.42	0.42	0.25	0.24	0.24	0.24	0.25	0.25	0.24	0.24	0.25	0.25	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.23	0.23	0.26	0.25	0.25
T	ACA	0.51	0.49	0.48	0.49	0.50	0.46	0.46	0.47	0.23	0.20	0.20	0.19	0.22	0.21	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.19	0.19	0.19	0.19	0.19	0.22	0.20
	ACC	0.02	0.03	0.02	0.02	0.05	0.03	0.05	0.04	0.32	0.35	0.35	0.35	0.34	0.35	0.36	0.36	0.36	0.35	0.33	0.33	0.33	0.33	0.34	0.33	0.33	0.35	0.35	0.38	0.38	0.37
	ACG	0.02	0.02	0.02	0.02	0.01	0.01	0.04	0.03	0.23	0.23	0.23	0.23	0.20	0.20	0.17	0.18	0.17	0.18	0.20	0.20	0.20	0.19	0.19	0.23	0.23	0.23	0.24	0.21	0.17	0.20
	ACU	0.45	0.47	0.48	0.48	0.44	0.49	0.46	0.45	0.22	0.23	0.23	0.23	0.25	0.25	0.23	0.23	0.23	0.23	0.23	0.24	0.24	0.23	0.24	0.24	0.24	0.24	0.22	0.22	0.23	0.23
V	GUA	0.37	0.37	0.40	0.40	0.35	0.37	0.37	0.36	0.25	0.22	0.22	0.22	0.25	0.25	0.25	0.25	0.25	0.24	0.23	0.23	0.23	0.23	0.23	0.24	0.24	0.23	0.21	0.22	0.24	0.22
	GUC	0.03	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.24	0.26	0.27	0.27	0.22	0.22	0.25	0.25	0.26	0.25	0.27	0.27	0.27	0.27	0.27	0.26	0.27	0.26	0.26	0.26	0.25	0.23
	GUG	0.04	0.05	0.04	0.04	0.05	0.04	0.07	0.07	0.18	0.19	0.19	0.18	0.15	0.16	0.14	0.15	0.14	0.15	0.16	0.16	0.16	0.16	0.16	0.15	0.15	0.19	0.21	0.19	0.16	0.17
	GUU	0.57	0.57	0.54	0.55	0.58	0.57	0.53	0.54	0.32	0.33	0.33	0.33	0.37	0.37	0.36	0.36	0.36	0.36	0.34	0.34	0.34	0.34	0.34	0.35	0.34	0.32	0.32	0.33	0.35	0.38
W	UGA	0.96	0.95	0.98	0.98	0.95	0.98	0.96	0.97	0.57	0.49	0.49	0.49	0.58	0.58	0.52	0.51	0.53	0.55	0.51	0.50	0.51	0.51	0.51	0.52	0.51	0.52	0.54	0.47	0.45	0.43
	UGG	0.04	0.05	0.02	0.02	0.05	0.02	0.04	0.03	0.43	0.51	0.51	0.51	0.42	0.42	0.48	0.49	0.47	0.45	0.49	0.50	0.49	0.49	0.49	0.48	0.49	0.48	0.46	0.53	0.55	0.57
Y	UAC	0.19	0.18	0.12	0.12	0.22	0.15	0.20	0.21	0.58	0.59	0.59	0.59	0.57	0.57	0.55	0.54	0.55	0.56	0.56	0.56	0.56	0.56	0.56	0.55	0.55	0.61	0.63	0.52	0.60	0.59
	UAU	0.81	0.82	0.88	0.88	0.78	0.85	0.80	0.79	0.42	0.41	0.41	0.41	0.43	0.43	0.45	0.46	0.45	0.44	0.44	0.44	0.44	0.44	0.44	0.45	0.45	0.39	0.37	0.48	0.40	0.41

Supplementary Figure 11: Codon usage in *Paramecium* species. Shown is the frequency of each codon for each amino acid across the Paramecium species included in this study.



Supplementary Figure 12: Decay of linkage disequilibrium across the mitochondrial genomes of *P. tetraurelia*, *P. sexaurelia*, and *P. caudatum*.

