A mutational gradient drives somatic mutation accumulation in mitochondrial DNA and influences germline polymorphisms and genome composition

Monica Sanchez-Contreras[1], Mariya T. Sweetwyne[1], Brendan F. Kohrn[1], Kristine A. Tsantilas[2], Michael J Hipp[1], Elizabeth K. Schmidt[1], Jeanne Fredrickson[1], Jeremy A. Whitson[1], Matthew D. Campbell[3], Peter S. Rabinovitch[1], David J. Marcinek[3], Scott R. Kennedy[1*]

[1]Department of Laboratory Medicine and Pathology, University of Washington, Seattle WA

[2]Department of Biochemistry, University of Washington, Seattle WA

[3]Department of Radiology, University of Washington, Seattle WA

*Corresponding Author: Scott R Kennedy

Email: scottrk@uw.edu

Abstract

Background

Mutations in the mitochondrial genome (mtDNA) can cause devastating maternally inherited diseases, while the accumulation of somatic mtDNA mutations is linked to common diseases of aging. Although mtDNA mutations impact human health, the process(es) that give rise to these mutations are unclear and are under considerable debate. We analyzed the distribution of naturally occurring somatic mutations across the mouse and human mtDNA obtained by Duplex Sequencing to provide clues to the mechanism by which *de novo* mutations arise as well as how the genome is replicated.

Results

We observe two distinct mutational gradients in G→A and T→C transitions, but not their complements, that are delimited by the light-strand origin and the control region (CR). The gradients increase with age and are lost in the absence of DNA polymerase γ proofreading activity. A nearly identical pattern is present in human mtDNA somatic mutations. The distribution of mtDNA single nucleotide polymorphisms (SNPs) in the human population and genome base composition across >3,000 vertebrate species mirror this gradient pattern, pointing to evolutionary conservation of this phenomenon. Lastly, high-resolution analysis of the mtDNA control region highlights mutational 'hot-spots' and 'cold-spots' that strongly align with important regulatory regions.

Conclusions

Collectively, these patterns support an asymmetric strand-displacement mechanism with key regulatory structures in the CR and argue against alternative replication models. The mutational gradient is a fundamental consequence of mtDNA replication that drives somatic mutation accumulation and influences inherited polymorphisms and, over evolutionary timescales, genome composition.

72 **Introduction**

73 Owing to their evolutionary origin, mitochondria have retained a small extra-nuclear genome
74 encoding essential components of the electron transport chain (ETC), as well as transfer and ribosomal
75 RNAs required for their translation (Fig. 1a). The ETC is responsible for producing cellular energy
76 through oxidative phosphorylation and maintaining a reducing chemical environment. As such, the
77 genetic information encoded in the mtDNA is essential for maintaining cellular homeostasis. However,
78 due to the absence of several DNA repair pathways, mtDNA exhibits mutation frequencies >100-fold
79 higher than the nuclear genome [1].

80 Mutations in the mtDNA cause a number of devastating maternally inherited diseases, while the
81 accumulation of mutations in the soma is linked to common diseases of the elderly, including cancer,
82 diabetes, and neurodegenerative diseases (Reviewed in [2]). While an important driver of human health,
83 the mutagenic processes that give rise to these mutations are under considerable debate [3,4]. As
84 originally posited by Denham Harmon, the proximity of mtDNA to the ETC should result in high levels of
85 oxidative damage (*i.e.* 8-oxo-dG), yielding predominantly G→T/C→A transversions [5,6]. Counter to this
86 prediction, low levels of G→T/C→A mutations and a preponderance of G→A/C→T and T→C/A→G
87 transitions are observed [7–10]. The presence of these mutations has been interpreted as arising from
88 either base selection errors by DNA polymerase γ (Pol-γ) or spontaneous deamination of deoxycytidine
89 and deoxyadenosine, and not due to reactive oxygen species (ROS) induced 8-oxo-dG adducts.

90 Regardless of the specific source of mutagenesis, the replication of the mtDNA by Pol-γ is
91 required for fixation of mutations into the genome. Thus, the distribution of mutations can provide clues
92 to the mechanism by which genome replication gives rise to *de novo* mutations. The mechanism of
93 mtDNA replication remains poorly understood, but, in vertebrates, is generally thought to occur via an
94 asynchronous strand displacement mechanism involving two separate, strand-specific, origins
95 [11,12](Fig. 1a-c). In this model, replication is initiated at the heavy-strand (H-strand) origin (Ori_H),
96 located in the non-coding CR, using a displacement loop (D-loop) as the replicon primer (Fig. 1b,d).
97 Synthesis of the nascent H-strand displaces the original H-strand into a single-stranded state. Upon

98 traversing the light-strand (L-strand) origin (Ori_L), located approximately 11,000 bp away from the CR, a
99 second replication fork is established and proceeds in the opposite direction, resulting in the original H-
100 strand becoming double-stranded (Fig. 1c,d). Replication is completed when both replication forks
101 complete their circumnavigation. Alternative vertebrate models have been proposed whereby the
102 displaced H-strand is annealed to RNA transcripts, termed RITOLS, that serve to prevent the single-
103 stranded state and act as intermittent priming sites for L-strand replication (Fig. 1e) [13,14]. Visualization
104 of replication intermediates by 2D-gel electrophoresis has also indicated the presence of coupled-strand
105 synthesis involving a more conventional leading/lagging strand replication fork initiating from a
106 bidirectional origin (Ori_b) in the mtDNA CR or potentially throughout a multi-kilobase "initiation zone"
107 (Fig. 1f) [15–17]. The asynchronous and coupled-strand mechanisms have been proposed to be present
108 at the same time, contingent on the physiological state of the cell [15,18]. Lastly, alternative tRNA genes
109 outside of the Ori_L tRNA cluster have been proposed to act as alternative L-strand origins [19,20].

110 Each of these replication models have significant implications for mtDNA mutagenesis. As
111 hypothesized in previous phylogenetic studies, an asymmetric mechanism of mtDNA replication could
112 explain the phenomena of G/C strand bias, A/T-skew, and mutational gradients seen across taxa [21–
113 23]. Specifically, the long-lasting “naked” ssDNA replication intermediate in the original model predicts
114 elevated levels of G→A/C→T mutations when the template dC is in the (single-stranded) H-strand due
115 to cytidine exhibiting substantially increased deamination rates when present in a single-stranded state
116 [24]. In this case, the mutational pressure is away from dC content in the H-strand and towards increased
117 dT content. Moreover, genes closer to the Ori_H are expected to be more mutation prone than those
118 farther away due to longer times in the single-stranded state. In contrast, both conventional
119 leading/lagging-strand synthesis and intermittent priming models could produce G/C strand bias and/or
120 A/T-skew arising from different mutation frequencies between the leading and lagging strands, a
121 phenomenon observed in bacteria and nuclear DNA (nDNA) replication [25,26], but a mutational
122 gradient stemming from deamination events should be weak or absent due to negligible amounts of
123 ssDNA. Using modern sequencing technologies, the strand asymmetry in transitions has been
124 described in somatic mtDNA mutations [7–10]. More recently, high accuracy sequencing of murine

125 oocytes shows a similar bias towards the strand-asymmetric accumulation of transitions, establishing a
126 mechanistic link between the dominant mutagenic process in somatic tissues and what is seen in
127 population genetics [27]. However, to date, no mutational gradient has been reported outside the context
128 of phylogenetic analyses and it remains an open question if it is an active process or a byproduct of
129 selective pressure over time.

130 In this report, we have taken advantage of several large high accuracy mtDNA mutation data
131 sets previously generated with Duplex Sequencing (Duplex-Seq) to examine the distribution of somatic
132 mutations in the mtDNA of mice and humans (Sanchez-Contreras & Sweetwyne et al., *in preparation*
133 and [10,28,29]). We find that G→A and T→C transitions, but not their complementary mutations, exhibit
134 a strand-asymmetric gradient delimited by the Ori_L and the CR. This gradient is evolutionarily conserved
135 between mouse and humans. The CR also exhibits a remarkably different mutational pattern compared
136 to the coding portion of the genome and is consistent with the presence of a stable D-loop structure
137 bounded by highly conserved regulatory sequence blocks (Fig. 1b). Comparison of the somatic muta-
138 tional gradient to the distribution of SNPs in the human population, as well as the distribution of bases
139 along the genome across species, shows remarkable concordance. Taken together, our findings demon-
140 strate that an active mutational gradient drives the unequal accumulation of mutations in mtDNA and is
141 most consistent with a strand-asymmetric replication model with an extensive ssDNA replication inter-
142 mediate. Moreover, this unusual mutagenic process influences population level haplotypes and likely
143 drives genome composition over evolutionary time scales.

144

145 **Results**

146 As part of a comprehensive analysis on the effects of aging and mitochondrial-targeted
147 interventions on somatic mtDNA mutation accumulation in eight different mouse tissues, we used
148 Duplex-Seq to collect 34,113 independent, high accuracy, somatic mutations spread across the entirety
149 of the mtDNA molecule (Additional File 1: Supplemental Data 1). In the course of initially analyzing our
150 data, we noted significant variability in the per gene mutation frequency (when looking at individual

151 mutation types). Ordering the genes by their location in the genome, instead of grouping by complex,
152 showed an increasing frequency in G→A mutations, reminiscent of what has been observed in
153 phylogenetic studies (Additional File 2: Supplemental Fig. 1) [23,30]. Intrigued by this observation, we
154 took advantage of the large number of mutations to obtain a higher resolution understanding of how
155 mutagenesis varies across the mtDNA. However, variants are spread out across 85 individual samples
156 with a mean of 401 (range: 48-1496) mutations per sample, corresponding to a mean density of 0.025
157 mutations/bp. Because mutations are spread across 12 different mutation classes, the mutation density
158 of individual samples would not provide a higher resolution than at a per gene level. To overcome this
159 issue, we combined the data from all tissues to produce the most robust data set possible. Specifically,
160 we divided the genome into 100bp bins (total of 163) and, for each mutation class (*i.e.* G→A, G→C,
161 G→T, *etc*), summed the mutation counts observed in each bin across our all samples, separated by age
162 cohort (young (n=40): 4.5 months; old (n=45): 26 months). We then normalized for both genome base
163 composition and variability in sequencing depth of each bin by dividing the mutation count by the total
164 number of wild-type mutable bases sequenced across the constituent samples (Additional File 1:
165 Supplementary Data 2 & 3). This effectively gives a weighted mean of the mutation frequency for each
166 bin for all samples.

