Published online 6 October 2021

https://doi.org/10.1093/nar/gkab901

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

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] and Scott R. Kennedy[1,*]

[1]Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA 98195, USA, [2]Department of Biochemistry, University of Washington, Seattle, WA 98195, USA and [3]Department of Radiology, University of Washington, Seattle, WA 98195, USA

Received July 23, 2021; Revised September 10, 2021; Editorial Decision September 20, 2021; Accepted September 22, 2021

ABSTRACT

Mutations in mitochondrial DNA (mtDNA) cause maternally inherited diseases, while somatic mutations are linked to common diseases of aging. Although mtDNA mutations impact health, the processes that give rise to them are under considerable debate. To investigate the mechanism by which *de novo* mutations arise, we analyzed the distribution of naturally occurring somatic mutations across the mouse and human mtDNA obtained by Duplex Sequencing. We observe distinct mutational gradients in G→A and T→C transitions delimited by the light-strand origin and the mitochondrial Control Region (mCR). The gradient increases unequally across the mtDNA with age and is lost in the absence of DNA polymerase γ proofreading activity. In addition, high-resolution analysis of the mCR shows that important regulatory elements exhibit considerable variability in mutation frequency, consistent with them being mutational 'hot-spots' or 'cold-spots'. Collectively, these patterns support genome replication via a deamination prone asymmetric strand-displacement mechanism as the fundamental driver of mutagenesis in mammalian DNA. Moreover, the distribution of mtDNA single nucleotide polymorphisms in humans and the distribution of bases in the mtDNA across vertebrate species mirror this gradient, indicating that replication-linked mutations are likely the primary source of inherited polymorphisms that, over evolutionary timescales, influences genome composition during speciation.

INTRODUCTION

Owing to their evolutionary origin, mitochondria have retained a small extra-nuclear genome encoding essential components of the electron transport chain (ETC), as well as transfer and ribosomal RNAs required for their translation (Figure 1A). The ETC is responsible for producing cellular energy through oxidative phosphorylation and maintaining a reducing chemical environment. As such, the genetic information encoded in the mtDNA is essential for maintaining cellular homeostasis. However, due to the absence of several DNA repair pathways, mammalian mtDNA exhibits mutation frequencies >100-fold higher than the nuclear genome (2).

Mutations in the mtDNA cause a number of devastating maternally inherited diseases, while the accumulation of mutations in the soma is linked to common diseases of the elderly, including cancer, diabetes, and neurodegenerative diseases (Reviewed in (3)). While an important driver of human health, the mutagenic processes that give rise to these mutations are under considerable debate (4,5). As originally posited by Denham Harmon, the proximity of mtDNA to the ETC should result in high levels of oxidative damage (i.e. 8-oxo-dG), yielding predominantly G→T/C→A transversions (6,7). Counter to this prediction, low levels of G→T/C→A mutations and a preponderance of G→A/C→T and T→C/A→G transitions are observed (8–11). The presence of these mutations has been interpreted as arising from either base selection errors by

[*]To whom correspondence should be addressed. Tel: +1 206 543 5452; Email: scottrk@uw.edu
Present address: Jeremy A Whitson, Department of Biology, Davidson College, Davidson, NC 28035, USA.

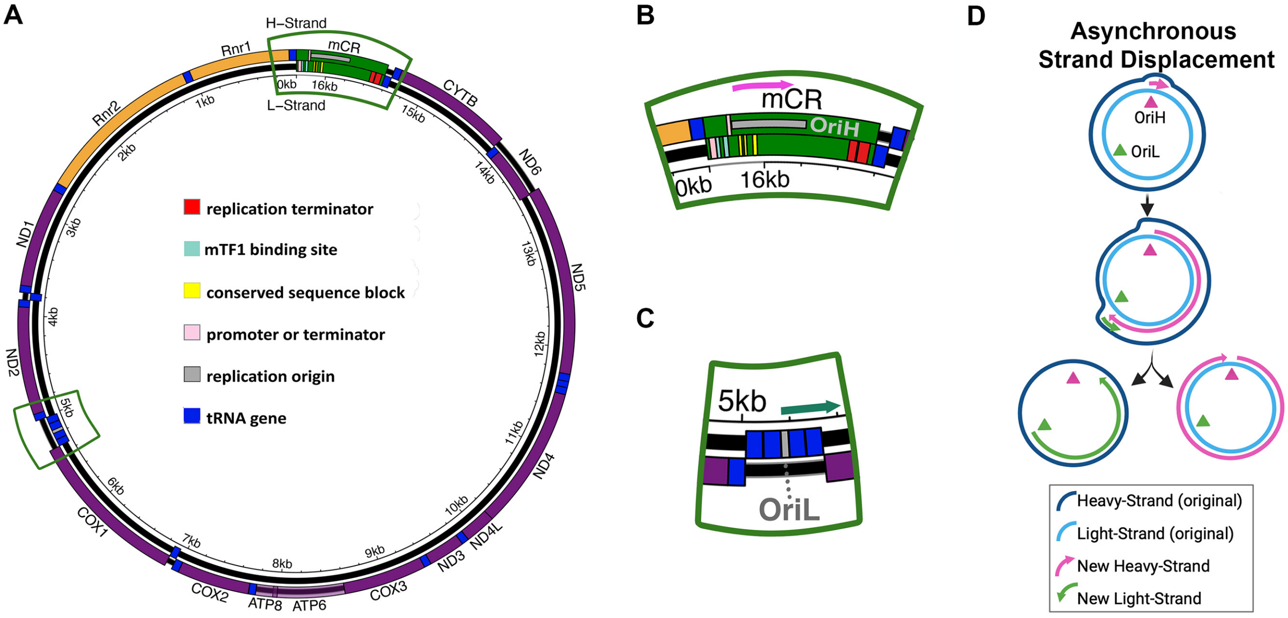

Figure 1. Schematic of mammalian mtDNA and proposed replication models. (**A**) Schematic of mouse mtDNA gene organization. Gene order and regulatory structures are preserved between humans and mice. Outer ring represents the (reference) light-strand and the inner ring represents the (anti-reference) heavy-strand. (*orange* = rRNA gene, *purple* = protein coding, *dark blue* = tRNA gene, *green* = control region). (**B**) Magnified area of the control region. *Pink arrow* = direction of replication. (**C**) Magnified area of the Ori$_L$. *Green arrow* = direction of replication. (**D**) Schematic of asynchronous strand-displacement model as originally proposed by Clayton and colleagues. Figure adapted from Lujan *et al.* and licensed under the CC BY 4.0 (1).

DNA polymerase γ (Pol-γ), spontaneous deamination of deoxycytidine and deoxyadenosine, or both, and not due to reactive oxygen species (ROS) induced 8-oxo-dG adducts.

Replication of the mtDNA by Pol-γ is required for fixation of mutations into the genome. Thus, the distribution of mutations can provide clues to the mechanism by which *de novo* mutations arise. The mechanism of mtDNA replication remains poorly understood, but, in vertebrates, is generally thought to occur via an asynchronous strand displacement mechanism involving two separate, strand-specific, origins (12,13). In this model, replication is initiated at the heavy-strand (H-strand) origin (Ori$_H$), located in the non-coding mCR, using a displacement loop (D-loop) as the replicon primer (Figure 1B,D). Synthesis of the nascent H-strand displaces the original H-strand into a single-stranded state. Upon traversing the light-strand (L-strand) origin (Ori$_L$), located ∼11,000 bp away from the mCR, a second replicon is established and proceeds in the opposite direction, resulting in the original H-strand becoming double-stranded (Figure 1C,D). Replication is completed when both replication forks complete their circumnavigation. Alternative vertebrate models have been proposed whereby the displaced H-strand is annealed to RNA transcripts, termed RITOLS, that serve to prevent the single-stranded state and act as intermittent priming sites for L-strand replication (Supplementary Figure S1A) (14,15). Visualization of replication intermediates by 2D-gel electrophoresis has also indicated the presence of coupled-strand synthesis involving a more conventional leading/lagging strand replication fork initiating from a bidirectional origin (Ori$_b$) in the mCR or potentially throughout a multi-kilobase 'initiation zone' (16–18) (Supplementary Figure S1B). The asynchronous and coupled-strand mechanisms have been proposed to be present at the same time, contingent on the physiological state of the cell (16,19). Lastly, alternative tRNA genes outside of the Ori$_L$ tRNA cluster have been proposed to act as alternative L-strand origins (20,21).

Each of these replication models have significant implications for mtDNA mutagenesis. As hypothesized in previous phylogenetic studies, an asymmetric mechanism of mtDNA replication could explain the phenomena of strand bias, G/C- and A/T-skew, and mutational gradients seen across related taxa (22–24). Specifically, the long-lasting 'naked' ssDNA replication intermediate in the original model predicts elevated levels of G→A/C→T mutations when the template dC is in the (single-stranded) H-strand due to cytidine exhibiting substantially increased deamination rates when present in a single-stranded state (25). In this case, the mutational pressure is away from dC content in the H-strand and towards increased dT content. Moreover, genes closer to the mCR/Ori$_H$ are expected to be more mutation prone than those farther away due to longer times in the single-stranded state. In contrast, both conventional leading/lagging-strand synthesis and intermittent priming models could produce G/C strand bias and/or A/T-skew arising from different mutation frequencies between the leading and lagging strands, a phenomenon observed in bacteria and nuclear DNA (nDNA) replication (26,27), but a mutational gradient stemming from deamination events should be weak or absent due to negligible amounts of ssDNA. Using modern sequencing technologies, the strand asymmetry in transitions has been described in somatic mtDNA mutations (8–11). More recently, high accuracy sequencing of murine oocytes shows a similar bias towards the strand-asymmetric accumulation of transitions, establishing a link between the dominant mutagenic process in somatic tissues and what is seen in population genetics (28). However, to date, no mutational gradient has been reported outside the context of phylogenetic analyses and it remains an open question if it is an active process or a byproduct of selective pressure over time.

In this report, we have taken advantage of several large high accuracy mtDNA mutation data sets previously generated with Duplex Sequencing (Duplex-Seq) to examine the distribution of somatic mutations in the mtDNA of mice and humans (Sanchez-Contreras & Sweetwyne *et al.*, in preparation and (11,29,30)). We find that G→A and T→C transitions, but not their complementary mutations, exhibit a strand-asymmetric gradient delimited by the $\mathrm{Ori_L}$ and the mCR. This gradient is evolutionarily conserved between mouse and humans. The mCR also exhibits a remarkably different mutational pattern compared to the coding portion of the genome and is consistent with the presence of a stable D-loop structure bounded by highly conserved regulatory sequence blocks (Figure 1B). Comparison of the somatic mutational gradient to the distribution of SNPs in the human population, as well as the distribution of bases along the genome across species, shows remarkable concordance. Taken together, our findings demonstrate that an active mutational gradient drives the unequal accumulation of mutations in mtDNA and is most consistent with being the result of a strand-asymmetric replication model with an extensive ssDNA replication intermediate. Moreover, this unusual mutagenic process influences population level haplotypes and likely drives genome composition over evolutionary time scales in vertebrates.

