Population genetics of *Paramecium* mitochondrial genomes: recombination, mutation spectrum, and efficacy of selection

by

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

[‡]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
Lynch Lab
Department of Biology
Jordan Hall 327, 1001 E 3rd St
Indiana University, Bloomington, IN 47405
(812) 856-0115
pjohri@indiana.edu

RUNNING TITLE: Evolution of mitochondrial genomes.

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

ABSTRACT

There is a dearth of understanding of evolution of mitochondrial genomes and their population-genetic environment among unicellular eukaryotes. 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 appear to 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 confirm a complete lack of recombination in *Paramecium* mitochondria and find that the mitochondrial genome appears to experience either similar or stronger levels of purifying selection than the nucleus. We discuss how our results might still be concordant with theoretical predictions.

63 **INTRODUCTION**

64

65 Mitochondria have played integral roles in furthering our understanding of relationships among
66 species as well as revealing population structure and demographic history. As a consequence,
67 we have obtained insights into the unique population-genetic properties of mitochondrial
68 genomes. In most species, mitochondria are inherited uniparentally (but see BARR *et al.* 2005).
69 Although mitochondrial genomes are known to frequently undergo recombination in plants
70 (STADLER and DELPH 2002; MACKENZIE 2007) as well as fungi (FRITSCH *et al.* 2014), no
71 recombination has been detected in animals (BALLARD and WHITLOCK 2004; PIGANEAU and
72 EYRE-WALKER 2004). Because of the unique mode of transmission and lack of recombination in
73 some species, mitochondria have been suggested to have lower effective population sizes than
74 their nuclear counterparts and therefore to accumulate more deleterious mutations (LYNCH and
75 BLANCHARD 1998; NEIMAN and TAYLOR 2009). In addition, mitochondrial genomes experience
76 much higher spontaneous rates of mutation than their corresponding nuclear genomes in animals
77 but exhibit the opposite trend in plants (LYNCH *et al.* 2006).

78

79 Unlike the relatively uniform and conserved properties of metazoan mitogenomes, mitochondrial
80 genomes in unicellular eukaryotes exhibit remarkable variation in genome structure and GC
81 content. Mitochondrial genome structures range from hundreds of short linear segments (0.3-8.3
82 kb) in *Amoebidium parasiticum* (BURGER *et al.* 2003), an opisthokont, to many small (< 10 kb)
83 circular genomes in the excavate *Diplonema papillatum* (VLCEK *et al.* 2011), to a single larger
84 linear or circular chromosome, and a variety of other states (reviewed in SMITH and KEELING
85 2015). In addition to variation in organization, mitochondria from unicellular lineages display

86 widely diverse GC compositions, ranging from as low as 10% in yeast (SMITH 2012) to 60% in
87 *Lobochlamys culleus* (BORZA *et al.* 2009), although most species are AT rich, with an average
88 GC content of 35% (SMITH 2012).
89
90 Much has been learned about the structure, evolution and population-genetic environment of
91 mitochondria in the main model systems, especially in plants and metazoans (LYNCH 2007;
92 SMITH 2016). However we lack such understanding in mitochondria of most unicellular
93 eukaryotes, which harbor most eukaryotic phylogenetic diversity. We address this gap by
94 surveying both within and between-species variation in mitochondrial genomes among highly
95 diverged ciliate species.
96
97 In the large and morphologically and ecologically diverse ciliate lineage, mitochondrial genomes
98 are organized into single large linear chromosomes, several tens of kilobases in length, with
99 telomeres at the ends (GODDARD and CUMMINGS 1975; MORIN and CECH 1988; SWART *et al.*
100 2012). Few mitochondrial genomes have been fully sequenced among the ciliates, with two in
101 the *Paramecium* genus (*P. tetraurelia* and *P. caudatum*; (BARTH and BERENDONK 2011)), and
102 only a few others: *Tetrahymena pyriformis* (BURGER *et al.* 2000), *Euplotes minuta* and *Euplotes*
103 *crassus* (DE GRAAF *et al.* 2009), *Oxytricha trifallax* (SWART *et al.* 2012), *Stentor coeruleus*
104 (SLABODNICK *et al.* 2017), *Ichthyophthirius multifiliis* (COYNE *et al.* 2011), and the anaerobic
105 ciliate *Nyctotherus ovalis* (DE GRAAF *et al.* 2011).
106
107 The *Paramecium* genus is a particularly interesting system in which to study the evolution of
108 mitochondrial genomes. While it contains a number of distinct morphospecies, it is especially

109 known for including a species complex consisting of multiple morphologically identical but
110 sexually isolated species – the *P. aurelia* complex (SONNEBORN 1975). Species in the *P. aurelia*
111 complex are extremely ancient, with the estimated time of divergence for the complex as a whole
112 being on the order of 300 million years ago (MCGRATH *et al.* 2014b), implying that the genus
113 *Paramecium* itself is even more ancient. Interestingly, among *Paramecium* species, there is a
114 doubling of GC content in the branch leading to the *P. aurelia* complex (BURGER *et al.* 2000;
115 BARTH and BERENDONK 2011). *Paramecium* therefore also offers an attractive system with
116 which to study the evolution of nucleotide composition across mitochondrial genomes that are
117 structurally very similar.
118
119 In this study we further the understanding of the biology and the population-genetic
120 environment of ciliate mitochondria by presenting the complete mitochondrial genomes of
121 nine species belonging to the *P. aurelia* complex, four outgroup species, and 5-10 isolates
122 each of four *Paramecium* species. Using phylogenetic and population-genetic analyses, we
123 investigate variation in protein-coding genes, tRNA content, and causative forces
124 governing the evolution of nucleotide composition across the phylogeny. In addition, we
125 estimate the recombination rate across the genome and address the controversy of whether
126 mitochondrial genomes experience reduced purifying selection in comparison to their
127 nuclear counterparts.
128
129
130
131

132 **RESULTS**

133

134 **Mitochondrial genome assembly and SNP detection**

135 We assembled complete mitochondrial genomes of nine species belonging to the *P. aurelia*
136 complex: *P. biaurelia*, *P. tetraurelia*, *P. sexaurelia*, *P. octaurelia*, *P. novaurelia*, *P.*
137 *decaurelia*, *P. dodecaurelia*, *P. quadecaurelia*, and *P. jenningsi* as well as four outgroup
138 species: *P. caudatum*, *P. caudatum-C026*, *P. multimicronucleatum*, and *P.*
139 *multimicronucleatum-Peniche3I* (Figure 1). We note that *P. caudatum-C026* and *P.*
140 *multimicronucleatum-Peniche3I* were initially sampled as individual isolates belonging to
141 the *P. caudatum* and *P. multimicronucleatum* species (based on morphological criteria),
142 respectively. However, the analysis of the resulting sequencing data revealed that they are
143 highly diverged from the reference strains (see below) and are therefore almost certainly
144 separate species and were treated as such in subsequent analyses. For all 13 mitochondrial
145 genomes in this study, Illumina reads were assembled using SPAdes (BANKEVICH *et al.*
146 2012), and mitochondrial contigs were identified by BLAST searches against the publicly
147 available *P. caudatum* and *P. tetraurelia* mitochondrial sequences (see the Materials and
148 Methods section for more details). The assembled sequences are largely complete with the
149 exception of the *P. multimicronucleatum* and *P. multimicronucleatum* Peniche3I
150 mitocontigs, which appear to be missing small portions of the 3′ end of the chromosome
151 (Figure 2).

152

153 In addition, we sequenced 10 isolates of both *P. tetraurelia* and *P. sexaurelia*, and 5 isolates
154 of each *P. caudatum* and *P. multimicronucleatum* (Supplementary Figure 1). Illumina

155 paired-end reads from these isolates were mapped to the assembled reference genomes and
156 SNPs were called as in JOHRI *et al.* (2017). These genomes were also assembled *de novo*
157 (Supplementary Figures 2, 3 and 4) in order to examine large-scale genome organization
158 mapping.
159

160 **Genome structure and organization**

161 All *Paramecium* mitochondrial genomes are linear; ~40 kb in the *P. aurelia* species and ~44
162 kb in *P. caudatum* and *P. caudatum-C026*. Telomeric repeats identified on both ends of the
163 assembled genomes imply that the full lengths of the linear contigs have been assembled in
164 most species. While we did not recover complete mitochondrial contigs for *P.*
165 *multimicronucleatum* and *P. multimicronucleatum-Peniche3I*, the length of the existing
166 contigs suggests an overall size closer to that observed in *P. aurelia* than to the larger
167 mitochondrial genomes in the *P. caudatum* lineage. We also note that the assemblies for two
168 of the *P. aurelia* species contain extensions (Supplementary Figure 5 and 6). In *P.*
169 *novaurelia*, an additional ~18 kb is present at the 5′ end of the mitocontig, while in *P.*
170 *quadecaurelia,* a small, ~1 kb extension is seen at the 3′ end. However, the read coverage
171 over these regions is very different from the rest of the mitocontigs, suggesting either
172 misassembly or heterogeneity within cell populations. We therefore ignored these
173 extensions in subsequent analysis.
174

175 The organization of the genomes is very similar, with gene order preserved almost perfectly
176 between all species (Figure 2). The GC content is relatively uniform along the length of
177 the mitochondrial chromosome in all *Paramecium* species, with the exception of the

178 immediate subtelomeric areas and the ribosomal RNAs in the non-*aurelia* species
179 (Supplementary Figures 7-9).
180
181 We also identified telomeric repeats based on the sequences at the end of assembled
182 mitocontigs (see Methods for more details). In most *P. aurelia* species, we identify
183 almost identical repeats, with a 23-bp consensus sequence
184 GCCCTGGTGGCCCTAGAAGCTCC (Supplementary Figure 10). However, in *P.*
185 *jenningsi* and *P. sexaurelia* the telomere repeat motif is the same length, but has 1 base
186 pair different from this consensus. Of note, these two species are the earliest diverging
187 ones within the *P. aurelia* complex species included in our analysis. We observed even
188 more divergent telomeric sequences in *P. caudatum*
189 (GCCCTGGTAACGCTGGTCGCCCTTTTAAAATA) and *P. multimicronucleatum*
190 (GCCCTTGTTACACTTGGTGGCTCTTTAAGCTCT). In these species, the core
191 telomeric repeat sequence has been expanded by additional 10 bp sequences that are not
192 present in *P. aurelia*. The *Paramecium* core telomeric repeat is broadly similar to that of
193 *Tetrahymena* (ACCCTCGTGTCTCTTTA), the other Oligohymenophorean genus for
194 which mitochondrial genomes are available, but distinct from the what is observed in
195 distantly branching ciliates such as the spirotrichean *Oxytricha trifallax* (Supplementary
196 Figure 10).
197

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

199 All *Paramecium* mitochondrial genomes contain a core set of 15 genes (Figure 2) involved
200 in electron transport and ATP synthesis (*atp9*, *cob*, *cox1*, *cox2*, *nadh1*, *nadh2*, *nadh3*,

