Unravelling the Mitochondrial Mutational Landscape in Chordates: damage-induced versus replication-induced signatures, their Etiologies, and dynamics

Dmitrii Iliushchenko[1], Bogdan Efimenko[1,5], Alina G. Mikhailova[1], Victor Shamanskiy[1,5], Murat K. Saparbaev[7], Ilya Mazunin[4], Dmitrii Knorre[8], Wolfram S. Kunz[6], Stepan Denisov[3], Konstantin Khrapko[9], Jacques Fellay[2], Konstantin Gunbin[1,5], Konstantin Popadin[1,2]

[1] Center for Mitochondrial Functional Genomics, Immanuel Kant Baltic Federal University, Kaliningrad, Russian Federation
[2] School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[3] School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, United Kingdom
[4] Center for Molecular and Cellular Biology, Skolkovo Institute of Science and Technology, Moscow, Russian Federation
[5] A.A. Kharkevich Institute for Information Transmission Problems, Moscow, Russian Federation
[6] Department of Epileptology, University Bonn Medical Center, Bonn, Germany
[7] Groupe «Mechanisms of DNA Repair and Carcinogenesis», Equipe Labellisée LIGUE 2016, CNRS UMR9019, Université Paris-Saclay, Gustave Roussy Cancer Campus, F-94805 Villejuif, France
[8] Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russian Federation
[9] Northeastern University, Boston, MA, USA

Abstract

To elucidate the primary factors shaping mitochondrial DNA (mtDNA) mutagenesis, we reconstructed a deep 192-component mtDNA mutational spectrum using 129,100 polymorphic synonymous mutations from the Cytb gene of 1040 chordate species. We then deconvoluted this spectrum into three key signatures: (i) symmetrical mutations, predominantly C>T and rare transversions, linked to pol-γ's replication errors; (ii) asymmetrical C>T mutations, indicative of single-stranded DNA damage; and (iii) asymmetrical A>G mutations, also resulting from single-stranded DNA damage but particularly influenced by metabolic and age-specific mitochondrial environment. Interestingly, the two asymmetrical signatures collectively surpass pol-γ mutations, indicating that damage is the primary factor in mtDNA mutagenesis over replication. The diverse contribution of all three signatures across different species and human tissues provides insights into a range of physiological and metabolic traits. This mtDNA spectrum-aware methodology enhances phylogenetic inferences, deepens our understanding of species-specific mutational and selection processes, and offers insights into the dynamics of mutations in human mtDNA.

Introduction

DNA mutations can be a result of either replication or damage [1], with a range of mutagens responsible for these changes. Reconstruction of a deep mutational spectrum, usually with 96 components, helps to decompose it into distinct mutational components (signatures), allowing us to trace the effects of various mutagens[2]. In the human nuclear genome, both germline and somatic variants, including cancerous ones, have been instrumental in reconstructing and deconvoluting these spectra, leading to major breakthroughs in both fundamental [3] and applied [4] research.

Despite significant advancements in comprehending the mutagenesis of the human nuclear genome, the mitochondrial genome is less well-characterised, yet playing a vital role in numerous human diseases and ageing. The mutagenesis of the mitochondrial genome is mysterious: being a hundred times faster than in the nuclear genome, it remains partially elucidated, since its spectra show neither expected mutational signatures of ROS, UV light nor tobacco smoke in associated cancer data. Furthermore, the mechanisms of mtDNA replication and repair have been under ongoing debates. Detailed analysis of the mitochondrial mutational spectrum could offer key insights into these processes. On an evolutionary scale, understanding mtDNA mutagenesis is crucial for accurate phylogenetic inferences and for uncovering variations in selection processes between species[25] as well as for understanding the dynamics of deleterious human variants.