MATERIALS AND METHODS

Duplex sequencing and data processing

The DNA libraries and sequencing of the Duplex-Seq libraries was performed as indicated in the original publications. Specifically, the human data was obtained from Baker *et al.* (SRA accession PRJNA449763 (11) and Hoekstra *et al.* (SRA accession PRJNA237667) (29) using only data from healthy controls in our analysis. Data from diseased samples were not included. Mouse data was generated from male C57Bl/6J mice at 4–5 ($N = 5$) and 26 ($N = 6$) months of age. Briefly, aged mice were obtained from the NIA aged rodent colony at an age of 22–23 months and then housed at the University of Washington animal facility until the desired age under approved conditions. Animals were euthanized at the indicated age and a ∼2 mm section of the heart (apex), liver (lobe VI posterior), kidney (outer cortex), skeletal muscle (proximal gastrocnemius), brain (both cerebellum and hippocampus), and eye (retina and eye cup) were flash frozen and stored at -80°C until processed for sequencing. A ∼1mm tissue punch was used to obtain a representative tissue sample from brain regions. Half of the retina was used for processing and eye cups were used in their entirety. Total DNA was purified from each tissue punch/sample using a QIAamp Micro DNA kit using the manufacturer's protocol. Total DNA was prepared for Duplex Sequencing using our previously published protocol with modifications as described in Hoekstra *et al.* (29,31). Pol-$\gamma^{\mathrm{exo-}}$ mouse data was generated from Pickrell *et al.* (SRA accension PRJNA729056) (30).

After obtaining the raw data, we processed all Duplex-Seq data using v1.1.4 of our Snakemake-based Duplex-Seq-Pipeline to ensure that all data were uniformly processed with the exception that different data sets had different unique molecular identifier (UMI) and read lengths (31). Briefly, we perform a reference free consensus approach for error correction similar to what is reported in Stoler *et al.* (32). Briefly, the UMI and any associated spacer sequence are parsed from the read, the read 1 and read 2 UMI from a read pair is sorted alphabetically and associated with the read's SAM record when converted to an unaligned BAM file (See Stoler *et al.* for details (32)). After the UMIs from read 1 and read 2 are sorted alphabetically such that all reads derived from the same strand of the same parental molecule are grouped together. UMI families from opposite strands of the same parental DNA fragment are grouped sequentially in the sorted file. A per-position consensus is generated for each single-strand is generated with a cut-off of 70% identity and a minimum of 3 reads with the same UMI being required to call a consensus, as previously described (31). The double-strand consensus is then generated from the two single-strand consensuses sharing the same UMI, if present, with the exception that the identity of the base must match between the two single-strand consensus. Reads with > 2% ambiguous bases are removed from further analysis. The resulting post-processed fastq files are aligned against the reference genome (hg38 chrM for the human data and mm10 chrM for mouse the data) using bwa v0.7.17 (33). The overlapping portions of reads and 10-cycles of the 5′ and 3′ ends of the reads are clipped using fgbio (https://github.com/fulcrumgenomics/fgbio) and adapter sequences are removed using Cutadapt (34). Insertion/deletion (in/del) realignment and in/del left alignment were performed by the Genome Analysis Toolkit (v3.8.x). A draft list of variants is generated using the pileup functionality of samtools (35). Reads containing non-SNP variants (defined as a variant allele fraction > 40%) are parsed out and subjected to BLAST-based alignment against a database containing potential common contaminants (dog: canFam3; bovine: bosTau9; nematode: ce11; mouse: mm10; human: hg38; rat: Rnor_6.0). The inclusion of our target genome in this database also allows for the identification of pseudogenes. Reads that unambiguously map to the same coordinates as the original alignment are kept and the remaining reads and associated variants are removed from further analysis. An exception to this process was made for mouse mtDNA due to the presence of a ∼5000 bp nuclear pseudogene with perfect identity to mm10 chrM (36). In this case, any ambiguous BLAST alignments mapping to this region were assumed to be mitochondrial in origin and kept. The mutated reads passing our BLAST filter are merged back with the non-mutated reads and mutation frequencies calculated based on the provided target coordinates.

To generate the bin data for each age cohort (i.e. 4.5mo and 24–26mo), we divided the genome up into the indicated bin size and then calculated the mutation frequency for each bin, i, and mutation 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 sample j and S_{ij} is the number of sequenced bases in bin i of sample j that are mutable for mutation type N. A mutation was only counted if its variant allele fraction (VAF) was $< 1\%$ to minimize the effects of inherited or early arising mutations.

For the D-loop focused analysis, we generated a new partial mm10 chrM reference consisting of the mCR portion with the flanking 1000 bases (chrM 14400–16299::chrM 1–1001) in order to allow for better alignment of reads across the entirety of the mtDNA. The binning process was performed as described, but with 50bp bins. Per bin mutation frequencies were calculated as described above.

Tumor sequencing data

Tumor single nucleotide variant data was generated using the methods outlined previously and obtained from The Cancer Mitochondria Atlas data portal (https://ibl.mdanderson.org/tcma/download.html) (37). Similar to our Duplex-Seq analysis, the human mtDNA was divided into 100bp bins and the number of variants of each of the 12 mutation classes was divided by the number of wild-type base of the respective mutation class within each bin (i.e. #G→A/#G's, etc).

SNP and genome composition data

The SNP data sets were obtained from Gu *et al*. and Bolze *et al*. using their respective procedures (38,39). Briefly, the Gu *et al.* data set is comprised of 44,334 SNPs reported in the MITOMAP database that was filtered for haplotype private variants. The Bolze *et al*. data comprises 14,283 homoplasmic variants from 196,324 unrelated individuals. Because this dataset is not limited to haplotype private SNPs, as the Gu *et al.* data, we limited our analysis to rare SNPs (population frequency $\leq$ 1:1000) in order to minimize population structure of the data from confounding our analysis. SNPs occurring more than once were assumed to have arisen from a single *de novo* event. We divided to human mtDNA into 200bp bins and then calculated SNP density by summing the number of variants of each of the 12 mutation classes and then dividing by the number of respective wild-type bases within each bin (i.e. #G→A/#G's, etc).

For the genome composition analysis, a complete set of curated mtDNA sequences and annotations were downloaded from the NCBI Reference Sequence project (https://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/) in GenBank format. Entries were parsed by taxonomic Class and filtered to separately keep *Mammalia* (mammals), *Archelosauria* (birds and crocodilians), *Lepidosauria* (lizards and snakes), and *Actinopterygii*i (ray-finned fish0. mtDNA sequence corresponding to the major arc of each species was extracted and divided into 100bp bins and the nucleotide composition calculated as a percent of each base type. The slope and/or correlation coefficient for the change in genome composition as a function of bin number (i.e. genome position) for each individual species was then calculated as described below.

Statistical analysis

Statistical analysis was performed in python using either Statsmodels (https://github.com/statsmodels/statsmodels) or SciPy (40), 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.

We used a permutation-based test to calculate a *P*-value for the observed regression slope that takes into account the sequence composition of each 100bp bin in the coding region of the genome (bins 1–153; genome positions 1–15,400). Briefly, a weighted probability for each mutation type (i.e. G→A, T→C, *etc*) to occur was calculated for each bin 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 of mutations of each type (excluding mutations in the mCR) 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. After reach permutation, linear regression of the distributed data was performed on the major and minor arc as described above. This procedure was repeated 10,000 times. The *P*-value for the permutation test was calculated by dividing the number of times the slope was greater than or equal to the observed regression slope by 10,000.

A similar permutation-based approach was used to determine over- or under-abundance of mutations in the mCR, with some modifications. Briefly, the mCR was divided into twenty 50bp bins and a weighted probability for each mutation type calculated as described above. The mutations were randomly permuted using the calculated weights and the per bin mutation frequency was calculated as described above and the values for each bin kept. This procedure was repeated 100,000 times and a Bonferroni corrected confidence interval for determining mutational over- or under-representation was set to 99.75% to account for the 20 bins that comprised the mCR region being evaluated. Datapoints outside this range were considered significantly different from random chance.

RESULTS

As part of a comprehensive analysis on the effects of aging and mitochondrial-targeted interventions on somatic mtDNA mutation accumulation in eight different mouse tissues, we used Duplex-Seq to collect 34,113 independent, high accuracy, somatic mutations spread across the entirety of the mtDNA molecule (Supplementary Data S1). During the course of initially analyzing our data, we noted significant variability in the per gene mutation frequency (when looking at individual mutation types). Ordering the genes by their location in the genome, instead of by complex, showed an increasing frequency in G→A mutations, reminiscent of what has been observed in phylogenetic studies (Supplementary Figure S2) (24). Intrigued by this observation, we took advantage of the large number of mutations to obtain a higher resolution understanding of how mutagenesis varies across the mtDNA.

Variants are spread out across 85 individual samples with a mean of 401 (range: 48–1,496) mutations per sample, corresponding to a mean density of 0.025 mutations/bp. Because mutations are spread across 12 different mutation classes, the mutation density of individual samples would not provide a higher resolution than at a per gene level. To overcome this issue, we combined the data from all tissues to produce the most robust data set possible. Specifically, we divided the genome into 100bp bins (total of 163) and, for each mutation class (i.e. G→A, G→C, G→T, *etc*), summed the mutation counts observed in each bin across our all samples, separated by age cohort (young ($n = 40$): 4.5 months; old ($n = 45$): 26 months). We then normalized for both genome base composition and variability in sequencing depth of each bin by dividing the mutation count by the total number of wild-type mutable bases sequenced across the constituent samples (Supplementary Data S2 and S3). This effectively gives a weighted mean of the mutation frequency for each bin for all samples.

Plotting the mutation frequency by genome position (i.e. bin) in our 26 month old cohort ($N = 25{,}020$ mutations) reveals an apparent discontinuous gradient bounded by the $\mathrm{Ori_L}$ and mCR for G→A and T→C transitions, but not their respective complementary mutation types (Figure 2A, B). An exception is the mCR (bins 154–163; genome positions 15,400–16,299) which exhibits a notable spike in C→T, but a decline in G→A mutations, whereas both T→C and A→G mutations show increases in the mCR, consistent with previous reports (8,28). Performing separate regressions of the minor arc (defined as the region between the end of the mCR and $\mathrm{Ori_L}$ (i.e. bins 1–47 or positions 1–4,800)) and major arcs (defined as the region between the $\mathrm{Ori_L}$ and the start of the mCR (i.e. bins 53–153 or positions 5,300–15,400) show highly significant increases in mutation frequency across their respective genomic coordinates (minor arc: G→A slope = $8.24 \pm 3.06 \times 10^{-8}$, $P = 0.007$; T→C slope = $1.08 \pm 0.35 \times 10^{-8}$, $P = 0.002$; major arc: G→A slope = $7.42 \pm 0.75 \times 10^{-8}$, $P = 5.35 \times 10^{-23}$, T→C slope = $1.23 \pm 0.10 \pm 10^{-8}$, $P = 1.33 \times 10^{-35}$). With the exception of G→C mutations in the major arc, no other mutation types exhibited a gradient (Supplementary Figure S3 and Supplementary Table S1). Notably, the G→C gradient is >10-fold smaller than the transition-based gradients and its relevance to mitochondrial biology, if any, is unclear. The strand bias (reference L-strand G→A and T→C mutations are equivalent to anti-reference H-strand C→T and A→G mutations, respectively) is consistent with the previous reports in somatic mtDNA mutations and the gradient is most consistent with the previously hypothesized strand-asynchronous replication mechanism with a deamination prone single-stranded replication intermediate involving only two origins of replication (8,12,13,28).