201 *nadh4*, *nadh4L*, *nadh5*, *nadh6*, *nadh7*, nadh9 and *nadh10)*, the heme maturase *YejR*, and 8
202 ribosomal protein genes (*rpL6*, *rpL12*, *rpL14*, *rpL16*, *rpS3*, *rpS12*, *rpS13*, and *rpS14*). One
203 ribosomal protein subunit *rpS19* was found to be absent from *P. tetraurelia*, *P. octaurelia*,
204 *P. novaurelia*, *P. dodecaurelia*, and *P. quadecaurelia* and therefore appears to have been
205 independently lost from the mitochondria at least 3 times along the *P. aurelia* phylogeny
206 (Figure 1). Remarkably, we also find a presence-absence polymorphism of *rpS19* in *P.*
207 *sexaurelia* isolates, with the ORF not being identifiable in 2 out of 11 isolates studied
208 (Supplementary Figure 11). Although this might suggest a mitochondrial gene in the
209 process of transfer to the nucleus, the closest hit in the nuclear genome has an E value of
210 0.27, implying either a complete loss or perhaps transfer to a mitochondrial plasmid, which
211 has not been captured by any study.
212
213 *Paramecium* genomes also contain 16 ciliate-specific genes of unknown function, called
214 *Ymf* genes (BURGER *et al.* 2000), named according to the Commission on Plant Gene
215 Nomenclature (PRICE and REARDON 2001). Each sequenced strain contains all of these
216 genes, with the exception of *P. multimicronucleatum* strain M13 in which *Ymf80* and
217 *Ymf83* appear to have fused with *nadh4* and *Ymf64*, respectively. In addition, several ORFs
218 originally identified in *P. tetraurelia*, are observed in only one or a few additional *P.*
219 *aurelia* mitogenomes, typically overlapping longer ORFs: *orf189*, *orf285*, and *orf314*
220 restricted to the three *P. tetraurelia* isolates; *orf221* present in *P. tetraurelia*, *P. octaurelia*
221 and *P. biaurelia*; *orf265* in *P. biaurelia*; and orf78 found only in *P. tetraurelia* isolate A.
222
223 We used HHPred (SODING *et al.* 2005) to predict homologs of the Ymf genes and found

224 that parts of Ymf genes display homology to standard mitochondrial genes, encoding
225 ribosomal subunits and proteins involved in electron transport. About 37% of sequence
226 from *Ymf59* was found to be homologous to *rpS10*, ~22% sequence of *Ymf61* to a 30S
227 ribosomal protein S24E, ~28% of *Ymf64* sequence to a 30 S *rpS3*, ~79% of *Ymf65*
228 sequence to the NADH subunit and ~11% of *Ymf67* to cytochrome c oxidase, subunit P.
229 Homologs of *Ymf59* and *Ymf64* were previously similarly identified in *Oxytricha* (SWART
230 *et al.* 2012) and reached the same conclusions.
231
232 In addition, we found a previously unidentified ORF larger than 100 AA, which appears to
233 be present in all *Paramecium* mitochondrial genomes and is located at the very 5′ end of the
234 chromosome, before the large ribosomal RNA. This ORF has homology to the CRISPR-
235 associated endoribonuclease Cas6; its functional significance is unclear at present.
236
237 A notable Ymf gene is *Ymf64*, which accounts for a significant part of the difference
238 between the length of *P. aurelia* and *P. caudatum* mitochondrial genomes, being much
239 longer in the latter (~1700 AA vs. ~600 AA) (BARTH and BERENDONK 2011). In two
240 isolates of *P. multimicronucleatum* (M04 and M05), the locus appears to have split into
241 multiple ORFs (Figure 2), with the one most clearly homologous to *Ymf64* being shortened
242 even relative to *P. aurelia*. However, in the *P. multimicronucleatum* isolate M13, as well as
243 *P. multimicronucleatum-Peniche3I*, *Ymf64* is relatively long (1000 and 863 AA
244 respectively).
245
246 Most other ORFs exhibit relatively constant length across all *Paramecium* mitochondrial

247 genomes, with the exceptions of *atp9*, which is elongated by ~30-60 AA in all *P. sexaurelia*
248 isolates, *rpS3* and *rps19*, which are specifically elongated in *P. caudatum* and *P.*
249 *caudatum-C026*, and *Ymf76*, which is very short in one of the *P. biaurelia* isolates.
250
251
252 **Rate of evolution of protein-coding genes across the phylogeny**

253 Most species were found to be fairly evolutionarily distant from each other, as
254 measured by the average number of nonsynonymous (*dN*) and synonymous (*dS*) site
255 substitutions per site (Supplementary Figure 12). *P. decaurelia* and *P. dodecaurelia*
256 (average *dS*: 0.85; average *dN*: 0.05, average total divergence: 0.23) are the closest
257 species pair, followed by *P. tetraurelia* and *P. octaurelia* (average *dS*: 0.93; average
258 *dN*: 0.09; average total divergence: 0.27). All other species pairs have on average *dS* >
259 1.0, i.e., synonymous sites in protein coding genes have on average undergone more
260 than 1 substitution each since the time of divergence. We note here that mean *dS*
261 between *P. caudatum* and *P. caudatum-C026* is ~2.0 (average *dN*: 0.07; average total
262 divergence: 0.28), implying that these two isolates represent separate species. A similar
263 observation confirms that *P. multimicronucleatum-Peniche3I* and *P.*
264 *multimicronucleatum* are probably separate species (average *dS* = 1.97; average *dN*:
265 0.10; average total divergence: 0.26). Our observations suggest the possibility of
266 cryptic species complexes in both *P. caudatum* and *P. multimicronucleatum*, as
267 previously suspected (HORI *et al.* 2006; TARCZ *et al.* 2012). Thus the phenomenon of
268 species complexes containing numerous morphologically identical species, already
269 known for the *P. aurelia* (SONNEBORN 1937) and *T. pyriformis* (GRUCHY 1955)

complex, might be much more widespread among ciliates than previously appreciated.

We next asked whether different functional groups of genes exhibit different sequence constraints by measuring number of nonsynonymous relative to synonymous substitutions (*dN/dS*) across the phylogeny. Overall, enzyme components of the respiratory chains exhibit the highest degree of conservation across the phylogeny (average $dN/dS \cong 0.04$), followed by ribosomal proteins (average $dN/dS \cong 0.06$) and then *Ymf* genes (average $dN/dS \cong 0.11$) (Supplementary Table 1, Supplementary Figure 13). Ciliate-specific *Ymf* genes therefore exhibit a relatively higher rate of evolution than other genes, with as little as 30% sequence identity between *P. aurelia*, *P. caudatum,* and *P. multimicronucleatum*, similar to observations in *Tetrahymena* (MORADIAN *et al.* 2007), as well as with higher values of π_n/π_s in *P. tetraurelia*, *P. sexaurelia*, *P. caudatum*, and *P. multimicronucleatum* (Supplementary Figure 14).

The rapid divergence of *Ymf* genes can either be explained by relaxed purifying selection at the sequence level, or recurrent positive selection over long periods of time, or a combination of both. To distinguish between these possibilities, we performed a McDonald Kreitman test (MK test) in the species *P. tetraurelia*, *P. sexaurelia*, *P. caudatum*, and *P. multimicronucleatum*. As most of these species are highly diverged from each other, we used ancestral reconstruction over the set of all 13 taxa to first infer ancestral nucleotides for each internal node. The numbers of synonymous (D_S) and nonsynonymous (D_N) changes were then inferred along each terminal branch leading to each of the 4 species mentioned above. We tested for positive selection by determining

293 whether the proportion of nonsynonymous fixed differences relative to synonymous
294 ones was significantly larger than the proportion of nonsynonymous relative to
295 synonymous polymorphic differences. None of the genes in *P. tetraurelia* were found to
296 have undergone positive selection by this criterion (Holm–Bonferroni corrected $p \geq 0.05$
297 in each case). In *P. sexaurelia*, *cox1*, *Ymf76*, and *Ymf63* were found to be the only genes
298 that showed significant positive selection, whereas in *P. caudatum*, only *Ymf64*
299 exhibited positive selection (Supplementary Table 2). It should be noted that *Ymf64* has
300 also increased in length in *P. caudatum* and might represent an example of gain of
301 function. With very few SNPs available in *P. multimicronucleatum*, the MK test was not
302 significant for any gene. Nevertheless, we find no statistical enrichment of number of
303 genes evolving under positive selection among the *Ymf* vs. other genes in *P. sexaurelia*
304 (p = 0.56; Fisher's exact test) and *P. caudatum* (p = 0.40; Fisher's exact test),
305 suggesting that majority of them are likely evolving fast due to relaxed sequence
306 constraints.
307

308 **Levels of expression of protein-coding genes in the mitochondria**

309 At present, very little is known about the regulation of transcription and of the
310 posttranscriptional stability of transcripts in ciliate mitochondria. Given the organization of
311 *Paramecium* mitogenomes, with a small group of genes on the forward strand in the middle
312 of the chromosome separating two other large groups of genes, at least three transcriptional
313 units are likely to exist (BURGER *et al.* 2000), but the number could be significantly higher.
314 To address this issue, we examined the abundance of transcripts (generated from PolyA+-
315 selected RNA samples) obtained from single strains for each of *P. biaurelia*, *P. sexaurelia*,

316 and *P. caudatum* (see the Materials and Methods section for details). We observe very high
317 levels of rRNA suggesting that rRNA might be expressed separately, but in addition, we
318 also notice large differences between the abundances of protein-coding transcripts
319 (Supplementary Figure 15, 16 and 17), with the most highly expressed genes often varying
320 both between species and within the same species in different conditions (Supplementary
321 Figure 18). These observations suggest a complex transcriptional organization and/or
322 posttranscriptional regulation of RNA levels in ciliate mitochondria. While the available
323 RNA-seq data are from samples enriched for polyadenylated transcripts, and mitochondrial
324 transcripts in *Paramecium* may not have polyA tails, this should not have affected the
325 ratios of mitochondrial transcripts.
326

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

328 Perhaps the most remarkable discontinuity in *Paramecium* mitochondrial genomes is the
329 difference between the high GC content observed in species of the *P. aurelia* complex (~40%)
330 and the low GC content in the outgroup species (~20%). Very low GC content is also observed
331 in mitochondrial genomes of other ciliates, for instance in the five *Tetrahymena* species
332 (~18-21%), in *Icthyophthirius multifiliis* (~16%) (COYNE *et al.* 2011), and in *Oxytricha*
333 *trifallax* (~24%). However, nuclear genomes have relatively similar levels of GC content across
334 the *Paramecium* species (MCGRATH *et al.* 2014a). The low GC content of *P. caudatum* and *P.*
335 *multimicronucleatum* mtDNA is marked by a highly biased codon usage (BARTH and
336 BERENDONK 2011), with most synonymous positions exhibiting a strong bias for A or T
337 nucleotides (Supplementary Figure 19). Indeed, the difference in GC content is most
338 pronounced at 4-fold degenerate sites, with the GC content at such sites being as low as

