Investigation

Mutational Biases and Selection in Mitochondrial Genomes: Insights from a Comparative Analysis of Natural and Experimental Populations of *Caenorhabditis elegans*

Alexandre Schifano, Ulfar Bergthorsson*, Vaishali Katju*

Program in Evolutionary Biology, Department of Ecology and Genetics (IEG), Uppsala University, 752 36 Uppsala, Sweden

**Joint corresponding authors:*
Vaishali Katju: vaishali.katju@ebc.uu.se
Ulfar Bergthorsson: ubergthorsson@gmail.com

Running Title:

mtDNA variation in *C. elegans* natural isolates

Keywords: C. elegans, mtDNA, natural isolates, selection, heteroplasmy, codon usage bias

ABSTRACT

Spontaneous mutations display biases in their relative frequencies with important consequences for genome structure and composition. While laboratory studies have provided important insights into the spontaneous mutation spectrum, laboratory environments for optimal growth may engender biases that are not representative of natural populations. We analyzed the mitochondrial genomes of 1,524 *Caenorhabditis elegans* natural isolates comprising 550 unique haplotypes to investigate mtDNA polymorphism in the wild. Ancestral reconstruction was used to polarize 2,464 variants (88 indels, 2,376 SNPs) and the results were compared to mutations identified in experimental lines under relaxed selection. MtDNA variant distribution in natural isolates is strongly dependent on site-degeneracy in a manner consistent with purifying selection. There is significant variation in the synonymous and nonsynonymous polymorphism between genes. Specifically, ETC complex I genes are enriched for nonsynonymous polymorphism. The probability of synonymous mutation is higher at sites with flanking G/C nucleotides and the per gene synonymous polymorphism is negatively correlated with A+T-content at the 1st and 2nd codon positions. Furthermore, the 5′ and 3′ ends of genes have both higher A+T-content and less synonymous polymorphism than central regions. There is evidence of natural selection for preferred codons. We identify the first cases of large heteroplasmic mtDNA structural variants in *C. elegans* natural isolates, comprising deletions and duplications. Although some patterns of mtDNA mutational bias are similar between laboratory and natural populations, there exist significant differences. In particular, transversions typically associated with oxidative damage are less common at four-fold degenerate sites in natural populations relative to the laboratory.

58 **SIGNIFICANCE**

59

60 Elements of genome structure, such as base composition and DNA strand bias are in large
61 parts caused by the spectrum of spontaneous mutations. However, laboratory estimates of the
62 mutation spectrum can be biased in an artificial environment that favors fast growth and short
63 generation times. Polarized synonymous and rare variants in the mitochondrial genome of
64 1,524 natural isolates of *Caenorhabditis elegans* were analyzed to test if laboratory estimates
65 of the mutation spectrum are representative of spontaneous mutations in the wild. The
66 genomic distribution of synonymous variants is influenced by (i) base composition
67 differences between and within genes, and (ii) natural selection for preferred codons. Certain
68 categories of mutations, especially G/C → T/A transversions, are underrepresented in natural
69 isolates relative to laboratory experiments. This discrepancy suggests that oxidative DNA
70 damage is not as important a source of mutations in the wild relative to laboratory
71 populations.

72 **INTRODUCTION**

73

74 The acquisition of mitochondria was a crucial stage in eukaryotic evolution,
75 facilitating gains in energy production through oxidative phosphorylation (Lane 2011). Most
76 animal mitochondria are maternally inherited and comprise small compact genomes with
77 higher mutation rates than nuclear DNA (Brown et al. 1979; Ladoukakis and Zouros 2017).
78 Furthermore, animal mtDNA seem to undergo very little to no recombination (Eyre-Walker
79 et al. 1999). Hence, their evolution is essentially clonal and new mitochondrial genotypes
80 (mitotypes) are created by the accumulation of mutations and not by recombination. The
81 study of mtDNA mutation spectra provide insights into the influence of spontaneous
82 mutations on mtDNA genome evolution by helping decipher the origin of observed DNA
83 base composition and strand biases, as well as the respective roles of both natural selection
84 and mutational biases in shaping genomes, genome structure and codon usage. Furthermore,
85 the base composition itself, a consequence of prior mutational biases and selection, can
86 influence mutation rates. Finally, the mutation spectra can enable inferences about possible
87 sources of mtDNA mutations and link them to both intrinsic and environmental factors
88 promoting such mutations. For example, the consequences of reactive oxygen species (ROS)
89 are predicted to have specific mutagenic effects which might be especially relevant when
90 studying mtDNA mutations since ROS are by-products of mitochondrial activity (Kow
91 2002).

92

93 Despite a plethora of mutation accumulation (MA henceforth) which serve to
94 minimize the influence of natural selection on newly originating variants, only a small subset
95 have analyzed the rate and spectrum of spontaneous mutations of mitochondrial origin
96 (Denver et al. 2000; Haag-Liautard et al. 2008; Lynch et al. 2008; Howe et al. 2009; Molnar
97 et al. 2011; Saxer et al. 2012; Sung et al. 2012; Xu et al. 2012; Konrad et al. 2017; reviewed
98 in Katju and Bergthorsson 2019). Because eukaryotic cells can host up to thousands of
99 mtDNA copies, spontaneous mtDNA mutations will therefore initially only be present at a
100 very low intracellular frequency compared to novel nuclear mutations. This difference makes
101 the detection of novel mtDNA variants particularly challenging with the use of common
102 DNA sequencing methods and necessitating extensive MA experiments over multiple
103 generations to allow new mtDNA variants to increase to detectable frequencies. The fate of
104 most spontaneous mutations is loss by genetic drift. Hence, even in MA experiments

105 maintained across multiple parallel lines for multiple generation, the number of detectable
106 mutations remains low with regular sequencing methods (Sanger or Illumina HiSeq),
107 compromising their statistical power to analyze the mtDNA mutation spectrum. Furthermore,
108 the diminutive size of animal mtDNA genomes exacerbates this challenge of detecting novel
109 mtDNA genomes in MA experiments even though animal mtDNA mutation rates can be
110 several orders of magnitude greater than their nuclear counterparts (Konrad et al. 2017).
111
112 Several studies have attempted to measure the mtDNA mutation spectrum in the
113 model nematode, *Caenorhabditis elegans*, often employing an MA approach in order to
114 minimize the influence of natural selection (Denver et al. 2000; Konrad et al. 2017; Waneka
115 et al. 2021; Leuthner et al. 2022), but the results have varied substantially between studies.
116 Highly sensitive sequencing methods such as duplex sequencing enable the detection of even
117 the rarest variants with high fidelity (Schmitt et al. 2012) and were employed by two recent
118 studies focusing on *C. elegans* mtDNA mutations (Waneka et al. 2021; Leuthner et al. 2022).
119 Although duplex sequencing allowed for a far greater yield of spontaneous mtDNA mutations
120 in these two recent studies, their results differed somewhat despite the employment of the
121 same sequencing technology. The results from duplex sequencing reported by Waneka et al.
122 (2021) are consistent with the MA experiment in Konrad et al. (2017), with both studies
123 detecting a strong G/C → A/T mutational bias, with G/C → A/T transitions being the most
124 common, followed by G/C → T/A transversions. Waneka et al. (2021) additionally observed
125 a strand bias for suspected oxidative damage on the exposed strand. In contrast, a duplex
126 sequencing study by Leuthner et al. (2022) found that mtDNA mutations increasing A+T-
127 content were predominantly transversions and G/C → C/G mutations, which were rare or
128 absent in preceding studies, were the second most common class of base substitutions. The
129 large number of mtDNA mutations detected in both duplex sequencing studies argues against
130 chance events as being the source of the differences between them. Konrad et al. (2017)
131 additionally analyzed the mtDNA genomes of 38 *C. elegans* wild isolates (Thompson et al.
132 2013) to investigate the spectrum of mtDNA mutations observed under selective conditions
133 in comparison to spontaneous mtDNA mutations originating in their MA experiment under a
134 regime of minimal selection.
135
136 The signature of oxidative damage across several laboratory-based *C. elegans* studies
137 may appear to support the hypothesis that mitochondrial ROS are a major contributor to

138 mtDNA mutations. Indeed, one hypothesis regarding oxidative damage posits that higher
139 metabolic demands leading to increase ATP production may contribute to the production of
140 ROS (Gillooly et al. 2005) and therefore increase G/C → T/A transversions. Can optimal
141 growth conditions for *C. elegans* in the laboratory be a factor in promoting oxidative damage
142 relative to wild populations? Our understanding of the mtDNA mutation rate in *C. elegans* is
143 primarily based on laboratory experiments with worms cultivated at the maximum rate in a
144 predictable environment (Denver et al. 2000; Konrad et al. 2017; Waneka et al. 2021;
145 Leuthner et al. 2022). In contrast, wild worms can have vastly different growth rates and
146 generation times, and experience a diversity of growth conditions and environmental stressors
147 not encountered under laboratory conditions. These in turn may influence DNA replication
148 and repair resulting in mutation patterns that differ from laboratory evolved lines. One of the
149 main objectives of this study is to test if the spectrum of mtDNA mutations in natural
150 populations of *C. elegans* differs from that of laboratory studies. To this end, we examined
151 the mtDNA genomic sequences of 1,524 natural isolates of *C. elegans* from the CaeNDR
152 database (Crombie et al. 2024) representing 550 genetically distinct isotypes clustered into
153 347 unique mitotypes. We employed a novel programming algorithm for high accuracy
154 ancestral sequence reconstruction (ASR) for base substitutions and small indel changes
155 (Jowkar et al. 2022) in order to identify and polarize mtDNA variants occurring in these *C.*
156 *elegans* natural isolates. Natural populations are subject to stronger purifying selection than
157 laboratory populations, but with a large data set which includes 900 mutations at four-fold
158 degenerate sites, we demonstrate significant differences in the mutation spectrum between
159 natural and laboratory populations. In addition to conducting tests for codon usage bias,
160 neighbor effect, and positive selection, we additionally identify the first instances of several
161 large mitochondrial deletions and duplications in natural populations of *C. elegans*.
162

163 ## RESULTS

164

165 We identified and polarized mtDNA mutations from 531 distinct mtDNA isotypes
166 (550 minus the 19 strains flagged for possible mtDNA contamination) corresponding to 1,524
167 natural isolates of *C. elegans* (supplementary file S1) using a method combining phylogeny
168 and ancestral sequence reconstruction. With the exception of two low-complexity regions
169 comprising 466 bp in the reference mitochondrial genome, the entire mitochondrial genome
170 (13,315 bp in the Bristol N2 laboratory strain) was analysed. We identified 2,464 mutations

171 comprising 88 small indels and 2,376 SNPs (supplementary file S2). Approximately 46%
172 (1,125) of these mutations were inferred between nodes within the phylogeny and the
173 remaining 54% (1,339) were identified in the terminal branches when compared to their most
174 recent reconstructed ancestor. The number of mutations per strain or node varied from 0 to
175 101 with a mean and median count of five and two mutations per branch, respectively. The
176 largest number of inferred mutations along a branch was located in the ancestor of the
177 Hawaiian cluster which includes the natural isolate CB4856. Transitions accounted for 81.8%
178 of all SNPs, which included 1,062 A/G → G/C and 881 G/C → A/G mutations (Table 1). The
179 transition to transversion (*Ts/Tv*) ratio for all SNPs is 4.49. The most common transversion
180 was A/T → T/A (226) followed by G/C → T/A (144), A/T → C/G (54) and G/C → C/G (9).
181
182 *Mutational spectra of mtDNA base substitutions in natural isolates differ from that of*
183 *experimental lines*
184
185 We inferred the mutation spectra using all the mtDNA base substitutions found in this
186 study (Fig. 1a) as well as one using only base substitutions found at four-fold degenerate sites
187 (Fig. 1b). We also adapted the data from two recent studies investigating *C. elegans*
188 mitochondrial mutations under experimental laboratory conditions employing the wildtype
189 Bristol (N2) strain (Waneka et al. 2021; Leuthner et al. 2022) (Fig. 1c and d). The number of
190 mutations was normalized by the base composition of the reference genome (or of the
191 reference genome's four-fold degenerate sites). In other words, the counts of G/C → N/N
192 mutations were divided by the number of G/C base pairs in the reference genome (and all
193 A/T → N/N mutations by the number of A/T base pairs) to account for the inherent A/T-bias
194 in the *C. elegans* mitochondrial genome.
195
196 A key difference between natural isolates and the results from duplex sequencing of
197 laboratory strains is that transitions are significantly more frequent in the wild compared to
198 the laboratory experiments (Table 1; Fig. 1a). A comparison of the number of transitions and
199 transversions between four-fold degenerate sites in natural isolates (Fig. 1b) and variants at
200 all sites in duplex sequencing from laboratory experiments (Waneka et al. 2021; Leuthner et
201 al. 2022; Fig. 1c and d) is highly significant (χ^2 test-statistic = 548.68; *d.f.* = 2; N = 1,913; p =
202 7.2×10^{-120}). The results from Leuthner et al. (2022) are clearly an outlier here with only 18%
203 of all mutations identified as transitions (Fig. 1c). If we only compare four-fold degenerate

204 sites in natural isolates (Fig. 1b) and the duplex sequencing variants at all sites of Waneka et
205 al. (2021) (Fig. 1d) where transitions comprised 75% and 57% of all mutations, respectively,
206 the difference is still highly significant (χ^2 test-statistic = 33.18; *d.f.* = 1; N = 1,153; $p = 8.4 \times$
207 10^{-9}).
208
209 G/C → A/T transitions are more frequent than A/T → G/C at all sites in duplex
210 sequencing experiments whereas the converse is observed for four-fold degenerate sites in the
211 natural isolates (Table 1; χ^2 test-statistic = 44.26; *d.f.* = 2; N = 960; $p = 2.5 \times 10^{-10}$). However,
212 these transitions still show an AT-bias as the ratio of $\frac{G/C \rightarrow A/T}{A/T \rightarrow G/C} > 1$ when the base composition
213 of the genome is accounted for in both the laboratory experiments and natural isolates (Table
214 1). The ratio between these two transitions is higher in the duplex sequencing experiments
215 where the G/C → A/T transitions per G/C base pair are 5.8 – 6.5× more likely than A/T →
216 G/C transitions per A/T base pair. In contrast, the G/C → A/T transitions are only 2.6× more
217 likely than A/T → G/C transitions when controlled for base composition (Table 1) in the
218 natural isolates. The $\frac{G/C \rightarrow A/T}{A/T \rightarrow G/C}$ ratio in the natural isolates is observed to be 4.8 at four-fold
219 degenerate sites which demonstrates that the high number of A/T → G/C transitions at these
220 sites is due to the high A+T-content at four-fold degenerate sites of the mitochondrial
221 genome (Table 1).
222
223 Furthermore, there are also significant differences between the natural isolates and the
224 duplex sequencing of laboratory strains in the relative numbers of different transversions
225 (Table 1; χ^2 test-statistic = 338.82; *d.f.* = 6; N = 953; $p = 3.9 \times 10^{-70}$). To investigate
226 differences in the relative proportions of the various transversion classes, our analysis was
227 restricted to four-fold degenerate sites from natural isolates as transversions at two-fold
228 degenerate sites of protein-coding sequences result in amino acid replacement which are
229 subject to stronger purifying selection pressure in the wild. Similar to that observed
230 previously in the case of transitions, one of the duplex sequencing studies (Leuthner et al.
231 2022) is a strong outlier here as it reports a high proportion of G/C → C/G transversions, a
232 substitution class that is usually amongst the rarest substitutions in all other studies of
233 mtDNA mutation rates in *C. elegans*, and ranging between 0% to 4% of all mutations. If we
234 restrict the comparison to four-fold degenerate sites in the natural isolates and the duplex
235 sequencing experiments of Waneka et al. (2021), the difference in the proportions of various