167 By plotting the mutation frequency by genome position (*i.e.* bin) in our 26 month old cohort
168 (25,020 mutations), an apparent discontinuous gradient bounded by the Ori_L and CR is observed for
169 G→A and T→C transitions, but not their respective complementary mutation types (Fig. 2a,b). An
170 exception is the CR (bins 154-163; genome positions 15400-16,299) which exhibits a notable spike in
171 C→A, but a decline in G→A mutations, whereas both T→C and A→G mutations show increases in the
172 CR, consistent with previous reports [7,27]. Performing separate regressions of the minor and major
173 arcs (bins 1-48 and 53-154 or genome positions 1-4800 and 5,300-15,400, respectively) show highly
174 significant increases in mutation frequency across their respective genomic coordinates (minor arc:
175 G→A slope=$8.24\pm3.06\times10^{-8}$, p=0.007; T→C slope=$1.08\pm0.35\times10^{-8}$, p=0.002; major arc: G→A
176 slope=$7.42\pm0.75\times10^{-8}$, p=5.35×10^{-23}, T→C slope=$1.23\pm0.10\pm10^{-8}$, p=1.33×10^{-35}). With the exception of
177 G→C mutations in the major arc, no other mutation types exhibited a gradient (Additional File 2:

178 Supplementary Fig. 2 and Additional File 3: Supplementary Table 1). Notably, the G→C gradient is >10-
179 fold smaller than the transition-based gradients and its relevance to mitochondrial biology, if any, is
180 unclear. The strand bias (reference L-strand G→A and T→C mutations are equivalent to anti-reference
181 H-strand C→T and A→G mutations, respectively) is consistent with the previous reports in somatic
182 mtDNA mutations and the gradient is most consistent with the previously hypothesized strand-
183 asynchronous replication mechanism with a deamination prone single-stranded replication intermediate
184 involving only two origins of replication [7,11,12,27].

185 Our mouse mtDNA data set combined mutation profiles of 8 unique tissue types from 6 organ
186 systems, therefore we sought to validate our analysis by accounting for the tissue-specific effects and
187 local differences in sequence contexts identified in these data (Sanchez-Contreras & Sweetwyne et al.,
188 *in preparation*). To address the possibility that one tissue type in our data was driving the observed
189 gradient, we performed a leave-one-out approach by eliminating one tissue and then performed the
190 same analysis on the reduced data set, repeating this analysis for each tissue type. As expected, the
191 removal of data of any one tissue type did not alter our findings (Additional File 3: Supplemental Table
192 2). These results point to the gradient not being an artifact of any single tissue type in our data. We next
193 addressed the potential impact of different local sequence contexts within each bin by performing Monte-
194 Carlo simulations that randomly redistributed each mutation observed in the 153 bins corresponding to
195 the non-CR portion of the genome (genome positions 1-15,400) using a weighted probability for each
196 bin based on its base composition. After redistribution, the mutation frequency was then recalculated for
197 each bin and the procedure repeated 10,000 times. As expected, we observe no gradient in either the
198 major or minor arcs (Fig. 2c,d; Additional File 2: Supplemental Fig. 3; Additional File 3: Supplemental
199 Table 3). Our analysis confirms that the strong positive mutational gradients in G→A and T→C
200 transitions in both the major and minor arcs of the mouse mtDNA are not artifacts and is most consistent
201 with a strand asynchronous replication mechanism and inconsistent with a conventional leading/lagging
202 strand mechanism.

203

Effects of age and Pol-γ fidelity on the mutational gradient support an asynchronous replication model.

A key question is the identity of the biological process giving rise to the observed gradient. As noted previously, the classic asynchronous replication model hypothesizes a long-lived ssDNA intermediate (Fig. 1c). The consequence of this model is that the portions of the mtDNA closest to their initiating origin should disproportionately accumulate G→A and T→C L-strand mutations during the aging process due to more time in the single-stranded state and should manifest as an increase in the gradient slope over time. To test this hypothesis, we made use of the young (4-5 mo; n=9,093 mutations) and old age (26 mo; n=25,020 mutations) cohorts in our data set to evaluate the interaction between aging and genome position on the gradient slope. Both major arc G→A and T→C L-strand gradients, as well as T→C mutations in the minor arc, exhibit a significant increase in their respective slopes during aging (Major arc: G→A interaction=$4.21 \pm 0.80 \times 10^{-8}$, $p=1.54 \times 10^{-7}$; T→C interaction=$1.03 \pm 0.11 \times 10^{-8}$; $p=1.31 \times 10^{-21}$; Minor arc: T→C interaction=$8.38 \pm 3.89 \times 10^{-9}$, p=0.031) (Fig. 3a,b; Additional File 3: Supplemental Table 4; Additional File 1: Supplemental Data 2 & 3). These findings, again, point to the asynchronous replication model as being most consistent with a deamination prone replication intermediate that experiences increased time in the single-stranded state. Furthermore, they demonstrate that this mutational gradient process is the primary driver of age-associated somatic mutations in mtDNA.

While the non-uniform increase in mutations with age is most consistent with deamination, it is possible that some other aspect of mtDNA replication could lead to this pattern. For example, Pol-γ is thought to exist both with and without its p55 accessory subunit, which has been reported to affect fidelity [31,32]. To test the effects of Pol-γ fidelity on the mutation gradient, we reanalyzed the distribution of 30,264 independent mutations obtained from a previous study using Duplex-Seq on mtDNA from mice homozygous for exonuclease deficient Pol-γ (Pol-γ^{exo-}) [29]. The loss of exonuclease activity in these mice results in a ~100-fold increase in mtDNA mutations [29,33]. If the mutational gradient is a fundamental aspect of Pol-γ base selectivity (regardless of the specific cause), we would expect the gradient to still be present or exacerbated in the absence of exonuclease activity. In contrast, if the

231 gradient is due to a non-polymerase source, such as DNA damage, then the frequent misincorporation
232 events of the Pol-γ^{exo-} enzyme should result in a more uniform distribution of mutations across the
233 mtDNA with little to no gradient present.

234 In contrast to our results in wild-type mice, the strong positive gradient in G→A and T→C
235 transitions is no longer present (Fig. 3c,d; Additional File 3: Supplemental Table 5; Additional File 1:
236 Supplemental Data 4). Instead, we note slight, but statistically significant, negative slopes in G→A and
237 T→C transitions, as well as T→A, C→G, and G→T transversions in the major arc and a slight positive
238 slope in minor arc A→T, but the relative effect size is substantially smaller than what is seen in G→A
239 and T→C mutations in wild-type mice and its relevance in mtDNA biology, if any, is unclear (Additional
240 File 2: Supplemental Fig. 4; Additional File 3: Supplemental Table 1&5). We did not evaluate the
241 distribution of mutations in the CR due to the likely presence of concatemers in the Pol-γ^{exo-} mouse CR
242 [34], the effects of which can be seen by the significantly lower mutation frequencies in bins containing
243 this region (Fig. 3c,d; Additional File 2: Supplemental Fig. 4). Taken together, our analysis showing that
244 the gradient unequally changes across the genome with age and the loss of the strong positive gradient
245 from replication by error prone Pol-γ^{exo-} points to a mechanism that is extrinsic to the polymerase itself
246 and is, again, most consistent with a DNA replication intermediate with a long-lived single-stranded
247 state.

248

249 ***Conserved regulatory elements exhibit mutagenic 'hot-spots' and 'cold-spots'***

250 The CR contains several important regulatory elements, including both transcriptional promoters,
251 the Ori_H, several highly conserved sequence blocks (CSB), and extended termination-associated
252 sequences (ETAS), whose specific regulatory functions are incompletely understood (Reviewed in [35])
253 (Fig. 1b). We and others have noted a distinctly different mutation frequency and spectrum in the CR
254 compared to the coding portion of the genome in both humans and mice [7,27,36], suggesting that the
255 unique function and structure may strongly influence CR mutagenesis, but high-resolution mapping of
256 mutations has not been reported.

257 The CR lies at the extreme 3' terminal end of the *M. musculus* mtDNA reference genome, which
258 presents issues during data alignment that gives rise to significant biases in sequence depth and
259 mutation calls. To address this potential bias, we modified the mtDNA reference to place the CR in the
260 middle of the sequence and realigned our data to this modified reference. In addition, we decreased our
261 bin size to 50bp to allow for a higher resolution mapping of mutations. The CR exhibits prominent spikes
262 and troughs that closely correspond to the ETAS, 7s DNA D-loop, CSBs, and the transcriptional
263 promoters (Fig. 4; Additional File 1: Supplemental Data 5) [37]. To confirm our findings, and to determine
264 if any of these conserved sequences comprise a mutational 'hot-spot' or 'cold-spot', we performed
265 Monte-Carlo simulations using the same strategy as described for our mutational gradient analysis, but
266 with 50bp bins, repeating the sampling 100,000 times, and setting the two-tailed Bonferroni corrected
267 significance to $p<0.0025$ (Fig. 4; Additional File 2: Supplemental Fig. 5, *black line & grey shading*). The
268 simulations confirm that C→T, T→C, G→A, and A→G, but not other mutation classes, show significant
269 deviations from random sampling in these conserved structures. Of particular interest is a consistent
270 mutational 'hotspot' for C→T, T→C, and A→G mutations, but a 'cold-spot' for G→A in the ETAS (Fig. 4,
271 *red blocks*). This observation suggests the presence of a structure that is highly prone to certain mutation
272 types and resistant to others or, alternatively, the loss of L-strand dG's prevents the maintenance of
273 mtDNA. Consistent with the possibility that mutations can be selected against, all four transition types
274 show a significant depletion of variants in the region between CSB3 and mt-tRNA^{Phe} that corresponds
275 to the transcription promoters and mitochondrial transcription factor A (TFAM) binding sites, which are
276 thought to be the source of the Ori_H replication primer (Fig. 4, genome position 16,100-16,299)[35].
277 Interestingly, no high level heteroplasmic or homoplasmic variants have been detected in the same
278 region in human population studies, suggesting that this region is extremely important for mtDNA
279 maintenance [36]. Lastly, all four transitions exhibit a significant spike in the region between CSB1 and
280 genome position ~15,900 consistent with this region harboring the 7s DNA/RNA D-loop. Taken together,
281 our high resolution analysis of CR mutations highlights the presence of both mutagenic hot-spots and
282 cold-spots that correspond to highly conserved regulatory elements responsible for the distinctive
283 mutational bias previously noted in the CR. Additionally, our data suggest the presence of unique DNA

284 structures within these sequence blocks that differently affect DNA damage and/or replication fidelity
285 and also suggest that some regions important for mtDNA replication may poorly tolerate mutagenesis.
286
287 ***A mutational gradient is conserved in human mtDNA.***
288 We next determined the evolutionary conservation of the patterns we observe in our mouse data.
289 To do so, we made use of prior reported Duplex-Seq data sets for human mtDNA [10,28]. As with the
290 mouse data, we performed a binned mutation frequency analysis with bin size of 200bp due to the
291 reduced number of mutations compared to our mouse data. Consistent with our mouse data, we observe
292 a gradient for both G→A and T→C mutations in the major arc (G→A: slope=4.86±0.93x10^{-7}, p=1.93x10^{-}
293 ^{7}; T→C: slope=1.84±0.26x10^{-7}, p=1.69x10^{-12}) that is bounded by the Ori_L and CR (Figure 4a,b; Additional
294 File 3: Supplemental Table 6; Additional File 1: Supplemental Data 6). Unlike the mouse data, the minor
295 arc did not exhibit an apparent gradient and no other mutation types exhibited a significant increase in
296 either the major or minor arcs (Additional File 2: Supplemental Fig. 6; Additional File 3: Supplemental
297 Table 6).
298 Our analysis points to a somatic mutational gradient as an evolutionarily conserved feature of
299 vertebrate mtDNA. However, all our data were collected using Duplex-Seq, leaving open the possibility
300 that the gradient pattern is an artifact of our Duplex-Seq protocol or our data analysis pipeline. While we
301 consider this scenario unlikely, we sought to observe this gradient in an independently generated data
302 using more conventional sequencing approaches. Somatic mtDNA mutations occur at very low
303 frequencies (~10^{-6}-10^{-5}), making their detection with conventional sequencing difficult [38]. To overcome
304 this limitation, we analyzed mtDNA mutation call data published by the Pan-Cancer Analysis of Whole
305 Genomes (PCAWG) Consortium [39]. This data set consists of 7,611 independent somatic variants
306 (variant allele fraction (VAF)>0.01; mean VAF=0.2) from 2,536 tumors across 38 different cancer types.
307 Because cancer is a clonal process arising from a single cell, the detected variants are largely a
308 snapshot of the mtDNA mutations present early in tumor formation and have much higher VAFs than
309 what is typically detected in Duplex-Seq data. Importantly for our purpose, this characteristic of the tumor

310 data is expected to largely eliminate the potential confounder of low frequency artifacts giving rise the
311 observed gradient.