339 3.5% in *P. multimicronucleatum* and as high as 54.5% in *P. decaurelia* (Figure 1). The 0-
340 fold redundant sites in the outgroup species have much higher GC content (~26-37%) and
341 the GC content (31-36%) of the rRNAs is similar across the entire genus suggesting strong
342 selection on rRNA nucleotide content. At least three forces may be responsible for
343 determining the GC content at 4-fold degenerate sites: (1) mutational processes, (2) codon
344 usage bias, and/or (3) genome-wide selection for higher/lower GC content. We therefore
345 looked at GC composition in other regions of the genomes like tRNAs and intergenic regions
346 that do not experience selection for codon usage. Both tRNAs and intergenic regions have
347 elevated GC content in the *P. aurelia* species (~40-45% and ~34-37%) relative to *P. caudatum*
348 and *P. multimicronucleatum* (~14-22% and ~16-20%; Figure 1), suggesting that either
349 mutational changes or selection for GC content is primarily responsible for changes in nucleotide
350 composition across the genome.
351
352 To distinguish between mutation and selection, we evaluated the mutational pressures acting on
353 *Paramecium* species, using two approaches. First, we used the mitochondrial mutation spectrum
354 obtained from mutation-accumulation (MA) studies of *P. tetraurelia* (SUNG *et al.* 2012) to
355 calculate the AT mutation bias, $m = v/u$, where v is the mutation rate from G/C to A/T and u is
356 the rate of mutation from A/T to G/C (Table 1). We found $m \cong 1.0$, showing little mutation bias
357 and corresponding to expected equilibrium GC content (SUEOKA 1993) of 51.3% (Table 1). This
358 is very close to the GC content of 4-fold redundant sites (54%) in *P. tetraurelia*. We also
359 analyzed existing sequencing data from MA experiments in *P. biaurelia* and *P.*
360 *sexaurelia* (LONG *et al.* 2018a; Supplementary Table 5). These data imply slight mutation
361 bias towards GC in *P. biaurelia* and *P. sexaurelia*, corresponding to expected equilibrium

362 GC percentages (SUEOKA 1993) of 55.8% and 67.0% respectively. Given the relatively
363 small number of mutations (a few tens) identified in each set of MA experiments, there is
364 likely to be considerable uncertainty in these estimates. Nonetheless these data do not
365 present evidence that 4-fold degenerate sites in *P. aurelia* mitochondria are evolving away
366 from mutation equilibrium via selection.
367
368 Next, derived singleton alleles at 4-fold degenerate sites were used to quantify the number of
369 G/C to A/T mutations relative to A/T to G/C mutations and thus infer an estimate of AT
370 mutation bias from population data. Again, we find that m is close to 1.0 (i.e. no mutation bias
371 towards A/T) in both *P. tetraurelia* and *P. sexaurelia*, with the predicted GC content under
372 mutation equilibrium being remarkably close to that of their 4-fold degenerate sites (Table 1).
373 These results suggest that the composition of 4-fold degenerate sites is mostly determined by
374 mutation. The same can also be seen by calculating the strength of selection ($S = 4N_e s$ or
375 $2N_e s$) favoring A/T at 4-fold degenerate sites. As S usually takes values between 0.1 and 4.0
376 (LYNCH 2007; LONG *et al.* 2018b) and we obtain much lower magnitudes (0.0-0.5) in the *P.*
377 *aurelia* species (Table 1), we infer negligible or weak selection at 4-fold degenerate sites,
378 consistent with absence of codon bias at the third position.
379
380 In contrast to the *P. aurelia* species, we infer a significant AT mutation bias in *P. caudatum*
381 (Table 1), and predicted GC content at equilibrium of ~17%, with the observed GC content in
382 intergenic and tRNA regions (16% and 20%) being remarkably close to this value. These
383 calculations suggest that genome-wide GC content in *P. caudatum* is mostly governed by
384 mutation bias, and that there has been a decrease in AT mutation bias along the branch leading to

385 the *P. aurelia* species.

386

387 **tRNA content variation across and within species**

388 While the number and identity of protein-coding genes is generally the same across the
389 genus, tRNAs display significant variation. All species in the *P. aurelia* complex have only
390 three tRNA loci, at highly conserved locations (Figure 2), most of the time consisting of a
391 Y, F, and W tRNA (the latter one recognizing the UGA stop codon that has been
392 reassigned in ciliate mitochondria to tryptophan). The one exception is *P. octaurelia*, in
393 which a T tRNA has been substituted for the Y tRNA that other species possess. In
394 stark contrast, mitochondrial genomes in the *P. caudatum* and *P. multimicronucleatum*
395 lineages contain large and highly variable sets of tRNAs (Figure 3), with as many as 38
396 tRNA predictions in *P. multimicronucleatum* M05, and with individual *P. caudatum* and *P.*
397 *multimicronucleatum* isolates having different tRNAs at multiple loci in their
398 mitogenomes. These tRNAs often overlap protein-coding genes (in contrast to what is
399 observed in *P. aurelia*), in particular *Ymf* genes. Of note, the three tRNAs found in the *P.*
400 *aurelia* species are also present in all outgroup species.

401

402 **Recombination in the mitochondria**

403 We performed multiple tests for the presence of recombination in the mitochondria. First, we
404 evaluated the relationship between linkage disequilibrium (LD), calculated by r^2 (HILL and
405 ROBERTSON 1968) and distance between sites; r^2 does not decrease with distance in any of the
406 four species, consistent with the expectation under no recombination. Next, we conducted the
407 four-gamete test (FGT) (HUDSON and KAPLAN 1985), which detects recombination by searching

408 for pairs of polymorphic sites with all four segregating haplotypes and assumes that such pairs of
409 sites must have arisen via recombination (under the infinite-sites model). We found all four
410 gametes at 2 pairs of sites (1.45×10^{-5} of all pairs) in *P. tetraurelia*, 312781 sites (2.14×10^{-2} of
411 all pairs) in *P. sexaurelia*, 42647 sites (1.21×10^{-2} of all pairs) in *P. caudatum*, and 0 in *P.*
412 *multimicronucleatum*. The observed variation between species correlates well with levels of total
413 sequence diversity and indicates the bias in power to detect recombination towards species with
414 more sequence variation. Although results from the FGT suggest the presence of some
415 recombination, the probability of finding all four gametes does not increase with the distance
416 between the pairs of sites, contrary to expectation under recombination, and is inconsistent with
417 other distance-based analyses. Violation of the infinite-sites model (i.e. recurrent mutation at the
418 same site) can result in the presence of pairs of sites with all four alleles. The FGT works best for
419 species and genomes whose recombination rate is much larger than the mutation rate, which may
420 not be the case in *Paramecium* mitochondria.
421
422 In order to test if the presence of four gametes was more likely to be caused by recombination or
423 mutation, we used LDhat (McVean *et al.* 2002) to test for recombination and estimate $4N_er$,
424 using values of $4N_e\mu$ as close to empirical data as possible. Here r is the recombination rate, μ is
425 the mutation rate and N_e is the effective population size of the species under consideration. $4N_er$
426 estimated under the finite-sites mutation model (Wakeley 1997; Hudson 2001) was found to be
427 0.0 for all four species, and is consistently 0.0 in all non-overlapping windows (1000 bp)
428 spanning the genome. We also used permutation tests (McVean *et al.* 2002) to detect
429 recombination. The idea behind these tests is that in the absence of recombination the relative
430 position of SNPs would not change inferred values of statistics used to measure recombination.

431 Thus, under absence of recombination, statistics like sum-of-distance between sites with all four
432 gametes (G4) (MEUNIER and EYRE-WALKER 2001), and correlation between r^2 or D' and
433 physical distance, would not be significantly different from an expectation obtained by randomly
434 shuffling SNPs and computing these statistics. We find no statistical significance for the
435 presence of recombination using permutation tests and thus conclude that there is no
436 recombination in *Paramecium* mitochondria.
437

438 **Efficacy of selection in the mitochondria vs. nucleus**

439 Finally, we asked if genes in the mitochondria compared to the nucleus, experienced
440 weaker purifying selection, as would be expected due to smaller effective population
441 size and the lack of recombination in the mitochondria. We did so by evaluating the
442 strength of selection in each of the two genomes using multiple statistics. The average
443 value of d_N/d_S over the mitochondrial phylogeny is 0.05 (SD = 0.05; median = 0.03),
444 while that over the nuclear phylogeny, comprising of the same taxa is 0.05 (SD = 0.04;
445 median = 0.04). The distributions of these d_N/d_S values are not significantly different (p
446 = 0.54, Student's t-test), suggesting that the average amount of purifying selection
447 experienced by genes in the nucleus is not significantly different from that experienced
448 by those in the mitochondria. We also observe no significant difference when
449 performing this comparison only examining genes encoding proteins that are part of the
450 oxidative phosphorylation pathway (mitochondrial: 0.02, nuclear: 0.03, p=0.28) or only
451 those that encode ribosomal subunits (mitochondrial: 0.05, nuclear: 0.04, p=0.74).
452

453 Using π_n/π_s as a proxy for recent purifying selection, we find that the distribution of

454 π_n/π_s of mitochondrial (mt) and nuclear genes (nuc) is not significantly different in *P.*
455 *tetraurelia* (mean mt = 0.174; mean nuc = 0.198; $p = 0.56$) and *P. caudatum* (mean mt =
456 0.135; mean nuc = 0.170; $p = 0.22$) respectively. However, mean π_n/π_s in the
457 mitochondrial genes is significantly lower than that of nuclear genes in *P. sexaurelia*
458 (mean mt = 0.051; mean nuc = 0.268; $p < 2.2 \times 10^{-16}$) and *P. multimicronucleatum* (mean
459 mt = 0.099; mean nuc = 0.178; $p = 1.87 \times 10^{-3}$) respectively, suggesting that
460 mitochondrial genes might be under similar or stronger purifying selection than nuclear
461 genes.
462
463 Next, we calculated the neutrality index (NI = $(P_n / P_s) / (D_n / D_s)$; P_n and P_s: number of
464 nonsynonymous and synonymous polymorphisms respectively; D_n and D_s: number of
465 nonsynonymous and synonymous fixed differences respectively) of each gene in the
466 mitochondrion and nucleus for *P. tetraurelia*, *P. sexaurelia*, and *P. caudatum*. The
467 number of branch-specific substitutions at synonymous (D_s) and nonsynonymous (D_n)
468 sites was calculated by ancestral reconstruction. Comparing the distributions of values
469 of NI in the mitochondrion and nucleus shows that overall there is no significant
470 difference between the two means (whether we include or exclude *Ymf* genes) in *P.*
471 *tetraurelia* and *P. caudatum* (Table 2). Consistent with the above, distributions of NI
472 values in *P. sexaurelia* are significantly different between the nucleus and
473 mitochondrion. We note that the absolute values of NIs seem to indicate pervasive
474 positive selection in both the nucleus and the mitochondrion. However, underestimation
475 of the number of synonymous changes (*Ds*) can result in lower NI, mimicking positive
476 selection. We therefore also calculated neutrality index as $(\pi_n/\pi_s)/ (d_N/d_S)$ (BETANCOURT