236 types of transversion remains highly significant (Table 1; χ^2 test-statistic = 94.23; *d.f.* = 3; *N*
237 = 330; $p = 2.7 \times 10^{-20}$). Although the four categories of transversions listed in Table 1 are in
238 strikingly different proportions between four-fold degenerate sites in natural isolates and the
239 duplex sequencing study of Waneka et al. (2021), we note that the most common
240 transversions in the laboratory study are G/C → T/A (72.5% of all transversions, 31.2% of all
241 base substitutions) which were attributed to oxidative damage. In contrast, the very same G/C
242 → T/A transversions are only 21.7% and 5.3% of all transversions and base substitutions at
243 four-fold degenerate sites in the natural isolates, respectively.
244
245 *Base composition of four-fold degenerate sites in natural isolates is not in mutational*
246 *equilibrium*
247
248 At mutational equilibrium, the number of mutations that increase the A+T-content of
249 the mitochondrial genome is expected to be equal to the number of G/C-increasing mutations.
250 In both of the duplex sequencing studies (Waneka et al. 2021; Leuthner et al. 2022), A/T-
251 increasing mutations vastly outnumber G/C-increasing mutations (Table 1). The ratios of
252 A/T-increasing to G/C-increasing mutations were 6.0 and 3.3 in the studies by Leuthner et al.
253 (2022) and Waneka et al. (2021), respectively. However, the base composition at
254 nondegenerate sites is not expected to be in mutational equilibrium due to selection on the
255 amino acid sequence. Consequently, nondegenerate sites in protein-coding regions of mtDNA
256 typically have higher G+C-content than two-fold and four-fold degenerate sites. The high
257 ratio of A/T-increasing to G/C-increasing mutations under relaxed selection in laboratory
258 experiments is a reflection of the deviation from mutational equilibrium at nondegenerate
259 sites. For instance, if we only consider the four-fold degenerate sites in Waneka et al. (2021),
260 the ratio of A/T-increasing mutations to G/C-increasing mutations is closer to one (1.18). In
261 the natural isolates, the numbers of A/T-increasing and G/C-increasing mutations are more
262 similar, with a ratio of 0.92. However, the ratio of A/T-increasing to G/C-increasing
263 mutations is only 0.80 at four-fold degenerate sites in the natural isolates. The difference
264 between the number of A/T-increasing and G/C-increasing mutations at four-fold degenerate
265 sites in natural isolates is significant (χ^2 test-statistic = 9.66; *d.f.* = 1; $N = 766$; $p = 0.0019$).
266 Furthermore, there is no significant difference between the ratio of A/T-increasing to G/C-
267 increasing mutations inferred at the four-fold degenerate sites within the internal nodes of the
268 phylogenetic tree (0.84; $N = 406$) versus single mitotypes (0.76; $N = 360$) within the natural

269 isolates (χ^2 test-statistic = 0.48; *d.f.* = 1; N = 766; p = 0.4851). The observation that G/C-
270 increasing mutations significantly outnumber A/T-increasing mutations at four-fold
271 degenerate sites in the wild suggests that these sites are not in mutational equilibrium.
272
273 *Strong signature of purifying selection in mtDNA protein-coding genes of natural isolates*
274
275 Polymorphic sites with respect to base substitutions in the mtDNA protein-coding
276 regions of the natural isolates show evidence of strong purifying selection acting on the first
277 and second codon positions (Fig. 2a). The second codon position, which is always
278 nondegenerate, has only ~7% of the total number of observed mtDNA base substitutions. The
279 first and third codon positions harbor approximately 19% and 74% of base substitutions in
280 the mtDNA protein-coding genes, respectively. In the duplex sequencing study by Waneka et
281 al. (2021), the proportions of mutations per codon position are closer to the neutral
282 expectation under a regime of relaxed selection (Fig. 2a). The slightly higher proportions of
283 mutations at the first and second codon positions in their study possibly reflects higher G/C-
284 content at these sites rendering them more prone to mutations. With regards to codon
285 degeneracy, we standardized the observed number of base substitutions in mtDNA protein-
286 coding regions by the total number of nondegenerate, two-fold degenerate and four-fold
287 degenerate sites in the *C. elegans* reference mtDNA genome. The relative proportion of base
288 substitutions observed in the natural isolates at four-fold degenerate sites was 0.64, compared
289 with 0.05 and 0.31 at nondegenerate and two-fold degenerate sites, respectively (Fig. 2b).
290 These contrasts are even more evident when comparing the natural isolates' data with the
291 results from Waneka et al. (2021), where four-fold and nondegenerate sites harbor
292 comparable proportions of mutations (Fig. 2b). Interestingly, the results adapted from the
293 study by Waneka et al. (2021) find two-fold degenerate sites exhibiting a reduction in the
294 proportion of mutations compared to four-fold and nondegenerate sites (Fig. 2b).
295
296 The *Ts/Tv* ratios of inferred base substitutions vary predictably between codon
297 positions in the natural isolates with the third and second codon position displaying the
298 highest (5.35) and lowest ratio (2.66), respectively (Fig. 3a). For comparison, in the duplex
299 sequencing results of Waneka et al. (2021) the mutation proportions per codon position are
300 much closer to the neutral expectation under relaxed selection (Fig. 3a). Overall, the *Ts/Tv*
301 ratios per codon position are lower in the duplex sequencing study (Waneka et al. 2021),

302 demonstrating that the bias towards transitions is less pronounced under a regime of relaxed
303 selection. Surprisingly, the *Ts/Tv* ratio at the third codon position in the Waneka et al. (2021)
304 study is more than 2× higher than that observed at the first and second codon positions (Fig.
305 3a). In the natural isolates, transversions comprise ~6% of the mutations at two-fold
306 degenerate sites while constituting approximately 25% and 31% of mutations at four-fold
307 degenerate and nondegenerate sites, respectively. The *Ts/Tv* ratio varies by site degeneracy in
308 the protein-coding genes, with two-fold degenerate sites having the highest *Ts/Tv* ratio as
309 expected (16.70), followed by four-fold degenerate sites (3.07) and nondegenerate sites
310 (2.25) (Fig. 3b). Although the *Ts/Tv* ratios are far lower in the duplex sequencing study of
311 rare new mutations under a regime of relaxed selection (Waneka et al. 2021), two-fold
312 degenerate sites still display a higher *Ts/Tv* ratio than four-fold degenerate and nondegenerate
313 sites (Fig. 3b), indicative of a modicum of purifying selection operating in this experimental
314 regime.
315
316 *Numbers of polymorphic base substitutions vary by mtDNA gene and ETC complex*
317
318 Of the 12 protein-coding genes comprising the mitochondrial genome of *C. elegans*,
319 *nd4L* has significantly fewer polymorphic substitutions (G test-statistic = 11.1, $d.f.$ = 1, p =
320 8.41×10^{-4}) whereas *nd1* has significantly more polymorphic substitutions (G test-statistic =
321 31.7, $d.f.$ =1, $p = 1.79\times10^{-8}$) relative to the other protein-coding genes (Fig. 4a). Furthermore,
322 the nonsynonymous to synonymous (*N/S*) ratios are significantly higher in *nd*2 (Fisher's
323 exact test: $p = 3.44 \times 10^{-3}$) and in *nd6* (Fisher's exact test: $p = 0.0351$) but significantly lower
324 in *cox1* (Fisher's exact test: $p = 4.06 \times 10^{-4}$) compared to the other protein-coding genes (Fig.
325 4b). In *nd2*, there were both fewer synonymous mutations (G test-statistic = 9.95, $d.f.$ = 1, p =
326 1.61×10^{-4}, Fig. 4c) and more nonsynonymous mutations (G test-statistic = 11.14, $d.f.$ = 1, p
327 $= 8.44 \times 10^{-4}$, Fig. 4d) compared to other mtDNA protein-coding genes. Similarly, the low
328 *N/S* in *cox1* is due to both fewer nonsynonymous mutations (G test-statistic = 18.87, $d.f.$ = 1,
329 $p = 1.40 \times 10^{-5}$, Fig. 4d) and more synonymous mutations than the genomic average. The
330 high *N/S* in *nd6* is due to significantly greater number of nonsynonymous mutations
331 compared to the genomic average (G test-statistic = 10.81, $d.f.$ = 1, $p = 1.01 \times 10^{-3}$; Fig. 4d).
332 The gene bearing the most mutations per site, *nd1* (Fig. 4a), has significantly greater number
333 of mutations at both synonymous (G test-statistic = 15.54, $d.f.$ = 1, p = 1.01×10^{-5}) and
334 nonsynonymous sites (G test-statistic = 21.09, $d.f.$ = 1, $p = 4.37 \times 10^{-6}$; Figs. 4c and d).

335

336 There were no significant differences in the total number of mutations per site when
337 the 12 mtDNA protein-coding genes were grouped into their respective electron transport
338 chain (ETC) complexes (*G* test-statistic = 2.37, *d.f.* = 1, $p = 0.498$; Fig. 5a). However, genes
339 in complex I display a significantly higher *N/S* ratio (Fisher's exact test: $p = 1.93 \times 10^{-3}$; Fig.
340 5b) due to both significantly fewer synonymous mutations (*G* test-statistic = 10.27, *d.f.* = 1, p
341 $= 1.35 \times 10^{-3}$; Fig. 5c) and higher nonsynonymous mutations (*G* test-statistic = 36.28, *d.f.* = 1,
342 $p = 1.71 \times 10^{-9}$; Fig. 5d) than genes in the other three mitochondrial respiratory complexes.
343 Furthermore, the seven highest *N/S* ratios, even if not significantly departing from the
344 genomic average, are found in ETC complex I genes. In contrast, mtDNA genes functioning
345 in complex IV have a significantly lower *N/S* ratio (Fisher's exact test: $p = 1.40 \times 10^{-5}$) owing
346 to significantly fewer nonsynonymous mutations (*G* test-statistic = 33.93, *d.f.* = 1, $p = 5.71 \times$
347 10^{-9}; Fig. 5d) relative to the genes in other ETC complexes.
348

349 *Local context-dependence of mutations*

350

351 Sites with G/C flanking nucleotides, either as 5′ or 3′ neighbors, have higher
352 probability of mutations than sites with A/T neighbors (Fig. 6). Significant departures from
353 equal mutation distribution at two- and four-fold degenerate sites further support this
354 observation (supplementary table S1). Two-fold degenerate sites have a 1.85× greater
355 probability of a mutation if their 5′ neighbor is G/C compared to A/T. Likewise, there is a
356 1.58× greater probability of mutations at two-fold degenerate sites with 3′ G/C neighbors
357 rather than A/T. Four-fold degenerate sites with 5′ G/C neighbors have a 1.7× greater
358 probability of mutations relative to sites with 5′ A/T neighbors. Lastly, four-fold degenerate
359 sites with 3′ G/C neighbors have a 1.91× greater probability of mutations compared to sites
360 with 3′ A/T. Context-dependent mutations can also contribute to the variation in synonymous
361 mutation rates between mtDNA genes. Because degenerate sites at 3rd base position are
362 flanked by nucleotides at 1st and 2nd codon positions, we tested for an association between
363 base composition at these positions with synonymous polymorphisms. The synonymous
364 polymorphisms in *C. elegans* mtDNA genes is positively correlated with the G+C-content at
365 the 1st and 2nd codon positions (Pearson's $r = 0.703$, *d.f.* = 10, $p = 0.0107$) but not at the 3rd
366 codon position (Pearson's $r = -0.002$, *d.f.* = 10, $p = 0.9958$).
367

368 *The distribution of synonymous polymorphisms depends on their location within genes*

369

370 There was no relationship between distance from the mtDNA control region and
371 synonymous polymorphism. However, the combined data from all 12 protein-coding genes
372 shows that the 5′ and 3′ ends of protein-coding genes contained fewer synonymous
373 polymorphisms than centrally located sites within the open-reading frame. For example, the
374 synonymous polymorphism count is two and five at the first and second 9-bp window (at the
375 5′ end), respectively (Fig. 7a). The last three 9-bp windows at the 3′ end averaged ~6.7
376 synonymous polymorphisms per window (Fig. 7b). In contrast, centrally located 9-bp
377 windows within the open-reading frame averaged ~20.3 synonymous polymorphisms per
378 windows (from the fourth to the sixth 9-bp window) (Fig. 7a). Similarly, A+T-content
379 decreases along the same axis with increasing distance from both the start and stop codons
380 (Figs. 7c and d).

381

382 *Small indel variants detected at low heteroplasmic frequencies with a clustered distribution*
383 *in several hotspots in the mtDNA genome*

384

385 There are two hypervariable mtDNA regions in the *C. elegans* natural isolates
386 harboring small indels, namely (i) the noncoding region containing the D-loop, and (ii) a
387 noncoding poly-A sequence between *atp6* and $tRNA_{Lys}$. Exempting these two regions, 88
388 polymorphic small indels were identified. The vast majority of these indels were either a 1 bp
389 insertion or a deletion and only two variants were 2 bp indels (an insertion and deletion each).
390 Indel polymorphisms were found at one in every 22 tRNA and 350 protein-coding positions.
391 It is worth noting that no indels were found in rRNA genes (supplementary file S2). In total,
392 29 (~33%) small indels were identified in protein-coding genes, corresponding to seven
393 deletions and 22 insertions. All but two of these 29 indels in protein-coding genes are
394 heteroplasmic frameshift mutations, ranging in heteroplasmic frequency from 11% to 52%.
395 Two indels were homoplasmic within the same isolate (XZ1516), a 1 bp deletion and a 1 bp
396 insertion five nucleotides apart in the *nd2* gene, thereby preserving the reading frame. The
397 tRNA genes harbored 56 (~64%) small indel polymorphisms, corresponding to 47 single bp
398 deletions, a single 2 bp deletion and eight 1 bp insertions. Apart from the two hypervariable
399 regions, three (~3%) indel polymorphisms were located in noncoding regions.