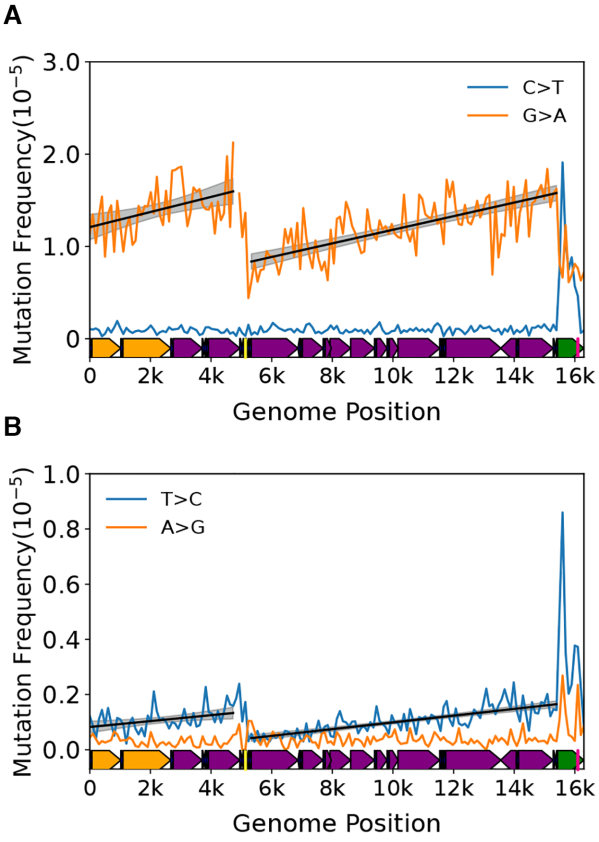

Figure 2. Somatic transitions mutations exhibit a mutational gradient in mouse mtDNA. Plot and linear regression (*black line*) of reference strand (i.e. L-strand) (**A**) C→T and G→A and (**B**) T→C and A→G mutation frequencies as a function of genome position. Each data point denotes a 100bp bin. *grey shading* = 95% confidence interval of linear regression (*black line*). 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, *yellow* = $\mathrm{Ori_L}$, *magenta* = $\mathrm{Ori_H}$).

Our mouse mtDNA data set combined mutation profiles of 8 unique tissue types from six organ systems, therefore we sought to account for potential tissue-specific effects and local differences in sequence contexts identified in these data (Sanchez-Contreras & Sweetwyne *et al.*, *in preparation*). To address the possibility that one tissue type in our data was driving the observed gradient, we performed a leave-one-out approach by eliminating one tissue type and then performed the same analysis on the reduced data set, repeating this analysis for each tissue type. As expected, the removal of data of any one tissue type did not alter our findings (Supplementary Table S2). These results point to the gradient not being an artifact of any single tissue type in our data. We next addressed the potential impact of different local sequence contexts within each bin by performing Monte-Carlo simulations that randomly redistributed each mutation observed in the 153 bins corresponding to the non-mCR portion of the genome (genome positions 1–15,400) using a weighted probability for each bin based on its base composition. After redistribution, the mutation frequency was then recalculated for each bin and a linear regression performed separately for the major and minor arcs. This procedure was repeated 10,000 times. As expected, the regression slopes of the observed G→A and T→C data significantly exceeded the simulated regression slopes of the permuted data in both the major or minor arcs, indicating that the positive mutational gradients are unlikely to be a consequence of nucleotide distribution or sequence (Supplementary Figure S4).

In sum, because the gradient is strand asymmetric, occurring when template dC and dA are in the (anti-reference) H-strand, and because cytidine and adenosine are more prone to deamination in the absence of base-pairing (41), our collective analysis is most consistent with a strand asynchronous replication mechanism with a long-lived single-stranded replication intermediate. Moreover, our data are inconsistent with a coupled leading/lagging strand mechanism, which does not have significant single-stranded regions that would systematically vary in duration along the genome.

Effects of age and Pol-γ fidelity point to the mutational gradient being the result of replication-linked deamination

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 (Figure 1D). The consequence of this model is that the portions of the mtDNA closest to their initiating origin should accumulate G→A and T→C mutations in the L-strand at a higher rate during aging due to more time in the single-stranded state and is expected manifest as an increase in the gradient slope over time. To test this hypothesis, we made use of the young (4–5 months; $N = 9{,}093$ mutations) and old age (26 months; $N = 25{,}020$ 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 $= 8.38 \pm 3.89 \times 10^{-9}$, $P = 0.031$) (Figure 3A, B; Supplementary Table S3; Supplementary Data S2 and S3). 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 replication-linked mutational gradient process is the primary driver of age-associated somatic mutations in mtDNA.

The presence of deoxyuracil in DNA is highly mutagenic because dU:dA base-pairs are properly hydrogen bonded and, consequently, are not readily proofread by the Pol-γ exonuclease domain, resulting in apparent C→T mutations (42,43). A similar scenario is relevant for deoxyinosine (42). As such, in the presence of proofreading, the majority of mutations have been proposed to be the result of replication past deaminated bases. However, base selectivity of the Pol-γ catalytic site has also been shown to have lower discrimination for dNTPs that result in transitions (44). In addition, Pol-γ has been shown to exist with and without its p55 accessory subunit that affects processivity and fidelity (45,46). Therefore, one explanation for the gradient is that it is the result of base selectivity errors by the Pol-γ catalytic site that stochastically escape proofreading.

To help distinguish between mutations driven by replication past deaminated bases versus poor base discrimination by the catalytic site, we analyzed 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-$\gamma^{\text{exo–}}$) (30). If the gradient and strand asymmetry are the result of base selectivity errors of the catalytic site that escape from proofreading, then the expectation would be for the gradient to remain in the absence of exonuclease activity. Conversely, the loss of the gradient and strand asymmetry would indicate that base discrimination is not strongly affected by genome location or base composition of the respective strands. Consistent with this last prediction, the strong positive gradient and strand asymmetry seen in G→A and T→C transitions from wild-type mice are no longer present (Figure 3C, D; Supplementary Table S4; Supplementary Data S4). The loss of the strong positive gradient, as well as strand bias, in the Pol-$\gamma^{\text{exo-}}$ mouse mtDNA points to a mechanism that is not the result of poor discrimination of non-damaged DNA. We did not evaluate the distribution of mutations in the mCR due to the likely presence of concatemers in the Pol-$\gamma^{\text{exo-}}$ mouse mCR (47), the effects of which can be seen by the significantly lower mutation frequencies in bins containing this region (Figure 3C,D; Supplementary Figure S5). We also note a slight, but statistically significant, negative slopes in G→A and T→C transitions, as well as T→A, C→G and G→T transversions in the major arc and a slight positive slope in minor arc A→T, but the relative effect size is substantially smaller than what is seen in G→A and T→C mutations in wild-type mice and its relevance in mtDNA biology, if any, is unclear (Supplementary Figure S5; Supplementary Table S1 and S4). Taken together with our analysis showing that the gradient slope increases with age, our data suggest a model where most mutations arise from deaminated bases that are a consequence of a long-lived single-stranded replication intermediate.

The distribution of variants in the mCR suggest regulatory elements exhibit altered mutational potentials

The mCR contains several important regulatory elements, including both transcriptional promoters, the Ori_{H}, several highly conserved sequence blocks (CSB), and extended termination-associated sequences (ETAS), whose specific regulatory functions are incompletely understood (Reviewed in Gustafsson *et al.* (48)) (Figure 1B). We and others have noted a distinctly different mutation frequency and spectrum in the mCR compared to the coding portion of the genome in both humans and mice (8,28,49), suggesting that the unique function and structure may strongly influence mCR mutagenesis, but high-resolution mapping of mutations has not been reported.

The mCR lies at the extreme 3′ terminal end of the *M. musculus* mtDNA reference genome, which presents is-

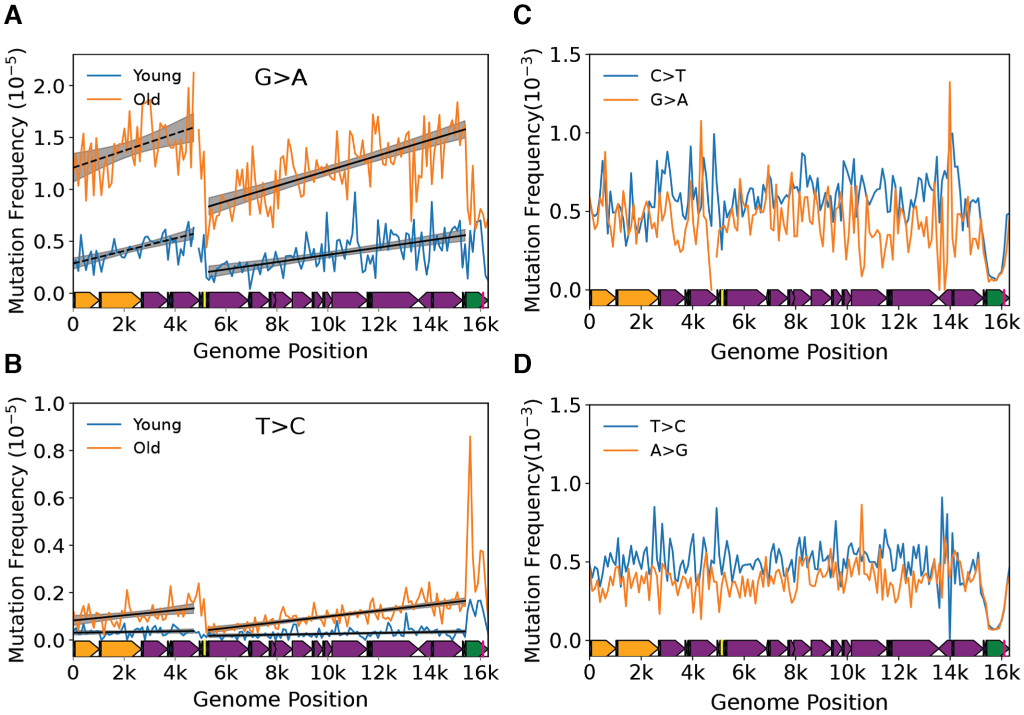

Figure 3. A mutational gradient is established over the course of natural aging and is not directly caused by polymerase-γ base selectivity. Changes in the gradient slope of the major arc between 4–6-month old (*blue*) and 26-month old (*orange*) mice for (**A**) G→A and (**B**) T→C mutations. *Solid black line* = linear regression slope $P < 0.05$; *dashed black line* = linear regression slope $P > 0.05$; *grey shading* = fitted regression; 95% confidence interval. (**C**) C→T and G→A and (**D**) T→C and 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, *yellow* = $\mathrm{Ori_L}$, *magenta* = $\mathrm{Ori_H}$.