477 *et al.* 2012; denoted by NIπ below) for each gene with d_N/d_S obtained from pairwise
478 comparison with respect to the closest species. As a close outgroup species was
479 required, this analysis was restricted to *P. tetraurelia* and *P. sexaurelia*. Again,
480 consistent with all analyses above, there is no difference between the mean NI_π value in
481 the mitochondrion and nucleus in *P. tetraurelia*, but mean NI_π is significantly higher in
482 the nucleus than the mitochondrion in *P. sexaurelia* (Table 2). This suggests that there
483 are more deleterious variants segregating in the nucleus than in the mitochondrion in *P.*
484 *sexaurelia*; these results do not change upon excluding the *Ymf* genes. Together, these
485 results suggest species-specific differences in the relative strength of purifying selection
486 between the mitochondrion and nucleus. Overall, mitochondria appear to experience
487 either similar or stronger efficacy of purifying selection than the nucleus.
488
489 A major caveat of comparing neutrality indices between all genes present in the
490 mitochondria vs. nucleus, is that we might be measuring differences is strength rather
491 than efficacy of purifying selection. In order to correct for that, we compared nuclear
492 genes with similar functions to those in the mitochondria. We thus performed all
493 analyses noted above in a restricted set of nuclear genes. We compared the 14
494 mitochondrial genes with ~13-27 nuclear genes involved in the oxidative
495 phosphorylation pathway, as well as the 8 ribosomal genes in the mitochondria with
496 ~77-430 nuclear genes that encode for structural components of the ribosomes.
497 Although we fail to get significant p values, as the sample sizes are very small, we
498 continue to observe similar trends as described above for all genes (Supplementary
499 Table 3).

500

501 As an aside, we tried to reduce the bias caused by saturation of DS by re-calculating NI
502 only in the mitochondrion: (i) using all 13 taxa (in order to break up longer branches),
503 (ii) restricting the analysis to the set of genes that have pairwise $d_S < 1.0$ (from their
504 closest outgroup), (iii) estimating different variations of neutrality indices (see Methods)
505 in order to reduce various statistical biases (GREENLAND 1982a; STOLETZKI and EYRE-
506 WALKER 2011). We continued to recover extremely low values of NIs (Supplementary
507 Table 4) and failed to obtain values close to NI_n, suggesting that the absolute values of
508 NI can be misleading and have to be interpreted with caution.

509

510

511 **DISCUSSION**

512

513 ***Ciliate-specific mitochondrial genes***

514 *Paramecium* mitochondrial genomes possess 16 lineage-specific open-reading frames (referred
515 to as *Ymf* genes) that have no known homologs in non-ciliate species and lack assigned
516 functions, but are nonetheless conserved across *Paramecium* and *Tetrahymena* species. Other
517 ciliates like *Oxytricha* have also been found to have unidentified open-reading frames (SWART *et*
518 *al.* 2012), but not all *Ymf* genes have homologs identified in ciliates outside of
519 *Oligohymenophora* (which contains both *Tetrahymena* and *Paramecium*). We find that *Ymf*
520 genes are evolving faster, mostly under relaxed purifying selection, and compared to other genes
521 are also more likely to undergo positive selection. Shedding light on the identity of the *Ymf* genes
522 could possibly indicate entirely new set of genes present in the ancestor of *Oligohymenophora*

523 mitochondria.

524

525 We found that parts of five of the 16 *Ymf* genes show homology to standard mitochondrial
526 proteins, especially ribosomal proteins. As ciliates are far from most model organisms in the
527 eukaryotic tree, it is very likely that some *Ymf* genes are ribosomal subunits or genes belonging
528 to the NADH complex and are simply not identifiable because of extreme divergence. Similar
529 findings are reported across protist and other mitochondrial genomes (e.g. DE GRAAF *et al.* 2009;
530 POMBERT *et al.* 2013; BURGER *et al.* 2016; SKIPPINGTON *et al.* 2017) that have diverged
531 substantially from the most well-studied mammalian mitochondrial genomes. In some species,
532 RNA editing (including insertions and deletions) can be substantial (GRAY 2003), which can
533 mask the identification of the proteins encoded, especially when mostly sequence analyses are
534 performed. RNA editing, however, has been reported to not occur in *P. tetraurelia* mitochondria
535 (ORR *et al.* 1997) and is therefore unlikely to be the primary cause of observed divergence. In
536 *Tetrahymena thermophila*, the ATP synthase forms an unusual structure and possesses
537 completely novel subunits whose orthologs are not identifiable in other organisms
538 (BALABASKARAN NINA *et al.* 2010). It is therefore possible that, through some combination of
539 the above phenomena, ciliates possess structurally unique proteins that perform similar functions
540 in the mitochondria, but are difficult to identify based on other known sequences. Although the
541 origin of these genes is unclear, it is intriguing that a set of such fast evolving genes is preserved
542 across highly diverged species.

543

544 Curiously, we find a previously unidentified ORF at the 5′ end of all *Paramecium*
545 mitogenomes, which displays homology to Cas6 found widely in bacteria and archaea

546 (HAFT *et al.* 2005). This could potentially represent a case of horizontal gene transfer from
547 the endosymbiotic bacteria (PREER 1969; FOKIN and GORTZ 2009) living in *Paramecium*.
548 Sequencing of more unicellular eukaryotic mitochondrial genomes would provide insight
549 into the origin of this ORF.
550
551 ***Change in mutation spectrum and nucleotide composition***
552 We used a combination of previous mutation-accumulation studies and our population genomic
553 data to suggest that the change in nucleotide composition of *P. aurelia* mitochondrial genomes
554 toward higher GC is a result of change in mutational bias. Unfortunately, due to our modest
555 sample sizes, some of the singleton SNPs in our analysis are present at fairly high frequencies in
556 the population and thus might be influenced by selection. Mutation-accumulation lines in *P.*
557 *caudatum* would help further disentangle these forces. Because ciliates have very low mutation
558 rates, their mutation-accumulation study requires a large number of generations to produce only
559 a handful of mutations, making it very difficult to obtain precise estimates of the mutation
560 spectrum. The most feasible strategy to more precisely estimate these spectra in *Paramecium*
561 would thus be to produce larger population samples in order to observe lower-frequency variants.
562 We also note that although an alternative explanation for higher GC content in the *P. aurelia*
563 species could be biased gene conversion, we found no evidence of mitochondrial recombination
564 (including non-crossovers) in the *P. aurelia* species.
565
566 An interesting question raised by our results is how fast the mutation spectrum in mitochondria
567 evolves across species (LYNCH *et al.* 2008; MONTOOTH and RAND 2008). Re-analyzing data from
568 previous work on mutation-accumulation lines in mitochondria of other model organisms

569 presents a number of interesting observations (Table 1). First, within opisthokonts, there is a
570 huge variation in mutation bias (*m*) in the mitochondria, ranging from nearly 0 (strongly biased
571 towards GC) to values much larger than 1 (biased towards AT; Table 1). However, most species
572 except *S. cerevisiae* have a mutation bias towards A/T, consistent with most mitochondrial
573 genomes being AT-rich. Second, across opisthokonts, we find that selection can have a
574 significant impact on mitochondrial GC content. The effect of selection (*S*) can be observed by
575 comparing observed genome-wide nucleotide composition (or specifically at 4-fold degenerate
576 sites) with the expected composition under mutation equilibrium (Table 1). Again, in most
577 species selection favors higher G/C, but in *S. cerevisiae* there appears to be strong selection
578 favoring A/T genome-wide. Thus there might not be a universal direction for mutation bias or
579 selection for nucleotide composition in mitochondria. Finally, closely related species can have
580 significant changes in mutation bias in the mitochondria (HOWE *et al.* 2010; KONRAD *et al.* 2017)
581 as observed between *C. elegans* and *C. briggsae*. All of these observations suggest that it is
582 biologically reasonable that there have been changes in mutational biases towards AT along the
583 long branch leading to the *P. aurelia* species.
584
585 It has been observed that higher AT content in mitochondrial genomes is usually accompanied
586 by the presence of more tRNA genes (SMITH 2012). This correlation holds in ciliates as well,
587 where the *P. aurelia* species with higher GC content only possess 3 tRNA genes relative to other
588 *Paramecium* species that possess many more tRNA genes. Although we have shown that
589 there have been multiple gains and losses of tRNA genes in *Paramecium* species, the
590 source of these new tRNA genes is not evident. It may be possible for new tRNA gene
591 copies to be transferred from the nucleus to the mitochondrial genome. Phylogenetic

592 analyses and mito-nuclear comparisons may illuminate the mutational and selective
593 forces driving mitochondrial tRNA gain and loss.
594
595 ***Similar efficacy of purifying selection experienced by the mitochondria and nucleus***
596 Mitochondrial genomes are often non-recombining and are usually passed on via
597 uniparental inheritance. Mitochondria are therefore expected to have lower effective
598 population sizes than that of the nucleus within the same organism (LYNCH and
599 BLANCHARD 1998; NEIMAN and TAYLOR 2009). One consequence of reduced effective
600 population sizes is an increased probability of segregation and even fixation of slightly
601 deleterious mutations. There have been multiple contradictory reports on whether
602 mitochondrial genes experience stronger or weaker purifying selection than the nuclear
603 genes. Studies examining a few protein-coding loci but with larger number of species
604 have concluded that mitochondria experience less effective purifying selection than the
605 nucleus (WEINREICH and RAND 2000; BETANCOURT *et al.* 2012; POPADIN *et al.* 2013).
606 However a recent study conducted on multiple individuals and whole genomes in
607 *Drosophila melanogaster* and humans found no significant difference between the
608 efficacy of purifying selection in the mitochondrial vs. nuclear protein-coding genes
609 (COOPER *et al.* 2015).
610
611 We confirm the absence of recombination in *Paramecium* mitochondria, which was
612 previously hinted at, but restricted to a study of a single gene, cytochrome b (BARTH *et*
613 *al.* 2008). We also evaluated the extent of purifying selection in ciliate mitochondria
614 relative to the nucleus for 4 species: *P. tetraurelia*, *P. sexaurelia*, *P. caudatum*, and *P.*

615 *multimicronucleatum*. We find that despite the lack of recombination in the
616 mitochondria, the efficacy of selection is either similar to or stronger in the
617 mitochondria relative to the nucleus.
618
619 Although our results seem to be in discordance with theoretical predictions, they are not
620 necessarily surprising. One possibility is that heteroplasmy and bottlenecks during
621 transmission of mitochondria can result in proliferation of a random subset of
622 deleterious mutations, which then get exposed to cell-level selection (STEWART *et al.*
623 2008; WAI *et al.* 2008; STEWART and LARSSON 2014), eventually leading to stronger
624 purifying selection observed. Another possibility is that we observe stronger efficacy of
625 selection in the mitochondria because the magnitude of negative selection itself is
626 stronger in the mitochondrial genes (POPADIN *et al.* 2013), because they differ from the
627 average nuclear gene in several key ways. Most mitochondrial genes are expressed at
628 high levels compared to the nucleus (HAVIRD and SLOAN 2016). Mitochondrial encoded
629 proteins (cox1 and cox2) that are part of the oxidative phosphorylation (OXPHOS)
630 pathway are core enzyme catalytic subunits (TSUKIHARA *et al.* 1996; ZHANG and
631 BROUGHTON 2013; HAVIRD and SLOAN 2016). And, most genes retained in the
632 mitochondria encode for highly hydrophobic proteins and high GC content (JOHNSTON
633 and WILLIAMS 2016). In order to test theoretical predictions successfully, we stress the
634 importance of comparing equivalent sets of genes—finding comparable sets of genes
635 between these two genomes may be difficult (but see LYNCH 1996; LYNCH 1997).
636 Although we made efforts on this front, by also conducting all analyses with nuclear
637 genes involved in the same pathway as the mitochondrial genes, most of the caveats

638 described above continue to affect our results.