The reconstruction of a detailed mutational spectrum of single base substitutions for each species demands extensive data. Traditionally, comparative species analyses employ the transitions to transversions ratio as a basic yet insightful characteristic of the mutational spectrum[6]. Availability of more sequence data, allowed researchers to use more complex 6-component (focusing only on pyrimidines in Watson-Crick base pairs: C>A, C>G, C>T, T>A, T>C, T>G, under the assumption of symmetrical mutagenesis on complementary strands)[7–9] and 12-component (doubling the 6-component spectrum to account for asymmetrical mutagenesis on complementary strands, inasmuch as it includes complementary substitutions, for instance, C>A and G>T as distinct events)[5] spectra in analyses for comparative studies. However, more comprehensive 96-component (expanding the 6-component spectrum with an inclusion of the nucleotide context - 4 bases immediately 5’ and 4 bases immediately 3’ into spectra) and 192-component (a doubled version of the 96-component spectrum, assuming asymmetry between complementary strands) spectra, which consider adjacent nucleotide context, have been limited to extensively studied species such as humans and SARS-CoV-2. For rarely-sequenced, non-model taxa, constructing these in-depth spectra is challenging.

Here, focusing on chordates, we overcame this limitation by integrating rare, species-specific normalised mtDNA polymorphisms into a comprehensive spectrum representative of all Chordata. Analysing 129,100 synonymous mutations from CytB sequences of 1040 chordate species, we compiled a 192-component mtDNA spectrum. This extensive spectrum enables the exploration of critical questions: What is the nature of the majority of

mutations in mtDNA? What are the similarities in mtDNA mutational spectra between different classes of chordates? Is there a distinction between germ-line and somatic variants? What are the conservative and variable components of the spectrum, and why? Which mutagens are responsible for the observed patterns? How pronounced are the replication- and damage-induced components? Therefore, our study revealed that the primary factor responsible for mtDNA mutations is single-stranded DNA damage, which varies as a consequence of the metabolic rates of different taxa.

METHODS

Heavy strand notation is everywhere

1. Reconstruction of the integral 192-component mutational spectrum for CytB for all chordates.

To reconstruct the mutational spectrum for CytB, we made the assumption that all synonymous variants in chordate mitochondrial genomes are possibly neutral. Although potential non neutrality has not been conclusively proven [10] we conducted limited testing to assess the impact of eliminating highly-constrained synonymous positions. Preliminary results indicated that their exclusion did not significantly alter the outcomes (data not shown). However, future analyses may unveil the nature of these highly-constrained positions, whether they result from selection and should be excluded in reconstructing the neutral mutational spectrum or if they represent mutational cold spots, which can be tested in longer nucleotide contexts.

We followed next steps:

1a) Observed mutations:

We analysed 1040 chordate species with at least five unique CytB sequences from GenBank. The mutational spectrum was derived through the following steps: (i) obtaining a codon-based multiple alignment using macse v2[11]; (ii) reconstructing the phylogenetic tree and rooting it with the closest relative species as the outgroup; (iii) reconstructing ancestral sequences (most probable nucleotide in each alignment position) at each internal node using RAxML[12]; (iv) identifying all single-nucleotide synonymous substitutions on each tree branch; (v) categorising these mutations into 192 groups based on the three-nucleotide context (12 substitution types × 4 nucleotide types at the 3' position × 4 nucleotide types at the 5' position).

1b) Expected mutations:

To consider differences in nucleotide and trinucleotide composition in different taxa, we calculated the expected mutational spectra. Focusing on reference CytB genes obtained from GenBank for chordata species with known observed mutations we executed a *in silico* saturation mutagenesis procedure. The central concept behind *in silico* saturation mutagenesis is to replace each nucleotide with one of three possible alternatives and select only synonymous substitutions. These substitutions were then recorded alongside their neighbouring nucleotides, and categorised according to the specific type of substitution they represented. Using this methodology, we identified a total of 850,508 synonymous mutations.

1c) Adjustment of observed by expected mutations to obtain species-specific mutational spectra:

For each species, we calculated the mutation rates by dividing the number of observed variants by the number of expected variants in the corresponding context, resulting in 192 mutation rates.