400

401 In addition to the hypervariable regions mentioned above, we identified additional
402 small indel hotspots in four genes, *tRNA$_{Pro}$, tRNA$_{Tyr}$, nd4* and *nd5* (supplementary file S2). A
403 *tRNA$_{Pro}$* sequence corresponding to mtDNA:14-22 comprised of $(A)_2$-T-$(A)_5$ in the reference
404 genome contained 32 and two cases of single A deletions and insertions, respectively. Natural
405 isolate ECA740 incurred two A insertions within this region, with one insertion inferred in an
406 ancestral reconstructed sequence and a subsequent second insertion leading to ECA740
407 having a double A insertion relative to the reference genome. The second hotspot occurs at
408 position mtDNA:1,750 located within *tRNA$_{Tyr}$* with 10 independent instances of an inferred T
409 deletion and one T insertion. The secondary structure of these two tRNAs appears to be
410 conserved. The third hotspot is found within a $(T)_8$ homopolymeric run in *nd4* at position
411 mtDNA:6,752. This site has three instances of a single T insertion and one case of a single T
412 deletion. All four frameshift variants were found in low heteroplasmic frequencies ranging
413 from 12–20%. The fourth and final indel hotspot occurs towards the 5′ end of *nd5* within a
414 $(T)_8$ homopolymeric sequence spanning positions mtDNA:11,722-11,729. We identified 11
415 cases of a single T insertion within this hotspot resulting in frameshifts spanning most of the
416 *nd5* coding sequence. Similar to the *nd4* variants, these *nd5* frameshift variants were found in
417 low heteroplasmic frequencies ranging from 11% to 25%. While insertions outnumbered
418 deletions overall, the insertion/deletion bias is strikingly different between the tRNA and
419 protein-coding genes. In tRNA genes, deletions were found to outnumber insertions by a
420 factor of six. Conversely, insertions outnumbered deletions by a factor of three in protein-
421 coding genes.
422
423 *First identification of novel heteroplasmic large mtDNA deletions and duplications in natural*
424 *isolates of C. elegans*
425
426 We identified eight heteroplasmic large structural variants (SVs), with each SV
427 identified by three SV callers (Delly, Manta and Pindel). No variants (post-filtering) were
428 called by only two of the callers employed. Of the eight variants, six were found to be large
429 deletions while two were found to be duplications spanning the control region and
430 neighboring genes (Table 2; Fig. 8; supplementary figs. S2 and S3). The heteroplasmic
431 variant frequencies determined by read-depth analyses ranged from a low 4% up to 79%
432 (Table 2). The two duplication variants found in strains PS2025 and MY23, while strikingly
433 similar with respect to both location and duplication span, did not share any notable SNPs

434 and do not appear to have the same origin. In particular, the precise breakpoint locations
435 differed between the duplications.
436
437 *Signature of purifying selection in heteroplasmic variants*
438
439 Variants ranging in frequency from 3% to 97% within a natural isolate were
440 characterized as heteroplasmic. In addition to the large heteroplasmic structural variants
441 discussed above, of the 2,464 base substitutions and small indels variants in the natural
442 isolates, only 6.6% (162) were heteroplasmic (Fig. 9a). Of these, 144 (~89%) were within
443 protein-coding genes. While only 4% and 11% of synonymous and nonsynonymous base
444 substitutions were heteroplasmic, 100% of the small indel frameshift variants in the protein-
445 coding genes were found to be heteroplasmic. The variant frequency distribution of small
446 indel mutations differs between protein-coding and tRNA genes (Fig. 9b). The vast majority
447 of small indels occurring in protein-coding genes are found at low heteroplasmic frequencies.
448 In contrast, small indel mutations occurring in tRNA genes tend to have significantly higher
449 frequencies (fixed or near fixation) ($U = 90$, $z = 6.6878$, $p < 0.00001$). The frequency
450 distribution of heteroplasmic mtDNA mutations in the natural isolates is bimodal (Fig. 9a)
451 and differs from the more uniform distribution observed for their counterparts in a long-term
452 *C. elegans* MA experiment under minimal selection (Konrad et al. 2017; Fig. 9c). In protein-
453 coding genes, the median frequency of heteroplasmic synonymous variants is significantly
454 greater than that of heteroplasmic nonsynonymous variants ($U = 1135$, $z = 2.9829$, $p =$
455 0.00288). However, the median frequency of heteroplasmic nonsynonymous variants is not
456 significantly greater than the median frequency of heteroplasmic frameshift variants ($U =$
457 534, $z = 0.7614$, $p = 0.4473$).
458
459
460 *McDonald-Kreitman tests detect no evidence of positive natural selection*
461
462 To complement the metrics on mtDNA polymorphism and heteroplasmy in the
463 natural isolates, we conducted McDonald-Kreitman tests (McDonald and Kreitman 1991)
464 using *Caenorhabditis brenneri* as an outgroup species to investigate what proportion, if any,
465 of the substitutions in the mtDNA protein-coding genes of the natural isolates are driven by
466 positive adaptive evolution. We compared the within-species (*C. elegans* wild isolates)
467 synonymous and nonsynonymous polymorphism to between-species synonymous and

468 nonsynonymous divergence (supplementary table S2). The neutrality index across the 12
469 concatenated protein-coding genes was 0.79 (Fisher's exact test: $p = 0.11$), which is not
470 indicative of positive natural selection in the coding sequence between the two species. The
471 neutrality index values fluctuate considerably between the different mtDNA genes analyzed,
472 ranging from 0.42 to 3.3, but none of the genes show an excess of nonsynonymous changes
473 between species (supplementary table S2).
474
475 *Codon usage bias has the potential to skew compositional and strand biases in the C. elegans*
476 *mitochondrial genome*
477
478 If the base composition of four-fold degenerate codon positions is strictly a function
479 of mutation pressure, and not natural selection on synonymous mutations or specific local
480 sequence context, these sites should have similar base composition between codon families.
481 Four of the seven four-fold degenerate codon families deviated significantly in their
482 nucleotide distribution at their third base position from other four-fold degenerate codon
483 families (supplementary table S3). Furthermore, five of the 13 two-fold degenerate codon
484 families also show significant variation in their nucleotide distribution at their 3rd base
485 position (supplementary tables S4 and S5). The most commonly used codons in each codon
486 family (preferred codons) contain either a T (most common) or an A in their 3rd base position.
487 The codon usage preferences in *C. elegans* mtDNA suggest natural selection for codon usage,
488 which in addition to the mutational spectrum, introduces a further base compositional bias
489 and strand asymmetry to the *C. elegans* mitochondrial genome. A comparison of synonymous
490 mutations present only in single mitotypes (rare) versus variants that are shared by more than
491 a single mitotype, lends further support for natural selection on codon preference in *C.*
492 *elegans* mtDNA. In rare synonymous variants, mutations from preferred codons outnumber
493 mutations to preferred codons by 357 to 245. In contrast, in variants that are shared by more
494 than a single mitotype, mutations from and to preferred codons are similar (364 and 333,
495 respectively). This difference between rare and shared synonymous variants is significant (G
496 test-statistic = 6.56, $d.f. = 1$, $p = 0.0104$). Rare variants are thought to be more representative
497 of the mutation spectrum and less affected by natural selection than more common (shared)
498 variants. The greater relative abundance of mutations to preferred synonymous codons in
499 shared variants compared to rare variants is consistent with natural selection for preferred
500 synonymous codons.

501

502 **DISCUSSION**

503

504 *Natural populations have fewer mtDNA mutations typically associated with oxidative damage*
505 *relative to laboratory populations*

506

507 Polarized DNA polymorphism from natural isolates was used to test if the mtDNA
508 mutational patterns observed in laboratory experiments recapture the mutational
509 spectrum of natural populations. Laboratory experiments exhibit varying patterns but still
510 converge on high proportions of G/C → T/A transversions, which have been interpreted as
511 signatures of oxidative damage (Waneka et al. 2021; Leuthner et al. 2022). This hypothesis
512 has been reinforced by observations of mutational strand bias (Cheng et al. 1992; Waneka et
513 al. 2021). One of the challenges when comparing natural populations to laboratory
514 experiments (even more so when these laboratory lines are the result of several generations of
515 mutation accumulation under minimal selection) is the difference in the strength of purifying
516 selection, which could have significant effects on the observed mutational spectrum. To
517 mitigate the effect of selection on the observed mutational spectrum in the natural isolates,
518 we focused on sites where selection is expected to be at a minimum, such as four-fold
519 degenerate sites. Furthermore, in the natural isolates' dataset, we compared the results
520 between (i) polymorphisms where polarity was inferred using ancestral reconstruction, and
521 (ii) rare variants that were only observed in single mitotypes. Purifying selection is a potent
522 force acting on mtDNA polymorphism in the natural isolates. This is evident in the
523 distribution of mtDNA substitutions which are strongly associated with codon position and
524 site degeneracy. Four-fold degenerate sites harbor the vast majority of mutations in the
525 natural isolates. Furthermore, the high *Ts*/*Tv* ratio at two-fold degenerate sites compared to
526 four-fold degenerate sites is a prime example of how strongly nonsynonymous mutations are
527 counter-selected in the wild.

528

529 Within transitions and transversions, there are similarities in the ranking of
530 proportions of different base substitutions between the natural isolates and laboratory
531 experiments (Waneka et al. 2021; Leuthner et al. 2022). G/C → A/T transitions are more
532 frequent than the reverse transition of A/T → G/C, contributing to an A+T-mutational bias in
533 both the natural isolates and laboratory experiments. The main source of transversions in both

534 natural isolates and laboratory lines are G/C → T/A mutations, further adding to the mtDNA
535 A+T-mutational bias. However, the natural isolates exhibit a significantly greater *Ts*/*Tv* ratio
536 than the laboratory populations. The main contributor to the difference in the *Ts*/*Tv* ratio
537 between the mutation spectra of the natural isolates *vs.* laboratory populations is a far greater
538 proportion of G/C → T/A transversions in the laboratory experiments (Waneka et al. 2021;
539 Leuthner et al. 2022). The proportion of G/C → T/A transversions is approximately 4×
540 greater at four-fold degenerate sites in the laboratory experiments relative to the natural
541 isolates. G/C → T/A transversions have been associated with oxidative damage of DNA
542 (Cheng et al. 1992; Kirkwood and Kowald 2012), and the striking difference in their
543 proportions between laboratory experiments and natural populations might suggest that
544 laboratory grown populations are more prone to mtDNA oxidative damage.
545
546 Metabolic rate has been suggested to be a major determinant of mtDNA evolution
547 (Martin and Palumbi 1993; Rand 1994; Gillooly et al. 2005). The rate-of-living theory
548 (Rubner 1908; Pearl 1928) and the free radical theory (Harman 1956) that expands upon it,
549 proposed that higher metabolic rate organisms have shorter life spans due to the faster
550 accumulation of damages from metabolic by-products and more specifically due to the
551 mutagenic nature of ROS species (Beckman and Ames 1998). Moreover, mtDNA maybe
552 especially prone to oxidative damage compared to nuclear DNA (Richter et al. 1988).
553 However, the relationships between basal metabolic rate and the rates of mutations and
554 molecular evolution in mtDNA have remained somewhat murky and adaptation associated
555 with other life-history traits such as life span may have been more important (Galtier et al.
556 2009). Many attributes of the base composition and strand bias in mitochondrial genomes
557 have been ascribed to DNA damage by ROS.
558
559 *C. elegans* is a widely used laboratory model and has been genetically modified
560 through decades of domestication to adapt to cultivation methods enforced upon it (Frézal
561 and Félix 2015). In the wild, *C. elegans* lives through boom-and-bust cycles depending on
562 temperature and food supplies, entering a diapause state called dauer through periods of
563 insufficient resources. The dauer larval state has been found to increase resistance to
564 oxidative damage, leading to increased life-span (Larsen 1993). The concentration of oxygen
565 on petri dishes is much higher than that in the preferred habitat of wild *C. elegans* and during
566 its domestication, laboratory strains have lost many hyperoxia avoidance behaviors that are

567 present in natural isolates (Chang et al. 2006). We hypothesize that laboratory growth
568 conditions for *C. elegans*, with unlimited food, relatively unrestricted space in the case of
569 MA experiments, and an absence of the multitude of environmental stressors present in the
570 wild lead to higher metabolic rates and greater oxidative damage over the course of a
571 laboratory experiment than that usually experienced in the wild.
572
573 *Number of polymorphic base substitutions varies by gene and ETC complex*
574
575 Purifying selection is the main evolutionary force acting on mtDNA polymorphism in
576 *C. elegans* natural isolates. The distribution of polymorphism is strongly linked to site
577 degeneracy, as shown, for example, by the high *Ts/Tv* at two-fold compared to four-fold
578 degenerate sites. However, the strength of purifying selection may not be homogeneously
579 distributed across all mtDNA protein-coding genes. ETC complex I genes have higher
580 nonsynonymous to synonymous polymorphism ratio (*N*:*S*), most apparent in *nd6* and *nd2.*
581 Indeed, five of the seven mtDNA genes in this complex exhibit higher than average *N*:*S*
582 ratios. Conversely, genes in complex IV tend to have a lower N:S, suggestive of strong
583 purifying selection. These patterns of mtDNA polymorphism are concordant with an analysis
584 of mtDNA divergence between *Caenorhabditis* species which found the highest and lowest
585 *dN/dS* in ETC complex I and IV genes, respectively (Nabholz et al. 2013). Furthermore, this
586 study found an inversely proportional relationship between *dN/dS* and mtDNA transcript
587 abundance of mitochondrial-encoded OXPHOS genes. Although the results were based on
588 transcript rather than protein abundance, this negative relationship is consistent with the
589 hypothesis that highly expressed proteins have slower rates of evolution (Drummond et al.
590 2005). However, an increasing number of studies have associated mtDNA mutations in ETC
591 complex I with life-history evolution (Garvin et al. 2015; Camus et al. 2017; Li et al. 2024),
592 or have identified ETC complex I as a target of directional selection (Mishmar et al. 2006).
593 For example, different mitotypes distinguished primarily by ETC complex I mutations are
594 associated with thermal adaptation in *Drosophila melanogaster* (Camus et al. 2017).
595 Similarly, in the agricultural pest *Spodoptera frugiperda*, the "rice" mitotype at the forefront
596 of pest invasions differed from the "corn" mitotype by 18 nonsynonymous mutations, 14 of
597 which are in ETC complex I (Li et al. 2024). ETC complex I is often associated with positive
598 selection and appears to be one of the primary targets of local adaptation in mitochondrial
599 genomes (Mishmar et al. 2006; Garvin et al. 2015; Camus et al. 2017; Li et al. 2024).
600

601 *The effect of A+T-content of flanking nucleotides on synonymous polymorphism*

602

603 There is a significant difference between A/T versus G/C neighbors on
604 synonymous polymorphism. The probability of a site exhibiting synonymous polymorphism
605 decreases with A+T-content of its adjacent nucleotides. Whereas variation in nonsynonymous
606 polymorphism between genes is frequently explained by differences in the relative strength of
607 natural selection, synonymous polymorphism is often presumed to be neutral. Population-
608 genetic processes such as genetic hitchhiking or background selection (Maynard Smith and
609 Haigh 1974; Charlesworth et al. 1993) which can introduce variation in synonymous
610 polymorphism in genomes undergoing genetic recombination, should not contribute to
611 between-gene variation in synonymous polymorphism in completely linked genomes such as
612 the *C. elegans* mitochondrial genome. There is a potential for selection on synonymous
613 mutations through codon usage and mRNA secondary structure, which can vary among
614 genes. However, local sequence context appears to have a significant contribution to
615 between-gene variation in synonymous polymorphism as demonstrated by a correlation
616 between A+T-content in the 1st and 2nd codon positions and synonymous polymorphism,
617 usually in 3rd codon position.