SUPPLEMENTARY TABLES

Supplementary Table 1: *dN*/*dS* across the *Paramecium* phylogeny (sorted by lowest to highest value of *dN*/*dS*).

gene	dN/dS
NADH_dehydrogenase_subunit_10	0.0071
NADH_dehydrogenase_subunit_1_b	0.0159
NADH_dehydrogenase_subunit_7	0.0177
ribosomal_protein_L14	0.0220
apocytochrome_b	0.0221
Ymf68	0.0234
NADH_dehydrogenase_subunit_3	0.0296
NADH_dehydrogenase_subunit_6	0.0305
cytochrome_c_oxidase_subunit_1	0.0337
NADH_dehydrogenase_subunit_4	0.0338
Ymf57	0.0342
Ymf80	0.0347
NADH_dehydrogenase_subunit_1_a	0.0396
ribosomal_protein_S12	0.0406
cytochrome_c_oxidase_subunit_2	0.0433
ribosomal_protein_L16	0.0482
ribosomal_protein_L12	0.0483
Ymf66	0.0491
ribosomal_protein_L6	0.0522
ribosomal_protein_S13	0.0544
NADH_dehydrogenase_subunit_4L	0.0559
NADH_dehydrogenase_subunit_5	0.0623
NADH_dehydrogenase_subunit_2	0.0661
Ymf79	0.0663
Ymf56	0.0703
Ymf78	0.0805
ribosomal_protein_S14	0.0818
Ymf65	0.0894
Ymf61	0.1018
Ymf67	0.1105
NADH_dehydrogenase_subunit_9	0.1158
Ymf84	0.1200
yejR	0.1241
Ymf59	0.1476
ribosomal_protein_S3	0.1574
Ymf63	0.1632
ATP_synthase_F0_subunit_9	0.1781
Ymf83	0.1803
Ymf76	0.1962
Ymf64	0.2154

Supplementary Table 2: List of genes under positive selection in *P. sexaurelia* and *P. caudatum*. All p values are corrected for multiple tests by Holm Bonferroni method.

P. sexaurelia				***P. caudatum***			
all sites included		**Using well conserved sites (see methods)**		**all sites included**		**Using well conserved sites (see methods)**	
gene	**p**	**gene**	**p**	**gene**	**p**	**gene**	**p**
Ymf64	1.69×10^{-13}	*Cox1*	5.01×10^{-3}	*Ymf63*	5.71×10^{-9}	*Ymf64*	3.48×10^{-7}
Ymf67	2.71×10^{-7}	*Ymf76*	9.49×10^{-3}	*Ymf64*	2.12×10^{-8}		
Ymf63	3.56×10^{-6}	*Ymf63*	2.99×10^{-2}	*rpS3*	2.78×10^{-4}		
Ymf84	1.84×10^{-5}			*Ymf59*	7.79×10^{-3}		
Cox1	2.28×10^{-4}			*Ymf61*	1.18×10^{-2}		
Nadh5	3.58×10^{-4}			*Ymf67*	2.65×10^{-2}		
rpL6	1.91×10^{-3}						
yejR	4.22×10^{-3}						
Ymf65	6.72×10^{-3}						
rpS13	8.23×10^{-3}						
cox2	1.05×10^{-2}						
Ymf76	2.72×10^{-2}						
Ymf61	2.79×10^{-2}						

Supplementary Table 3: Mitochondrial mutations identified in *P. biaurelia* and *P. sexaurelia* mutation accumulation experiments.

P. biaurelia				***P. sexaurelia***		
position	**genome**	**variant**		**position**	**genome**	**variant**
490	C	A		1577	A	G
606	C	A		1645	T	C
2646	T	A		3466	A	T
2887	T	G		3790	A	G
3155	A	G		4674	T	A
3155	A	T		4742	T	A
4779	C	A		4770	G	A
4794	G	A		4847	A	T
4811	T	G		5065	T	A
7442	A	T		5417	A	G
7498	A	G		6096	G	A
7529	T	A		6454	T	A
7728	T	A		6491	T	A
7753	T	A		6635	G	A
7979	T	A		6642	T	A
7998	A	G		6903	A	C
8071	T	A		7007	A	G

8076	T	A		7424	A	T
8077	A	T		7522	G	A
13086	T	A		7538	G	A
13140	A	C		7568	G	A
17709	T	A		7721	T	A
17709	T	C		7734	T	A
17761	T	A		7858	A	C
17787	G	A		8152	T	A
17790	C	A		8172	T	G
17841	T	A		8405	T	C
17841	T	G		8439	A	C
18090	A	T		8620	A	G
22843	T	A		8850	T	A
22843	T	G		9044	G	A
22955	A	T		9955	A	G
26509	T	A		10023	T	C
26657	T	C		10382	A	G
26668	A	G		10444	A	G
26702	T	A		10919	T	A
26703	T	A		11283	T	A
26744	C	A		11900	T	A
26766	G	A		12902	A	G
26768	T	A		13649	T	C
28321	T	A		15256	T	A
28378	T	A		15771	A	T
28498	T	G		15786	T	C
32849	T	A		15799	A	G
32849	T	G		16239	A	G
33094	T	G		16394	T	C
33096	T	G		17103	A	T
33139	A	G		18194	A	G
35080	T	C		19729	T	G
35485	T	A		20044	T	G
35512	A	G		20398	A	G
35564	T	C		20529	T	C
35723	C	T		21060	A	G
37061	G	A		21204	A	T
37104	T	G		21423	A	T
				21615	A	T
				22111	T	C
				23241	A	G
				24268	C	T
				24806	A	G
				25196	T	C
				25199	A	C