sues during data alignment that gives rise to significant biases in sequence depth and mutation calls. To address this potential bias, we modified the mtDNA reference to place the mCR in the middle of the sequence and realigned our data to this modified reference. In addition, we decreased our bin size to 50bp to allow for a higher resolution mapping of mutations. The mCR exhibits prominent spikes and troughs that closely correspond to the ETAS, 7s DNA D-loop, CSBs, and the transcriptional promoters (Figure 4; Supplemental Data S5) (50). To confirm our findings, and to determine if any of these conserved sequences are over/under-represented compared to random chance, indicative of mutational 'hot-spots' or 'cold-spots' within the mCR, we performed Monte-Carlo simulations using a similar strategy as described for our mutational gradient analysis, but with 50bp bins, repeating the sampling 100,000 times, and setting the two-tailed Bonferroni corrected significance to $P < 0.0025$ to account for dividing the mCR into 20 bins (Figure 4; Supplementary Figure S6, *black line & grey shading*). The simulations confirm that C→T, T→C, G→A, and A→G, but not other mutation classes, show significant deviations from random sampling in these conserved structures. Of particular interest is a consistent mutational over-representation for C→T, T→C, and A→G mutations, but an under-representation for G→A in the ETAS (Figure 4, *red blocks*). This observation suggests the presence of a structure that is highly prone to certain mutation types and resistant to others or, alternatively, the loss of L-strand dG's prevents the maintenance of mtDNA. Consistent with the possibility that mutations can be selected against, all four transition types show a significant depletion of variants in the region between CSB3 and mt-$\mathrm{tRNA^{Phe}}$ that corresponds to the transcription promoters and mitochondrial transcription factor A (TFAM) binding sites, which are thought to be the source of the $\mathrm{Ori_H}$ replication primer (Figure 4, genome position 16,100–16,299) (48). Interestingly, no high level heteroplasmic or homoplasmic variants have been detected in the same region in human population studies, suggesting that this region is extremely important for mtDNA maintenance (49). Lastly, all four transitions exhibit a significant spike in the region between CSB1 and genome position ∼15,900 consistent with this region harboring the 7s DNA/RNA D-loop. Taken together, our high-resolution analysis of mCR mutations highlights that conserved regulatory elements in the mCR exhibit mutational profiles consistent with 'hot-spots' and 'cold-spots' that suggest the presence of unique DNA structures that differently affect DNA damage and/or replication fidelity or may poorly tolerate mutagenesis due to their essential nature in genome replication.

A mutational gradient is conserved in human mtDNA

We next determined the evolutionary conservation of the patterns we observe in our mouse data. To do so, we made use of prior reported Duplex-Seq data sets for human mtDNA (11,29). As with the mouse data, we performed a binned mutation frequency analysis with bin size of 200bp

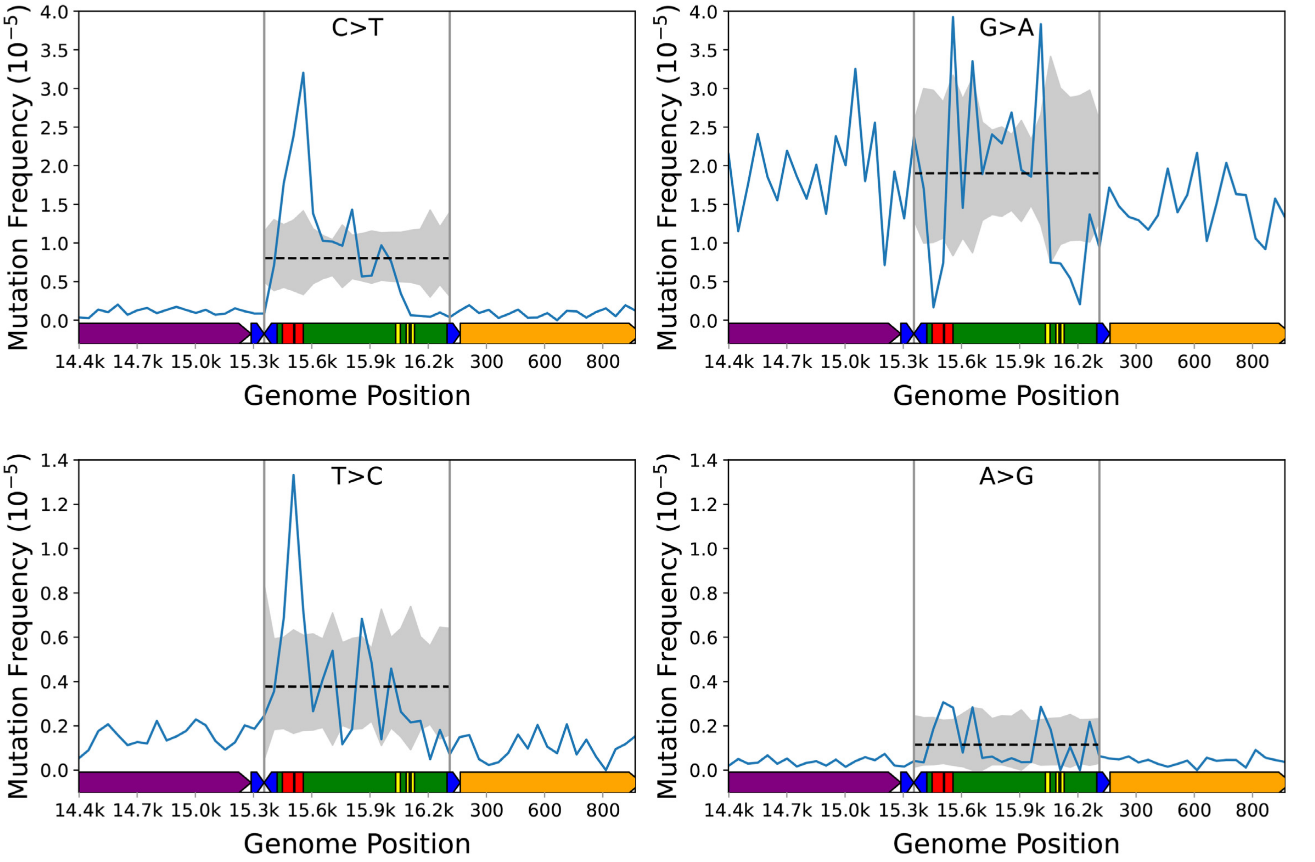

Figure 4. Mutations in the mtDNA control region display a non-uniform distribution and constraints at some loci. The observed per bin mutation frequency (*blue line*) and simulated distribution of data (*dotted black line* = mean, gray shading = 99.975% confidence interval) for C→T, G→A, T→C and A→G 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).

due to the reduced number of mutations compared to our mouse data. Consistent with our mouse data, we observe a gradient for both G→A and T→C mutations in the major arc (G→A: slope = $4.86 \pm 0.93 \times 10^{-7}$, $P = 1.93 \times 10^{-7}$; T→C: slope = $1.84 \pm 0.26 \times 10^{-7}$, $P = 1.69 \times 10^{-12}$) that is bounded by the Ori$_L$ and mCR (Figure 5A,B; Supplementary Table S5; Supplementary Data S6). Unlike the mouse data, the minor arc did not exhibit an apparent gradient and no other mutation types exhibited a significant increase in either the major or minor arcs (Supplementary Figure S7; Supplementary Table S5).

Our analysis points to a somatic mutational gradient as an evolutionarily conserved feature of vertebrate mtDNA. However, all our data were collected using Duplex-Seq, leaving open the possibility that the gradient pattern is an artifact of our Duplex-Seq protocol or our data analysis pipeline. While we consider this scenario unlikely, we sought to observe this gradient in an independently generated data using more conventional sequencing approaches. Somatic mtDNA mutations occur at very low frequencies (~10^{-6}–10^{-5}), making their detection with conventional sequencing difficult (51). To overcome this limitation, we analyzed mtDNA mutation call data published by the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium (37). This data set consists of 7,611 independent somatic variants (variant allele fraction (VAF) > 0.01; mean VAF = 0.2) from 2,536 tumors across 38 different cancer types. Because cancer is a clonal process arising from a single cell, the detected variants are largely a snapshot of the mtDNA mutations present early in tumor formation and have much higher VAFs than what is typically detected in Duplex-Seq data. Importantly for our purpose, this characteristic of the tumor data is expected to largely eliminate the potential confounder of low frequency artifacts giving rise the observed gradient.

We divided the genome into 100bp bins and, for each mutation type, calculated the mutation density (i.e. mean number of detected mutations per wild-type base) in each bin (Supplementary Data S7). Consistent with our Duplex-Seq data, we observe a clear gradient in both G→A and T→C transitions, but not their complement, that increases along the major arc (G→A $P = 7.43 \times 10^{-8}$; T→C $P = 1.16 \times 10^{-5}$) (Figure 5C, D; Supplementary Table S6). Both T→A and C→G transversions report a negative slope in the major arc and C→T and G→T exhibit a positive slope in the minor arc, but the magnitudes are extremely small and are likely a regression artifact. No other mutation types exhibit a gradient (Supplementary Figure S8; Supplementary Table S6). These data confirm both the presence of a mutational gradient in the major arc and that our results

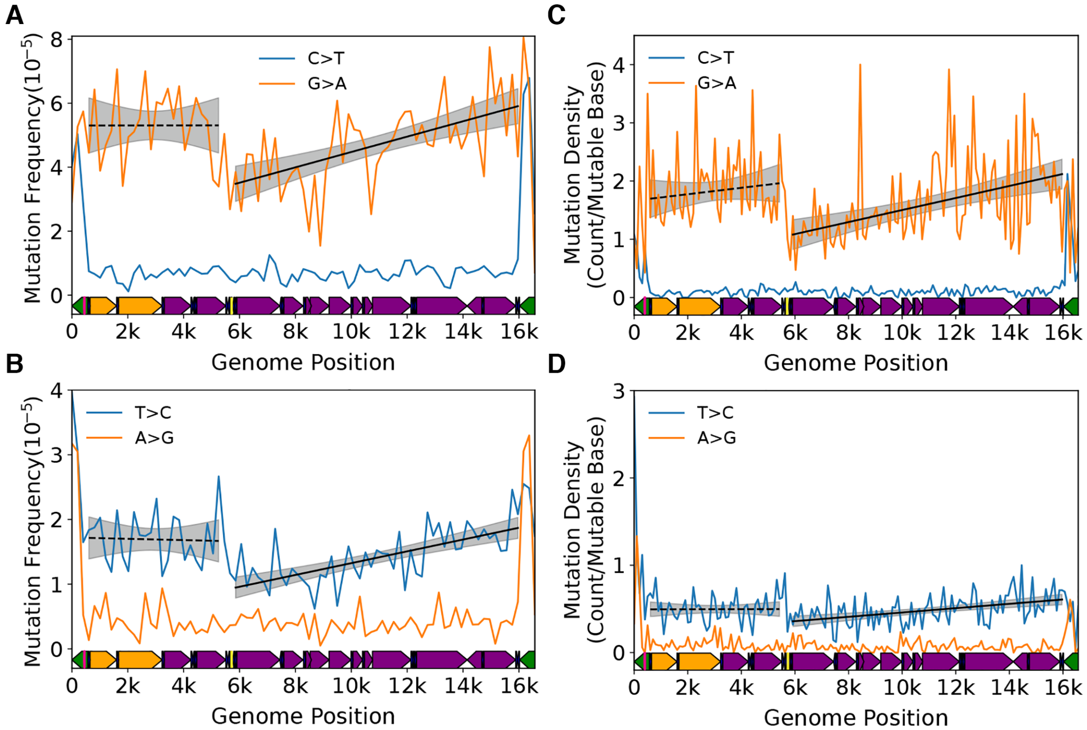

Figure 5. Somatic mutational gradient is conserved in human mtDNA. Plot and linear regression (*black line*) of reference strand (i.e. L-strand) (**A**) C→T and G→A and (**B**) T→C and A→G mutation frequencies as a function of genome position from prior published Duplex-Seq data. Each data point denotes a 200bp bin. Plot and linear regression (*black line*) of reference strand (i.e. L-strand) (**C**) C→T and G→A and (**D**) T→C and A→G mutation densities as a function of genome position in human tumor data from PCAWG dataset. Each data point denotes a 100 bp 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, *yellow* = $\mathrm{Ori_L}$, *magenta* = $\mathrm{Ori_H}$. *Solid black line* = linear regression slope $P < 0.05$; *dashed black line* = linear regression slope $P > 0.05$; *grey shading* = regression 95% confidence interval.

are unlikely to be due to an unknown issue with Duplex-Seq. Taken together, both our Duplex-Seq data and the PCAWG data recapitulate our findings in mouse mtDNA, pointing to the strong evolutionary conservation of G→A and T→C gradients among vertebrate species.