618

619 The effect of local sequence context also extends to within-gene variation in
620 synonymous polymorphism. The 5′ ends of *C. elegans* mtDNA genes, and 3′ ends to some
621 extent as well, exhibit less synonymous polymorphism than centrally located sites. This
622 coincides with a higher A+T-content at the 5′ and 3′ ends across the 12 mtDNA protein-
623 coding genes. Does this within-gene variation in base composition and mutation
624 rate/synonymous polymorphism indicate some structural constraints on mRNA structure? A
625 possible neutral explanation for this pattern is that amino acid composition is influenced by
626 mutation pressure in such a manner that the consequences of high A+T mutation pressure are
627 not just confined to synonymous sites but nonsynonymous sites as well (Foster et al. 1997).
628 Nonsynonymous sites within genes that are less important for optimal protein function might
629 therefore be more affected by A+T mutation pressure and code for amino acids with high
630 A+T-content at their nonsynonymous sites, usually at the 1st and 2nd codon positions.
631 Consequently, because of the influence of local sequence context on the frequency of
632 mutations, the probability of synonymous polymorphism would be lower in these high A+T-
633 regions of genes. However, it is also possible that this pattern of high A+T-content at the 5′

634 end of genes shares functional similarities to many prokaryotic genomes that also have higher
635 A+T-content at their genic 5′ ends (Allert et al. 2010; Tuller et al. 2010). In *E. coli*, protein
636 expression levels have been shown to be dependent on high A+U-content and low secondary
637 structure of mRNA, particularly at the gene's 5′ region, and to a lesser extent, on high A+U-
638 content in the 3′ region (Allert et al. 2010; Goodman et al. 2013). Considering the prokaryotic
639 origin of mitochondria, perhaps selection to maintain (i) high A+T-content, and/or (ii) low
640 complexity to prevent mRNA secondary structures that might interfere with optimal
641 translation efficiency is still important in mitochondrial genomes.
642
643 *Natural selection on codon usage bias*
644
645 Unequal use of synonymous codons can stem from compositional and strand biases
646 caused by the mutational spectrum and natural selection for particular codons within a family
647 of synonymous codons (Hershberg and Petrov 2008). The observed deviation in base
648 composition and strand symmetry at four-fold degenerate sites from the expected equilibrium
649 value based on the mutation spectrum in these *C. elegans* natural isolates could be explained
650 by codon usage selection. Furthermore, natural selection for optimal codon usage could
651 potentially influence the frequency spectrum of polymorphisms at degenerate sites. Families
652 of synonymous codons in the *C. elegans* mitochondrial genome always favor an A or T at
653 degenerate sites in the coding sequence, which is expected given the prevailing G/C → A/T
654 mutational bias. Furthermore, most codon families with four-fold degenerate sites are also
655 biased towards having a T instead of an A in their 3rd base position, which is also expected
656 given the observed compositional and mutational strand asymmetry. An example of a strong
657 preference for a T is the family of arginine codons as 29 of 31 arginine codons in the *C.*
658 *elegance* reference mtDNA contain a T in their 3rd base position. However, some four-fold
659 degenerate codon families are biased towards an A in their 3rd base position which runs
660 counter to the compositional and mutational strand asymmetry suggesting a role for natural
661 selection for codons ending in A in these families. An example of a strong preference for A at
662 3rd base position is the family of serine codons that start with AG. In this family, codon AGA
663 is present 126 times compared to 61 instances of codon AGT in the reference genome. If
664 mutational bias alone is responsible for codon usage bias, we expect that four-fold degenerate
665 codon families have the same distribution of bases in their 3rd base position, which they do
666 not. Furthermore, the significantly higher proportion of mutations to preferred codons among

667 inferred ancestral variants compared to rare variants is consistent with natural selection for
668 preferred codons. The spectrum of synonymous polymorphism in *C. elegans* mtDNA is
669 therefore a function of mutational patterns, which include base compositional biases, strand
670 asymmetry and local sequence context, as well as natural selection for specific synonymous
671 codons.
672
673 *First identification of novel heteroplasmic structural variants in wild populations of C.*
674 *elegans*
675
676 Large structural mtDNA variants have previously only been described for laboratory
677 populations of *C. elegans* (Tsang and Lemire 2002; Liau et al. 2007; Dubie et al. 2020, 2024;
678 Sequeira et al. 2024). The natural isolates analyzed here add eight new heteroplasmic
679 structural variants comprising six deletions and two duplications. As all of the mtDNA genes
680 are essential, genic mtDNA deletions are always heteroplasmic and appear to be maintained
681 in populations by the balance between replication or transmission advantage and fitness cost
682 (Liau et al. 2007; Gitschlag 2016, 2020; Dubie et al. 2020, 2024; Sequeira et al. 2024). The
683 six deletions described here affect nine protein-coding genes and multiple tRNA genes.
684 Notably, regions in the *C. elegans* mtDNA have not been found to be deleted in either
685 laboratory or natural populations include the first few genes downstream of the D-loop (*nd6*,
686 *nd4L*, $tRNA_{Pro,Val,Trp,Glu}$, and 12S rRNA) and a noncoding region between *nd4* and *cox1* which
687 has been suggested to harbor a second replication origin. These regions that are not known to
688 have been affected by deletions are possibly essential for mtDNA maintenance. In addition to
689 the six deletions, duplications encompassing the D-loop were identified in two natural
690 isolates. Duplications of the mtDNA replication origin can possibly be favored in intra-
691 individual selection by providing additional replication initiation targets (Blanc and Dujon
692 1980). However, some studies have suggested that mtDNA structural variants, including
693 duplications, can be a conduit for molecular adaptation (Minhas et al. 2023).
694
695 *Conclusions*
696
697 Polarized polymorphisms in *C. elegans* natural isolates allowed us to study a larger
698 sample of mtDNA mutations than previously achieved in laboratory settings (Denver et al
699 2001; Konrad et al 2017; Waneka et al 2021; Leuthner et al 2022). While this approach may
700 lack the controls possible in laboratory experiments, we show that the laboratory environment

701 can introduce a unique set of biases to the mutational spectrum of mtDNA, possibly due to
702 higher levels of mtDNA oxidative damage in experimental populations. ETC complex I
703 genes have greater levels of nonsynonymous polymorphism than other protein-coding
704 mtDNA genes in *C. elegans*. Whereas this could suggest less stringent selection on ETC
705 complex I genes, several studies have linked complex I mutations with local adaptation
706 (Mishmar et al. 2006; Garvin et al. 2015; Camus et al. 2017; Li et al. 2024). Local sequence
707 context, especially A+T-content of flanking nucleotides, significantly influences synonymous
708 polymorphism between- and within-genes. Comparisons between rare polymorphisms found
709 only in single mitotypes and inferred ancestral polymorphism present in more than a single
710 mitotype suggest that natural selection contributes to codon usage bias in *C. elegans*
711 mitochondrial genomes. Finally, we identify the first large heteroplasmic mtDNA structural
712 variants in natural isolates of *C. elegans*. Large, and likely highly disruptive, these structural
713 variants in wild populations possibly possess selfish properties akin to those observed in
714 similar variants arising in experimental laboratory populations. Inferences of mutation spectra
715 using data from natural populations is not without caveats, particularly regarding the possible
716 effects of selection on synonymous sites and possible errors in inferring ancestral states.
717 Nevertheless, the study of mutation spectra in natural isolates is highly relevant for the
718 interpretation of molecular evolution data, especially in organisms such as *C. elegans* whose
719 ecological habitat and environment differ significantly their rearing conditions in
720 laboratories.
721

722 ## MATERIALS AND METHODS

723

724 *C. elegans natural isolates used in this study*

725

726 Genomic DNA sequences of 550 isotypes corresponding to 1,524 natural isolates of
727 *C. elegans* were downloaded from the CaeNDR database (Crombie et al. 2024). A complete
728 list of strains used in this study is provided in supplementary file S1. Reads mapping to the
729 mitochondrial genomes were retained from the downloaded alignment files of the 550
730 isotypes. The reference mitochondrial genomes of two congeneric species, *C. brenneri* and *C.*
731 *tropicalis,* (Yang et al. 2016) were downloaded from NCBI for use as outgroups. During a
732 preliminary analysis, we identified 19 lines (supplementary fig. S1) exhibiting an unusual
733 pattern of heteroplasmic variant frequencies. This pattern comprised of several parallel

variants (shared with other isotypes) in very similar heteroplasmic frequencies. The combination of multiple parallel mutations in similar frequencies within a single isotype led us to consider contamination from another isotype as the most likely explanation. The mtDNA sequences of these 19 lines/isotypes were removed from further analysis, resulting in a final set of 531 isotypes.

Phylogenetic reconstruction

A consensus sequence for each natural isolate's alignment data was built using Samtools (Li et al. 2009). The mtDNA consensus sequences for all 531 isotypes and the outgroup reference sequences were aligned using MAFFT (Katoh and Standley 2013). All sequences were trimmed at two locations using SEQKIT (Shen et al. 2016). The first trim removed a variable homopolymeric run between *atp6* and $tRNA_{Lys}$. The second trim removed a highly variable noncoding sequence between $tRNA_{Ala}$ and $tRNA_{Pro}$ thought to contain the origin of replication. Identical mitochondrial consensus sequences (100% match) were clustered using CD-HIT (Li et al. 2001) and one sequence for each cluster was retained for further analysis, reducing the original number of 531 isotype sequences to 347 unique mitotypes. A maximum likelihood tree was built with the 347 unique mitotype sequences and the two reference sequences from the other *Caenorhabditis* species using IQ-TREE2 (Minh et al. 2020) with TIM2 + F + I + R4 as suggested by Model Finder Plus (Kalyaanamoorthy et al. 2017). The ML phylogenetic tree was used to identify the most basal strains or subtree of natural isolates which comprised six strains from Kauai Island in the Hawaiian archipelago. The strain XZ1516 from this basal group was used as an outgroup for subsequent analysis of the natural isolates. A second maximum likelihood tree was subsequently built with TN + F + I + R4 as suggested by Model Finder Plus (Kalyaanamoorthy et al. 2017) using XZ1516 as an outgroup.

Ancestral Sequence Reconstruction (ASR) and inference of the ancestral versus derived variant state

We used ARPIP (Jowkar et al. 2022) to reconstruct the ancestral sequence for each node in the phylogenetic tree. A custom R script using SEQINR (Charif and Lobry 2007) was used to catalog the ancestral mutations inferred by ARPIP along the phylogenetic tree by comparing each node to its two descendants and iterating this pattern until a leaf was reached.

768 This method can take into account parallel mutations and only counts mutations shared by a
769 clade as a single event. To polarize the variants found in the natural isolates, the variants of
770 each strain were compared to their most immediate ancestors for which a reference had been
771 previously reconstructed with ARPIP. To call the variants, the BAM files downloaded from
772 the CaeNDR were first converted back to FASTQ with SAMtools (Li et al. 2009), the reads
773 were aligned to the reconstructed ancestor's sequence using BWA (Li and Durbin 2009), and
774 optical duplicates were marked using Picard MarkDuplicates (Van der Auwera and O'Connor
775 2020). Finally, the variant calling itself was performed using GATK HaplotypeCaller (Van
776 der Auwera and O'Connor 2020) and variant frequency information was added using
777 BCFTOOLS (Li 2011). Subsequent variant annotation was performed in a custom script. For
778 variant annotation, the start and stop positions for each reference's genes were determined
779 using BLAST (Camacho et al. 2009) against the N2 reference sequence. A custom R script
780 using the VCFR (Knaus and Grünwald 2017) and SEQINR (Charif and Lobry 2007)
781 packages was used to annotate the variants. There were several instances of a cluster of lines
782 descended from the same ancestor that had a number of shared variants originally identified
783 as independent parallel mutations. In these instances, we only retained a single representative
784 mitotype with the highest number of mutations from the cluster.
785
786 *Association between flanking nucleotides and mutations*
787
788 The frequencies of the 5′ and 3′ immediate neighbors of synonymous polymorphisms
789 were analyzed to assess a potential mutational bias associated with flanking nucleotides. The
790 neighboring bases of synonymous mutations were analyzed separately for four-fold
791 degenerate and two-fold degenerate sites because codons with four-fold and two-fold
792 degenerate sites in the 3rd base position are more likely to have a G/C and A/T in their 2nd
793 base position, respectively. The counts of bases flanking synonymous mutations at two-fold
794 or four-fold degenerate sites were normalized by the base composition of 5′ or 3′ nucleotides
795 flanking all two-fold or four-fold degenerate sites in the reference mtDNA. The observed
796 counts were then statistically compared to the expected counts based on the base
797 composition.
798
799 *Distribution of synonymous mutations and A+T-content within mtDNA genes*
800

801 In order to analyze the distribution of synonymous mutations within the 12 protein-
802 coding mtDNA genes, a sliding window analysis (nonoverlapping three codon windows) was
803 performed on both cumulative synonymous mutation counts and A+T-content within genes.
804
805 *Codon Usage Bias*
806
807 Four-fold degenerate sites were used to draw conclusions about the mutation
808 spectrum in natural populations of *C. elegans*. However, selection on codon usage in many
809 organisms has the potential to bias the observed base substitution pattern. In order to
810 investigate the possible effect of natural selection on the preferred use of certain codons that
811 could lead to the observed bias in the base composition and mutation spectrum in the
812 mitochondrial genome, we compared the observed proportion of each amino acid codon to its
813 expected value based on the inferred mutation spectrum. We used the observed base
814 composition at the third base position of codons to compute a RSCU accounting for
815 compositional bias (Sharp et al. 1986). We first tested if the observed partition of codons for
816 each amino acid with four-fold degenerate sites showed a significant deviation from the
817 overall distribution of nucleotides at these sites. Post-hoc tests were performed comparing
818 each amino acid's third base composition to that of all other amino acids. For two-fold
819 degenerate sites, the same analyses were performed but the A/G and T/C ending amino acids
820 were analyzed separately. All tests of statistical significance were conducted using Pearson's
821 χ^2 test with 10,000 simulations and Bonferroni corrections.
822
823 The method used to polarize mutations resulted in two categories of polymorphisms,
824 namely, (i) ancestral variants that were inferred between nodes in the reconstructed
825 phylogeny, and (ii) rare variants which were only found in a single mitotype or strain, and
826 inferred to have occurred between a node and a leaf in the phylogeny. Variants shared by
827 large phylogenetic units, inferred to be ancestral, were expected to be, on average, older than
828 variants only found in a single strain or mitotype. Conversely, rare variants that are only
829 found in a single mitotype are expected to be, on average, younger, or perhaps less favored
830 by natural selection than common variants (Zhu et al. 2017). We compared mutations to and
831 from preferred synonymous codons between rare and ancestral synonymous polymorphisms
832 to test for natural selection on synonymous codon usage. All amino acids were assigned their
833 preferred codon as being the most used codon for that amino acid within the *C. elegans* N2

834 mtDNA reference genome (WS283). A *G*-test was performed to test for significance between
835 ancestral and rare polymorphisms.
836
837 *Structural variants (SV)*
838
839 Three SV callers were used to analyze the mtDNA genomes of the 531 isotypes,
840 namely Manta (Chen et al. 2016), DELLY (Rausch et al. 2012) and Pindel (Ye et al. 2009)
841 using the WS283 mitochondrial genome of *C. elegans* for reference. The variants were
842 filtered to (i) remove variants falling within the region between $tRNA_{Ala}$ and $tRNA_{Pro}$, a highly
843 variable region, and (ii) remove variants supported by a low number of reads. This second
844 filter was set at "PAIR_COUNT > 12" for the variants obtained through Manta, "PE >= 10
845 && SR >= 5" for the variants obtained with DELLY and "DP > 500 && VAF > 0.03" for the
846 variants obtained with Pindel. The filtering itself was done using BCFtools (Li et al. 2011).
847 Variants detected by two or more callers were visually confirmed in Integrated Genome
848 Viewer (IGV) (Robinson et al. 2011) before further analysis. During preliminary analyses,
849 two mtDNA structural variants were called as deletions spanning almost the entirety of the
850 mitochondrial genome. Since these callers treat the genome as linear with artificial ends
851 downstream of the D-loop rather than circular, duplications spanning the ends of the
852 reference genome can lead to variants being incorrectly designated as deletions. In order to
853 circumvent this limitation, a modified reference genome with its ends placed in a noncoding
854 region between *nd4* and *cox1* at mtDNA:7,773 was used. For each variant, the heteroplasmic
855 frequency was calculated as the ratio of the relative mean read depth of the variant (the mean
856 read depth within the variant's region relative to the mean read depth outside of the variant's
857 region) over the relative mean read depth of the region across all other lines without SVs. Per
858 position read-depth was obtained using Samtools (Li et al. 2009). Due to less effective read-
859 mapping in the highly variable D-loop region, reads between mtDNA:12,637-13,794 were
860 not used in the variant frequency calculations. Variants spanning this region had their
861 frequencies computed separately upstream and downstream within the duplicated region and
862 averaged.
863
864 **SUPPLEMENTARY MATERIAL**
865 Supplementary material is available at *Genome Biology and Evolution* online.
866

ACKNOWLEDGEMENTS

The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at Uppmax, [partially funded by the Swedish Research Council through grant agreements no. 2022-06725 and no. 2018-05973].