312 We divided the genome into 100bp bins and, for each mutation type, calculated the mutation
313 density (i.e. mean number of detected mutations per wild-type base) in each bin (Additional File 1:
314 Supplemental Data 7). Consistent with our Duplex-Seq data, we observe a clear gradient in both G→A
315 and T→C transitions, but not their complement, that increases along the major arc (G→A $p=7.43x10^{-8}$;
316 T→C $p=1.16x10^{-5}$)(Fig. 5c,d; Additional File 3: Supplemental Table 7). Both T→A and C→G
317 transversions report a negative slope in the major arc and C→T and G→T exhibit a positive slope in the
318 minor arc, but the magnitudes are extremely small and are likely a regression artifact. No other mutation
319 types exhibit a gradient (Additional File 2: Supplemental Fig. 7; Additional File 3: Supplemental Table
320 7). These data confirm both the presence of a mutational gradient in the major arc and that our results
321 are unlikely to be due to an unknown issue with Duplex-Seq. Taken together, both our Duplex-Seq data
322 and the PCAWG data recapitulate our findings in mouse mtDNA, pointing to the strong evolutionary
323 conservation of G→A and T→C gradients among vertebrate species.

324

325 ***A mutational gradient is mirrored in germline SNPs and genome base composition.***

326 Previous work has noted similarities in the strand orientation and simple mutational spectra
327 between somatic mtDNA mutations and population level SNPs, suggesting a similar causative driver of
328 population level mtDNA sequence diversity [7,9,27]. We sought to further explore this relationship by
329 determining if the mutation gradient is reflected in the distribution of inherited single nucleotide variants,
330 as would be expected if this process is active in the germline. We initially sought to test this hypothesis
331 by mapping mutations obtained with Duplex-Seq of mouse oocytes [27], but the total number of
332 mutations (N=691) was insufficient to detect a gradient. We next evaluated the distribution of
333 homoplasmic SNPs in the human mtDNA by downloading a recently published list of 44,494 SNPs
334 obtained from MITOMAP and phylogenetically corrected such that each SNP was likely the result of an
335 independent *de novo* event [40,41]. Using the same binning approach as our human somatic data, we

336 calculated the mutation density (*i.e.* number of *de novo* SNPs per mutable base) in each bin. For this
337 analysis, we limited our analysis to the major arc due to 1) the absence of a clear minor arc gradient in
338 our human somatic data and 2) evidence of regions with an underrepresentation of SNPs in rDNA genes
339 [41]. Consistent with our somatic data, we observe a significant positive gradient in G→A and T→C (Fig.
340 6a,b; Additional File 2: Supplemental Fig. 8; Additional File 3: Supplemental Table 8). Notably,
341 complement SNP types (C→T and A→G, respectively), as well as G→C SNP, show significant
342 gradients, but the magnitude of their slope, especially relative to G→A SNPs is substantially smaller.
343 We sought to further validate this observation by performing this same analysis on a recently reported
344 database of homoplasmic SNPs from 196,983 individuals [41]. As with our initial data set, we observe
345 a significant correlation between SNP density and genome position of G→A SNPs (ρ=0.26; p=0.046;
346 Spearman correlation). We also observe a significant correlation between G→C SNP and genome
347 position (ρ=0.307; p=0.019; Spearman correlation) similar to the MITOMAP based dataset. No other
348 significant correlations were observed (Additional File 3: Supplemental Table 9). Thus, we are able to
349 confirm that, at the very least, a G→A gradient is present in human polymorphisms, consistent with our
350 somatic mtDNA data, and further supports the idea that the mechanism of mutagenesis in the somatic
351 tissue is likely the direct driver of human mtDNA variation.

352 The strong conservation of the somatic gradient between mice and humans and the presence of
353 the gradient in human SNP data suggest that this unusual mutational pressure is likely a major driver of
354 sequence diversity across species. Our somatic data point to a sustained G→A and T→C mutational
355 pressure of the L-strand with relatively little reversion. Over the long term, the L-strand is expected to
356 exhibit a spatially dependent depletion of dG and dT bases along the major arc and a concomitant
357 increase in dA and dC bases until some selective equilibrium is reached (Fig. 7a). Phylogenetic analyses
358 on the sequence differences between related species such as primates has been shown to exhibit a
359 gradient effect in T→C transitions [30]. Analysis of a relatively small number of vertebrate species
360 (N=118) has also suggested that this phenomenon is likely a general aspect of vertebrate mtDNA biology
361 [23].

362 We sought determine the generality of the gradient phenomenon by expanding these findings to
363 include the significantly increased number of vertebrate mtDNA sequences now available (N=3,614).
364 Performing this analysis on all available mammalian mtDNA sequences in the NCBI RefSeq database
365 (N=717) shows that the majority of sequences exhibit a significant spatially dependent depletion of dG
366 and dC (*i.e.* negative correlation coefficient) and a similar enrichment (*i.e.* positive correlation coefficient)
367 in dC and dT(Fig. 7b,d), confirming that this is a general phenomenon in mammalian mtDNA. While
368 consistent with our hypothesis, the correlation 1) does not inform on the magnitude of the correlation
369 and 2) does not explicitly link the change in the abundance of one base type with another. Specifically,
370 the magnitudes of the dG and dT composition slopes should be anti-correlated with the respective dA
371 and dC slope magnitudes within the same species. As can be seen in Figure 7c & 7e, with a few
372 exceptions, the slopes of dG and dA content, as well as dT and dC content, are strongly anti-correlated
373 (dG/dA Spearman's $\rho=0.43$, $p=3.8\times10^{-33}$; dT/dC Spearman's $\rho=0.51$, $p=1.4\times10^{-49}$) across currently
374 available mammalian mtDNA sequences with the direction of the anti-correlation consistent with a
375 graduated G→A and T→C mutation pressure. We next extended this approach to other vertebrate
376 classes, including birds (N=656), reptiles (N=212), and fish (N=2029). We did not evaluate non-
377 vertebrate mtDNA sequences due to higher levels of structural heterogeneity and gene composition in
378 these phyla. Like mammals, the majority of species within each vertebrate class show significant
379 gradients in mtDNA composition that are strongly anti-correlated in their dG/dA content, as well as dT/dC
380 content, indicating that this graduated mutation pressure is highly conserved across widely divergent
381 species that inhabit significantly different ecological niches and are subjected to very different selective
382 pressures (Additional File 2: Supplemental Fig. 9 & 10). Interestingly, several species strongly deviate
383 in either gradient direction and/or correlation strength, suggesting that these species are subject to
384 different selective pressures on their mtDNA (Additional File 2: Supplemental Fig. 9 & 10). Taken
385 together, our data point to the mutational process driving the accumulation mutations in somatic tissues
386 being the likely mechanistic driver of population level polymorphisms and sequence composition in
387 vertebrates.
388

389 **Discussion**

390 The advent of ultra-high accuracy sequencing methodologies have opened up the possibility of
391 studying the mutagenic processes in mtDNA in greater detail. Both we and others have used Duplex-
392 Seq, a method with an error background of $<1x10^{-7}$, to study somatic mtDNA mutations [7,10,27–
393 29,43,44]. These studies have broadly shown that mutations are heavily weighted towards G→A/C→T
394 and T→C/A→G transitions with very low levels of transversions, including the canonical ROS-
395 associated G→T/C→A mutations. In addition, these studies have shown a strong strand bias, with
396 G→A/C→T mutation being more prevalent when the dG base is in the L-strand. A notable difference in
397 the mutational frequency and spectrum in the CR is also reported. While these studies have provided a
398 broad understanding of mtDNA mutagenesis, the very low frequency of mutations ($<1x10^{-5}$) means that,
399 for any given sample, only a few dozen to a few hundred mutations are typically detected, leaving
400 conclusions about how these mutations are distributed unclear beyond broad regional differences (*i.e.*
401 CR vs coding or between genes). In this study, we aggregated several pre-existing Duplex-Seq data
402 sets to better asses the distribution of mutations across the mtDNA molecule at significantly higher
403 resolution than what has been previously reported.

404 In addition to recapitulating previous findings showing a strong bias towards transition mutations
405 over transversions and a higher mutation load in the CR, our analysis shows a strikingly non-uniform
406 gradational distribution of G→A and T→C transitions, but not their complement, along the coding portion
407 of the mtDNA. The totality of our data is most consistent with an asynchronous strand displacement
408 mechanism with a long lived, deamination prone, single-stranded H-strand. A key aspect of our data
409 that supports this hypothesis is the increased slope of G→A and T→C mutations with advancing age.
410 Any alternative replication model without a ssDNA intermediate would need to account for how
411 deamination-linked mutations could disproportionately increase as a function of genome position over
412 time beyond what is present in the gradient. The RITOL and strand-synchronous models lacks any
413 substantial ssDNA, with >80% of the displaced H-strand estimated to be annealed with RNA in the
414 RITOL model (Fig. 1d,e)[13,14]. Our Pol-γ^{exo-} data suggest that the gradient is not due to a simple
415 interaction between mtDNA base composition and polymerase base selectivity.

416 Holt and colleagues have reported that the synchronous and asynchronous mechanisms can
417 exist simultaneously, with the balance between these two mechanisms the result of the cell's
418 physiological state [15]. While our data do not support a classic leading/lagging strand mechanism, they
419 do not entirely refute its existence in all cases. Leading/lagging strand synthesis may be part of a stress
420 response pathway to quickly reestablish copy number levels. In support of this possibility, withdrawing
421 mtDNA depleting ethidium bromide from cells results in a burst of mtDNA synthesis with fully double-
422 stranded replication intermediates, which is interpreted as being due to a leading/lagging strand
423 replication fork [1]. Consistent with this idea, modulating the level of the mitochondrial transcripts via
424 changes in Twinkle helicase levels has been reported to switch between strand asynchronous and
425 lead/lagging strand synthesis [18]. However, our data are from tissues of unstressed wild-type animals
426 without known perturbations to mtDNA gene expression, pointing to the asymmetric model being the
427 predominant mtDNA synthesis mechanism under normal physiological conditions.