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

Previous work has noted similarities in the strand orientation and simple mutational spectra between somatic mtDNA mutations and population level SNPs, suggesting a similar causative driver of population level mtDNA sequence diversity (8,10,28). We sought to further explore this relationship by determining if the mutation gradient is reflected in the distribution of inherited single nucleotide variants, as would be expected if this process is active in the germline. We initially sought to test this hypothesis by mapping mutations obtained with Duplex-Seq of mouse oocytes (28), but the total number of mutations ($N = 691$) was insufficient to detect a gradient. We next evaluated the distribution of homoplasmic SNPs in the human mtDNA by downloading a recently published list of 44,494 SNPs obtained from MITOMAP and phylogenetically corrected such that each SNP was likely the result of an independent *de novo* event (38,52). Using the same binning approach as our human somatic data, we calculated the mutation density (i.e. number of *de novo* SNPs per mutable base) in each bin. For this analysis, we limited our analysis to the major arc due to (i) the absence of a clear minor arc gradient in our human somatic data and (ii) evidence of regions with an underrepresentation of SNPs in rDNA genes (39). Consistent with our somatic data, we observe a significant positive gradient in G→A and T→C (Figure 6A, B; Supplementary Figure 9; Supplementary Table S7). Notably, complement SNP types (C→T and A→G, respectively), as well as G→C SNP, show significant gradients, but the magnitude of their slope, especially relative to G→A SNPs, is substantially smaller. We sought to further validate this observation by performing this same analysis on a recently reported database of homoplasmic SNPs from 196,983 individuals (39). As with our initial data set, we observe a significant correlation between SNP density and genome position of G→A SNPs ($\rho = 0.26$; $P = 0.046$; Spearman correlation). We also observe a significant correlation between G→C SNP and genome position ($\rho = 0.307$; $P = 0.019$; Spearman correlation) similar to the MITOMAP based dataset. No other significant correlations were observed (Supplementary Table S8). Thus, we confirmed that, at the very least, a G→A gradient is present in human polymorphisms, consistent with our somatic mtDNA data, and further supports the idea that the mechanism of mutagenesis in the somatic tissue is likely the direct driver of human mtDNA variation.

The strong conservation of the somatic gradient between mice and humans and the presence of the gradient in human SNP data suggest that this unusual mutational pressure is

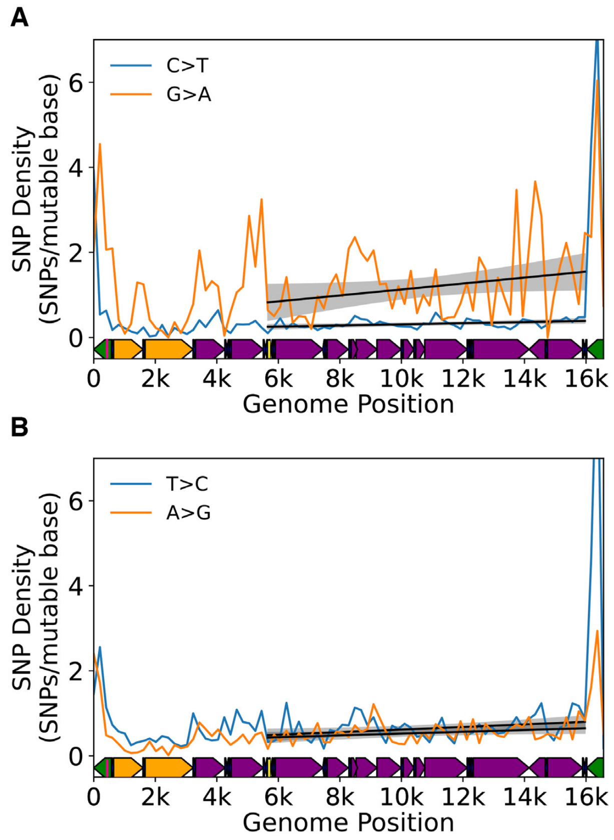

Figure 6. Mutational gradient is detected in major arc in human population SNPs. (**A**) Density of C→T and G→A SNPs on the L-strand. (**B**) Density of T→C and A→G SNPs on the L-strand. Data are from Gu *et al.* [40]. *Solid black line* = linear regression slope P < 0.05; *grey shading* = regression 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, *yellow* = Ori_L, *magenta* = Ori_H.

likely a major driver of sequence diversity across species. Our somatic data point to a sustained G→A and T→C mutational pressure of the L-strand with relatively little reversion. Over the long term, the L-strand is expected to exhibit a spatially dependent depletion of dG and dT bases along the major arc and a concomitant increase in dA and dC bases until some selective equilibrium is reached (Supplementary Figure S10A). Phylogenetic analyses on the sequence differences between related species such as primates has been shown to exhibit a gradient effect in T→C transitions (53). Analysis of a relatively small number of vertebrate species (N = 118) has also suggested that this phenomenon is likely a general aspect of vertebrate mtDNA biology (24).

We sought determine the generality of the gradient phenomenon by expanding these findings to include the significantly increased number of vertebrate mtDNA sequences now available (N = 5,326). Performing this analysis on all available mammalian mtDNA sequences in the NCBI RefSeq database (N = 1,266) shows that the majority of sequences exhibit a significant spatially dependent depletion of dG and dC (i.e. negative correlation coefficient) and a similar enrichment (i.e. positive correlation coefficient) in dC and dT (Figure 7A, B), confirming that this is a general phenomenon in mammalian mtDNA. While consistent with our hypothesis, the correlation (i) does not inform on the magnitude of the correlation and (ii) does not explicitly link the change in the abundance of one base type with another. Specifically, the magnitudes of the dG and dT composition slopes should be anti-correlated with the respective dA and dC slope magnitudes within the same species. As can be seen in Figure 7C, D, with a few exceptions, the slopes of dG and dA content, as well as dT and dC content, are strongly anti-correlated (dG/dA Spearman's $\rho = -0.45$, $P = 1.9 \times 10^{-64}$; dT/dC Spearman's $\rho = -0.57$, $P = 6.3 \times 10^{-111}$) across currently available mammalian mtDNA sequences with the direction of the anti-correlation consistent with a graduated G→A and T→C mutation pressure. We next extended this approach to other vertebrate classes, including *Archelosauria* (birds and crocodilians) (N = 1,015), *Lepidosauria* (lizards and snakes; N = 247), and *Actinopterygii* (ray-finned fish; N = 2,798). We did not evaluate non-vertebrate mtDNA sequences due to higher levels of structural heterogeneity and gene composition in these phyla. Like mammals, the majority of species within each vertebrate class show significant gradients in mtDNA composition that are strongly anti-correlated in their dG/dA content, as well as dT/dC content, indicating that this graduated mutation pressure is highly conserved across widely divergent species that inhabit significantly different ecological niches and are subjected to very different selective pressures (Supplementary Figures S11 and S12). Interestingly, several species strongly deviate in either gradient direction and/or correlation strength, suggesting that these species are subject to different selective pressures on their mtDNA (Supplementary Figures S11 and S12). Taken together, our data point to the mutational process driving the accumulation mutations in somatic tissues being the likely mechanistic driver of population level polymorphisms and sequence composition in vertebrates.

DISCUSSION

The advent of ultra-high accuracy sequencing methodologies have opened up the possibility of studying the mutagenic processes in mtDNA in greater detail. Both we and others have used Duplex-Seq, a method with an error background of $<1 \times 10^{-7}$, to study somatic mtDNA mutations (8,11,28–30,54,55). These studies have broadly shown that mutations are heavily weighted towards G→A/C→T and T→C/A→G transitions with very low levels of transversions, including the canonical ROS-associated G→T/C→A mutations. In addition, these studies have shown a strong strand bias, with G→A/C→T mutation being more prevalent when the dG base is in the L-strand. A notable difference in the mutational frequency and spectrum in the mCR is also reported. While these studies have provided a broad understanding of mtDNA mutagenesis, the very low frequency of mutations ($<1 \times 10^{-5}$) means that, for any given sample, only a few dozen to a few hundred mutations are typically detected, leaving conclusions about how these mutations are distributed unclear beyond regional differences (*i.e.* mCR versus coding or between genes). In this study, we aggregated several pre-