AUTHOR CONTRIBUTIONS

V.K. and U.B. conceived, designed and jointly supervised the study. A.S. conducted the bioinformatic phylogenetic analyses and generated the variant calls. A.S. and U.B. analyzed the data. All authors contributed to manuscript preparation.

FUNDING

This work was partially supported by a Swedish Research Council Grant to V.K. (2023-03860), a National Science Foundation grant (MCB-1817762) to V.K. and U.B., and Uppsala University start-up funds to V.K. A.S. was partially supported by a Tullberg scholarship for biological research from the T. Tullbergs Stipendiestiftelse and a Liljewalch travel scholarship from the Liljewalchs Stipendiestiftelse.

DATA AVAILABILITY

All data associated with this study have been made available in the supplementary materials.

888 **REFERENCES**

889

890 Allert M, Cox JC, Hellinga HW. 2010. Multifactorial determinants of protein expression in
891 prokaryotic open reading frames. *J Mol Biol.* 402:905–918.
892 Beckman KB, Ames BN. 1998. The free radical theory of aging matures. *Physiol*
893 *Rev.* 78:547–581.
894 Blanc H, Dujon B. 1980. Replicator regions of the yeast mitochondrial DNA responsible for
895 suppressiveness. *Proc Natl Acad Sci U S A*. 77:3942–3946.
896 Brown WM, George Jr M, Wilson AC. 1979. Rapid evolution of animal mitochondrial DNA.
897 *Proc Natl Acad Sci U S A*. 76:1967–1971.
898 Brys K, Vanfleteren JR, Braeckman BP. 2007. Testing the rate-of-living/oxidative damage
899 theory of aging in the nematode model *Caenorhabditis elegans*. *Exp Gerontol.*
900 42:845–851.
901 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
902 BLAST+: architecture and applications. *BMC Bioinformatics*. 10:421.
903 Camus MF, Wolff JN, Sgrò CM, Dowling DK. 2017. Experimental support that natural
904 selection has shaped the latitudinal distribution of mitochondrial haplotypes in
905 Australian *Drosophila melanogaster*. *Mol Biol Evol.* 34:2600–2612.
906 Chang AJ, Chronis N, Karow DS, Marletta MA, Bargmann CI. 2006. A distributed
907 chemosensory circuit for oxygen preference in *C. elegans*. *PLoS Biol.* 4:e274.
908 Charif D, Lobry JR. 2007. SeqinR 1.0-2: A contributed package to the R project for statistical
909 computing devoted to biological sequences retrieval and analysis. In: Bastolla U,
910 Porto M, Roman HE, Vendruscolo M, editors. *Structural approaches to sequence*
911 *evolution*. Biological and medical physics, biomedical engineering. New York:
912 Springer Verlag. p. 207–232.
913 Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on
914 neutral molecular variation. *Genetics*. 134:1289–1303.
915 Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ,
916 Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and
917 indels for germline and cancer sequencing applications. *Bioinformatics*. 32:1220-
918 1222.
919 Cheng KC, Cahill DS, Kasai H, Nishimura S, Loeb LA. 1992. 8-Hydroxyguanine, an
920 abundant form of oxidative DNA damage, causes G-T and A-C substitutions. *J Biol*
921 *Chem.* 267:166–172.
922 Crombie TA, McKeown R, Moya ND, Evans KS, Widmayer SJ, LaGrassa V, Roman N,
923 Tursunova O, Zhang G, Gibson SB, et al. 2024. CaeNDR, the *Caenorhabditis* natural
924 diversity resource. *Nucleic Acids Res.* 52:D850-D858.
925 Crombie TA, Zdraljevic S, Cook DE, Tanny RE, Brady SC, Wang Y, Evans KS, Hahnel S,
926 Lee D, Rodriguez BC, et al. 2019. Deep sampling of Hawaiian *Caenorhabditis*
927 *elegans* reveals high genetic diversity and admixture with global populations. *eLife*.
928 8:e50465.
929 Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T,

930 McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools.
931 *Gigascience*. 10:giab008.
932 Denver DR, Morris K, Lynch M, Vassilieva LL, Thomas WK. 2000. High direct estimate of
933 the mutation rate in the mitochondrial genome of *Caenorhabditis elegans*. *Science*.
934 289:2342–2344.
935 Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed
936 proteins evolve slowly. *Proc Natl Acad Sci U S A*. 102:14338–14343.
937 Dubie JJ, Caraway AR, Stout MM, Katju V, Bergthorsson U. 2020. The conflict within:
938 origin, proliferation and persistence of a spontaneously arising selfish mitochondrial
939 genome. *Philos Trans R Soc Lond B Biol Sci.* 375:20190174.
940 Dubie JJ, Katju V, Bergthorsson U. 2024. Dissecting the sequential evolution of a selfish
941 mitochondrial genome in *Caenorhabditis elegans*. *Heredity (Edinb).* 133:186–197.
942 Eyre-Walker A, Smith NH, Smith JM. 1999. How clonal are human mitochondria? *Proc Biol*
943 *Sci.* 266:477–483.
944 Foster PG, Jermiin LS, Hickey DA. 1997. Nucleotide composition bias affects amino acid
945 content in proteins coded by animal mitochondria. *J Mol Evol.* 44:282–288.
946 Frézal L, Félix M-A. 2015. *C. elegans* outside the Petri dish. *eLife*. 4:e05849.
947 Galtier N, Jobson RW, Nabholz B, Glémin S, Blier P. 2009. Mitochondrial whims: metabolic
948 rate, longevity, and the rate of molecular evolution. *Biol Lett.* 5:413–416.
949 Garvin MR, Bielawski JP, Sazanov LA, Gharrett AJ. 2015. Review and meta-analysis of
950 natural selection in mitochondrial complex I in metazoans. *J Zoolog Syst Evol Res.*
951 53:1–17.
952 Gillooly JF, Allen AP, West GB, Brown JH. 2005. The rate of DNA evolution: effects of
953 body size and temperature on the molecular clock. *Proc Natl Acad Sci U S A*.
954 102:140–145.
955 Gitschlag BL, Kirby CS, Samuels DC, Gangula RD, Mallal SA, Patel MR. 2016.
956 Homeostatic responses regulate selfish mitochondrial genome dynamics in *C.*
957 *elegans*. *Cell Metab.* 24:91–103.
958 Gitschlag BL, Tate AT, Patel MR. 2020. Nutrient status shapes selfish mitochondrial genome
959 dynamics across different levels of selection. *Elife.* 9:e56686.
960 Goodman DB, Church GM, Kosuri S. 2013. Causes and effects of N-terminal codon bias in
961 bacterial genes. *Science.* 342:475–479.
962 Harman D. 1956. Aging: a theory based on free radical and radiation chemistry. *J Gerontol.*
963 11:298–300.
964 Hershberg R, Petrov DA. 2008. Selection on codon bias. *Annu Rev Genet.* 42:287–299.
965 Honda Y, Honda S. 1999. The *daf-2* gene network for longevity regulates oxidative stress
966 resistance and Mn-superoxide dismutase gene expression in *Caenorhabditis elegans*.
967 *FASEB J.* 13:1385–1393.
968 Howe DK, Baer CF, Denver DR. 2009. High rate of large deletions in *Caenorhabditis*
969 *briggsae* mitochondrial genome mutation processes. *Genome Biol Evol.* 2:29–38.
970 Jowkar G, Pečerska J, Maiolo M, Gil M, Anisimova M. 2023. ARPIP: ancestral sequence
971 reconstruction with insertions and deletions under the Poisson indel process. *Syst Biol.*
972 72:307–318.

973 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder:
974 fast model selection for accurate phylogenetic estimates. *Nat Methods.* 14:587–589.
975 Katju V, Bergthorsson U. 2019. Old trade, new tricks: insights into the spontaneous mutation
976 process from the partnering of classical mutation accumulation experiments with
977 high-throughput genomic approaches. *Genome Biol Evol.* 11:136–165.
978 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
979 improvements in performance and usability. *Mol Biol Evol.* 30:772–780.
980 Kimura M. 1968. Evolutionary rate at the molecular level. *Nature.* 217:624–626.
981 Kirkwood TBL, Kowald A. 2012. The free-radical theory of ageing - older, wiser and still
982 alive: modelling positional effects of the primary targets of ROS reveals new support.
983 *Bioessays*. 34:692–700.
984 Knaus BJ, Grünwald NJ. 2017. vcfr: a package to manipulate and visualize variant call
985 format data in R. *Mol Ecol Resour.* 17:44–53.
986 Konrad A, Thompson O, Waterston RH, Moerman DG, Keightley PD, Bergthorsson U, Katju
987 V. 2017. Mitochondrial mutation rate, spectrum and heteroplasmy in *Caenorhabditis*
988 *elegans* spontaneous mutation accumulation lines of differing population size. *Mol*
989 *Biol Evol.* 34:1319–1334.
990 Kow YW. 2002. Repair of deaminated bases in DNA. *Free Radic Biol Med.* 33:886–893.
991 Ladoukakis ED, Zouros E. 2017. Evolution and inheritance of animal mitochondrial DNA:
992 rules and exceptions. *J Biol Res (Thessalon).* 24:2.
993 Lane N. 2011. Energetics and genetics across the prokaryote-eukaryote divide. *Biol Direct.*
994 6:35.
995 Larsen PL. 1993. Aging and resistance to oxidative damage in *Caenorhabditis elegans*. *Proc*
996 *Natl Acad Sci U S A*. 90:8905–8909.
997 Leuthner TC, Benzing L, Kohrn BF, Bergemann CM, Hipp MJ, Hershberger KA, Mello DF,
998 Sokolskyi T, Stevenson K, Merutka IR, et al. 2022. Resistance of mitochondrial DNA
999 to cadmium and Aflatoxin B1 damage-induced germline mutation accumulation in *C.*
1000 *elegans*. *Nucleic Acids Res*. 50:8626–8642.
1001 Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping
1002 and population genetical parameter estimation from sequencing data. *Bioinformatics*.
1003 27:2987–2993.
1004 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
1005 transform. *Bioinformatics.* 25:1754–1760.
1006 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin
1007 R; 1000 Genome Project Data Processing Subgroup. 2009. The sequence
1008 alignment/map format and SAMtools. *Bioinformatics.* 25:2078–2079.
1009 Li H, Liang X, Peng Y, Liu Z, Zhang L, Wang P, Jin M, Wilson K, Garvin MR, Wu K, et al.
1010 2024. Novel mito-nuclear combinations facilitate the global invasion of a major
1011 agricultural crop pest. *Adv Sci (Weinh)*. 11:e2305353.
1012 Li W, Jaroszewski L, Godzik A. 2001. Clustering of highly homologous sequences to reduce
1013 the size of large protein databases. *Bioinformatics.* 17:282–283.
1014 Liau W-S, Gonzalez-Serricchio AS, Deshommes C, Chin K, LaMunyon CW. 2007. A
1015 persistent mitochondrial deletion reduces fitness and sperm performance in
1016 heteroplasmic populations of *C. elegans. BMC Genet.* 8:8.

1017 Lu L, Zhao X, Zhang J, Li M, Qi Y, Zhou L. 2017. Calycosin promotes lifespan in
1018 *Caenorhabditis elegans* through insulin signaling pathway via *daf-16*, *age-1* and *daf-*
1019 *2*. *J Biosci Bioeng.* 124:1–7.
1020 Martin AP, Palumbi SR. 1993. Body size, metabolic rate, generation time, and the molecular
1021 clock. *Proc Natl Acad Sci U S A.* 90:4087–4091.
1022 Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res.*
1023 23:23–35.
1024 McDonald JH, Kreitman M. 1991. Adaptive evolution at the *Adh* locus in *Drosophila*.
1025 *Nature.* 351:652–654.
1026 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A,
1027 Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic
1028 inference in the genomic era. *Mol Biol Evol.* 37:1530–1534.
1029 Minhas BF, Beck EA, Cheng C-HC, Catchen J. 2023. Novel mitochondrial genome
1030 rearrangements including duplications and extensive heteroplasmy could underlie
1031 temperature adaptations in Antarctic notothenioid fishes. *Sci Rep.* 13:6939.
1032 Mishmar D, Ruiz-Pesini E, Mondragon-Palomino M, Procaccio V, Gaut B, Wallace DC.
1033 2006. Adaptive selection of mitochondrial complex I subunits during primate
1034 radiation. *Gene*. 378:11-18.
1035 Nabholz B, Ellegren H, Wolf JBW. 2013. High levels of gene expression explain the strong
1036 evolutionary constraint of mitochondrial protein-coding genes. *Mol Biol Evol.*
1037 30:272–284.
1038 Pearl R. 1928. The rate of living. New York: Knopf.
1039 Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural
1040 variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*.
1041 28:i333–i339.
1042 Rand DM. 1994. Thermal habit, metabolic rate and the evolution of mitochondrial DNA.
1043 *Trends Ecol Evol.* 9:125–131.
1044 Richter C, Park JW, Ames BN. 1988. Normal oxidative damage to mitochondrial and nuclear
1045 DNA is extensive. *Proc Natl Acad Sci U S A.* 85:6465–6467.
1046 Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP.
1047 2011. Integrative genomics viewer. *Nat Biotechnol.* 29:24–26.
1048 Rubner M. 1908. Das problem der lebensdauer und seine beziehungen zu wachstum und
1049 ernährung. Berlin, Boston: Oldenbourg Wissenschaftsverlag.
1050 Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012. Detection of ultra-rare
1051 mutations by next-generation sequencing. *Proc Natl Acad Sci U S A*. 109:14508–
1052 14513.
1053 Sequeira AN, O'Keefe IP, Katju V, Bergthorsson U. 2024. Friend turned foe: selfish behavior
1054 of a spontaneously arising mitochondrial deletion in an experimentally evolved
1055 *Caenorhabditis elegans* population. *G3 (Bethesda).* 14:jkae018.
1056 Sharp PM, Tuohy TM, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly
1057 differentiates highly and lowly expressed genes. *Nucleic Acids Res.* 14:5125–5143.
1058 Shen W, Le S, Li Y, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q
1059 file manipulation. *PLoS One.* 11:e0163962.

1060 Thompson O, Edgley M, Strasbourger P, Flibotte S, Ewing B, Adair R, Au V, Chaudhry I,
1061 Fernando L, Hutter H, et al. 2013. The million mutation project: a new approach to
1062 genetics in *Caenorhabditis elegans*. *Genome Res.* 23:1749–1762.
1063 Tsang WY, Lemire BD. 2002. Stable heteroplasmy but differential inheritance of a large
1064 mitochondrial DNA deletion in nematodes. *Biochem. Cell Biol.* 80:645–654.
1065 Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I,
1066 Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency
1067 protein translation. *Cell.* 141:344–354.
1068 Van der Auwera GA, O'Connor BD. 2020. Genomics in the cloud: ssing Docker, GATK, and
1069 WDL in Terra. O'Reilly Media, Inc.
1070 Van Raamsdonk JM, Meng Y, Camp D, Yang W, Jia X, Bénard C, Hekimi S. 2010.
1071 Decreased energy metabolism extends life span in *Caenorhabditis elegans* without
1072 reducing oxidative damage. *Genetics.* 185:559–571.
1073 Waneka G, Svendsen JM, Havird JC, Sloan DB. 2021. Mitochondrial mutations in
1074 *Caenorhabditis elegans* show signatures of oxidative damage and an AT-bias.
1075 *Genetics.* 219:iyab116.
1076 Yang SS, Li S, Wang GX. 2016. The complete mitochondrial genome of *Caenorhabditis*
1077 *tropicalis* n. sp. (Rhabditida: Rhabditidae). *Mitochondrial DNA A DNA Mapp Seq*
1078 *Anal.* 27:1763–1764.
1079 Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to
1080 detect break points of large deletions and zshort reads. *Bioinformatics*. 25:2865–2871
1081 Zhu YO, Sherlock G, Petrov DA. 2017. Extremely rare polymorphisms in *Saccharomyces*
1082 *cerevisiae* allow inference of the mutational spectrum. *PLoS Genet*. 13:e1006455.