428 A lingering question in the field of mtDNA replication concerns the conservation of the mtDNA
429 replication mechanism across taxa. The mitochondrial genome exhibits a wide range of sizes,
430 structures, and noncoding regulatory regions between phyla and kingdoms, suggesting that different
431 replication mechanisms were retained or acquired since the initial endosymbiosis event that gave rise
432 to mitochondria. For example, while vertebrates make use of a relatively compact CR with an initiating
433 origin and distal counter-directional origin, invertebrates tend to make use of a large highly AT-rich region
434 with only one confirmed origin and one likely late-firing proximal counter-directional origin [45]. Plant
435 mtDNA likely uses an entirely different recombination-dependent and/or rolling circle mechanism without
436 clearly defined replication origins [46]. Our data indicate that mapping of somatic mutations provides an
437 alternative approach to mapping origins of replication and other potential regulatory structures that is
438 free of the complications inherent to interpreting 2D-gels and electron micrographs. Indeed, an
439 analogous strategy has been used to map origins of replication in the human genome by taking
440 advantage of ultra-mutated tumors [47].

441 We can clearly discern the location of the Ori_L in both mouse and human data sets. These data
442 also argue against the proposed use of other tRNAs as L-strand priming sites, as well as a large

443 'initiation zone' for replication, as these models predict either multiple discontinuous gradients or lack a
444 gradient entirely. Instead, our data are consistent with a single L-strand origin in mammalian mtDNA.
445 Moreover, with our high-density mouse data set, we mapped areas of mutation over- and under-
446 abundance in the CR that correspond to sequence blocks essential for mtDNA H-strand replication.
447 Significant deviations are not obvious in the regions flanking the CR other than the Ori_L, as would be
448 expected if other sequences in these areas were essential for intermittent priming. Notably, avian mtDNA
449 lacks the predicted stem-loop structure of the mammalian Ori_L and 2D-gels point to initiation sites across
450 the entirety of the mtDNA, providing a potential model system to further investigate these alternative
451 origin models in vertebrates [17]. In line with this idea, we attempted to analyze Duplex-Seq data for
452 similar patterns in non-vertebrate organisms, *D. melanogaster* [43,48] and *A. thaliana* [44], but the
453 number of mutations were too low and the density too sparse to observe a clear signal, leaving this for
454 future work.

455 Human population studies have previously identified a bias in the occurrence of G→A and T→C
456 SNPs of the L-strand, as has comparison of human mtDNA sequences with those of evolutionary related
457 species[49,50]. We previously noted that this bias mirrors the strand asymmetry seen in somatic
458 mutagenesis of mtDNA through a process that is continuous throughout life and hypothesized that this
459 pattern was consistent with mtDNA replication via an asymmetric model [7]. Our data extend this
460 observation to also include a gradient in SNP distribution along the genome, as well as genome base
461 composition, further strengthening the link between somatic and germline processes. Our analysis in
462 genome composition point to this gradient being largely, but not universally, conserved in vertebrates.
463 The strength of this gradient, as highlighted by the variation in anticorrelation between complementary
464 bases and the presence of species that deviate significantly from the trend line, can vary significantly.
465 An important aspect of these observations is that they provide a feasible opportunity to mechanistically
466 study the processes that give rise to genetic variation at the population and taxonomic levels. This is
467 especially pertinent in species with very high or very low mutation rates or show unusual biases in
468 genetic variability. Indeed, recent work in *A. thaliana* linked mismatch repair with the low mutation rate
469 seen in this species, and three species of angiosperm genus *Silene* with notably different mutation rates

470 showed corresponding patterns in somatic mtDNA mutations [44,51], lending credence to the idea that
471 studying somatic mutagenic processes can inform on evolutionary and population level patterns. A
472 systematic analysis of somatic mutations in species with unusual genome composition or SNP patterns
473 may provide insight into how and why these species deviate so significantly from related species and
474 clues to their natural history.
475
476 **Conclusions**
477 The growing quantities of high-accuracy sequencing data generated from such technologies as
478 Duplex Sequencing has provided the ability to elucidate several mutagenic patterns and biases
479 previously unobserved at the somatic level. Taken together, these patterns argue for an asymmetric
480 strand-displacement replication model, as originally posited by Clayton and colleagues [11,12] and
481 against a more conventional leading/lagging strand replication fork. The mutations that arise as a
482 consequence of genome replication are likely from deamination events of the single-stranded
483 intermediate which affect genetic variation and, ultimately, genome composition.
484
485 **Methods and Materials**
486 Duplex Sequencing Data and Processing
487 The DNA libraries and sequencing of the Duplex-Seq libraries was performed as indicated in the
488 original publications. Specifically, the human data was obtained from normal tissue data from Baker et
489 al. (SRA accession PRJNA449763) [10] and Hoesktra et al. (SRA accension PRJNA237667) [28]. Five
490 young and six old wild-type mouse data was generated from male C57Bl/6J at 4-5 and 26 months of
491 age, respectively. Aged mice were obtained from the NIA aged rodent colony at an age of 22-23 months
492 and then housed at the University of Washington animal facility until the desired age under approved
493 conditions. Animals were euthanized at the indicated age and a ~2 mm section of the heart (apex), liver
494 (lobe VI posterior), kidney (outer cortex), skeletal muscle (proximal gastrocnemius), brain (both
495 cerebellum and hippocampus), and eye (retina and eye cup) were flash frozen and stored at -80°C until

496 processed for sequencing. A ~1mm tissue punch was used to obtain a representative tissue sample
497 from brain regions. Half of the retina was used for processing and eye cups were used in their entirety.
498 Total DNA was purified from each tissue punch/sample using a QIAamp Micro DNA kit using the
499 manufacturer's protocol. Total DNA was prepared for Duplex Sequencing using our previously published
500 protocol with modifications as described in Hoekstra et al. [28,52]. Pol-γ^{exo-} mouse data was generated
501 from Pickrell et al. (SRA accension PRJNA729056)[29].

502 After obtaining the raw data, we processed all Duplex-Seq data using v1.1.4 of our in house
503 developed Snakemake-based Duplex-Seq-Pipeline to ensure that all data were uniformly processed
504 with the exception that different data sets had different unique molecular identifier (UMI) and read
505 lengths[53]. Briefly, we perform a reference free consensus approach for error correction similar to that
506 reported in Stoler et al. [54]. The UMI and any associated spacer sequence are parsed from the read,
507 the read 1 and read 2 UMI from a read pair is sorted alphabetically and associated with the read's SAM
508 record when converted to an unaligned BAM file (See Stoler et al. [54] for details). After the UMIs from
509 read 1 and read 2 are sorted alphabetically such that all reads derived from the same strand of the same
510 parental molecule are grouped together. UMI families from opposite strands of the same parental DNA
511 fragment are grouped sequentially in the sorted file. A per-position consensus is generated for each
512 single-strand is generated with a cut-off of 70% identity and a minimum of 3 reads with the same UMI
513 being required to call a consensus, as previously described [52]. The double-strand consensus is then
514 generated from the two single-strand consensuses sharing the same UMI, if present, with the exception
515 that the identity of the base must match between the two single-strand consensus. Reads with >2%
516 ambiguous bases are removed from further analysis. The resulting post-processed fastq files are
517 aligned against the reference genome (hg38 chrM for the human data and mm10 chrM for mouse the
518 data) using bwa v0.7.17 [56]. The overlapping portions of reads and 10-cycles of the 5' and 3' ends of
519 the reads are clipped using fgbio (https://github.com/fulcrumgenomics/fgbio) and adapter sequences
520 are removed using Cutadapt [57]. Insertion/deletion (in/del) realignment and in/del left alignment were
521 performed by the Genome Analysis Toolkit (v3.8.x). A draft list of variants is generated using the pileup
522 functionality of samtools [58]. Reads containing non-SNP variants (defined as a variant allele

523 fraction >40%) are parsed out and subjected to BLAST-based alignment against a database containing
524 potential common contaminants (dog: canFam3; bovine: bosTau9; nematode: ce11; mouse: mm10;
525 human: hg38; rat: Rnor_6.0). The inclusion of our target genome in this database also allows for the
526 identification of pseudogenes. Reads that unambiguously map to the same coordinates as the original
527 alignment are kept and the remaining reads and associated variants are removed from further analysis.
528 An exception to this process was made for mouse mtDNA due to the presence of a ~5000bp nuclear
529 pseudogene with perfect identity to mm10 chrM [59]. In this case, any ambiguous BLAST alignments
530 mapping to this region were assumed to be mitochondrial in origin and kept. The mutated reads passing
531 our BLAST filter are merged back with the non-mutated reads and mutation frequencies calculated
532 based on the provided target coordinates.

533 To generate the bin data for each age cohort (*i.e.* 4.5mo and 24-26mo), we divided the genome
534 up into the indicated bin size and then calculated the mutation frequency for each bin, i, and mutation
535 type, N (*i.e.* G→A, T→C, *etc*), by $F_i^N = \frac{\sum_j M_{ij}}{\sum_j S_{ij}}$, where M_{ij} is the mutation count of type N in bin i of
536 sample j and S_{ij} is the number of sequenced bases in bin i of sample j that are mutable for mutation
537 type N. A mutation was only counted if its variant allele fraction (VAF) was <1% to minimize the effects
538 of inherited or early arising mutations.

539 For the D-loop focused analysis, we generated a new partial mm10 chrM reference consisting of
540 the CR portion with the flanking 1000 bases (chrM 14,400-16299::chrM 1-1001) in order to allow for
541 better alignment of reads across the entirety of the mtDNA. The binning process was performed as
542 described, but with 50bp bins. Per bin mutation frequencies were calculated as described above.