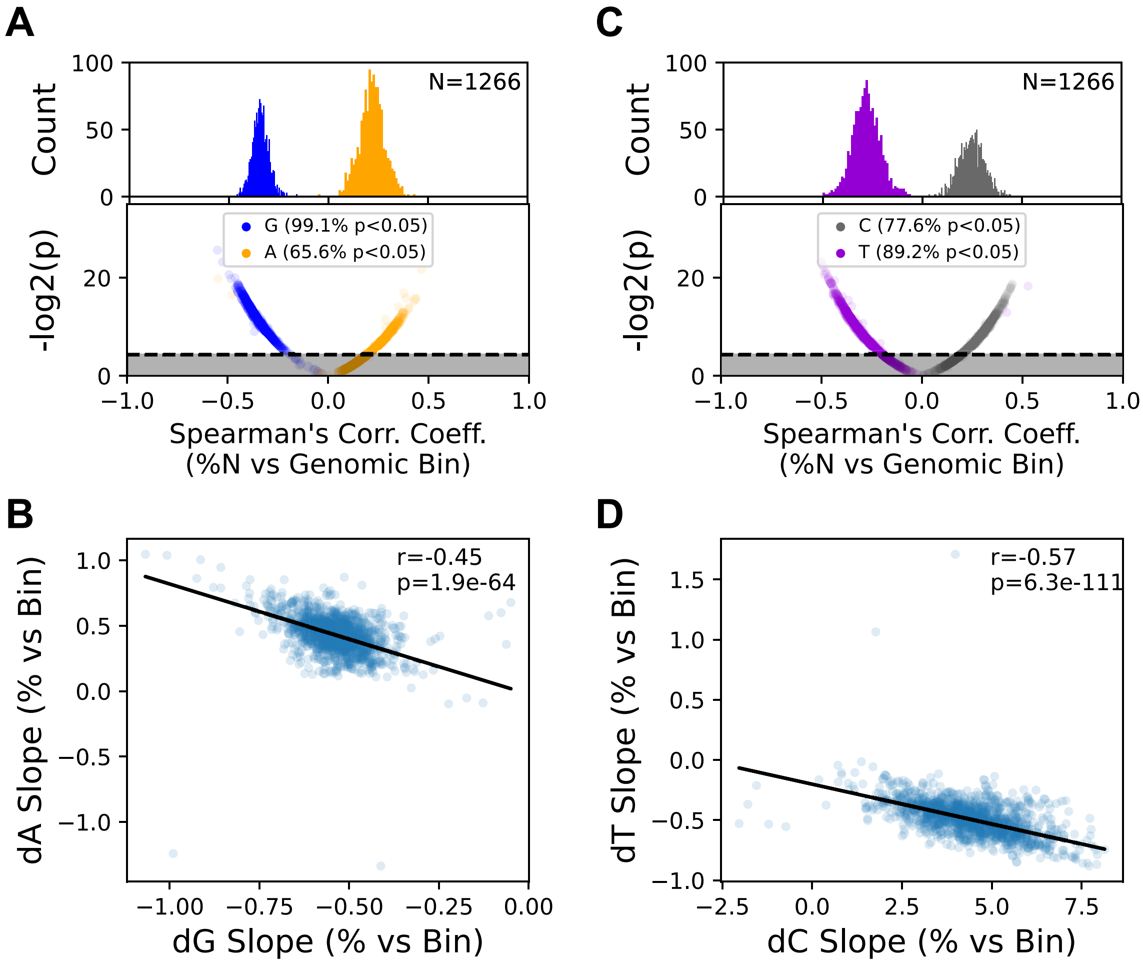

Figure 7. Genome composition bias mirrors the somatic gradient in mammals. (**A**) Most mammalian genomes show a statistically significant spatially dependent depletion of dG and enrichment of dA in the major arc. (**B**) The strength of dG depletion is directly proportional to dA enrichment in a species dependent manner. (**C**) and (**D**) dT and dC show a similar anti-correlative pattern as dG and dA.

existing Duplex-Seq data sets to better asses the distribution of mutations across the mtDNA molecule at significantly higher resolution than what has been previously reported.

Distribution of point mutations supports the strand displacement mode of mtDNA replication

In addition to recapitulating previous findings showing a strong bias towards transition mutations over transversions and a higher mutation load in the mCR, our analysis shows a strikingly non-uniform gradational distribution of G→A and T→C transitions, but not their complement, along the coding portion of the mtDNA. The totality of our data is most consistent with an asynchronous strand displacement mechanism with a long lived, deamination prone, single-stranded H-strand. A key aspect of our data that supports this hypothesis is the increased slope of G→A and T→C mutations with advancing age. Any alternative replication model without a ssDNA intermediate would need to account for how deamination-linked mutations could increase at a higher rate as a function of genome position over time. The RITOL and strand-synchronous models lacks any substantial ssDNA, with >80% of the displaced H-strand estimated to be annealed with RNA in the RITOL model (14,15). Comparison of Pol-γ and Pol-γ^{exo-} data suggest that the gradient is not due to a simple interaction between mtDNA base composition and polymerase base selectivity. Instead, the loss of the gradient and strand asymmetry in the absence of proofreading points to Pol-γ fidelity not being strongly affected by the underlying genome composition and indicates that point mutations are not arising simply due to base discrimination errors escaping proofreading. Given that the exonuclease domain is effectively blind to dU:dA base-pairs and that the ssDNA is highly prone to deamination, our data are most consistent with a model where most mutations arise from deaminated bases that are a consequence of a long-lived single-stranded replication intermediate.

We note that Holt *et al.* have reported that the synchronous and asynchronous mechanisms can exist simultaneously, with the balance between these two mechanisms the result of the cell's physiological state (16). While our data do not support a classic leading/lagging strand mechanism, they do not entirely refute its existence in all cases. Leading/lagging strand synthesis may be part of a stress response pathway to quickly reestablish copy number levels. In support of this possibility, withdrawing mtDNA depleting ethidium bromide from cells results in a burst of mtDNA synthesis with fully double-stranded replication intermediates, which is interpreted as being due to a leading/lagging strand replication fork (56). Consistent with this idea, modulating the level of the mitochondrial transcripts via changes in Twinkle helicase levels has been reported to switch between strand asynchronous and lead/lagging strand synthesis (19). However, our data are from tissues of unstressed wild-type animals without known perturbations to mtDNA gene expression, pointing to the asymmetric model being the predominant mtDNA synthesis mechanism under normal physiological conditions.

Alternative mechanisms for the mutational gradient

Several alternative models could account for the gradient phenomenon, including transcription-linked damage/mutagenesis, differential DNA repair, and somatic selection. Transcription has been linked to mutagenesis by several different mechanisms. In one scenario, the transcription bubble induces negative supercoiling of the DNA, resulting in transient strand-asymmetric single-stranded regions in the non-template strand that would be prone to damage in a manner similar to the asymmetric strand-displacement model (57). Two aspects argue against this possibility. First, genes are transcribed polycistronically such that all parts of the genome are transcribed by the same transcription complex. Consequently, no gene experiences differing effects arising from varying levels of transcription. Second, the direction of the mutational pressure (i.e. G→A on the L-strand and C→T on the H-strand) is in the opposite direction of what is predicted by the DNA strand from which the genes are transcribed. In mammalian DNA, all non-tRNA genes, except ND6, are templated by the C→T prone H-strand, thus the single-strand portion of the transcription bubble would be the opposite L-strand. Transcription of the L-strand could result in elevated C→T mutations on the H-strand, but would require a significant asymmetry in the frequency of transcription between the two promoters in order to induce such a large difference in mutagenesis between the two strands. A more plausible transcription-linked scenario, replication/transcription conflict, is well-described in bacteria to increase mutagenesis in a strand-oriented manner (58). Mitochondria have evolved to temporally separate these two processes, thereby reducing the potential for conflict. Specifically, mtDNA replication initiates at the Ori_H by priming off the L-strand transcript originating from the L-strand promoter. Thus, genome replication precludes transcription of the L-strand and vice versa, with the balance between replication versus transcription being regulated by the binding of mitochondrial transcription elongation factor (TEFM) to conserved sequence block 2 in the mCR (59). Importantly, if replication does not initiate from the Ori_H, then the Ori_L is not thought to initiate counter directional L-strand replication, thereby reducing conflict between L-strand synthesis and a transcription bubble originating from the H-strand promoter. (Figure 1A). However, this leaves the possible scenario where transcription from the H-strand promoter initiates transcription subsequent to the firing of the Ori_L, which would result in conflict if not properly regulated. A premature transcription termination site located at the 16S rRNA proximal tRNA^Leu binds mitochondrial transcription termination factor 1 (MTERF1) largely prevents transcription beyond this point. Biochemical analysis of the interaction of MTERF bound to tRNA^Leu sequence and the replisome has led to the model whereby an advancing replisome dissociates MTERF and the transcription complex, which allows for replication to complete its circumnavigation (60).

In addition to transcription, mutations could be induced by differences in DNA repair. Mitochondrial make use of several mechanisms that prevent the occurrence of mutations, including DNA proofreading during replication and base-excision repair (BER). We note that our analysis comparing proofreading proficient and deficient Pol-γ only directly addresses the contribution of base discrimination to the gradient and does not directly address the possibility that proofreading varies between the L- and H-strands, as well as linearly along the genome. While this remains a formal possibility, several reasons suggest that an exonuclease centric explanation is unlikely. First, the exonuclease domain is spatially distinct from the catalytic domain and makes no contact with the template DNA. As such, it is not 'aware' of the template sequence at which a mismatch has occurred. Given that misincorporations by the catalytic site are unbiased between the two strands and spatially along the genome, there is not a clear mechanism by which the exonuclease would change its proofreading activity in a spatially dependent manner. Secondly, only G→A and T→C L-strand mutations show a gradient, but the sequence contexts for all other mutation types (e.g. G→C, G→T) are the same, so the any bias would necessarily depend on both sequence context and the type of base mispair. In the case of BER, UNG1, the mitochondrial targeted isoform of uracil deglycosylase, is prevented from acting on single-stranded DNA by mtSSB, thereby preventing its repair until a double-stranded state is reestablished (61). Interestingly, the occupancy of mtSSB on the H-strand reflects the same gradient pattern as our data, suggesting a plausible mechanism by which mtSSB prevents repair of dU in the H-strand during replication which results in dU:dA that is unrecognized by the Pol-γ exonuclease during counter-directional L-strand synthesis (62).

Lastly, selection of mutations in the soma is an increasingly documented phenomenon in nuclear genes (63). As such, there remains the possibility that differential selection of mutations in mtDNA (both positive or negative) could, in principle, give rise to the observed gradient. Several studies have noted likely positive selection of small numbers of genome positions as evidenced by the same variant being recurrently observed in specific tissues across many individuals (64,65). However, these positions appear to be very limited in number and are hypothesized to arise from differing bioenergetic requirements of the tissues. Ideally, our analysis would address this issue by testing for a gradient in four-fold degenerate sites, which are classically used to study mutational patterns under the assumption that they are not under selective pressure. Unfortunately, this approach is infeasible in mouse mtDNA due only 71 out of 16,299 sites being four-fold degenerate L-strand deoxyguanosine, not all of which are seen to be mutated in our data set. However, several factors suggest that selection is unlikely to be the driver of the gradient pattern. First, our data makes use of data from eight different tissues (Supplementary Table S2), none of which are solely driving the gradient, suggesting that selection would need to be near universal across the assayed tissues and with similar strength. In addition, the gradient is the same between mice and humans, which have different genome sequences. Mitophagy, the autophagic removal of mitochondria, has been previously hypothesized to target mitochondria harboring mutations that cause dysfunction (66). Unequal selection of genes that happen to line up with their organization in the mtDNA could lead to over- or under-representation of mutations in those genes over time. However, a number of studies have explicitly tested for this hypothesis and found no evidence of positive or negative selection of point mutations (10,30,37). In sum, while our data does not conclusively eliminate the possibility of alternative models for the induction of mtDNA mutations, the underlying biology of mitochondria and the mtDNA suggests that, with the exception of strand-asymmetric BER, these are less parsimonious explanations for the observed mutational gradient when compared to replication-linked deamination events.

Linking of somatic mutational processes to population and evolutionary patterns in mtDNA

A lingering question in the field of mtDNA replication concerns the conservation of the mtDNA replication mechanism across taxa. The mitochondrial genome exhibits a wide range of sizes, structures, and noncoding regulatory regions between phyla and kingdoms, suggesting that different replication mechanisms were retained or acquired since the initial endosymbiosis event that gave rise to mitochondria. For example, while vertebrates make use of a relatively compact genome structure and mCR with an initiating origin and distal counter-directional origin, invertebrates tend to have a much more diverse repertoire of genome topologies and replication mechanisms (67). Meanwhile, plants utilize a substantially larger mtDNA structure that likely uses an entirely different recombination-dependent and/or rolling circle mechanism without clearly defined replication origins (68). Our data indicate that mapping of somatic mutations provides an alternative approach to mapping origins of replication and other potential regulatory structures that is free of the complications inherent to interpreting 2D-gels and electron micrographs that would likely perform poorly on these more complicated genomes. Indeed, an analogous strategy has been used to map origins of replication in the human genome by taking advantage of ultra-mutated tumors (69).