Table 1. The number and normalized proportions of base substitutions in this study and comparisons to three preceding experimental studies in a laboratory setting. The normalized proportions of each mutation were computed by first dividing the counts of a given mutation type (for example, A/T → G/C) by the number of the mutated nucleotides (A/T) in the genome. The results were then divided by the sum of these fractions for all six mutational categories.

Study	Mutation	Transversions				Transitions	
		A/T → C/G	A/T → T/A	G/C → T/A	G/C → C/G	A/T → G/C	G/C → A/T
This study	Counts	54	224	144	9	1057	875
	Proportions	0.023	0.095	0.061	0.004	0.447	0.370
	Normalised proportions	0.012	0.050	0.098	0.006	0.233	0.601
This study Four-fold degenerate sites	Counts	39	129	48	5	385	290
	Proportions	0.044	0.144	0.054	0.006	0.430	0.324
	Normalised proportions	0.014	0.047	0.111	0.012	0.141	0.675
Waneka *et al.* (2021)	Counts	2	20	79	8	50	94
	Proportions	0.008	0.079	0.312	0.032	0.198	0.372
	Normalised proportions	1.25E-05	0.032	0.388	0.039	0.079	0.462
Waneka *et al.* (2021) Four-fold degenerate sites	Counts	1	5	4	0	10	9
	Proportions	0.034	0.172	0.138	0	0.345	0.310
	Normalised proportions	0.010	0.051	0.258	0	0.101	0.580
Leuthner *et al.* (2022)	Counts	22	76	300	225	44	93
	Proportions	0.029	0.100	0.395	0.296	0.058	0.122
	Normalised proportions	1.42E-05	0.037	0.457	0.343	0.022	0.142
Konrad et al. (2017)	Counts	0	1	3	0	0	5
	Proportions	0	0.111	0.333	0	0	0.556
	Normalised proportions	0	0.039	0.360	0	0	0.600

Table 2. Large deletions identified by three SV callers (Delly, Manta and Pindel). The heteroplasmic frequencies were computed separately based on the ratio of the relative read depth of the deleted region in the focal strain relative to the read depth of the corresponding region across all other strains.

Strain	Position	Type	Size (bp)	Genes affected	Heteroplasmic frequency
ECA243	mtDNA:2,365-3,015	Deletion	651	*nd1*, *atp6*	0.41
JU1249	mtDNA:1,839-2,329	Deletion	491	*nd1*	0.79
JU2001	mtDNA:4,412-6,018	Deletion	1,607	$tRNA_{Gln}$, $tRNA_{Phe}$, *ctb1*, $tRNA_{Leu}$, *cox3*	0.45
NIC258	mtDNA:3,341-4,419	Deletion	1,079	$tRNA_{Leu}$, $tRNA_{Ser}$, *nd2*, $tRNA_{Ile}$, $tRNA_{Arg}$, $tRNA_{Gln}$	0.55
QG2075	mtDNA:9,797-11,815	Deletion	2,019	*cox2*, $tRNA_{His}$, 16S rRNA, *nd3, nd5*	0.04
QG2932	mtDNA:8,327-12,320	Deletion	3,994	*cox1*, $tRNA_{Cys}$, $tRNA_{Met}$,$tRNA_{Asp}$, $tRNA_{Gly}$, *cox2*, $tRNA_{His}$, 16S rRNA, *nd3, nd5*	0.095
MY23	mtDNA:12,024-452	Duplication	2,222	*nd5*, $tRNA_{Ala}$, D-loop, $tRNA_{Pro}$, tRNA_{Val}, *nd6*	0.10
PS2025	mtDNA:12,015-444	Duplication	2,223	*nd5*, $tRNA_{Ala}$, D-loop, $tRNA_{Pro}$, tRNA_{Val}, *nd6*	0.54

FIGURE LEGENDS

Fig. 1. The mutational spectrum of base substitutions in the mitochondrial genome of *C. elegans*. The mutation counts were normalized by the reference genome's base composition and scaled such that the sum of the frequencies in each experiment equals one. Transitions and transversions are denoted in grey and black, respectively. a) All 2,376 base substitutions inferred in this study of natural isolates. b) Base substitutions at four-fold degenerate sites in the natural isolates (900 SNPs). c) Base substitutions reported in a duplex sequencing study by Leuthner et al. 2022 (760 SNPs). d) Base substitutions reported in a duplex sequencing study by Waneka et al. 2021 (253 SNPs).

Fig. 2. The proportion of polymorphic base substitutions categorized by codon position and site degeneracy. a) The proportion of polymorphic base substitutions at the first, second and third codon position in *C. elegans* natural isolates (left) and in duplex sequencing results under relaxed selection (Waneka et al. 2021, right). b) The proportion of polymorphic base substitutions at nondegenerate, two-fold and four-fold degenerate sites in the *C. elegans* natural isolates were standardized by the total number of nondegenerate, two-fold and four-fold degenerate sites in the *C. elegans* mtDNA reference genome (left) and duplex sequencing results under relaxed selection (Waneka et al. 2021, right). The red horizontal lines represent the expected proportions if all sites accumulate base substitutions at equal rates.

Fig. 3. Transition to transversion (*Ts/Tv*) ratios of base substitutions by codon position and site degeneracy. a) The *Ts/Tv* ratios at different codon positions in *C. elegans* natural isolates (left) and in duplex sequencing results under relaxed selection (Waneka et al. 2021, right). b) The *Ts/Tv* ratios at all sites, non-, two-fold, and four-fold degenerate sites in *C. elegans* natural isolates (left) and in duplex sequencing results under relaxed selection (Waneka et al. 2021, right).

Fig. 4. The frequency of polymorphic base substitutions per site in the 12 mtDNA protein-coding genes in the *C. elegans* natural isolates. a) The frequency of all polymorphic base substitutions. b) The nonsynonymous to synonymous ratio (*N*/*S*) per gene. c) The number of synonymous substitutions per synonymous site. d) The number of nonsynonymous substitutions per nonsynonymous site. The red horizontal lines represent the average for the

protein-coding mitochondrial genome. *G*-tests comparing the observed versus expected mutation counts under homogeneous mutation frequency were performed using a Bonferroni correction for multiple comparisons (p-value < 0.004) except for the nonsynonymous to synonymous ratios which utilized Fisher's exact tests. Significant deviations from the expected frequencies are denoted with an asterisk.

Fig. 5. The frequency of polymorphic base substitutions per site with the mtDNA protein-coding genes grouped by their association with four electron transport chain (ETC) complexes. a) The frequency of all polymorphic base substitutions. b) The nonsynonymous to synonymous ratio (N/S) per ETC complex. c) The number of synonymous substitutions per synonymous site. d) The number of nonsynonymous substitutions per nonsynonymous site. The red horizontal lines represent the average for the protein-coding mitochondrial genome. *G*-tests comparing the observed versus expected mutation counts under homogeneous mutation frequency were performed using a Bonferroni correction for multiple comparisons (p-value < 0.0125) except for the nonsynonymous to synonymous ratios which utilized Fisher's exact tests. Significant deviations from the expected frequencies are denoted with an asterisk.

Fig. 6. Frequency of nucleotides flanking synonymous polymorphisms at four-fold and two-fold degenerate sites normalized by base composition.

Fig. 7. Number of synonymous polymorphisms and A+T-content in relation to the distance from a gene's start and stop codons. a) Synonymous polymorphism as a function of distance from start codons. b) Synonymous polymorphism as a function of distance from stop codons. c) A+T-content as a function of distance from start codons. d) A+T-content as a function of distance from stop codons. The metrics were compiled across the 12 protein-coding genes for the first and last 180 bp with the blue line representing a LOESS curve fitted to the data with a 95% CI shown in grey.

Fig. 8. The distribution of read depth for the four mitochondrial genomes with the highest heteroplasmic frequencies of mtDNA deletions among the natural isolates. The horizontal axis represents the position in the reference strain mtDNA molecule and the vertical axis represents the relative read depth at any given position. The red and blue line are the distribution of relative read depth in the mtDNA harboring the deletion and mtDNA in all

other natural isolates, respectively. The two grey vertical lines indicate the position of the deletion breakpoints. a) Deletion in natural isolate JU1249. b) Deletion in natural isolate NIC258. c) Deletion in natural isolate JU2001. d) Deletion in natural isolate ECA243.

Fig. 9. Variant frequency distribution for a) all mutations identified in the natural isolates comprising this study, b) all small indel variants identified in the natural isolates comprising this study, and c) all variants identified in a long-term *C. elegans* mutation accumulation experiment over 409 consecutive generations (adapted from Konrad et al. 2017). Variants with frequencies ranging from 0.03 to 0.97 were designated as heteroplasmic variants.

FIGURE 1

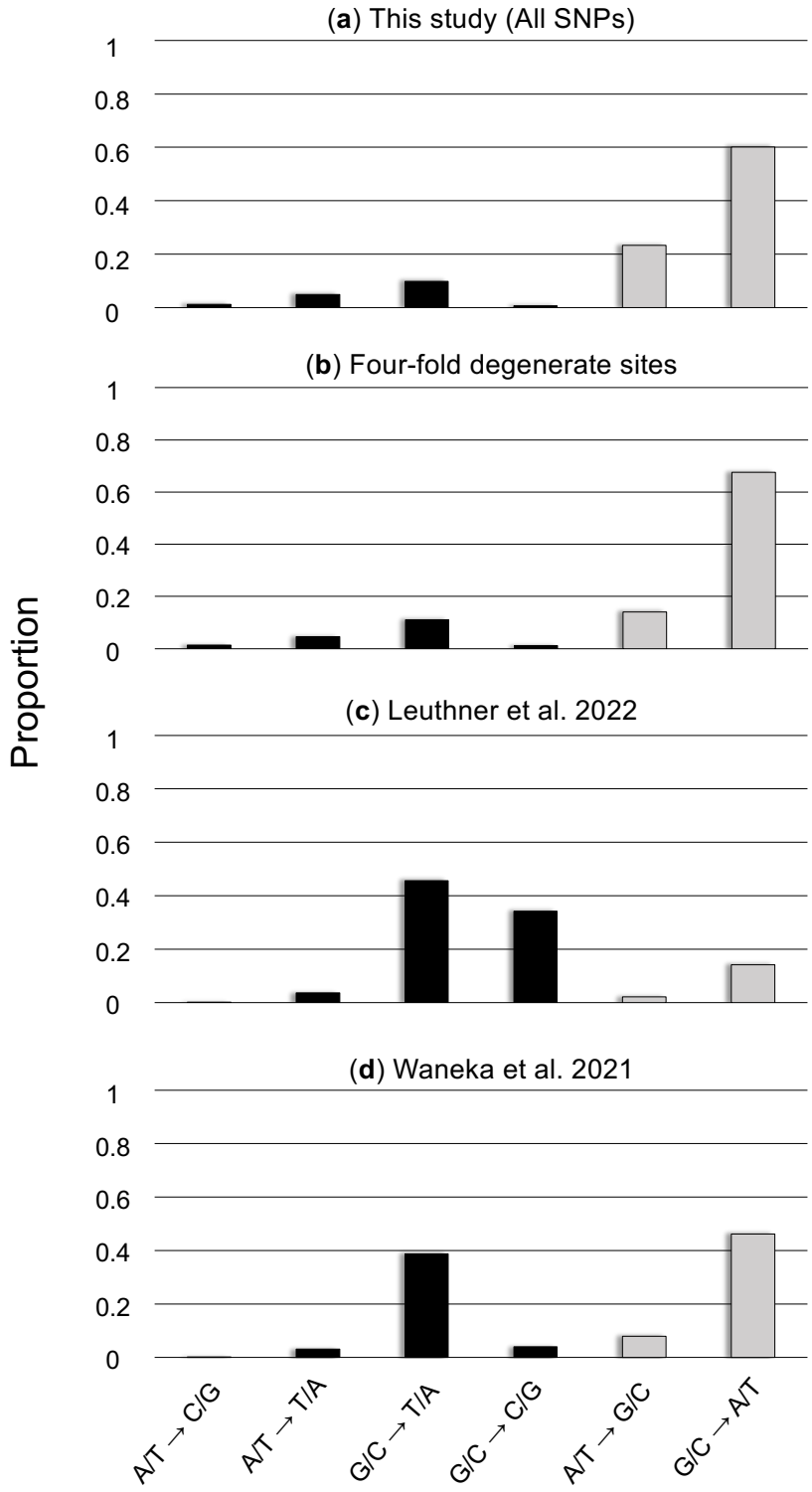

(a) This study (All SNPs)
(b) Four-fold degenerate sites
(c) Leuthner et al. 2022
(d) Waneka et al. 2021
Proportion
1
0.8
0.6
0.4
0.2
0
A/T → C/G
A/T → T/A
G/C → T/A
G/C → C/G
A/T → G/C
G/C → A/T