639

640 Lastly, our results need to be evaluated in the light of *Paramecium*-specific life cycle. In
641 *Paramecium*, there is cytoplasmic inheritance during conjugation (KOIZUMI and
642 KOBAYASHI 1989) and there is almost no exchange of cytoplasm during conjugation
643 (reviewed in MEYER and GARNIER 2002). Thus it appears that both parents pass on their
644 mitochondria to their offspring, without the degradation of mitochondrial genomes from
645 any one lineage. On the other hand *Paramecium* species frequently undergo asexual
646 reproduction, thus reducing the nuclear effective population size. Therefore, effective
647 population sizes of mitochondria might only be slightly lower or perhaps similar to that
648 of the nucleus. Due to asexual reproduction, under equilibrium conditions, expected
649 heterozygosity in the nucleus would fall between $4N_e\mu$ - $2N_e\mu$. Similarly in the
650 mitochondria, under equilibrium conditions, expected hetrozygosity would be $\sim 2N_e\mu$.
651 Assuming that the ratio of divergence at silent sites can be used as a proxy for the ratio
652 of mutation rate between the two compartments, it is possible to approximately estimate
653 the ratio of effective population sizes of the two compartments as $N_{e(m)}/N_{e(n)} = y \times$
654 $(\pi_m/\pi_n)/(d_m/d_n)$, where y would be some number between 1 and 2. For both *P.*
655 *tetraurelia* and *P. sexaurelia*, for which we have relatively closer outgroup species,
656 $N_{e(m)}/N_{e(n)}$ turns out to be between 0.94-1.88 and 6.58-13.16 respectively. Although
657 underestimation of neutral divergence in the mitochondria relative to the nucleus could
658 skew our inference, the above calculations suggest that perhaps effective population
659 sizes of mitochondria are larger than that of the nucleus in *Paramecium*.

660

661 We therefore conclude, that our results might continue to be consistent with theoretical
662 predictions. Unicellular eukaryotes, like *Paramecium* with their unique life cycle, might
663 represent a case where the effective population size of the mitochondria is much larger
664 than that of the corresponding nucleus and concordantly we observe stronger purifying
665 selection in the mitochondria relative to the nucleus.
666
667
668 **METHODS:**
669
670 **Genome sequencing and assembly**
671 DNA extraction, sequencing library preparation and genome sequencing were previously
672 described (Johri et al. 2017). Sequencing reads were assembled using SPAdes (BANKEVICH *et al.*
673 2012; version 3.5.0) after removing potential adapter sequence with Trimmomatic (BOLGER *et al.*
674 2014; version 0.33). Mitochondrial contigs were identified from the resulting assemblies by
675 BLAST (ALTSCHUL *et al.* 1997) searches against the published *P. caudatum* and *P. tetraurelia*
676 mitochondrial genomes.
677
678 **Mitochondrial genome annotation**
679 Genome annotation was carried out as follows. Protein coding genes were identified by
680 generating all ORFs longer than 60aa in all six frames, using the Mold, Protozoan, and
681 Coelenterate Mitochondrial Code (i.e., UGA codes for W instead of being a stop codon) and all
682 alternative start codons specific to *Paramecium* (AUU, AUA, AUG, AUC, GUG, and GUA),
683 and retaining the longest ORFs associated with each stop codon. BLASTP was then used to

684 identify homologs of annotated mitochondrial proteins in *P. tetraurelia* and *P. caudatum*.
685 Additional ORFs were identified by imposing the requirement that their length exceeds 100aa,
686 and subsequently annotated using BLASTP against the Non-redundant protein sequences (nr)
687 database and HMMER3.0 (EDDY 2011) scans against the PFAM 27.0 database (FINN *et al.*
688 2014). tRNA genes were annotated with tRNAscan-SE (SCHATTNER *et al.* 2005; version 1.21),
689 using the "Mito/Chloroplast" source. rRNA genes were identified using Infernal (NAWROCKI
690 *et al.* 2009; version 1.1.1).

691

692 **Identification of telomeres**

693 Telomeric repeats were identified as follows. The first and the last 200bp of each mitogenome
694 were used as input to the MEME *de novo* motif finding program (BAILEY *et al.* 2009; version
695 4.6.1), which was run with the following parameters: "*-maxw 25 -dna -nmotifs 5 -mod anr*". The
696 repetitive units defined that way were then manually aligned to each other and refined to arrive at
697 final telomeric repeats comparable across all species.

698

699 **Expression level in mitochondria**

700 For each species, sequencing reads were aligned against a combined Bowtie (LANGMEAD *et al.*
701 2009) index containing both the nuclear and mitochondrial genomes using TopHat2 (KIM *et al.*
702 2013; version 2.0.8) with the following settings: "*--bowtie1 --no-discordant --no-mixed --*
703 *microexon-search --read-realign-edit-dist 0 --read-edit-dist 4 --read-mismatches 4 --min-intron-*
704 *length 10 --max-intron-length 1000000 --min-segment-intron 10 --min-coverage-intron 10*".
705 Custom python scripts were then used to generate normalized coverage tracks over the
706 mitogenomes.

707

708 **Estimation of dN/dS, D_n, D_s, P_n, P_s, and π_n/π_s**

709 dN/dS was estimated either across the phylogeny using CODEML, PAML (YANG 2007;
710 version 4.9a) and for the *P. aurelia* species it was also estimated pairwise w.r.t the
711 closest outgroup species using yn00, PAML. π_n/π_s was obtained for all protein coding
712 genes in the 4 species- *P. tetraurelia*, *P. sexaurelia*, *P. caudatum*, and *P.*
713 *multimicronucleatum*, where total number of changes in synonymous sites was greater
714 than 1. This filter was executed in order to reduce errors in π_n/π_s due to very low values
715 of synonymous polymorphisms in a gene.

716

717 D_n and D_s, the number of nonsynonymous and synonymous changes, were inferred by
718 performing ancestral reconstruction at each site, and then counting branching-specific
719 substitutions. Ancestral reconstruction (PAML, baseml, GTR model) was conducted
720 over the phylogeny of all available taxa. For comparing statistics between mitochondrial
721 and nuclear genes, the ancestral reconstruction was performed over the same set of taxa
722 for both, i.e. over *P. tetraurelia*, *P. sexaurelia*, *P. caudatum*, and *P.*
723 *multimicronucleatum*. For all analyses that involved ancestral reconstruction, only sites
724 whose posterior probability of the inferred ancestral state ≥ 0.85 were used in the
725 analyses—this filter was imposed for counting synonymous and nonsynonymous
726 polymorphisms (P_s and P_n, respectively) as well as divergent sites (D_n and D_s). There
727 may be some concern that ancestral reconstruction could end up biasing the ratio of
728 D_n/D_s as many more changes at synonymous sites might result in lower confidence in
729 inferring ancestral states at synonymous but not nonsynonymous sites. Such a bias

730 would increase D_n relative to D_s and thus decrease values of NI. Therefore, all analyses
731 involving D_n and D_s were also performed including all sites, with no filter, and results
732 remained unchanged.
733
734 **Calculation of multiple estimators of Neutrality Index (NI) and statistical tests**
735 Several estimators of neutrality index have been proposed in order to counter different biases.
736 The estimators we used were calculated as follows:
737
738 Simple neutrality index, NI = $P_n/P_s/D_n/D_s$ (RAND and KANN 1996).
739 NI_{π} = $\pi_n/\pi_s/d_N/d_S$, where d_N/d_S was calculated pairwise, w.r.t closest outgroup species
740 (BETANCOURT *et al.* 2012).
741 $NI_{TG} = \Sigma_i[D_{si}.P_{ni}/(P_{si} + D_{si})] / \Sigma_i [P_{si}.D_{ni}/(P_{si} + D_{si})]$, where *i* is the *i*th gene (TARONE 1981;
742 GREENLAND 1982b).
743 $NI_{DoS} = D_n/(D_n+D_s) - P_n/(P_n+P_s)$ (STOLETZKI and EYRE-WALKER 2011).
744
745 For McDonald Kreitman test (MCDONALD and KREITMAN 1991)**,** Fisher's exact test in R (R-
746 CORE-TEAM 2014) was used to test significance. In all cases p value was corrected by
747 Bonferonni-Holm method (HOLM 1979) for multiple tests, using R.
748
749 **Mutation spectrum from mutation accumulation lines**
750 MA line experiments for *P. tetraurelia* had previously been published (SUNG *et al.* 2012) and
751 results from their analysis were used directly in this study.
752

753 MA experiments to obtain nuclear mutation rates in *P. biaurelia* and *P. sexaurelia* are currently
754 unpublished (LONG *et al.* 2018a). Data from their analysis were used to obtain mitochondrial
755 mutations as follows:
756
757 Sequencing reads were assembled for each MA line individually and mitochondrial contigs
758 identified as described above. A composite consensus mitogenome sequence was then created
759 from the individual assemblies by creating multiple sequence alignments of all mitochondrial
760 contigs using MAFFT (KATOH and STANDLEY 2013; version 7.221) and retaining the most
761 frequent base for each alignment column (with the exception of telomeres, which were
762 manually curated). Reads were then aligned in a 2×100bp format against a combined Bowtie
763 index, containing a combination of the nuclear and consensus mitochondrial genomes, allowing
764 for up to 3 mismatches and retaining only unique reads. Putative mutations were identified by
765 requiring that any variant is supported by at least 3 non-redundant read pairs on each strand, by
766 not more than 4 times more reads on one strand than on the other, and is also observed in ≥5%
767 of reads covering a given position were included. Telomeric sequences were excluded due to an
768 excessively high number of sequence variants observed in those regions. Mutation rate of A/T
769 → G/C (u) and G/C → A/T (v) mutations was calculated as follows:
770 u = (number of A/T → G/C mutations) / (total number of utilizable A/T sites)
771 v = (number of G/C → A/T mutations) / (total number of utilizable G/C sites)
772
773 **Mutation spectrum from population genomics data**
774 For each SNP, the ancestral allele was inferred by performing ancestral reconstruction on the
775 13-taxa phylogeny to predict the nucleotides on internal nodes (see above). The ancestral allele

776 was used to infer the derived allele segregating in *P. tetraurelia*, *P. sexaurelia*, *P. caudatum* and
777 *P. multimicronucleatum*. This analysis was restricted to sites where the ancestral nucleotide was
778 inferred with confidence score >= 0.90, where the derived allele was present in only a single
779 individual, i.e. was a singleton, and was at 4-fold degenerate site. Of these, we counted all
780 mutations that were from G/C to A/T or from A/T to G/C, with respect to the total number of
781 utilizable sites that were counted according to the same criteria as above. Care was also taken to
782 remove all sites that were part of overlapping ORFs.