We can clearly discern the location of the $\mathrm{Ori_L}$ in both mouse and human data sets. These data also argue against the proposed use of other tRNAs as L-strand priming sites, as well as a large 'initiation zone' for replication, as these models predict either multiple discontinuous gradients or lack a gradient entirely. Instead, our data are consistent with a single L-strand origin in mammalian mtDNA. Moreover, with our high-density mouse data set, we mapped areas of mutation over- and under-abundance in the mCR that correspond to sequence blocks essential for mtDNA H-strand replication. Significant deviations are not obvious in the regions flanking the mCR other than the $\mathrm{Ori_L}$, as would be expected if other sequences in these areas were essential for intermittent priming. Notably, avian mtDNA lacks the predicted stem-loop structure of the mammalian $\mathrm{Ori_L}$ and 2D-gels point to initiation sites across the entirety of the mtDNA, providing a potential model system to further investigate these alternative origin models in vertebrates (18). In line with this idea, we attempted to analyze Duplex-Seq data for similar patterns in non-vertebrate organisms, *D. melanogaster* (54,70) and *A. thaliana* (55), but the number of mutations were too low and the density too sparse to observe a clear signal, leaving this for future work. However, analysis of the genome composition of these two species shows a different gradient in dG composition for *D. melanogaster* and no apparent gradient in *A. thaliana*, suggesting that these organisms likely make use of alternative mechanisms of replication, as has been previously hypothesized (Supplementary Figure S10B, C) (68,71).

Human population studies have previously identified a bias in the occurrence of G→A and T→C SNPs of the L-strand, as has comparison of human mtDNA sequences with those of evolutionary related species (72,73). We previously noted that this bias mirrors the strand asymmetry seen in somatic mutagenesis of mtDNA through a process that is continuous throughout life and hypothesized that this pattern was consistent with mtDNA replication via an asymmetric model (8). Our data extend this observation to also include a gradient in SNP distribution along the genome, as well as genome base composition, further strengthening the link between somatic and germline processes. Our analysis in genome composition point to this gradient being largely conserved in vertebrates. The strength of this gradient, as highlighted by the variation in anticorrelation between complementary bases and the presence of species that deviate significantly from the trend line, can vary significantly. An important aspect of these observations is that they provide a feasible opportunity to mechanistically study the processes that give rise to genetic variation at the population and taxonomic levels. This is especially pertinent in species with very high or very low mutation rates or show unusual biases in genetic variability. Indeed, recent work in *A. thaliana* linked mismatch repair with the low mutation rate seen in this species, and three species of angiosperm genus *Silene* with notably different mutation rates showed corresponding patterns in somatic mtDNA mutations (55,74), lending credence to the idea that studying somatic mutagenic processes can inform on evolutionary and population level patterns. A systematic analysis of somatic mutations in species with unusual genome composition or SNP patterns may provide insight into how and why these species deviate so significantly from related species and clues to their natural history.

The growing quantities of high-accuracy sequencing data generated from such technologies as Duplex Sequencing has provided the ability to elucidate several mutagenic patterns and biases previously unobserved at the somatic level. Taken together, these patterns argue for an asymmetric strand-displacement replication model, as originally posited by Clayton and colleagues (12,13) and against a more conventional leading/lagging strand replication fork. Furthermore, our data point to an intimate link between replication, DNA damage, and mutagenesis whereby the majority of mutations are the consequence of deamination events on a single-stranded replication intermediate that are fixed during the completion of replication. Ultimately, this replication-linked process affects genetic variation and genome composition.

DATA AVAILABILITY

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 accension PRJNA449763 and SRA accension PRJNA237667. Only normal control samples were analyzed.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Dr Scott Lujan for providing the Circos templates used as the basis for Figure 1A.

FUNDING

Congressionally Directed Medical Research Programs [W81XWH-16-1-0579, W81XWH-18-1-0339]; National Human Genome Research Institute [R21 HG011229]; National Institute on Aging [K01 AG062757-01, P01 AG001751, T32AG000057, T32AG066574]; National Institute of Diabetes and Digestive and Kidney Diseases [R21DK128540]. Funding for open access charge: National Institute on Aging [P01 AG001751]; National Institute of Diabetes and Digestive and Kidney Diseases [R21DK128540]; National Human Genome Research Institute [R21 HG011229].

Conflict of interest statement. S.R.K. 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.

REFERENCES

1. Lujan,S.A., Longley,M.J., Humble,M.H., Lavender,C.A., Burkholder,A., Blakely,E.L., Alston,C.L., Gorman,G.S., Turnbull,D.M., McFarland,R. *et al.* (2020) Ultrasensitive deletion detection links mitochondrial DNA replication, disease, and aging. *Genome Biol.*, **21**, 248.
2. Marcelino,L.A. and Thilly,W.G. (1999) Mitochondrial mutagenesis in human cells and tissues. *Mutat. Res.*, **434**, 177–203.
3. Taylor,R.W. and Turnbull,D.M. (2005) Mitochondrial DNA mutations in human disease. *Nat. Rev. Genet.*, **6**, 389–402.
4. Sevini,F., Giuliani,C., Vianello,D., Giampieri,E., Santoro,A., Biondi,F., Garagnani,P., Passarino,G., Luiselli,D., Capri,M. *et al.* (2014) mtDNA mutations in human aging and longevity: Controversies and new perspectives opened by high-throughput technologies. *Exp. Gerontol.*, **56**, 234–244.
5. Chocron,E.S., Munkácsy,E. and Pickering,A.M. (2019) Cause or casualty: the role of mitochondrial DNA in aging and age-associated disease. *Biochim. Biophys. Acta Mol. Basis Dis.*, **1865**, 285–297.
6. Cheng,K.C., Kasais,H., Nishimuras,S. and Loeb,L.A. (1992) 8-hydroxyguanine, an abundant form of oxidative DNA damage, causes G→T and A→C substitutions. *J. Biol. Chem.*, **267**, 166–172.
7. Harman,D. (1956) Aging: A theory based on free radical and radiation chemistry. *J. Gerontol.*, **11**, 298–300.
8. Kennedy,S.R., Salk,J.J., Schmitt,M.W. and Loeb,L.A. (2013) Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. *PLoS Genet.*, **9**, e1003794.
9. Williams,S.L., Mash,D.C., Züchner,S. and Moraes,C.T. (2013) Somatic mtDNA mutation spectra in the aging human putamen. *PLoS Genet.*, **9**, e1003990.
10. Ju,Y.S., Alexandrov,L.B., Gerstung,M., Martincorena,I., Nik-Zainal,S., Ramakrishna,M., Davies,H.R., Papaemmanuil,E., Gundem,G., Shlien,A. *et al.* (2014) Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *Elife*, **3**, e02935.
11. Baker,K.T., Nachmanson,D., Kumar,S., Emond,M.J., Ussakli,C., Brentnall,T.A., Kennedy,S.R. and Risques,R.A. (2019) Mitochondrial DNA mutations are associated with ulcerative colitis preneoplasia but tend to be negatively selected in cancer. *Mol. Cancer Res.*, **17**, 488–498.
12. Brown,T.A. (2005) Replication of mitochondrial DNA occurs by strand displacement with alternative light-strand origins, not via a strand-coupled mechanism. *Genes Dev.*, **19**, 2466–2476.
13. Clayton,D.A. (1982) Replication of animal mitochondrial DNA. *Cell*, **28**, 693–705.
14. Reyes,A., Kazak,L., Wood,S.R., Yasukawa,T., Jacobs,H.T. and Holt,I.J. (2013) Mitochondrial DNA replication proceeds via a 'bootlace' mechanism involving the incorporation of processed transcripts. *Nucleic Acids Res.*, **41**, 5837–5850.
15. Yasukawa,T., Reyes,A., Cluett,T.J., Yang,M.-Y., Bowmaker,M., Jacobs,H.T. and Holt,I.J. (2006) Replication of vertebrate mitochondrial DNA entails transient ribonucleotide incorporation throughout the lagging strand. *EMBO J.*, **25**, 5358–5371.
16. Holt,I.J., Lorimer,H.E. and Jacobs,H.T. (2000) Coupled leading- and lagging-strand synthesis of mammalian mitochondrial DNA. *Cell*, **100**, 515–524.
17. Bowmaker,M., Yang,M.Y., Yasukawa,T., Reyes,A., Jacobs,H.T., Huberman,J.A. and Holt,I.J. (2003) Mammalian mitochondrial DNA replicates bidirectionally from an initiation zone. *J. Biol. Chem.*, **278**, 50961–50969.
18. Reyes,A., Yang,M.Y., Bowmaker,M. and Holt,I.J. (2005) Bidirectional replication initiates at sites throughout the mitochondrial genome of birds. *J. Biol. Chem.*, **280**, 3242–3250.
19. Cluett,T.J., Akman,G., Reyes,A., Kazak,L., Mitchell,A., Wood,S.R., Spinazzola,A., Spelbrink,J.N. and Holt,I.J. (2018) Transcript availability dictates the balance between strand-asynchronous and strand-coupled mitochondrial DNA replication. *Nucleic Acids Res.*, **46**, 10771–10781.
20. Seligmann,H. (2010) Mitochondrial tRNAs as light strand replication origins: similarity between anticodon loops and the loop of the light strand replication origin predicts initiation of DNA replication. *Biosystems*, **99**, 85–93.
21. Seligmann,H., Krishnan,N.M. and Rao,B.J. (2006) Possible multiple origins of replication in primate mitochondria: alternative role of tRNA sequences. *J. Theor. Biol.*, **241**, 321–332.
22. Asakawa,S., Kumazawa,Y., Araki,T., Himeno,H., Miura,K. and Watanabe,K. (1991) Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. *J. Mol. Evol.*, **32**, 511–520.
23. Reyes,A., Gissi,C., Pesole,G. and Saccone,C. (1998) Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol. Biol. Evol.*, **15**, 957–966.