FIGURE 2

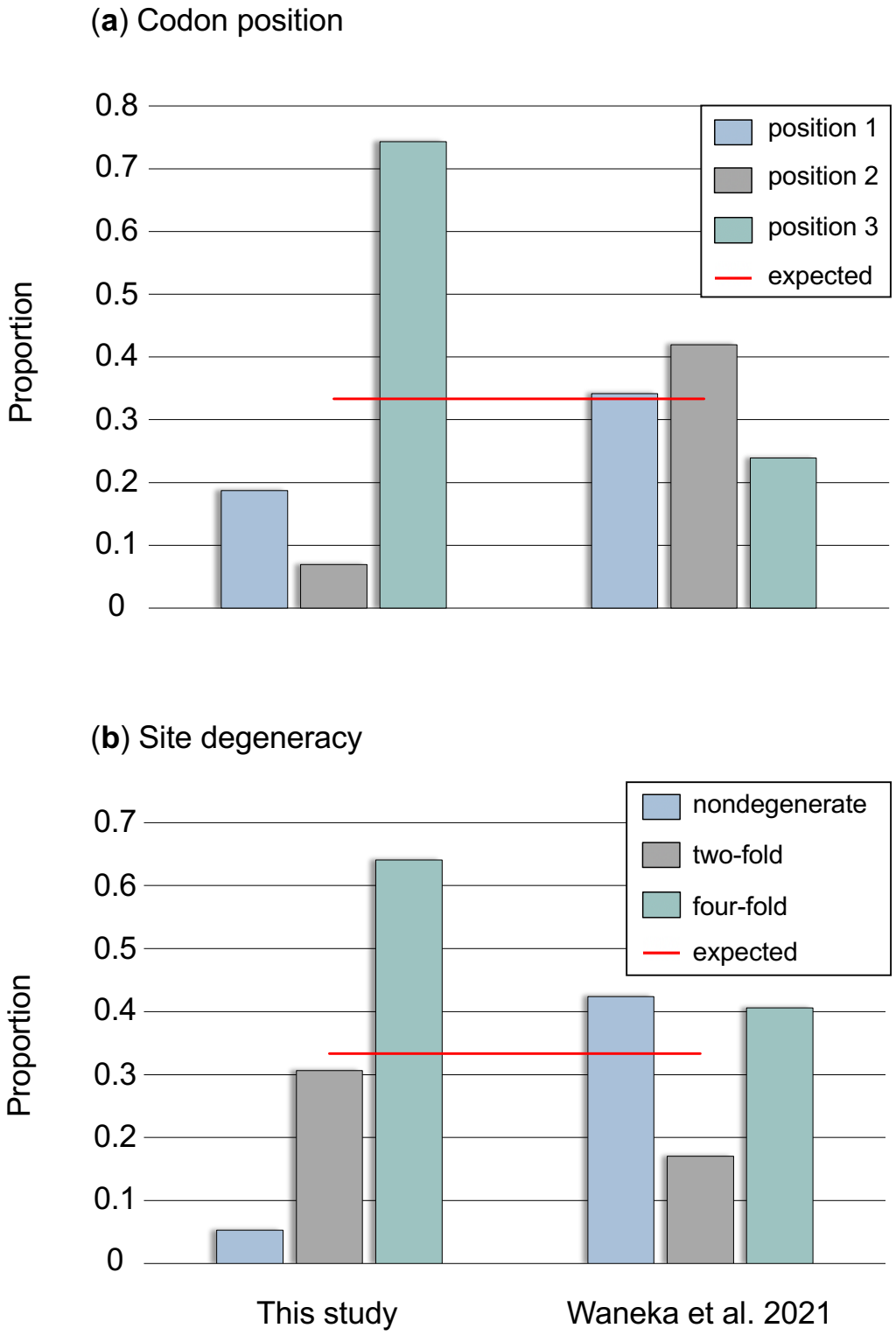

(a) Codon position
Proportion
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
position 1
position 2
position 3
expected
(b) Site degeneracy
Proportion
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
nondegenerate
two-fold
four-fold
expected
This study
Waneka et al. 2021

FIGURE 3

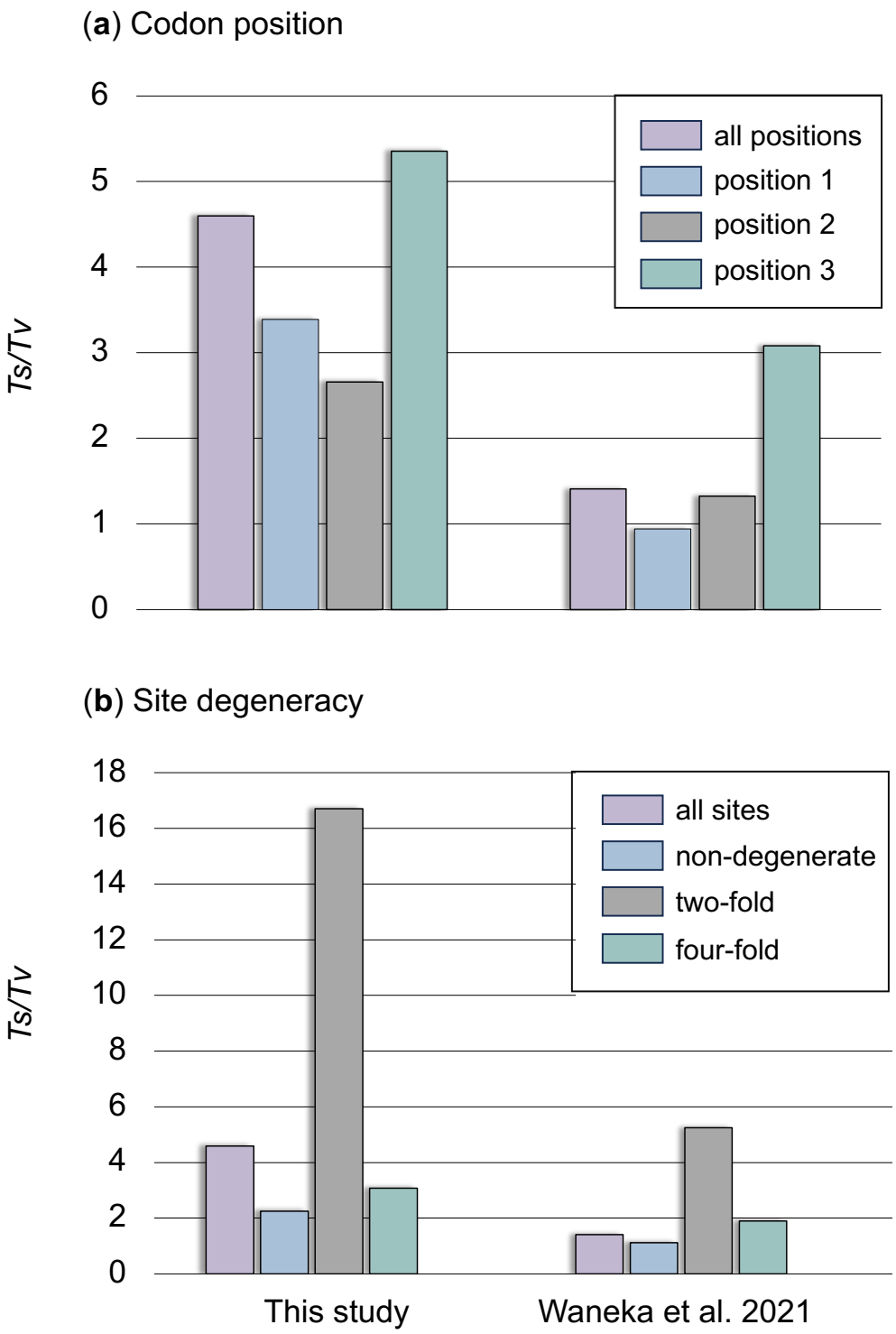

(a) Codon position
Ts/Tv
6
5
4
3
2
1
0
all positions
position 1
position 2
position 3
(b) Site degeneracy
Ts/Tv
18
16
14
12
10
8
6
4
2
0
all sites
non-degenerate
two-fold
four-fold
This study
Waneka et al. 2021

FIGURE 4

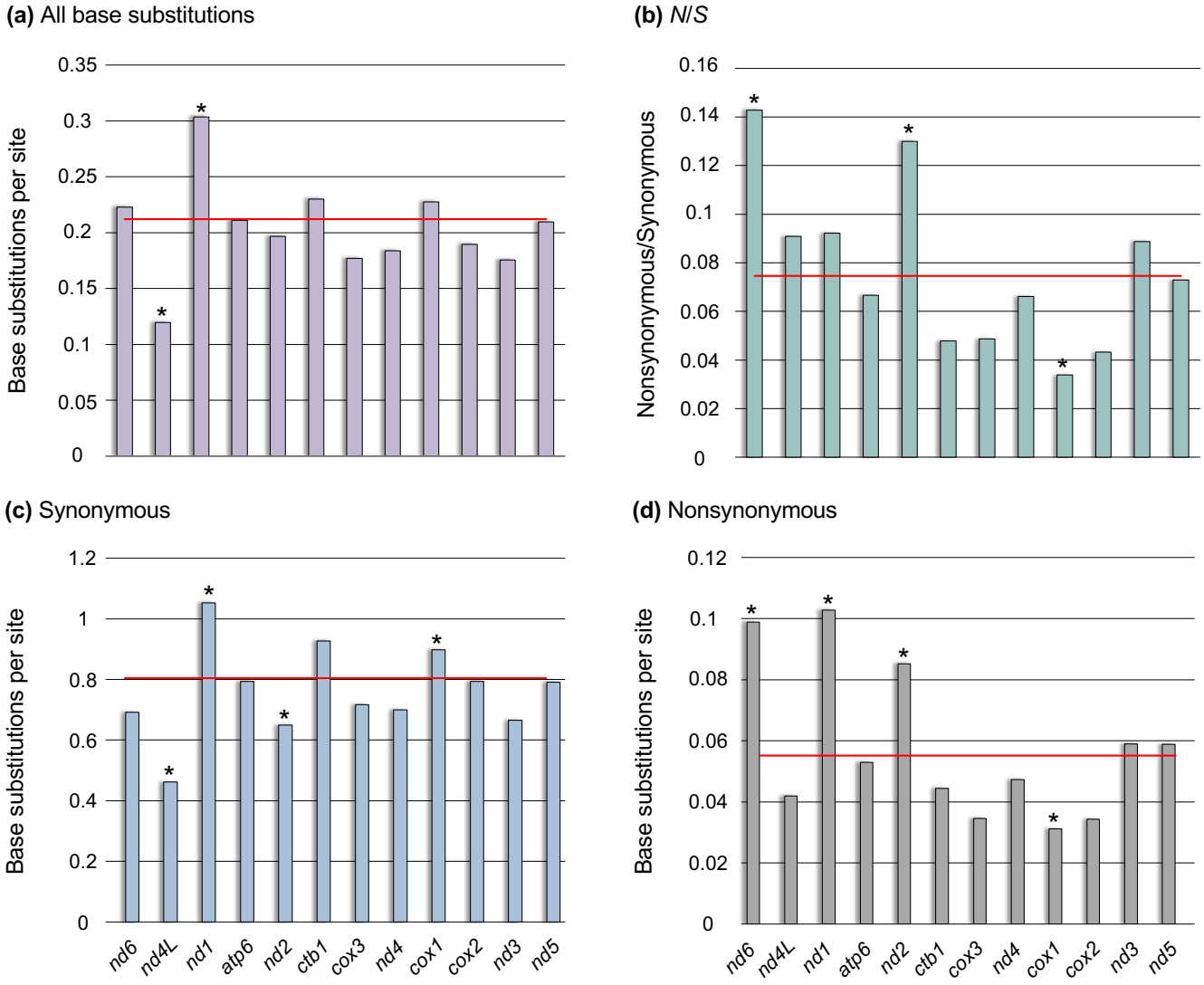

(a) All base substitutions
Base substitutions per site
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
(b) N/S
Nonsynonymous/Synonymous
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
(c) Synonymous
Base substitutions per site
1.2
1
0.8
0.6
0.4
0.2
0
(d) Nonsynonymous
Base substitutions per site
0.12
0.1
0.08
0.06
0.04
0.02
0
nd6
nd4L
nd1
atp6
nd2
ctb1
cox3
nd4
cox1
cox2
nd3
nd5

FIGURE 5

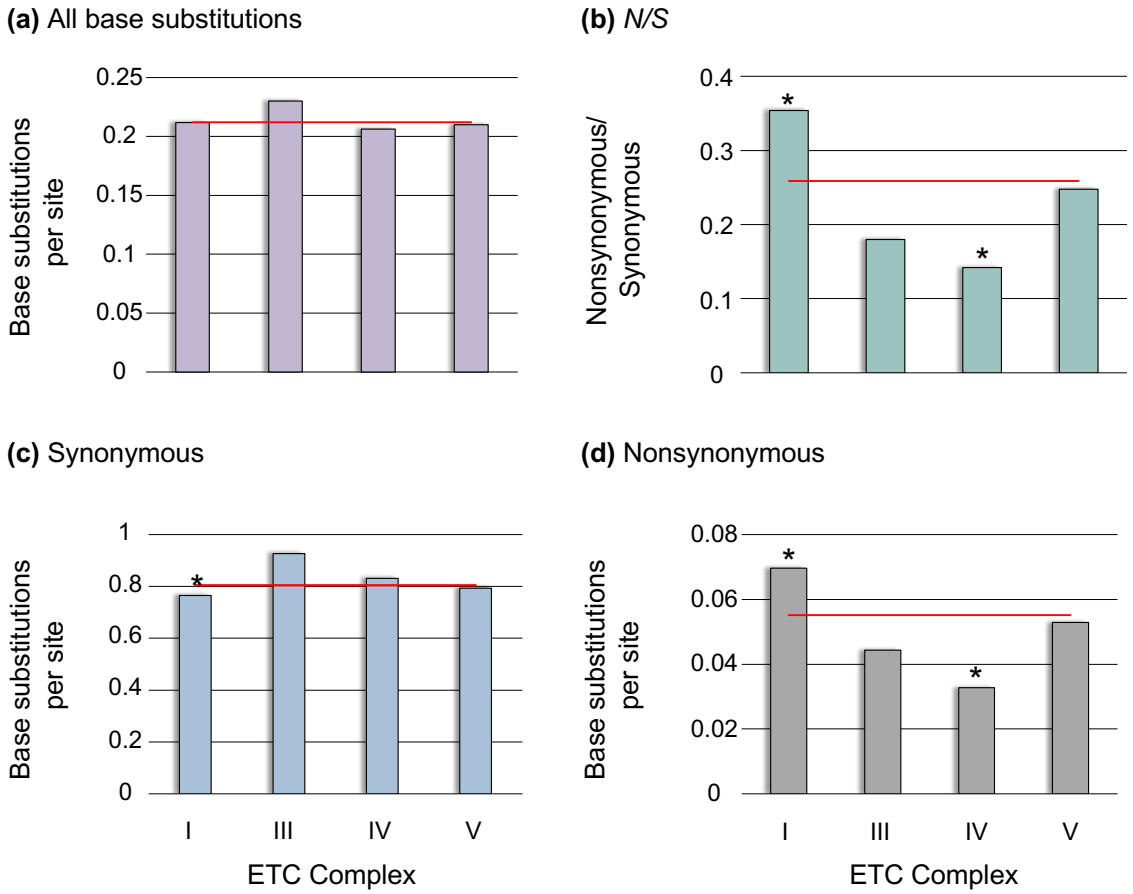

(a) All base substitutions
Base substitutions per site
0.25
0.2
0.15
0.1
0.05
0
(b) N/S
Nonsynonymous/ Synonymous
0.4
0.3
0.2
0.1
0
*
*
(c) Synonymous
Base substitutions per site
1
0.8
0.6
0.4
0.2
0
*
I
III
IV
V
ETC Complex
(d) Nonsynonymous
Base substitutions per site
0.08
0.06
0.04
0.02
0
*
*
I
III
IV
V
ETC Complex