543

544 <u>Tumor sequencing data</u>

545 Tumor single nucleotide variant data was generated using the methods outlined previously and
546 obtained from The Cancer Mitochondria Atlas data portal
547 (https://ibl.mdanderson.org/tcma/download.html) [39]. Similar to our Duplex-Seq analysis, the human

548 mtDNA was divided into 100bp bins and the number of variants of each of the 12 mutation classes was
549 divided by the number of wild-type base of the respective mutation class within each bin (i.e.
550 #G→A/#G's, etc).
551
552 SNP and Genome Composition data
553 The SNP data sets were obtained from Gu *et al* and Bolze *et al* using their respective procedures
554 [41,42]. Briefly, the Gu *et al.* data set is comprised of 44,334 SNPs reported in the MITOMAP database
555 that was filtered for haplotype private variants. The Bolze *et al* data comprises 14,283 homoplasmic
556 variants from 196,324 unrelated individuals. Because this dataset is not limited to haplotype private
557 SNPs, as the Gu *et al.* data, we limited our analysis to rare SNPs (population frequency ≤1:1,000) in
558 order to minimize population structure of the data from confounding our analysis. SNPs occurring more
559 than once were assumed to have arisen from a single *de novo* event. We divided to human mtDNA into
560 200bp bins and then calculated SNP density by summing the number of variants of each of the 12
561 mutation classes and then dividing by the number of respective wild-type bases within each bin (i.e.
562 #G→A/#G's, etc).
563 For the genome composition analysis, a complete set of curated mtDNA sequences and
564 annotations were downloaded from the NCBI Reference Sequence project
565 (https://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/) in GenBank format. Entries were parsed by
566 taxonomic Class and filtered to separately keep mammals, birds, reptiles, and fish. mtDNA sequence
567 corresponding to the major arc of each species was extracted and divided into 100 bp bins and the
568 nucleotide composition calculated as a percent of each base type. The slope and/or correlation
569 coefficient for the change in genome composition as a function of bin number (*i.e.* genome position) for
570 each individual species was then calculated as described below.
571
572 Statistical Analysis

Statistical analysis was performed in python using either Statsmodels (https://github.com/statsmodels/statsmodels) or SciPy [60], where indicated. Linear regression analysis was performed with the Python Statsmodels library using a robust linear model with Huber's T function as the M-estimator for down-weighting outliers. A robust linear regression model was used due to the violation of the assumptions of normality or homoscedasticity in some data sets that is required for ordinary linear regression models. To establish the effect of aging on the gradient slope, a robust linear regression model with the addition of an interaction term between age and bin number ($Y = \alpha + \beta_{bin} * bin + \beta_{age} * age + \beta_{binxage}(bin * age)$), with age being the categorical classifier with value 0 (young) or 1 (old), was used.

Monte-Carlo modeling of random mutagenesis was performed by first dividing the indicated genome interval (*i.e.* coding region vs CR) into the indicated bin size. A weighted probability for each mutation type (*i.e.* G→A, T→C, *etc*) to occur in each bin was calculated by dividing the cumulative depth of the wild-type mutable base in a bin by the cumulative depth of the wild-type of the same mutable base across the indicated genome interval. For each mutation type, the total number were randomly distributed across the bins using the calculated weights and then a per bin mutation frequency was calculated by dividing the number of mutations distributed in a bin by the cumulative sequencing depth of the mutable base within the same bin. This procedure was repeated 100,000 times for the D-loop analysis and 10,000 times for the coding region gradient analysis. An experimental Bonferonni corrected confidence interval for determining mutational hot-spots and cold-spots was set to 99.75%. Datapoints outside this range were considered significantly different from random chance.

Author Contributions

Conceptualization: MS-C, MTS, and SRK; Software: BFK and SRK; Sample and Data collection: MS-C, MTS, KAT, MJH, EKS, JF, JAW, MDC; Formal analysis: SRK; Visualizations: MTS and SRK; Writing-initial draft: SRK; Writing-review and editing: MS-C, MTS, BFK, KAT, PSR, DJM, and SRK; Funding

acquisition: MTS, PSR and SRK; Supervision: MS-C, MTS, PSR, DJM and SRK; Resources: PSR, DJM, and SRK. All authors read and approved the final manuscript.

Funding

We would like to acknowledge the following agencies/awards for supporting this project: K01 AG062757-01 to MTS; P01 AG001751 to PSR; R21 HG011229 and W81XWH-18-1-0339 to SRK.

Availability of data and materials

The Duplex-Seq-Pipeline is written in Python and R, but has dependencies written in other languages. The Duplex-Seq-Pipeline software has been tested to run on Linux, Windows WSL1, Windows WSL2, and Apple OSX. The software can be obtained at https://github.com/Kennedy-Lab-UW/Duplex-Seq-Pipeline and https://doi.org/10.5281/zenodo.5084120 under the BSD license. The normal mouse data is available at SRA accension PRJNA727407. Only wild-type, non-intervention samples were used. The Pol-γ^{exo-} mouse data is available at SRA accension PRJNA729056). The human data is available at SRA accession PRJNA449763 and SRA accension PRJNA237667. Only normal control samples were analyzed.

Ethics approval

Wild-type mouse tissues were collected from mice at the University of Washington under IACUC approved protocols. Previously published human and mouse data sets were collected under the terms described in their respective publications.

Competing interests

SRK is an equity holder and paid consultant for Twinstrand Biosciences, a for-profit company commercializing Duplex Sequencing. No Twinstrand products were used in the generation of the data

for this paper. The views expressed in this publication are those of the author(s) and not necessarily those of the NIH, CDMRP, or DOD.

Reference

1. Yasukawa T, Yang M-Y, Jacobs HT, Holt IJ. A Bidirectional origin of replication maps to the major noncoding region of human mitochondrial DNA. Mol Cell. 2005;18:651–62.

2. Taylor RW, Turnbull DM. Mitochondrial DNA mutations in human disease. Nat Rev Genet. 2005;6:389–402.

3. Sevini F, Giuliani C, Vianello D, Giampieri E, Santoro A, Biondi F, et al. mtDNA mutations in human aging and longevity: Controversies and new perspectives opened by high-throughput technologies. Exp Gerontol. 2014;56:234–44.

4. Chocron ES, Munkácsy E, Pickering AM. Cause or casualty: The role of mitochondrial DNA in aging and age-associated disease. Biochim Biophys Acta Mol Basis Dis. 2019;1865:285–97.

5. Cheng KC, Kasais H, Nishimuras S, Loeb LA. 8-hydroxyguanine, an abundant form of oxidative DNA damage, causes G➞T and A➞C substitutions. J Biol Chem. 267:166–72.

6. Harman D. Free radical theory of aging. Mutat Res. 275:257–66.

7. Kennedy SR, Salk JJ, Schmitt MW, Loeb LA. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. Van Houten B, editor. PLoS Genet. 2013;9:e1003794.

8. Williams SL, Mash DC, Züchner S, Moraes CT. Somatic mtDNA mutation spectra in the aging human putamen. Van Houten B, editor. PLoS Genet. 2013;9:e1003990.

9. Ju YS, Alexandrov LB, Gerstung M, Martincorena I, Nik-Zainal S, Ramakrishna M, et al. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. Elife. 2014;3:e02935.

10. Baker KT, Nachmanson D, Kumar S, Emond MJ, Ussakli C, Brentnall TA, et al. Mitochondrial DNA mutations are associated with ulcerative colitis preneoplasia but tend to be negatively selected in cancer. Mol Cancer Res. 2019;17:488–98.

11. Brown TA. Replication of mitochondrial DNA occurs by strand displacement with alternative light-strand origins, not via a strand-coupled mechanism. Genes Dev. 2005;19:2466–76.

12. Clayton DA. Replication of animal mitochondrial DNA. Cell. 1982;28:693–705.

13. Reyes A, Kazak L, Wood SR, Yasukawa T, Jacobs HT, Holt IJ. Mitochondrial DNA replication proceeds via a ‘bootlace’ mechanism involving the incorporation of processed transcripts. Nucleic Acids Res. 2013;41:5837–50.

656 14. Yasukawa T, Reyes A, Cluett TJ, Yang M-Y, Bowmaker M, Jacobs HT, et al. Replication of
657 vertebrate mitochondrial DNA entails transient ribonucleotide incorporation throughout the lagging
658 strand. EMBO J. 2006;25:5358–71.

659 15. Holt IJ, Lorimer HE, Jacobs HT. Coupled leading- and lagging-strand synthesis of mammalian
660 mitochondrial DNA. Cell. 2000;100:515–24.

661 16. Bowmaker M, Yang MY, Yasukawa T, Reyes A, Jacobs HT, Huberman JA, et al. Mammalian
662 mitochondrial DNA replicates bidirectionally from an initiation zone. J Biol Chem. 2003;278:50961–9.

663 17. Reyes A, Yang MY, Bowmaker M, Holt IJ. Bidirectional replication initiates at sites throughout the
664 mitochondrial genome of birds. J Biol Chem. 2005;280:3242–50.

665 18. Cluett TJ, Akman G, Reyes A, Kazak L, Mitchell A, Wood SR, et al. Transcript availability dictates
666 the balance between strand-asynchronous and strand-coupled mitochondrial DNA replication. Nucleic
667 Acids Res. 2018;46:10771–81.

668 19. Seligmann H. Mitochondrial tRNAs as light strand replication origins: Similarity between
669 anticodon loops and the loop of the light strand replication origin predicts initiation of DNA replication.
670 Biosystems. 2010;99:85–93.

671 20. Seligmann H, Krishnan NM, Rao BJ. Possible multiple origins of replication in primate
672 mitochondria: Alternative role of tRNA sequences. J Theor Biol. 2006;241:321–32.

673 21. Asakawa S, Kumazawa Y, Araki T, Himeno H, Miura K, Watanabe K. Strand-specific nucleotide
674 composition bias in echinoderm and vertebrate mitochondrial genomes. J Mol Evol. 1991;32:511–20.

675 22. Reyes A, Gissi C, Pesole G, Saccone C. Asymmetrical directional mutation pressure in the
676 mitochondrial genome of mammals. Mol Biol Evol. 1998;15:957–66.

677 23. Faith JJ, Pollock DD. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate
678 mitochondrial genomes. Genetics. 2003;165:735–45.

679 24. Frederico LA, Kunkel TA, Shaw BR. A sensitive genetic assay for the detection of cytosine
680 deamination: determination of rate constants and the activation energy. Biochemistry. 1990;29:2532–7.

681 25. Lujan SA, Williams JS, Pursell ZF, Abdulovic-Cui AA, Clark AB, Nick McElhinny SA, et al.
682 Mismatch repair balances leading and lagging strand DNA replication fidelity. Pearson CE, editor.
683 PLoS Genet. 2012;8:e1003016.

684 26. Maslowska KH, Makiela-Dzbenska K, Mo J-Y, Fijalkowska IJ, Schaaper RM. High-accuracy
685 lagging-strand DNA replication mediated by DNA polymerase dissociation. Proc Natl Acad Sci USA.
686 2018;115:4212–7.

687 27. Arbeithuber B, Hester J, Cremona MA, Stoler N, Zaidi A, Higgins B, et al. Age-related
688 accumulation of de novo mitochondrial mutations in mammalian oocytes and somatic tissues. Hurst
689 LD, editor. PLoS Biol. 2020;18:e3000745.

690 28. Hoekstra JG, Hipp MJ, Montine TJ, Kennedy SR. Mitochondrial DNA mutations increase in early
691 stage Alzheimer disease and are inconsistent with oxidative damage: Mitochondrial Mutations in AD.
692 Ann Neurol. 2016;80:301–6.

693 29. Pickrell AM, Huang C-H, Kennedy SR, Ordureau A, Sideris DP, Hoekstra JG, et al. Endogenous
694 Parkin preserves dopaminergic substantia nigral neurons following mitochondrial DNA mutagenic
695 stress. Neuron. 2015;87:371–81.

696 30. Raina SZ, Faith JJ, Disotell TR, Seligmann H, Stewart C-B, Pollock DD. Evolution of base-
697 substitution gradients in primate mitochondrial genomes. Genome Res. 2005;15:665–73.