These substitution rates were then transformed into frequencies, ensuring that the sum of all 192 normalised substitution rates equaled 1 for each species.

1d) Integral taxa-specific mutational spectrum.
Finally, we averaged the mutational spectra of species samples to derive class-specific or phylum-specific mutational spectra.

2) Comparison of the mutational spectra.
In our comparison of the mutational spectra, we primarily used cosine similarity as the main metric. Our analysis comprised two fundamental aspects. Firstly, we conducted pairwise comparisons of mutational spectra at the species-specific level within and between classes, considering all conceivable combinations. Then we focused on comparison of the mutational spectra at the class-specific level. To facilitate this, we implemented jackknife resampling. In this process, we randomly selected 20 species from each pair of classes, calculated the 192-component mutational spectrum for both classes, and computed the cosine similarity. We repeated this process in 1000 iterations for every conceivable class combination. Through the use of jackknife resampling, we effectively adjusted the influence of varying class sizes.

3) Signatures' analysis.
To decompose class-specific mtDNA mutational spectra into COSMIC signatures we used SigProfilerAssignment tool[13]. Since COSMIC only contains symmetrical signatures (SBS96) that do not account for strand-specific mutagenesis, we split the asymmetrical mitochondrial 192-component spectra into "low" and "high" 96-component spectra based on the abundance of specific transitions. The "high" spectra include more frequent $C_H>T_H$ and $A_H>G_H$ transitions, while the "low" spectra include $G_H>A_H$ and $T_H>C_H$ transitions, where H signifies heavy strand notation of substitution. Moreover, we included in the decomposition analysis the artificial spectra "diff", that was calculated by subtraction of "low" from "high" spectra. Complementary transversions rates (for example, $A_H>C_H$ and $T_H>G_H$) were either averaged and equally added to the "low", "diff" and "high" spectra or eliminated from analysis. In addition, SigProfilerAssignment, like any other SigProfiler tool, is capable of processing distinct substitution counts on reference genomes including human nuclear genome. Accordingly, to ensure correspondence and to mimic the substitution counts observed in the human genome, we rescaled spectra of classes multiplying them by the trinucleotide frequencies of the human nuclear genome. We applied SigProfilerAssignment v0.0.30 using the cosmic_fit function, with the following parameters: genome_build='GRCh37', nnls_add_penalty=0.01 (reduce number of derived noisy signatures that explain low number of mutations), cosmic_version=3.3. Furthermore, to reduce the effect of the unexpected signatures in decomposition of average mammals spectra, we excluded the following subgroups of signatures from the analysis: immunosuppressants, treatment, colibactin, lymphoid, and artefact.

4) Abasic sites patterns in mtDNA.

Cai et al.[14] provided a detailed analysis of abasic (AP) sites, annotated with single-nucleotide resolution, in the mouse mitochondrial genome (mm10), focusing on both the heavy and light strands. We quantified the AP site occurrences within all 64 trinucleotide sequences of the mouse mtDNA, excluding the control region. These values were then adjusted in relation to the number of trinucleotide sequences present within the genome. This method was applied to both strands, allowing for a comparative analysis of strand-specific normalised counts. For the heavy strand analysis, the reverse complements of the trinucleotides were utilised.
We created trinucleotide logos using the Python library logomaker [15]. These logos used the normalised AP sites counts for each trinucleotide as the nucleotide weights. These weights were averaged across all trinucleotides and normalised in accordance with the unity sum rule.