FIGURE 6

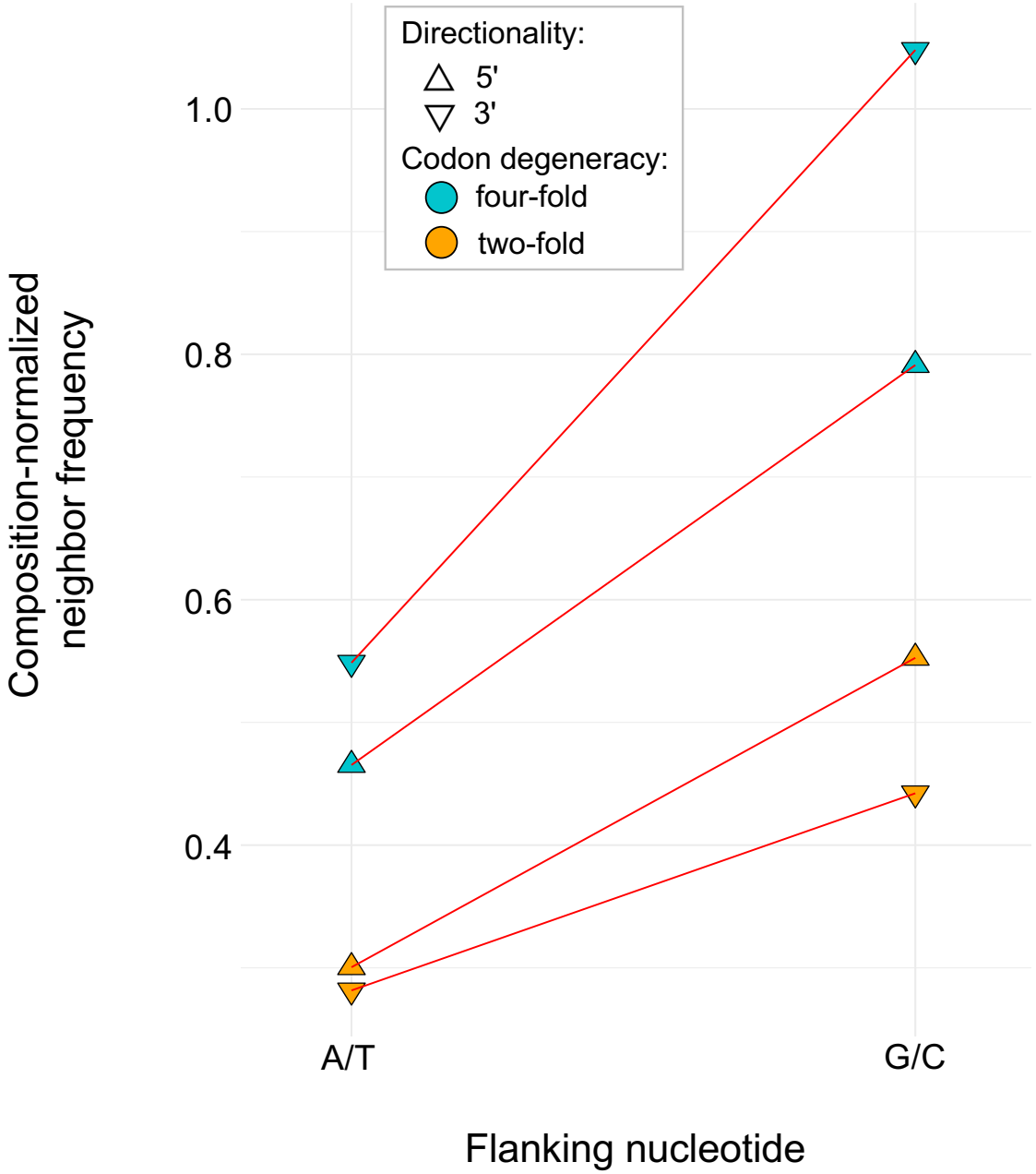

Directionality:
5'
3'
Codon degeneracy:
four-fold
two-fold
1.0
0.8
0.6
0.4
Composition-normalized neighbor frequency
A/T
G/C
Flanking nucleotide

FIGURE 7

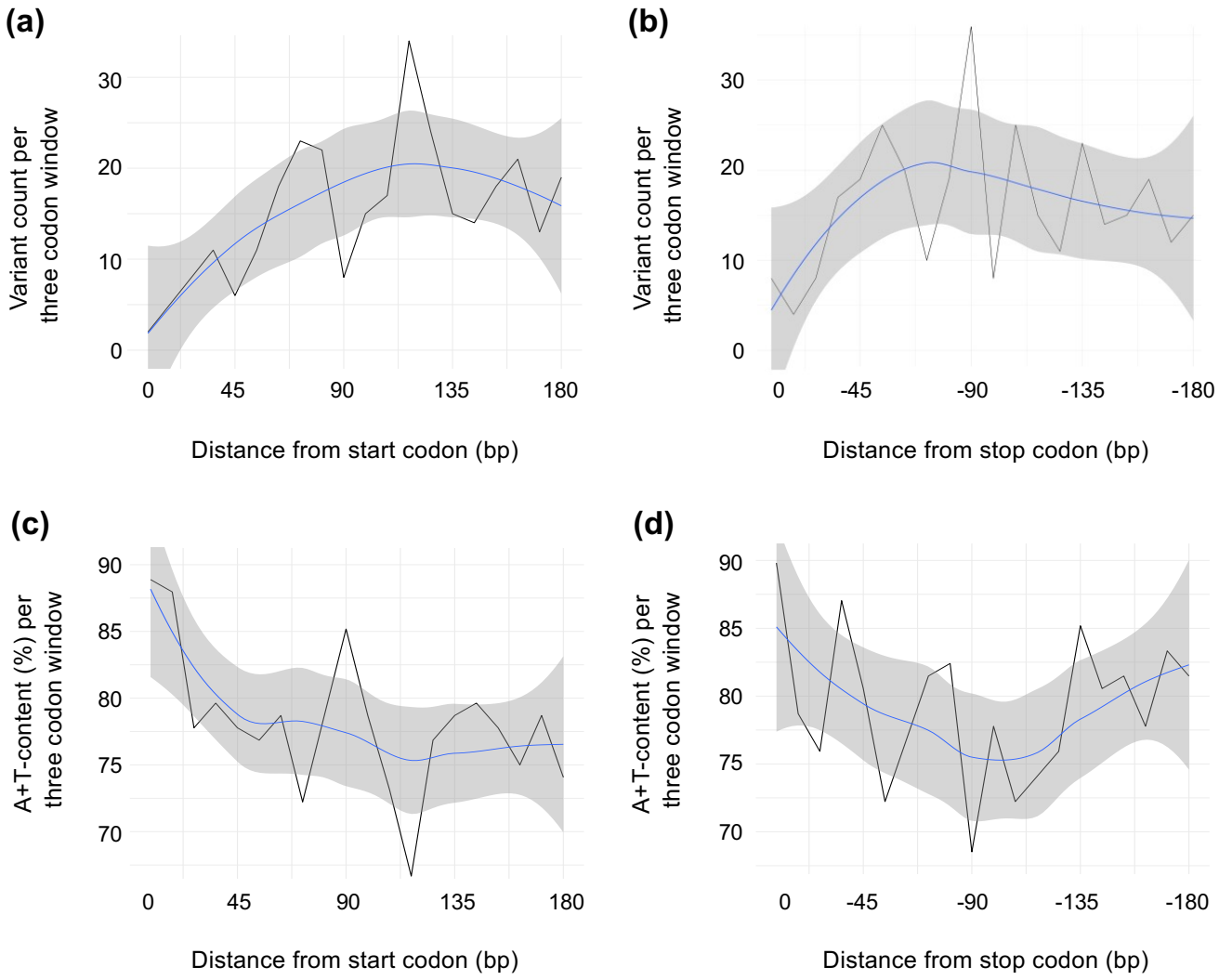

(a)
Variant count per three codon window
30
20
10
0
0
45
90
135
180
Distance from start codon (bp)
(b)
Variant count per three codon window
30
20
10
0
0
-45
-90
-135
-180
Distance from stop codon (bp)
(c)
A+T-content (%) per three codon window
90
85
80
75
70
0
45
90
135
180
Distance from start codon (bp)
(d)
A+T-content (%) per three codon window
90
85
80
75
70
0
-45
-90
-135
-180
Distance from stop codon (bp)

FIGURE 8

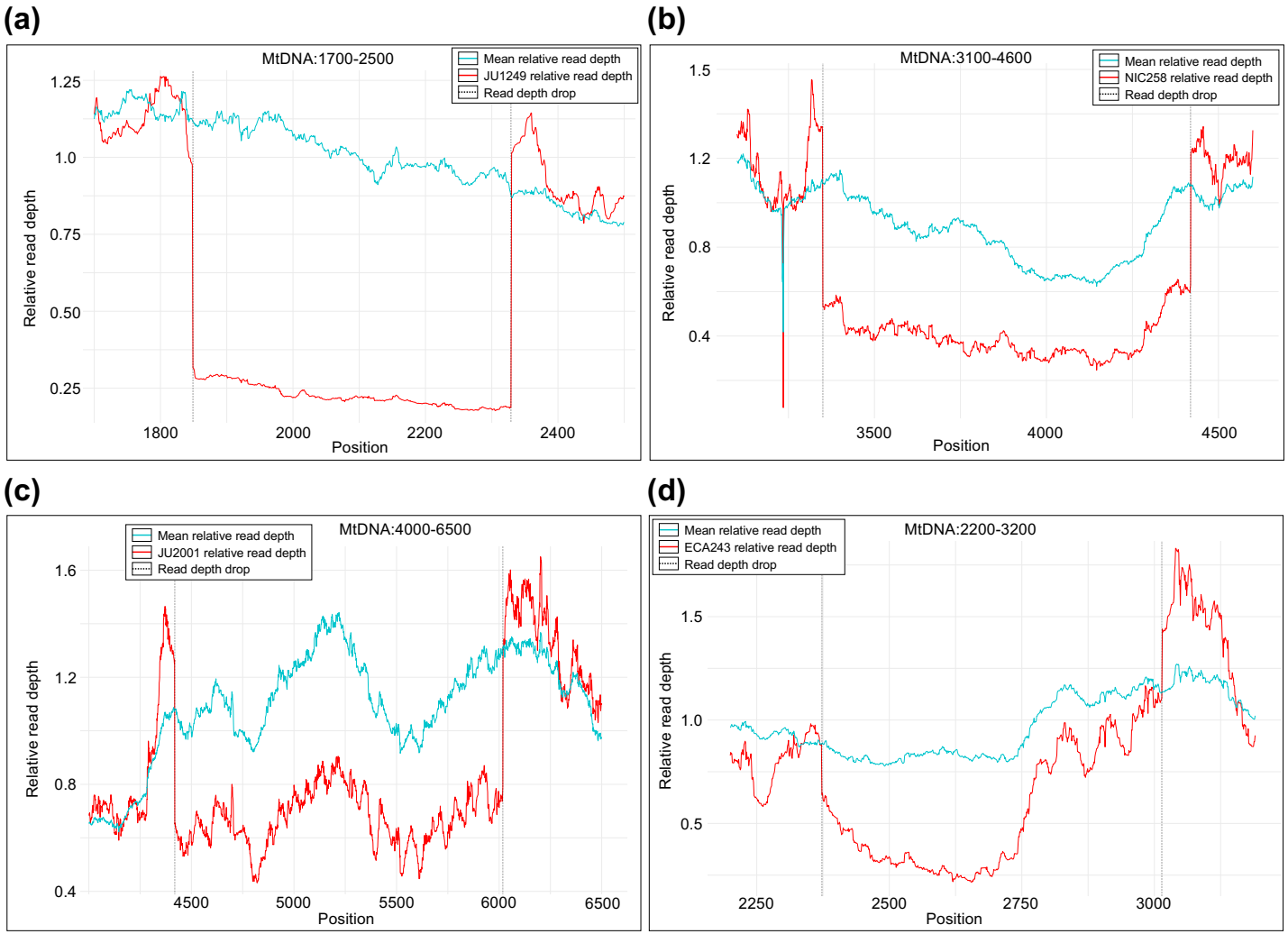
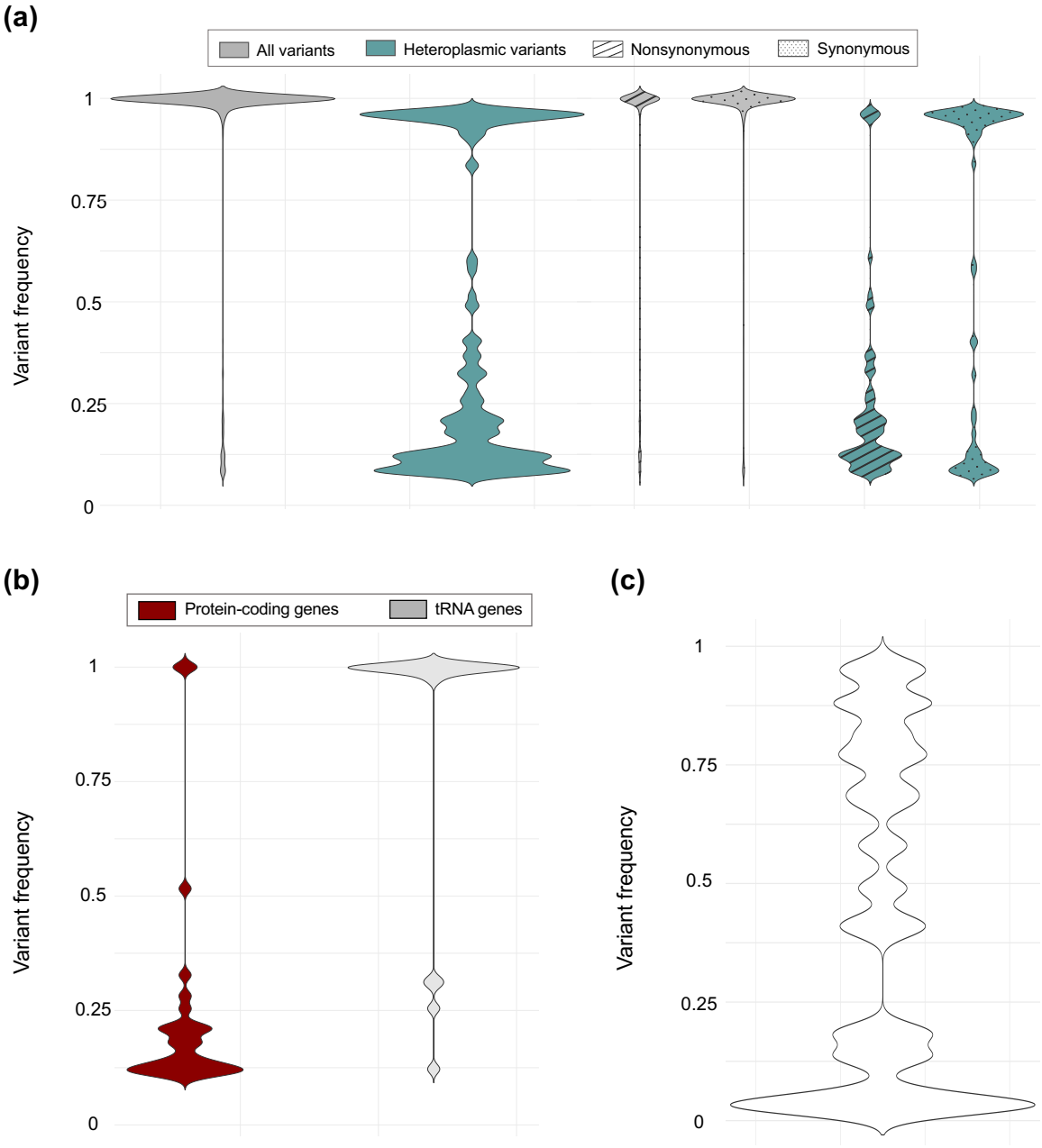

(a)
MtDNA:1700-2500
Mean relative read depth
JU1249 relative read depth
Read depth drop
Relative read depth
1.25
1.0
0.75
0.50
0.25
1800
2000
2200
2400
Position
(b)
MtDNA:3100-4600
Mean relative read depth
NIC258 relative read depth
Read depth drop
Relative read depth
1.5
1.2
0.8
0.4
3500
4000
4500
Position
(c)
MtDNA:4000-6500
Mean relative read depth
JU2001 relative read depth
Read depth drop
Relative read depth
1.6
1.2
0.8
0.4
4500
5000
5500
6000
6500
Position
(d)
MtDNA:2200-3200
Mean relative read depth
ECA243 relative read depth
Read depth drop
Relative read depth
1.5
1.0
0.5
2250
2500
2750
3000
Position

FIGURE 9


(a)
All variants
Heteroplasmic variants
Nonsynonymous
Synonymous
Variant frequency
1
0.75
0.5
0.25
0
(b)
Protein-coding genes
tRNA genes
Variant frequency
1
0.75
0.5
0.25
0
(c)
Variant frequency
1
0.75
0.5
0.25
0