698 31. Lim SE, Longley MJ, Copeland WC. The mitochondrial p55 accessory subunit of human DNA
699 polymerase γ enhances DNA binding, promotes processive DNA synthesis, and vonfers N-
700 ethylmaleimide resistance. J Biol Chem. 1999;274:38197–203.

701 32. Longley MJ, Nguyen D, Kunkel TA, Copeland WC. The fidelity of human DNA polymerase γ with
702 and without exonucleolytic proofreading and the p55 accessory subunit. J Biol Chem.
703 2001;276:38555–62.

704 33. Vermulst M, Bielas JH, Kujoth GC, Ladiges WC, Rabinovitch PS, Prolla TA, et al. Mitochondrial
705 point mutations do not limit the natural lifespan of mice. Nat Genet. 2007;39:540–3.

706 34. Williams SL, Huang J, Edwards YJK, Ulloa RH, Dillon LM, Prolla TA, et al. The mtDNA mutation
707 spectrum of the progeroid *Polg* mutator mouse includes abundant control region multimers. Cell
708 Metabolism. 2010;12:675–82.

709 35. Gustafsson CM, Falkenberg M, Larsson N-G. Maintenance and expression of mammalian
710 mitochondrial DNA. Annu Rev Biochem. 2016;85:133–60.

711 36. Wei W, Tuna S, Keogh MJ, Smith KR, Aitman TJ, Beales PL, et al. Germline selection shapes
712 human mitochondrial DNA diversity. Science. 2019;364:eaau6520.

713 37. Sbisà E, Tanzariello F, Reyes A, Pesole G, Saccone C. Mammalian mitochondrial D-loop region
714 structural analysis: identification of new conserved sequences and their functional and evolutionary
715 implications. Gene. 1997;205:125–40.

716 38. Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for
717 detecting rare and subclonal mutations. Nat Rev Genet. 2018;19:269–85.

718 39. PCAWG Consortium, Yuan Y, Ju YS, Kim Y, Li J, Wang Y, et al. Comprehensive molecular
719 characterization of mitochondrial genomes in human cancers. Nat Genet. 2020;52:342–52.

720 40. Brandon MC. MITOMAP: a human mitochondrial genome database--2004 update. Nucleic Acids
721 Res. 2004;33:D611–3.

722 41. Gu X, Kang X, Liu J. Mutation signatures in germline mitochondrial genome provide insights into
723 human mitochondrial evolution and disease. Hum Genet. 2019;138:613–24.

724 42. Bolze A, Mendez F, White S, Tanudjaja F, Isaksson M, Jiang R, et al. A catalog of homoplasmic and
725 heteroplasmic mitochondrial DNA variants in humans [Internet]. BioRxiv: Genetics; 2019 Oct.
726 Available from: https://doi.org/10.1101/798264

727 43. Samstag CL, Hoekstra JG, Huang C-H, Chaisson MJ, Youle RJ, Kennedy SR, et al. Deleterious
728 mitochondrial DNA point mutations are overrepresented in Drosophila expressing a proofreading-
729 defective DNA polymerase γ. Chinnery P, editor. PLoS Genet. 2018;14:e1007805.

730 44. Wu Z, Waneka G, Broz AK, King CR, Sloan DB. *MSH1* is required for maintenance of the low
731 mutation rates in plant mitochondrial and plastid genomes. Proc Natl Acad Sci USA. 2020;117:16448–
732 55.

733 45. Saito S, Tamura K, Aotsuka T. Replication origin of mitochondrial DNA in insects. Genetics.
734 2005;171:1695–705.

735 46. Cupp JD, Nielsen BL. Minireview: DNA replication in plant mitochondria. Mitochondrion.
736 2014;19:231–7.

737 47. Shinbrot E, Henninger EE, Weinhold N, Covington KR, Göksenin AY, Schultz N, et al.
738 Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns
739 and human origins of replication. Genome Res. 2014;24:1740–50.

740 48. Andreazza S, Samstag CL, Sanchez-Martinez A, Fernandez-Vizarra E, Gomez-Duran A, Lee JJ, et
741 al. Mitochondrially-targeted APOBEC1 is a potent mtDNA mutator affecting mitochondrial function
742 and organismal fitness in Drosophila. Nat Commun. 2019;10:3280.

743 49. Tanaka M, Ozawa T. Strand asymmetry in human mitochondrial DNA mutations. Genomics.
744 1994;22:327–35.

745 50. Belle EMS, Piganeau G, Gardner M, Eyre-Walker A. An investigation of the variation in the
746 transition bias among various animal mitochondrial DNA. Gene. 2005;355:58–66.

747 51. Broz AK, Waneka G, Wu Z, Gyorfy MF. Detecting de novo mitochondrial mutations in
748 angiosperms with highly divergent evolutionary rates. Genetics. 2021;29.

749 52. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, et al. Detecting ultralow-frequency
750 mutations by Duplex Sequencing. Nat Protoc. 2014;9:2586–606.

751 53. Koster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics.
752 2012;28:2520–2.

753 54. Stoler N, Arbeithuber B, Guiblet W, Makova KD, Nekrutenko A. Streamlined analysis of duplex
754 sequencing data with Du Novo. Genome Biol. 2016;17:180.

755 55. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations
756 by next-generation sequencing. Proc Natl Acad Sci USA. 2012;109:14508–13.

757 56. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
758 Bioinformatics. 2009;25:1754–60.

759 57. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
760 EMBnet.journal. 2011;17:3.

761 58. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map
762 format and SAMtools. Bioinformatics. 2009;25:2078–9.

763 59. Calabrese FM, Simone D, Attimonelli M. Primates and mouse NumtS in the UCSC Genome
764 Browser. BMC Bioinformatics. 2012;13:S15.

765 60. SciPy 1.0 Contributors, Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, et al. SciPy
766 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72.

767 61. Lujan SA, Longley MJ, Humble MH, Lavender CA, Burkholder A, Blakely EL, et al. Ultrasensitive
768 deletion detection links mitochondrial DNA replication, disease, and aging. Genome Biol. 2020;21:248.

769

770

771

772

773

774

775

776

777

778

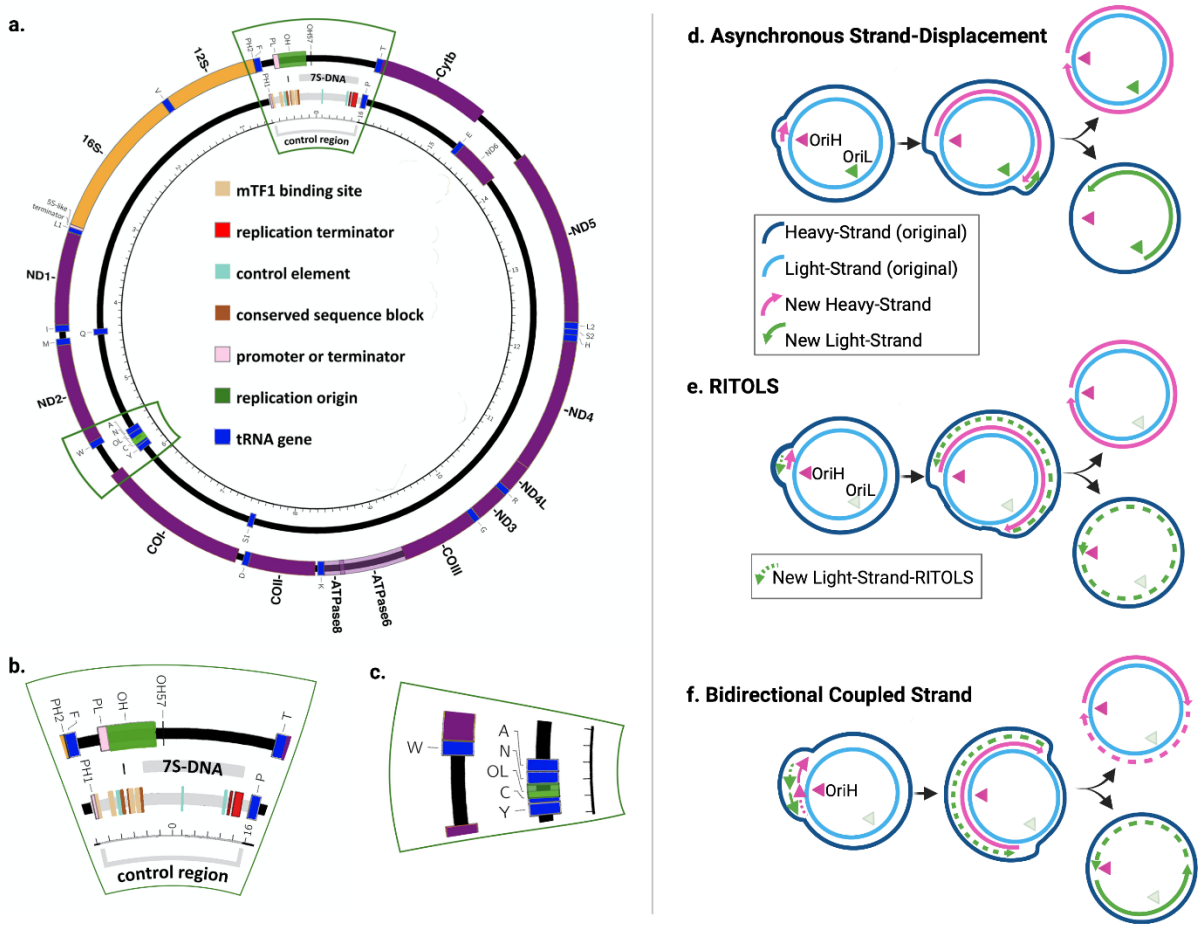

Fig 1. a Schematic of mammalian mtDNA and proposed replication models. Gene order and regulatory structures are preserved between humans and mice. Outer ring represents the light strand and the inner ring represents the heavy strand. Genes colored by complex is as follows: *blue*=ribosomal genes; *yellow*=Complex I; *orange*=Complex III; *red*=Complex IV; *purple*=Complex V. **b** Magnified area of the control region. **c** Magnified area of the Ori_L. **d** Schematic of asynchronous strand-diplacement model as originally proposed by Clayton and colleagues. **e** Schematic of RITOL model. **f** Schematic of strand-synchronous bidirectional replication model. Figure adapted from Lujan *et al*. and licensed under the CC BY 4.0 [61].