783

784 **Recombination analyses**

785 All analyses to detect recombination were restricted to SNPs that were biallelic, homozygous,
786 and had a known ancestral state. The statistic (r^2) to measure linkage disequilibrium was
787 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
788 (McVean *et al.* 2002) was used to infer recombination rates using the permuted composite
789 likelihood test as well other permutation tests. These tests were performed under both the gene
790 conversion (average tract length: 500) and cross-over models with 2 values of θ (= $4N_e\mu$) for
791 each species: the closest allowed θ value lower than that estimated from nucleotide diversity
792 values, and the closest higher value.

793

794 **Identifying nuclear genes belonging to Oxidative Phosphorylation pathway and Ribosomal**
795 **complex**

796 KEGG (Kanehisa *et al.* 2017) was used to obtain the full list of genes that are part of complexes
797 involved in Oxidative Phosphorylation (Complexes I-V) for *P. tetraurelia*. A total number of 87
798 genes in *P. tetraurelia* were obtained, and their corresponding orthologs were identified in other

species. For genes encoding proteins that are part of the ribosomal complex, we used the PANTHER (MI *et al.* 2017) annotation obtained in a previous study (MCGRATH *et al.* 2014b) for all species and selected all genes that were structural constituents of ribosome. This allowed us to start with a set of 585 genes in *P. tetraurelia*, 564 in *P. sexaurelia*, and 213 genes in *P. caudatum*. However, analyses requiring orthologs from all 3 species were conducted with a smaller subset of genes.

ACKNOWLEDGEMENTS:

We thank Jean-Francois Gout to provide us with DNA sequences, as well as for a critical reading of the manuscript. We thank Hongan Long for providing us with sequences from *P. biaurelia* and *P. sexaurelia* nuclear mutation-accumulation experiments. We also thank Jeff Palmer for constructive ideas and discussions about the project. This work was financially supported by the National Science Foundation (MCB-1050161 and DEB-1257806).

DATA AVAILABILITY:

Mitochondrial genome assemblies will be made publicly available upon acceptance of the manuscript. Supplementary material is available at FigShare.

REFERENCES

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389-3402.

BAILEY, T. L., M. BODEN, F. A. BUSKE, M. FRITH, C. E. GRANT *et al.*, 2009 MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res **37:** W202-208.

BALABASKARAN NINA, P., N. V. DUDKINA, L. A. KANE, J. E. VAN EYK, E. J. BOEKEMA *et al.*, 2010 Highly divergent mitochondrial ATP synthase complexes in *Tetrahymena thermophila*. PLoS Biol **8:** e1000418.

BALLARD, J. W., and M. C. WHITLOCK, 2004 The incomplete natural history of mitochondria. Mol Ecol **13:** 729-744.

BANKEVICH, A., S. NURK, D. ANTIPOV, A. A. GUREVICH, M. DVORKIN *et al.*, 2012 SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **19:** 455-477.

BARR, C. M., M. NEIMAN and D. R. TAYLOR, 2005 Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. New Phytol **168:** 39-50.

BARTH, D., and T. U. BERENDONK, 2011 The mitochondrial genome sequence of the ciliate *Paramecium caudatum* reveals a shift in nucleotide composition and codon usage within the genus *Paramecium*. BMC Genomics **12:** 272.

BARTH, D., E. PRZYBOS, S. I. FOKIN, M. SCHLEGEL and T. U. BERENDONK, 2008 Cytochrome b sequence data suggest rapid speciation within the *Paramecium aurelia* species complex. Mol Phylogenet Evol **49:** 669-673.

BETANCOURT, A. J., B. BLANCO-MARTIN and B. CHARLESWORTH, 2012 The relation between the neutrality index for mitochondrial genes and the distribution of mutational effects on fitness. Evolution **66:** 2427-2438.

BOLGER, A. M., M. LOHSE and B. USADEL, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30:** 2114-2120.

BORZA, T., E. K. REDMOND, M. LAFLAMME and R. W. LEE, 2009 Mitochondrial DNA in the *Oogamochlamys* clade (*Chlorophyceae*): high Gc content and unique genome architecture for Green algae(1). J Phycol **45:** 1323-1334.

BURGER, G., L. FORGET, Y. ZHU, M. W. GRAY and B. F. LANG, 2003 Unique mitochondrial genome architecture in unicellular relatives of animals. Proc Natl Acad Sci U S A **100:** 892-897.

BURGER, G., S. MOREIRA and M. VALACH, 2016 Genes in Hiding. Trends Genet **32:** 553-565.

BURGER, G., Y. ZHU, T. G. LITTLEJOHN, S. J. GREENWOOD, M. N. SCHNARE *et al.*, 2000 Complete sequence of the mitochondrial genome of *Tetrahymena pyriformis* and comparison with *Paramecium aurelia* mitochondrial DNA. J Mol Biol **297:** 365-380.

COOPER, B. S., C. R. BURRUS, C. JI, M. W. HAHN and K. L. MONTOOTH, 2015 Similar efficacies of selection shape mitochondrial and nuclear genes in both *Drosophila melanogaster* and *Homo sapiens*. G3 **5:** 2165-2176.

COYNE, R. S., L. HANNICK, D. SHANMUGAM, J. B. HOSTETLER, D. BRAMI *et al.*, 2011 Comparative genomics of the pathogenic ciliate *Ichthyophthirius multifiliis*, its free-living relatives and a host species provide insights into adoption of a parasitic lifestyle and prospects for disease control. Genome Biol **12:** R100.

DE GRAAF, R. M., G. RICARD, T. A. VAN ALEN, I. DUARTE, B. E. DUTILH *et al.*, 2011 The organellar genome and metabolic potential of the hydrogen-producing mitochondrion of *Nyctotherus ovalis*. Mol Biol Evol **28:** 2379-2391.

DE GRAAF, R. M., T. A. VAN ALEN, B. E. DUTILH, J. W. KUIPER, H. J. VAN ZOGGEL *et al.*, 2009 The mitochondrial genomes of the ciliates *Euplotes minuta* and *Euplotes crassus*. BMC Genomics **10:** 514.

DENVER, D. R., K. MORRIS, M. LYNCH, L. L. VASSILIEVA and W. K. THOMAS, 2000 High direct estimate of the mutation rate in the mitochondrial genome of *Caenorhabditis elegans*. Science **289:** 2342-2344.

EDDY, S. R., 2011 Accelerated profile HMM searches. PLoS Comput Biol **7:** e1002195.

FINN, R. D., A. BATEMAN, J. CLEMENTS, P. COGGILL, R. Y. EBERHARDT *et al.*, 2014 Pfam: the

874 protein families database. Nucleic Acids Res **42:** D222-230.
875 FOKIN, S. I., and H. D. GORTZ, 2009 Diversity of *Holospora* bacteria in *Paramecium* and their
876 characterization. Microbiol Monogr **12:** 161-199.
877 FRITSCH, E. S., C. D. CHABBERT, B. KLAUS and L. M. STEINMETZ, 2014 A genome-wide map of
878 mitochondrial DNA recombination in yeast. Genetics **198:** 755-771.
879 GODDARD, J. M., and D. J. CUMMINGS, 1975 Structure and replication of mitochondrial DNA
880 from *Paramecium aurelia*. J Mol Biol **97:** 593-609.
881 GRAY, M. W., 2003 Diversity and evolution of mitochondrial RNA editing systems. IUBMB
882 Life **55:** 227-233.
883 GREENLAND, S., 1982a Interpretation and estimation of summary ratios under heterogeneity. Stat
884 Med **1:** 217-227.
885 GREENLAND, S., 1982b Interpretation of summary measures when interaction is present. Am J
886 Epidemiol **116:** 587-587.
887 GRUCHY, D. F., 1955 The breeding system and distribution of *Tetrahymena pyriformis*. J
888 Protozool **2:** 178-185.
889 HAAG-LIAUTARD, C., N. COFFEY, D. HOULE, M. LYNCH, B. CHARLESWORTH *et al.*, 2008 Direct
890 estimation of the mitochondrial DNA mutation rate in Drosophila melanogaster. PLoS
891 Biol **6:** e204.
892 HAFT, D. H., J. SELENGUT, E. F. MONGODIN and K. E. NELSON, 2005 A guild of 45 CRISPR-
893 associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic
894 genomes. PLoS Comput Biol **1:** e60.
895 HAVIRD, J. C., and D. B. SLOAN, 2016 The roles of mutation, selection, and expression in
896 determining relative rates of evolution in mitochondrial versus nuclear genomes. Mol
897 Biol Evol **33:** 3042-3053.
898 HILL, W. G., and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. Theor Appl
899 Genet **38:** 226-231.
900 HOLM, S., 1979 A simple sequentially rejective multiple test procedure. Scand J Stat **6:** 65-70.
901 HORI, M., I. TOMIKAWA, E. PRZYBOS and M. FUJISHIMA, 2006 Comparison of the evolutionary
902 distances among syngens and sibling species of *Paramecium*. Mol Phylogenet Evol **38:**
903 697-704.
904 HOWE, D. K., C. F. BAER and D. R. DENVER, 2010 High rate of large deletions in *Caenorhabditis*
905 *briggsae* mitochondrial genome mutation processes. Genome Biol Evol **2:** 29-38.
906 HUDSON, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:**
907 1805-1817.
908 HUDSON, R. R., and N. L. KAPLAN, 1985 Statistical properties of the number of recombination
909 events in the history of a sample of DNA sequences. Genetics **111:** 147-164.
910 JOHNSTON, I. G., and B. P. WILLIAMS, 2016 Evolutionary inference across eukaryotes identifies
911 specific pressures favoring mitochondrial gene retention. Cell Syst **2:** 101-111.
912 JOHRI, P., S. KRENEK, G. K. MARINOV, T. G. DOAK, T. U. BERENDONK *et al.*, 2017 Population
913 genomics of *Paramecium* species. Mol Biol Evol **34:** 1194-1216.
914 KANEHISA, M., M. FURUMICHI, M. TANABE, Y. SATO and K. MORISHIMA, 2017 KEGG: new
915 perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res **45:** D353-
916 D361.
917 KATOH, K., and D. M. STANDLEY, 2013 MAFFT multiple sequence alignment software version
918 7: improvements in performance and usability. Mol Biol Evol **30:** 772-780.
919 KIM, D., G. PERTEA, C. TRAPNELL, H. PIMENTEL, R. KELLEY *et al.*, 2013 TopHat2: accurate