				25227	A	G
				25290	T	A
				25347	T	A
				26548	T	A
				26672	T	G
				26934	T	A
				27174	T	C
				27714	T	G
				27924	A	C
				29490	A	T
				30667	T	C
				31217	C	T
				31870	A	G
				32870	T	G
				33090	T	A
				33794	A	G
				33800	T	G
				34300	C	T
				34334	G	A
				34348	G	A
				34736	T	A
				35045	T	A
				35700	C	T
				36667	T	A
				38460	T	A

Supplementary Table 4: Summary of results obtained from LDhat (pairwise) to detect the presence of recombination in mitochondrial genomes of *Paramecium* species. The *p* values (not corrected for multiple tests) indicate whether values of statistics calculated on observed data are significantly different from the same values obtained for permuted SNPs.

Species	**Model**	θ	***p* (Composite Likelihood Test)**	***p* (G4)**	***p* (correlation between r^2 and distance)**	***p* (correlation between D' and distance)**	**$4N_er$ (Hudson's method with finite sites)**	**$4N_er$ (estimated by Wakely, 1997)**
P. tetraurelia	cross-over	0.001	0.333	0.000	0.194	0.000	0	0
		0.01	1.000	0.000	0.211	0.000	0	0
	gene conversion	0.001	0.322	0.000	0.559	0.043	0	0

		0.01	1.000	0.000	0.549	0.043	0	0
P. sexaurelia	cross-over	0.01	1.000	0.965	0.980	1.000	0	0
		0.1	1.000	0.974	0.971	1.000	0	0
	gene conversion	0.01	1.000	0.107	0.043	0.990	0	0
		0.1	1.000	0.099	0.050	0.998	0	0
P. caudatum	cross-over	0.01	0.273	1.000	0.995	0.998	0	0
		0.1	0.000	1.000	0.998	0.999	0	0
	gene conversion	0.01	0.276	1.000	1.000	1.000	0	0
		0.1	0.000	1.000	1.000	1.000	0	0
P. multimicronucleatum	cross-over	0.01	0.236	0.811	0.782	0.466	0	0
	gene conversion	0.01	0.225	0.879	0.806	0.502	0	0

Supplementary Table 5: Comparison of neutrality indices between the mitochondrial and nuclear genes that encode for proteins that are part of the oxidative phosphorylation pathway and those that form the structural components of the ribosome.

	Oxidative phosphorylation			**Ribosomal proteins**		
	π_n/π_s	$P_N/P_S/D_N/D_S$	π_n/π_s/dN/dS	π_n/π_s	$P_N/P_S/D_N/D_S$	π_n/π_s/dN/dS
***P. tetraurelia*:**						
Number of nuclear genes	27	13	14	336	77	88
Mitochondrial	0.127	0.781	2.384	0.062	1.935	2.248
Nuclear	0.200	2.896	6.712	0.092	0.737	1.653
p	0.223	0.285	0.422	0.387	0.352	0.262
***P. sexaurelia*:**						
Number of nuclear genes	22	18	16	429	119	138
Mitochondrial	0.124	0.328	5.180	0.035	0.089	0.754
Nuclear	0.146	1.243	7.579	0.158	1.047	3.759
p	0.535	0.444	0.145	0.093	0.124	0.308
***P. caudatum*:**						
Number of nuclear genes	12	3	-	422	9	-
Mitochondrial	0.105	0.364	-	0.067	0.051	-
Nuclear	0.133	0.211	-	0.133	0.260	-
p	0.726	0.107	-	0.513	0.650	-

Supplementary Table 6: Neutrality indices (defined in the Methods section) in the mitochondria of *Paramecium* species (divergence estimated using all 13 taxa).

	P. tetraurelia		***P. sexaurelia***		***P. caudatum***	
	NI [SD]	**NI_{TG}**	**NI [SD]**	**NI_{TG}**	**NI [SD]**	**NI_{TG}**
All mitochondrial genes	1.521 [2.031]	0.765	0.281 [0.475]	0.213	0.500 [0.592]	0.314
Excluding Ymf genes	1.794 [2.515]	0.690	0.339 [0.601]	0.265	0.645 [0.706]	0.403
All mitochondrial genes with dS < 1.0	1.379 [1.454]	0.829	0.539 [0.843]	0.371	0.513 [0.618]	0.314
All mitochondrial genes with dS < 1.0, excluding Ymf genes	1.607 [1.890]	0.715	0.859 [1.211]	0.495	0.674 [0.894]	0.253