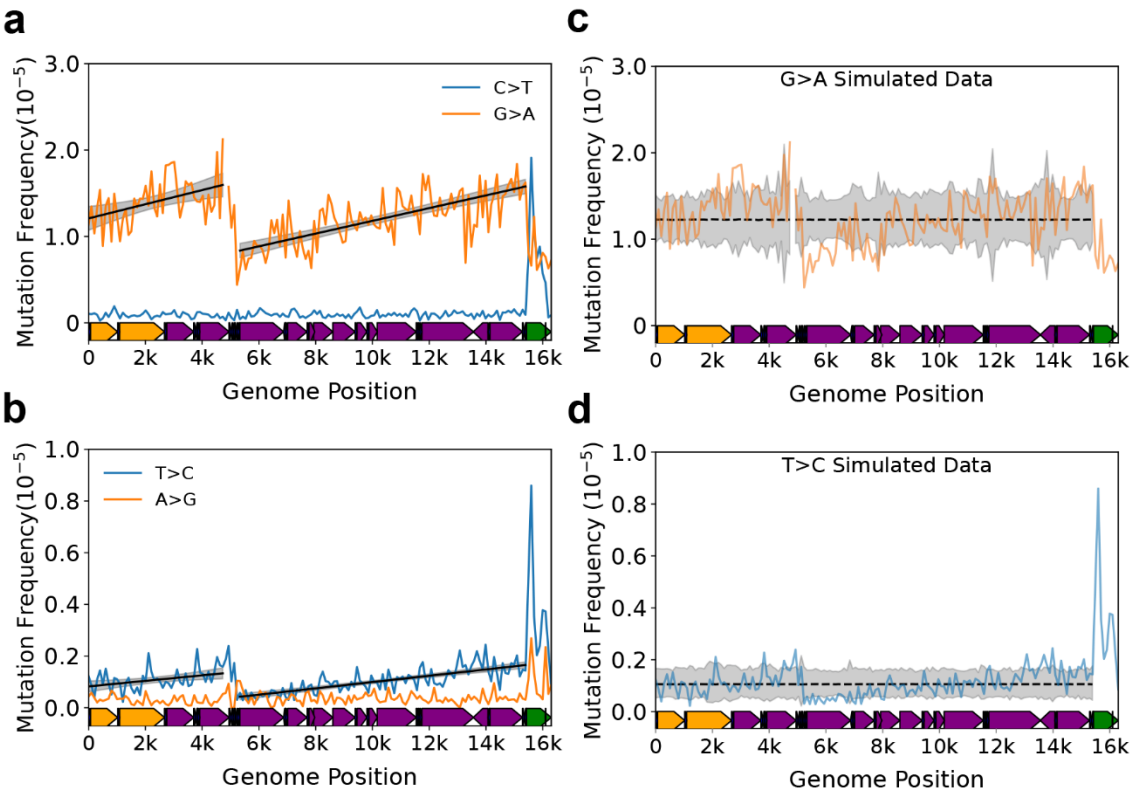

Fig 2. Somatic transitions mutations exhibit a mutational gradient in mouse mtDNA. **a-b** Plot and linear regression (*black line*) of reference strand (*i.e.* L-strand) C→T/G→A and T→C/A→G mutation frequencies as a function of genome position. Each data point denotes a 100bp bin. **c-d** Distribution of simulated mutation frequencies of G→A and T→C mutations along the mouse mtDNA. Simulations are based on the data in **a** and **b**. *dotted black line*=bin specific mean; *grey shading*=empirical 95% confidence interval; *red line*=fitted regression, dotted denotes p>0.05 for slope and solid denotes p≤0.05; *red shading*=95% confidence interval of linear regression. Mouse mtDNA structure and coordinates are shown on the x-axis and are the same for all panels (*orange*=rRNA gene, *purple*=protein coding, *dark blue*=tRNA gene, *green*=control region).

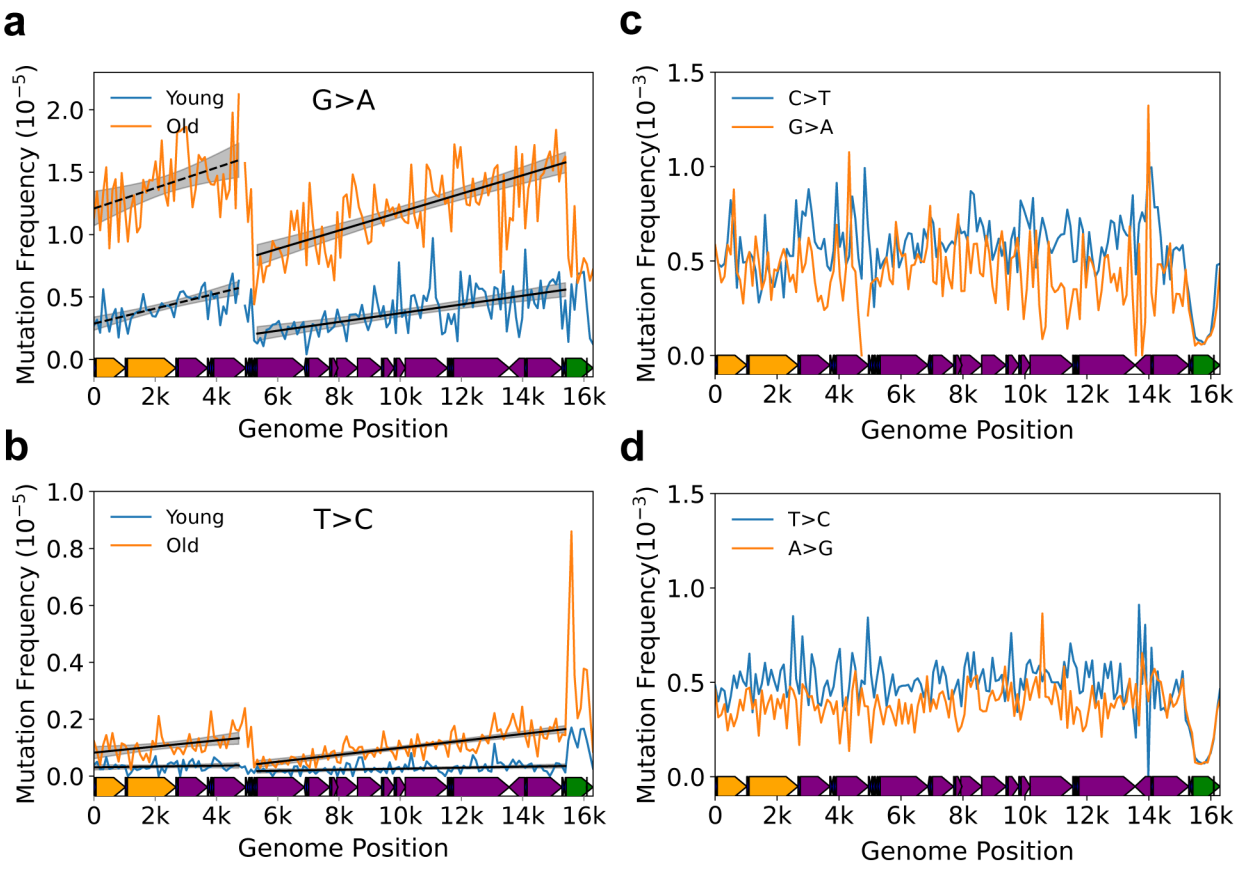

Fig 3. A mutational gradient is established over the course of natural aging and is not directly caused by polymerase-γ base selectivity. **a-b** Changes in the gradient slope of the major arc between 4-6 month old (*blue*) and 26-month old (*orange*) mice. *Black line* and *grey shading*=fitted regression and 95% confidence interval. **c-d** Plot and linear regression (*black line*) of reference strand (*i.e.* L-strand) C→T/G→A and T→C/A→G mutation frequencies show an absence of mutational gradient in Pol-γ^{exo-} mice, which lack a functional exonuclease activity in DNA Polymerase γ. Mouse mtDNA structure and coordinates are shown on the x-axis and are as follows: *orange*=rRNA gene, *purple*=protein coding, *dark blue*=tRNA gene, *green*=control region.

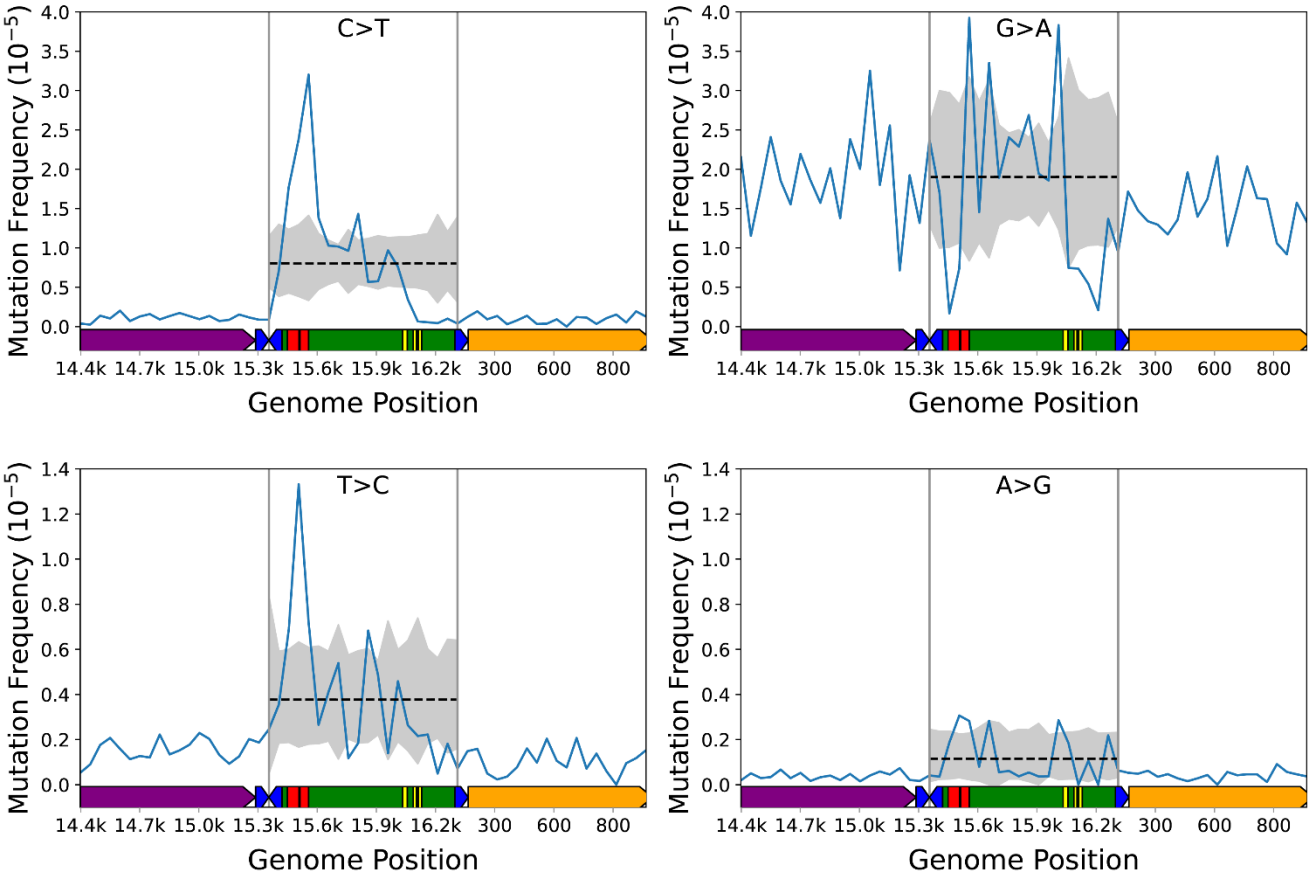

Fig 4. Mutations in the mtDNA control region display a non-uniform distribution and constraints at some loci. The observed per bin mutation frequency (*blue line*) and simulated distribution of data (*dotted black line*=mean, gray shading=experimental 99.975% confidence interval) C→T, G→A, T→C, and A→G mutations. Each data point denotes a 50bp bin. Data points outside the shaded areas are either over- or under-represented compared to random chance. Mouse mtDNA structure and coordinates are shown on the x-axis and are the same for all panels (orange=*mt-Rnr1* gene*, purple*=*Cytb* gene*, dark blue*=tRNA genes*, green*=control region, *red*=ETAS1&2; *yellow*=CSB1-3).