920 alignment of transcriptomes in the presence of insertions, deletions and gene fusions.
921 Genome Biol **14:** R36.
922 KOIZUMI, S., and S. KOBAYASHI, 1989 Microinjection of plasmid DNA encoding the A surface
923 antigen of *Paramecium tetraurelia* restores the ability to regenerate a wild-type
924 macronucleus. Mol Cell Biol **9:** 4398-4401.
925 KONRAD, A., O. THOMPSON, R. H. WATERSTON, D. G. MOERMAN, P. D. KEIGHTLEY *et al.*, 2017
926 Mitochondrial mutation rate, spectrum and heteroplasmy in *Caenorhabditis elegans*
927 spontaneous mutation accumulation lines of differing population size. Mol Biol Evol **34:**
928 1319-1334.
929 LANGMEAD, B., C. TRAPNELL, M. POP and S. L. SALZBERG, 2009 Ultrafast and memory-efficient
930 alignment of short DNA sequences to the human genome. Genome Biol **10:** R25.
931 LONG, H., T. G. DOAK and M. LYNCH, 2018a Consistently low base-substitution mutation rates
932 within the *Paramecium aurelia* species complex.
933 LONG, H., W. SUNG, S. KUCUKYILDIRIM, E. WILLIAMS, S. F. MILLER *et al.*, 2018b Evolutionary
934 determinants of genome-wide nucleotide composition. Nature ecology & evolution.
935 LYNCH, M., 1996 Mutation accumulation in transfer RNAs: molecular evidence for Muller's
936 ratchet in mitochondrial genomes. Mol Biol Evol **13:** 209-220.
937 LYNCH, M., 1997 Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA
938 genes. Mol Biol Evol **14:** 914-925.
939 LYNCH, M., 2007 *The origins of genome architecture*. Sinauer Associates, Sunderland (MA).
940 LYNCH, M., and J. L. BLANCHARD, 1998 Deleterious mutation accumulation in organelle
941 genomes. Genetica **102-103:** 29-39.
942 LYNCH, M., B. KOSKELLA and S. SCHAACK, 2006 Mutation pressure and the evolution of
943 organelle genomic architecture. Science **311:** 1727-1730.
944 LYNCH, M., W. SUNG, K. MORRIS, N. COFFEY, C. R. LANDRY *et al.*, 2008 A genome-wide view
945 of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A **105:** 9272-
946 9277.
947 MACKENZIE, S. A., 2007 The unique biology of mitochondrial genome instability in plants, pp.
948 36–46 in *Plant Mitochondria*, edited by D. LOGAN. Blackwell Publishing, Oxford.
949 MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the Adh locus in
950 *Drosophila*. Nature **351:** 652-654.
951 MCGRATH, C. L., J. F. GOUT, T. G. DOAK, A. YANAGI and M. LYNCH, 2014a Insights into three
952 whole-genome duplications gleaned from the *Paramecium caudatum* genome sequence.
953 Genetics **197:** 1417-1428.
954 MCGRATH, C. L., J. F. GOUT, P. JOHRI, T. G. DOAK and M. LYNCH, 2014b Differential retention
955 and divergent resolution of duplicate genes following whole-genome duplication.
956 Genome Res **24:** 1665-1675.
957 MCVEAN, G., P. AWADALLA and P. FEARNHEAD, 2002 A coalescent-based method for detecting
958 and estimating recombination from gene sequences. Genetics **160:** 1231-1241.
959 MEUNIER, J., and A. EYRE-WALKER, 2001 The correlation between linkage disequilibrium and
960 distance: implications for recombination in hominid mitochondria. Mol Biol Evol **18:**
961 2132-2135.
962 MEYER, E., and O. GARNIER, 2002 Non-Mendelian inheritance and homology-dependent effects
963 in ciliates. Advances in genetics **46:** 305-337.
964 MI, H., X. HUANG, A. MURUGANUJAN, H. TANG, C. MILLS *et al.*, 2017 PANTHER version 11:
965 expanded annotation data from Gene Ontology and Reactome pathways, and data

966 analysis tool enhancements. Nucleic Acids Res **45:** D183-D189.
967 MONTOOTH, K. L., and D. M. RAND, 2008 The spectrum of mitochondrial mutation differs across
968 species. PLoS Biol **6:** e213.
969 MORADIAN, M. M., D. BEGLARYAN, J. M. SKOZYLAS and V. KERIKORIAN, 2007 Complete
970 mitochondrial genome sequence of three *Tetrahymena* species reveals mutation hot spots
971 and accelerated nonsynonymous substitutions in Ymf genes. PLoS ONE **2:** e650.
972 MORIN, G. B., and T. R. CECH, 1988 Mitochondrial telomeres: surprising diversity of repeated
973 telomeric DNA sequences among six species of *Tetrahymena*. Cell **52:** 367-374.
974 NAWROCKI, E. P., D. L. KOLBE and S. R. EDDY, 2009 Infernal 1.0: inference of RNA alignments.
975 Bioinformatics **25:** 1335-1337.
976 NEIMAN, M., and D. R. TAYLOR, 2009 The causes of mutation accumulation in mitochondrial
977 genomes. Proc. R. Soc. B **276:** 1201-1209.
978 ORR, A. T., J. C. RABETS, T. L. HORTON and L. F. LANDWEBER, 1997 RNA editing missing in
979 mitochondria. Rna **3:** 335-336.
980 PIGANEAU, G., and A. EYRE-WALKER, 2004 A reanalysis of the indirect evidence for
981 recombination in human mitochondrial DNA. Heredity **92:** 282-288.
982 POMBERT, J.-F., A. SMIRNOV, E. R. JAMES, J. JANOUŠKOVEC, M. W. GRAY *et al.*, 2013 The
983 complete mitochondrial genome from an unidentified *Phalansterium* species. Protist
984 Genomics **1:** 25-32.
985 POPADIN, K. Y., S. I. NIKOLAEV, T. JUNIER, M. BARANOVA and S. E. ANTONARAKIS, 2013
986 Purifying selection in mammalian mitochondrial protein-coding genes is highly effective
987 and congruent with evolution of nuclear genes. Mol Biol Evol **30:** 347-355.
988 PREER, L. B., 1969 Alpha, an infectious macronuclear symbiont of *Paramecium aurelia*. J
989 Protozool **16:** 570-578.
990 PRICE, C. A., and E. M. REARDON, 2001 Mendel, a database of nomenclature for sequenced plant
991 genes. Nucleic Acids Res **29:** 118-119.
992 R-CORE-TEAM, 2014 R: A language and environment for statistical computing. **R Foundation**
993 **for Statistical Computing, Vienna, Austria.**
994 RAND, D. M., and L. M. KANN, 1996 Excess amino acid polymorphism in mitochondrial DNA:
995 contrasts among genes from *Drosophila*, mice, and humans. Mol Biol Evol **13:** 735-748.
996 SCHATTNER, P., A. N. BROOKS and T. M. LOWE, 2005 The tRNAscan-SE, snoscan and snoGPS
997 web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res **33:** W686-689.
998 SKIPPINGTON, E., T. J. BARKMAN, D. W. RICE and J. D. PALMER, 2017 Comparative
999 mitogenomics indicates respiratory competence in parasitic Viscum despite loss of
1000 complex I and extreme sequence divergence, and reveals horizontal gene transfer and
1001 remarkable variation in genome size. BMC Plant Biol **17:** 49.
1002 SLABODNICK, M. M., J. G. RUBY, S. B. REIFF, E. C. SWART, S. GOSAI *et al.*, 2017 The
1003 macronuclear genome of *Stentor coeruleus* reveals tiny introns in a giant cell. Current
1004 biology : CB **27:** 569-575.
1005 SMITH, D. R., 2012 Updating our view of organelle genome nucleotide landscape. Front Genet **3:**
1006 175.
1007 SMITH, D. R., 2016 The past, present and future of mitochondrial genomics: have we sequenced
1008 enough mtDNAs? Brief Funct Genomics **15:** 47-54.
1009 SMITH, D. R., and P. J. KEELING, 2015 Mitochondrial and plastid genome architecture:
1010 Reoccurring themes, but significant differences at the extremes. Proc Natl Acad Sci U S
1011 A **112:** 10177-10184.

1012 SODING, J., A. BIEGERT and A. N. LUPAS, 2005 The HHpred interactive server for protein
1013 homology detection and structure prediction. Nucleic Acids Res **33:** W244-248.
1014 SONNEBORN, T. M., 1937 Sex, sex inheritance and sex determination in *Paramecium aurelia*.
1015 Proc Natl Acad Sci U S A **23:** 378-385.
1016 SONNEBORN, T. M., 1975 *Paramecium aurelia* complex of 14 sibling species. Trans Am Microsc
1017 Soc **94:** 155-178.
1018 STADLER, T., and L. F. DELPH, 2002 Ancient mitochondrial haplotypes and evidence for
1019 intragenic recombination in a gynodioecious plant. Proc Natl Acad Sci U S A **99:** 11730-
1020 11735.
1021 STEWART, J. B., C. FREYER, J. L. ELSON, A. WREDENBERG, Z. CANSU *et al.*, 2008 Strong
1022 purifying selection in transmission of mammalian mitochondrial DNA. PLoS Biol **6:** e10.
1023 STEWART, J. B., and N. G. LARSSON, 2014 Keeping mtDNA in shape between generations. PLoS
1024 Genet **10:** e1004670.
1025 STOLETZKI, N., and A. EYRE-WALKER, 2011 Estimation of the neutrality index. Mol Biol Evol
1026 **28:** 63-70.
1027 SUEOKA, N., 1993 Directional mutation pressure, mutator mutations, and dynamics of molecular
1028 evolution. J Mol Evol **37:** 137-153.
1029 SUNG, W., A. E. TUCKER, T. G. DOAK, E. CHOI, W. K. THOMAS *et al.*, 2012 Extraordinary
1030 genome stability in the ciliate *Paramecium tetraurelia*. Proc Natl Acad Sci U S A **109:**
1031 19339-19344.
1032 SWART, E. C., M. NOWACKI, J. SHUM, H. STILES, B. P. HIGGINS *et al.*, 2012 The *Oxytricha*
1033 *trifallax* mitochondrial genome. Genome Biol Evol **4:** 136-154.
1034 TARCZ, S., A. POTEKHIN, M. RAUTIAN and E. PRZYBOS, 2012 Variation in ribosomal and
1035 mitochondrial DNA sequences demonstrates the existence of intraspecific groups in
1036 *Paramecium multimicronucleatum* (Ciliophora, Oligohymenophorea). Mol Phylogenet
1037 Evol **63:** 500-509.
1038 TARONE, R. E., 1981 On summary estimators of relative risk. J Chronic Dis **34:** 463-468.
1039 TSUKIHARA, T., H. AOYAMA, E. YAMASHITA, T. TOMIZAKI, H. YAMAGUCHI *et al.*, 1996 The
1040 whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A. Science **272:**
1041 1136-1144.
1042 VLCEK, C., W. MARANDE, S. TEIJEIRO, J. LUKES and G. BURGER, 2011 Systematically
1043 fragmented genes in a multipartite mitochondrial genome. Nucleic Acids Res **39:** 979-
1044 988.
1045 WAI, T., D. TEOLI and E. A. SHOUBRIDGE, 2008 The mitochondrial DNA genetic bottleneck
1046 results from replication of a subpopulation of genomes. Nat Genet **40:** 1484-1488.
1047 WAKELEY, J., 1997 Using the variance of pairwise differences to estimate the recombination rate.
1048 Genet Res **69:** 45-48.
1049 WEINREICH, D. M., and D. M. RAND, 2000 Contrasting patterns of nonneutral evolution in
1050 proteins encoded in nuclear and mitochondrial genomes. Genetics **156:** 385-399.
1051 XU, S., S. SCHAACK, A. SEYFERT, E. CHOI, M. LYNCH *et al.*, 2012 High mutation rates in the
1052 mitochondrial genomes of Daphnia pulex. Mol Biol Evol **29:** 763-769.
1053 YANG, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24:**
1054 1586-1591.
1055 ZHANG, F., and R. E. BROUGHTON, 2013 Mitochondrial-nuclear interactions: compensatory
1056 evolution or variable functional constraint among vertebrate oxidative phosphorylation
1057 genes? Genome Biol Evol **5:** 1781-1791.