24. Faith,J.J. and Pollock,D.D. (2003) Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics*, **165**, 735–745.
25. Frederico,L.A., Kunkel,T.A. and Shaw,B.R. (1990) A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry*, **29**, 2532–2537.
26. Lujan,S.A., Williams,J.S., Pursell,Z.F., Abdulovic-Cui,A.A., Clark,A.B., Nick McElhinny,S.A. and Kunkel,T.A. (2012) Mismatch repair balances leading and lagging strand DNA replication fidelity. *PLos Genet.*, **8**, e1003016.
27. Maslowska,K.H., Makiela-Dzbenska,K., Mo,J.-Y., Fijalkowska,I.J. and Schaaper,R.M. (2018) High-accuracy lagging-strand DNA replication mediated by DNA polymerase dissociation. *Proc. Natl Acad. Sci. U.S.A.*, **115**, 4212–4217.
28. Arbeithuber,B., Hester,J., Cremona,M.A., Stoler,N., Zaidi,A., Higgins,B., Anthony,K., Chiaromonte,F., Diaz,F.J. and Makova,K.D. (2020) Age-related accumulation of de novo mitochondrial mutations in mammalian oocytes and somatic tissues. *PLoS Biol.*, **18**, e3000745.
29. Hoekstra,J.G., Hipp,M.J., Montine,T.J. and Kennedy,S.R. (2016) Mitochondrial DNA mutations increase in early stage Alzheimer disease and are inconsistent with oxidative damage. *Ann. Neurol.*, **80**, 301–306.
30. Pickrell,A.M., Huang,C.-H., Kennedy,S.R., Ordureau,A., Sideris,D.P., Hoekstra,J.G., Harper,J.W. and Youle,R.J. (2015) Endogenous Parkin preserves dopaminergic substantia nigral neurons following mitochondrial DNA mutagenic stress. *Neuron*, **87**, 371–381.
31. Kennedy,S.R., Schmitt,M.W., Fox,E.J., Kohrn,B.F., Salk,J.J., Ahn,E.H., Prindle,M.J., Kuong,K.J., Shen,J.-C., Risques,R.-A. *et al.* (2014) Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.*, **9**, 2586–2606.
32. Stoler,N., Arbeithuber,B., Guiblet,W., Makova,K.D. and Nekrutenko,A. (2016) Streamlined analysis of duplex sequencing data with Du Novo. *Genome Biol.*, **17**, 180.
33. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
34. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 3.
35. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000,Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
36. Calabrese,F.M., Simone,D. and Attimonelli,M. (2012) Primates and mouse NumtS in the UCSC Genome Browser. *BMC Bioinformatics*, **13**, S15.
37. Yuan,Y., Ju,Y.S., Kim,Y., Li,J., Wang,Y., Yoon,C.J., Yang,Y., Martincorena,I., Creighton,C.J., Weinstein,J.N. *et al.* (2020) Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nat. Genet.*, **52**, 342–352.
38. Gu,X., Kang,X. and Liu,J. (2019) Mutation signatures in germline mitochondrial genome provide insights into human mitochondrial evolution and disease. *Hum. Genet.*, **138**, 613–624.
39. Bolze,A., Mendez,F., White,S., Tanudjaja,F., Isaksson,M., Jiang,R., Rossi,A.D., Cirulli,E.T., Rashkin,M., Metcalf,W.J. *et al.* (2020) A catalog of homoplasmic and heteroplasmic mitochondrial DNA variants in humans genetics. bioRxiv doi: https://doi.org/10.1101/798264, 26 June 2020, preprint: not peer reviewed.
40. Contributors,SciPy 1.0, Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272.
41. Gates,K.S. (2009) An overview of chemical processes that damage cellular DNA: spontaneous hydrolysis, alkylation, and reactions with radicals. *Chem. Res. Toxicol.*, **22**, 1747–1760.
42. Bessman,M.J., Lehman,I.R., Adler,J., Zimmerman,S.B., Simms,E.S. and Kornberg,A. (1958) Enzymatic synthesis of deoxyribonucleic acid. III. The incorporation of pyrimidine and purine analogues into deoxyribonucleic acid. *Proc. Natl Acad. Sci. U.S.A.*, **44**, 633–640.
43. Warner,H.R., Duncan,B.K., Garrett,C. and Neuhard,J. (1981) Synthesis and metabolism of uracil-containing deoxyribonucleic acid in *Escherichia coli*. *J. Bacteriol.*, **145**, 687–695.
44. Lee,H.R. and Johnson,K.A. (2006) Fidelity of the human mitochondrial DNA polymerase. *J. Biol. Chem.*, **281**, 36236–36240.
45. Longley,M.J., Nguyen,D., Kunkel,T.A. and Copeland,W.C. (2001) The fidelity of human DNA polymerase γ with and without exonucleolytic proofreading and the p55 accessory subunit. *J. Biol. Chem.*, **276**, 38555–38562.
46. Lim,S.E., Longley,M.J. and Copeland,W.C. (1999) The mitochondrial p55 accessory subunit of human DNA polymerase γ enhances DNA binding, promotes processive DNA synthesis, and vonfers N-ethylmaleimide resistance. *J. Biol. Chem.*, **274**, 38197–38203.
47. Williams,S.L., Huang,J., Edwards,Y.J.K., Ulloa,R.H., Dillon,L.M., Prolla,T.A., Vance,J.M., Moraes,C.T. and Züchner,S. (2010) The mtDNA mutation spectrum of the progeroid *Polg* mutator mouse includes abundant control region multimers. *Cell Metab.*, **12**, 675–682.
48. Gustafsson,C.M., Falkenberg,M. and Larsson,N.-G. (2016) Maintenance and expression of mammalian mitochondrial DNA. *Annu. Rev. Biochem.*, **85**, 133–160.
49. Wei,W., Tuna,S., Keogh,M.J., Smith,K.R., Aitman,T.J., Beales,P.L., Bennett,D.L., Gale,D.P., Bitner-Glindzicz,M.A.K., Black,G.C. *et al.* (2019) Germline selection shapes human mitochondrial DNA diversity. *Science*, **364**, eaau6520.
50. Sbisà,E., Tanzariello,F., Reyes,A., Pesole,G. and Saccone,C. (1997) Mammalian mitochondrial D-loop region structural analysis: identification of new conserved sequences and their functional and evolutionary implications. *Gene*, **205**, 125–140.
51. Salk,J.J., Schmitt,M.W. and Loeb,L.A. (2018) Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat. Rev. Genet.*, **19**, 269–285.
52. Brandon,M.C. (2004) MITOMAP: A human mitochondrial genome database-2004 update. *Nucleic Acids Res.*, **33**, D611–D613.
53. Raina,S.Z., Faith,J.J., Disotell,T.R., Seligmann,H., Stewart,C.-B. and Pollock,D.D. (2005) Evolution of base-substitution gradients in primate mitochondrial genomes. *Genome Res.*, **15**, 665–673.
54. Samstag,C.L., Hoekstra,J.G., Huang,C.-H., Chaisson,M.J., Youle,R.J., Kennedy,S.R. and Pallanck,L.J. (2018) Deleterious mitochondrial DNA point mutations are overrepresented in Drosophila expressing a proofreading-defective DNA polymerase γ. *PLoS Genet.*, **14**, e1007805.
55. Wu,Z., Waneka,G., Broz,A.K., King,C.R. and Sloan,D.B. (2020) *MSH1* is required for maintenance of the low mutation rates in plant mitochondrial and plastid genomes. *Proc. Natl Acad. Sci. U.S.A.*, **117**, 16448–16455.
56. Yasukawa,T., Yang,M.-Y., Jacobs,H.T. and Holt,I.J. (2005) A Bidirectional origin of replication maps to the major noncoding region of human mitochondrial DNA. *Mol. Cell*, **18**, 651–662.
57. Ma,J. and Wang,M.D. (2016) DNA supercoiling during transcription. *Biophys. Rev.*, **8**, 75–87.
58. Lang,K.S. and Merrikh,H. (2018) The Clash of macromolecular titans: replication-transcription conflicts in bacteria. *Annu. Rev. Microbiol.*, **72**, 71–88.
59. Agaronyan,K., Morozov,Y.I., Anikin,M. and Temiakov,D. (2015) Replication-transcription switch in human mitochondria. *Science*, **347**, 548–551.
60. Shi,Y., Posse,V., Zhu,X., Hyvärinen,A.K., Jacobs,H.T., Falkenberg,M. and Gustafsson,C.M. (2016) Mitochondrial transcription termination factor 1 directs polar replication fork pausing. *Nucleic Acids Res.*, **44**, 5732–5742.
61. Steen,K.W., Doseth,B., P. Westbye,M., Akbari,M., Kang,D., Falkenberg,M. and Slupphaug,G. (2012) mtSSB may sequester UNG1 at mitochondrial ssDNA and delay uracil processing until the dsDNA conformation is restored. *DNA Repair (Amst.)*, **11**, 82–91.
62. Fusté,J., Shi,Y., Wanrooij,S., Zhu,X., Jemt,E., Persson,Ö., Sabouri,N., Gustafsson,C.M. and Falkenberg,M. (2014) *In vivo* occupancy of mitochondrial single-stranded DNA binding protein supports the strand displacement mode of DNA replication. *PLoS Genet.*, **10**, e1004832.
63. Martincorena,I., Raine,K.M., Gerstung,M., Dawson,K.J., Haase,K., Van Loo,P., Davies,H., Stratton,M.R. and Campbell,P.J. (2017) Universal patterns of selection in cancer and somatic tissues. *Cell*, **171**, 1029–1041.
64. Samuels,D.C., Li,C., Li,B., Song,Z., Torstenson,E., Boyd Clay,H., Rokas,A., Thornton-Wells,T.A., Moore,J.H., Hughes,T.M. *et al.* (2013) Recurrent tissue-specific mtDNA mutations are common in humans. *PLoS Genet.*, **9**, e1003929.

65. Li,M., Schröder,R., Ni,S., Madea,B. and Stoneking,M. (2015) Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations. *Proc. Natl Acad. Sci. U.S.A.*, **112**, 2491–2496.
66. Kim,I., Rodriguez-Enriquez,S. and Lemasters,J.J. (2007) Selective degradation of mitochondria by mitophagy. *Arch. Biochem. Biophys.*, **462**, 245–253.
67. Lavrov,D.V. and Pett,W. (2016) Animal mitochondrial DNA as we do not know it: mt-genome organization and evolution in nonbilaterian lineages. *Genome Biol. Evol.*, **8**, 2896–2913.
68. Cupp,J.D. and Nielsen,B.L. (2014) Minireview: DNA replication in plant mitochondria. *Mitochondrion*, **19**, 231–237.
69. Shinbrot,E., Henninger,E.E., Weinhold,N., Covington,K.R., Göksenin,A.Y., Schultz,N., Chao,H., Doddapaneni,H., Muzny,D.M., Gibbs,R.A. *et al.* (2014) Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome Res.*, **24**, 1740–1750.
70. Andreazza,S., Samstag,C.L., Sanchez-Martinez,A., Fernandez-Vizarra,E., Gomez-Duran,A., Lee,J.J., Tufi,R., Hipp,M.J., Schmidt,E.K., Nicholls,T.J. *et al.* (2019) Mitochondrially-targeted APOBEC1 is a potent mtDNA mutator affecting mitochondrial function and organismal fitness in Drosophila. *Nat. Commun.*, **10**, 3280.
71. Jõers,P. and Jacobs,H.T. (2013) Analysis of replication intermediates indicates that *Drosophila melanogaster* mitochondrial DNA replicates by a strand-coupled theta mechanism. *PLoS One*, **8**, e53249.
72. Tanaka,M. and Ozawa,T. (1994) Strand asymmetry in human mitochondrial DNA mutations. *Genomics*, **22**, 327–335.
73. Belle,E.M.S., Piganeau,G., Gardner,M. and Eyre-Walker,A. (2005) An investigation of the variation in the transition bias among various animal mitochondrial DNA. *Gene*, **355**, 58–66.
74. Broz,A.K., Waneka,G., Wu,Z., Gyorfy,M.F. and Sloan,D.B. (2021) Detecting de novo mitochondrial mutations in angiosperms with highly divergent evolutionary rates. *Genetics*, **218**, iyab039.