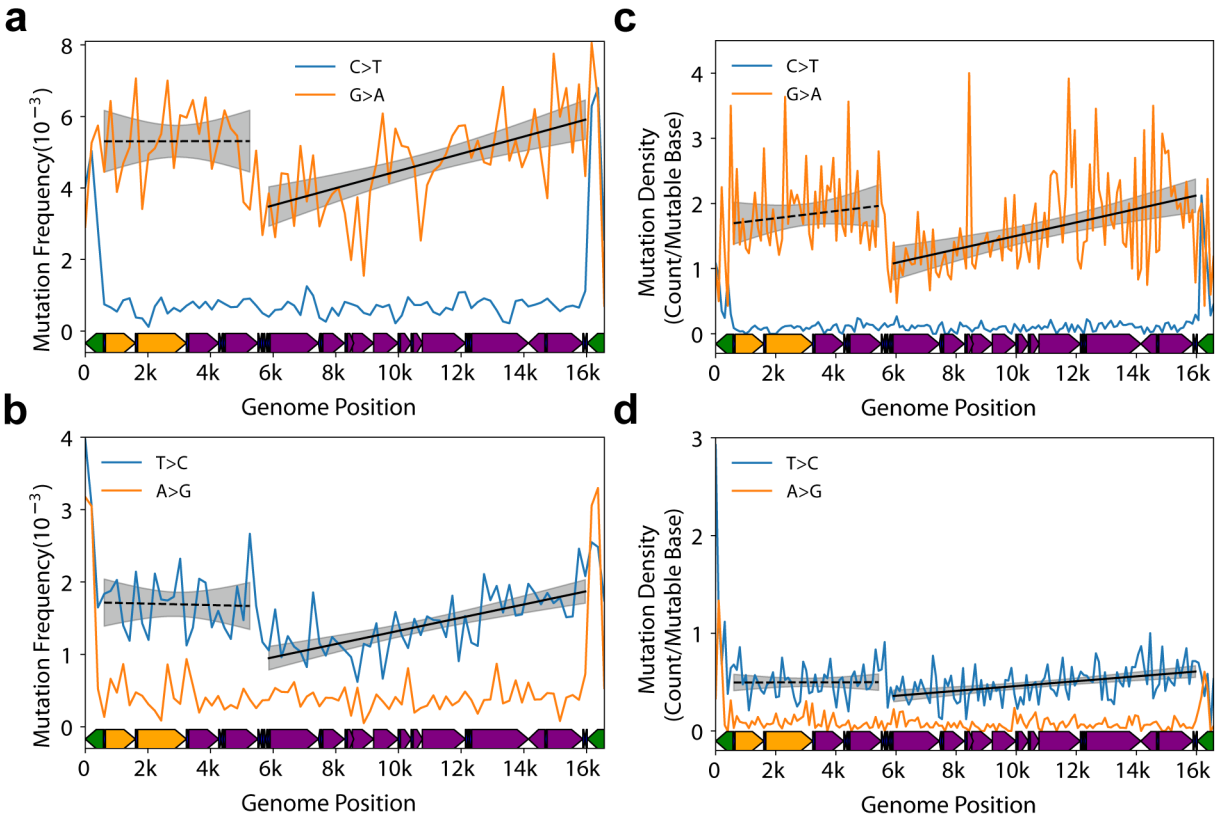

Fig 5. Somatic mutational gradient is conserved in human mtDNA. a-b Plot and linear regression (*black line*) of reference strand (*i.e.* L-strand) C→T/G→A and T→C/A→G mutation frequencies as a function of genome position from prior published Duplex-Seq data. Each data point denotes a 200bp bin. **c-d** Plot and linear regression (*black line*) of reference strand (*i.e.* L-strand) C→T/G→A and T→C/A→G mutation densities as a function of genome position in human tumor data from PCAWG dataset. Each data point denotes a 100bp bin. Human mtDNA structure and coordinates are shown on the x-axis and are the same for all panels and are as follows: *orange*=rRNA gene, *purple*=protein coding, *dark blue*=tRNA gene, *green*=control region. *Black dotted line*= linear regression slope p>0.05; *dashed black line*=linear regression slope p<0.05; *grey shading*= regression 95% confidence interval.

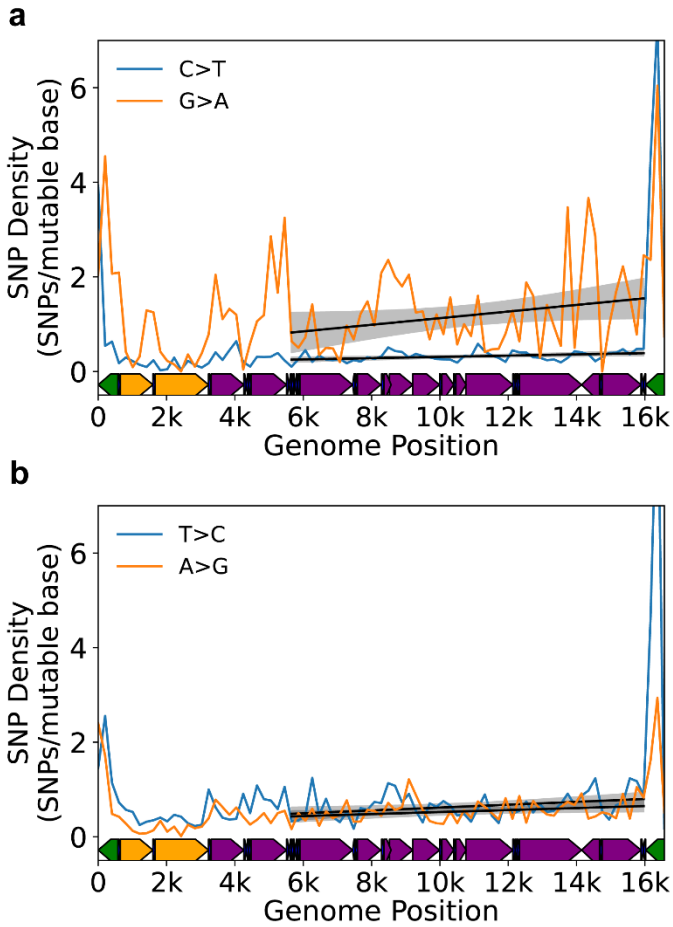

Fig 6. Mutational gradient is detected in major arc in human population SNPs. **a** Density of C→T and G→A SNPs on the L-strand. **b** Density of T→C and A→G SNPs on the L-strand. Data are from Gu *et al*. [40]. *Black dotted line*= linear regression slope p>0.05; *dashed black line*=linear regression slope p<0.05; *grey shading*= regression 95% confidence interval. Human mtDNA structure and coordinates are shown on the x-axis and are the same for all panels and are as follows: *orange*=rRNA gene, *purple*=protein coding, *dark blue*=tRNA gene, *green*=control region.

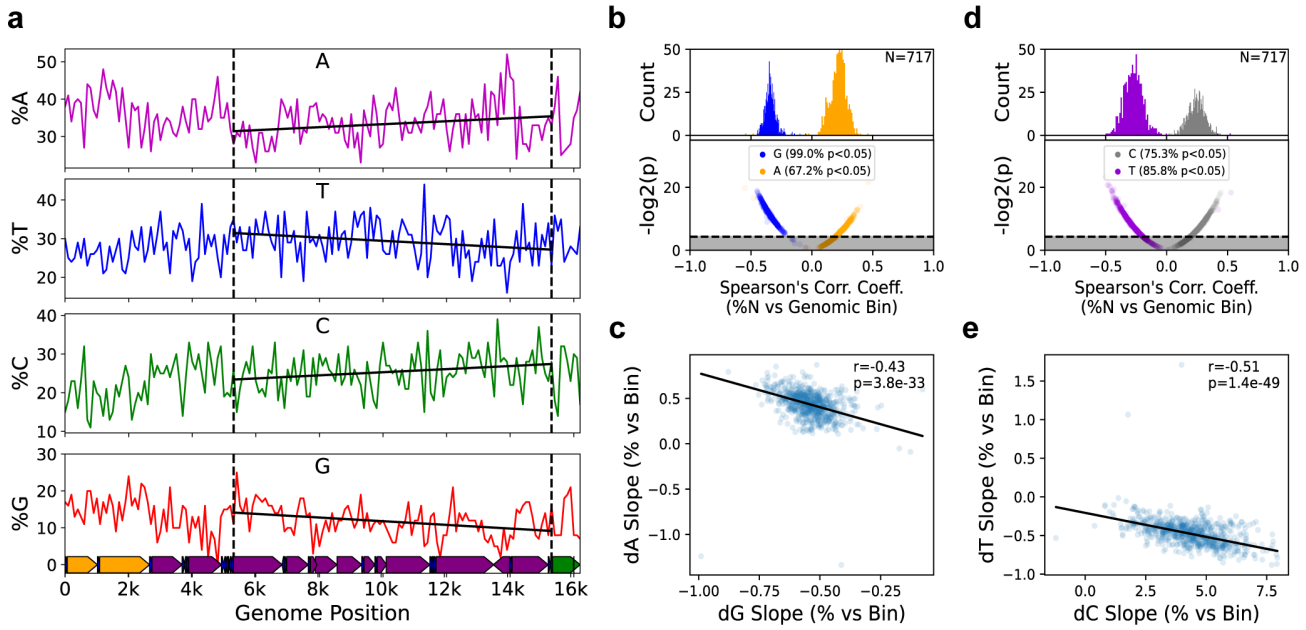

Fig 7. Genome composition bias mirrors the somatic gradient in mammals. a Base composition gradient as exemplified by the murine mtDNA. Vertical dashed lines delimit the major arc. Solid black lines are the best fit regression by robust linear regression. Slopes are significantly different from zero in all cases. Gene coloring: *orange*=rRNA gene, *purple*=protein coding, *dark blue*=tRNA gene, *green*=control region. **b,c** Anticorrelation of dG and dA base composition in mammals. Most mammalian genomes show a statistically significant spatially dependent depletion of dG and enrichment of dA in the major arc and the strength of dG depletion is directly proportional to dA enrichment in a species dependent manner. **c,d** dT and dC show a similar anti-correlative pattern as dG and dA.