FIGURES and TABLES

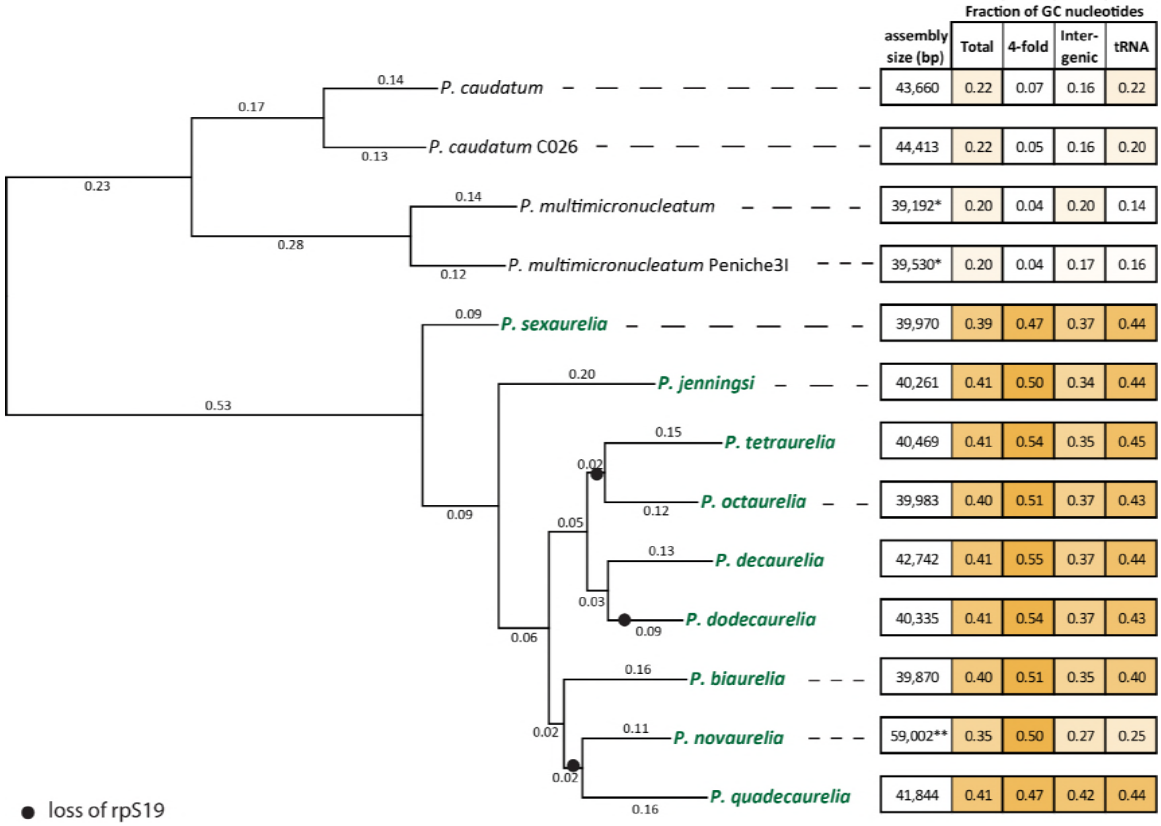

Figure 1: Mitochondrial phylogeny and change in nucleotide composition across *Paramecium* species. The species in green belong to the *P. aurelia* complex. Numbers on branches indicate total number of substitutions per site. Black solid circles show the inferred event of loss of the ribosomal protein *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.

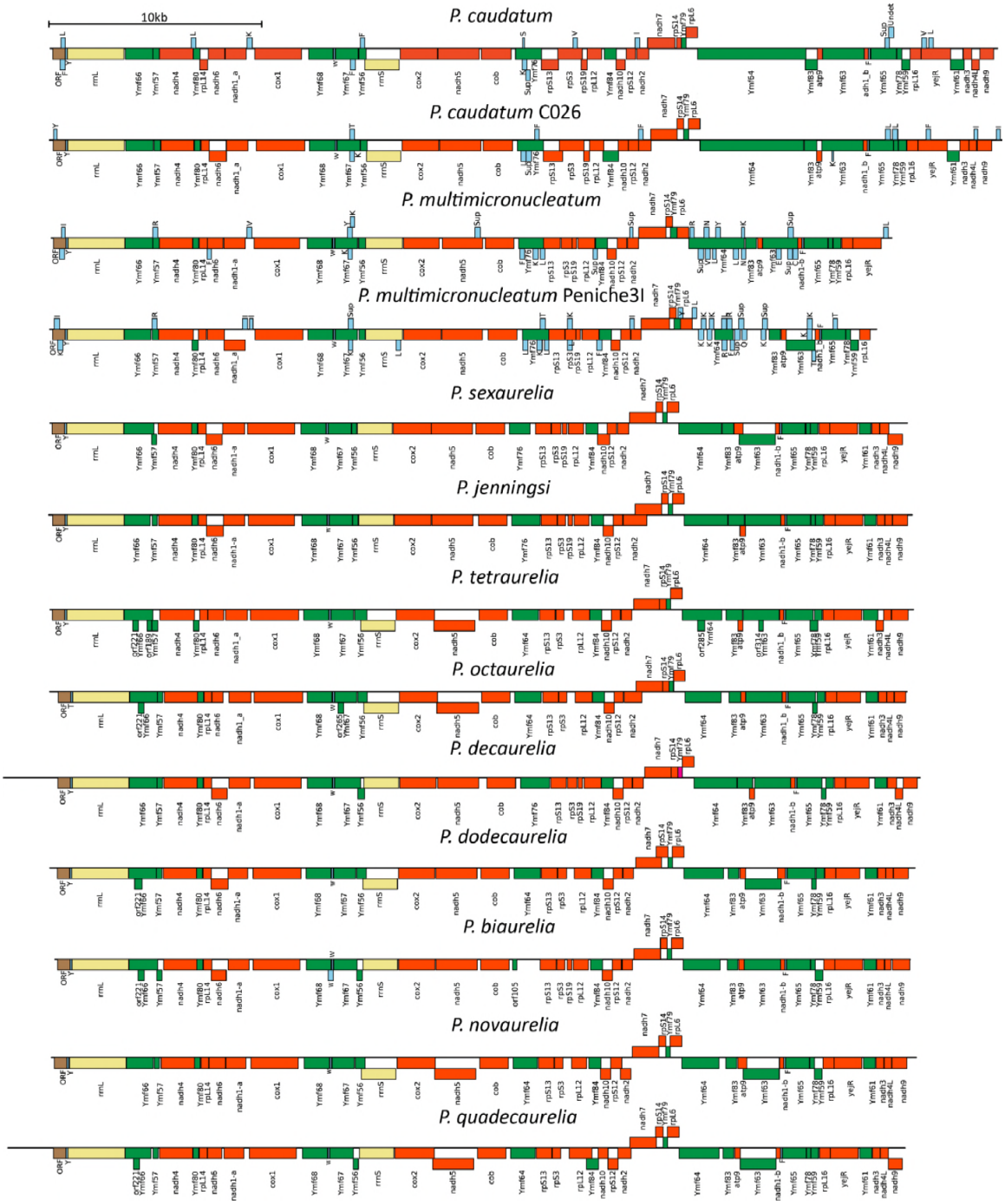

Figure 2: Structure of mitochondrial genomes in *Paramecium*. All *Paramecium* mitochondrial genomes are linear with 24 protein-coding genes (shown in reddish orange), 2 rRNA genes (yellow), lineage-specific genes referred to as *Ymf* genes (in green) and varying numbers and content of tRNA genes (shown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

AA	Anticodon	*Tetrahymena* species	*P. multimicronucleatum* (reference)	*P. multimicronucleatum* (strain: M05)	*P. multimicronucleatum* (strain: M13)	*P. multimicronucleatum-Peniche3I*	*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 3: 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 Bulmer equation, $P_{AT} = 1/(1 + me^{-S})$, 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	MA	55	0.79	0.56	0.40	0.51	0.278	0.085
P. sexaurelia	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[2]	MA	13	0.00	1.00	0.17	0.08	>> 0	>> 0
Caenorhabditis elegans[3,4]	MA	25	3.79	0.21	0.24	0.14	-0.073	0.224
Daphnia pulex[5]	MA	6	2.48	0.29	0.38	0.30	-0.182	-0.026
Caenorhabditis briggsae[6]	MA	15	18.10	0.05	0.25	0.14	-0.780	-0.451
Drosophila melanogaster[7]	MA	28	381.25	0.00	0.18	0.06	-1.923	-1.386

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

Table 2: Neutrality index (NI) of mitochondrial and nuclear genes, calculated using multiple statistics. All *p* values represent comparisons between nuclear and mitochondrial NIs within the same species. Statistically significant scores ($p < 0.05$) are underlined.

	P. tetraurelia	***P. sexaurelia***	***P. caudatum***
NI [$(P_N/P_S)/(D_N/D_S)$]			
nuclear	2.145	0.810	0.314
mitochondrial	1.528	0.351	0.324
mitochondrial excluding Ymf	1.855	0.430	0.281
p	0.272	<u>0.005</u>	0.141
p (excluding Ymf)	0.817	<u>0.007</u>	0.047
NI DoS [$(D_N/(D_N + D_S)) - (P_N/(P_N + P_S))$]			
nuclear	-0.041	0.125	0.286
mitochondrial	0.026	0.268	0.373
mitochondrial excluding Ymf	0.024	0.229	0.428
p	0.272	<u>0.001</u>	0.260
p (excluding Ymf)	0.738	0.272	0.094
NIπ [$(\pi_N/\pi_S)/(d_N/d_S)$]			
nuclear	4.018	4.639	-
mitochondrial	2.294	2.353	-
mitochondrial excluding Ymf	2.495	3.419	-
p	0.124	<u>3.52×10^{-10}</u>	-
p (excluding Ymf)	0.537	<u>2.89×10^{-5}</u>	-