5) Calculation of the mtDNA mutation spectrum asymmetry

The assessment of mtDNA asymmetry involved the transformation of a 192-component mutation spectrum into a 96-component spectrum. This was achieved by selecting frequencies of 96 single base substitutions of pyrimidines only (C>A, C>G, C>T, T>A, T>C and T>G) from the mitochondrial DNA mutational spectrum (see COSMIC) and dividing them by complementary substitutions frequencies.
The total mitochondrial asymmetry was determined by summing the differences between mutations presented in the 96-component mitochondrial mutational spectrum and their complementary mutations**.**

6) Analysis of mtDNA mutation spectrum in human cancers

Mutations in the full human mtDNA were derived from comprehensive analysis of the human mitochondrial genome by Yuan et al.[16] The mutation spectrum was calculated as described above, using the CRS reference sequence NC_012920.1. We assumed that almost all mtDNA mutations in cancer are nearly neutral and employed all mutations, including non-synonymous ones, to calculate the nearly neutral spectrum.
In our analyses, we calculated spectra for different parts of mtDNA. To compare the mutation spectra of genome regions that have low and high time spent single-stranded (TSSS) due to asynchronous replication, we separately calculated spectra for the region with low TSSS (first half of the major arc, 5,800 - 10,800) and the region with high TSSS (second half of the major arc, 11,000 - 16,000).

7) Data and code availability. All analyses we performed in python and R. Scripts and data are available on GitHub: https://github.com/mitoclub/mtdna-192component-mutspec-chordata

RESULTS

192-component mitochondrial mutational spectrum of all chordates

The 192-component mitochondrial mutational spectrum was generated by analysing 129,100 synonymous mutations derived from CytB sequences of 1040 chordate species. This process involved construction of species-specific sparse spectra, normalisation, and aggregation them into comprehensive 192-component mutational spectra for all chordates (Methods, Fig. 1, Fig. 2, Supplementary Table 1). To assess the spectrum's robustness with respect to the species composition, we calculated the standard deviation for each category of substitutions, using 1000 bootstrap species samples for each substitution type individually (indicated by whiskers in Fig. 2).

The resulting spectrum reveals a notable predominance of transitions, with $C_H>T_H$ and $A_H>G_H$ being the most common ones. This mitochondrial mutagenesis pattern aligns with observations in mammalian germline mutations[5], somatic mutations in human cancers[16], and healthy tissues[17]. The complete 192-component mitochondrial mutational spectrum for chordates is available in numeric format, akin to a COSMIC database, in Supplementary Table 2.

Variation in mtDNA Mutational Spectrum Across Species and Classes, Germline and Somatic cells: Implications for Damage

How does the 192-component mitochondrial mutational spectrum vary across species? To assess inter-species differences, we calculated cosine similarity for all species pairs (Methods). The resulting heatmap reveals minimal or absent clusters within different species classes (Supplementary Fig. 1). This limited class-specificity in mutational spectra suggests that external mutagens, i.e damage-related mutations, play a substantial and variable role in addition to replication-driven mutations that are expected to be more similar within each class due to shared evolutionary history. These external factors, in the case of mtDNA, can be linked to varying levels of damage associated with differences in longevity in mammals[5] and in temperature-associated basal metabolic rates across all chordates [5,18]. This limited class-specificity in mutational spectra suggests that varied and non-uniform damage-related mutations (due to variation in life-history traits and relevant damage) play a substantial role in addition to unvaried replication-driven mutations, which are expected to be more uniform within each class due to shared evolutionary history. Thus there can be the link between varying levels of damage and differences in life-history traits, i. e. longevity in mammals[5] or temperature-associated basal metabolic rates across all chordates [5,18]. These eco-physiological traits can lead to the closer resemblance of the long-lived mammals' mutational spectra with the warm-water fishes' spectra rather than with the short-lived mammals' spectra. Investigating the

interplay between external eco-physiological and internal genetic factors in shaping mtDNA mutational spectra across species presents a compelling avenue for future research.

Despite the observed limited class-specificity (Supplementary Fig. 1), we found it worthwhile to consider this natural grouping factor during derivation of the integral class-specific spectra. Employing the methodology outlined in the previous section , we reconstructed separate spectra for each of the five chordate classes. These class-specific spectra (Supplementary Fig. 2) exhibited notable similarities to the integral spectrum (Fig. 2b), indicating a conservation of basic mtDNA mutational processes across chordates. To assess class-specific differences in mtDNA mutational spectra, we compared them, computing median pairwise cosine similarities between different classes (Methods; Fig. 3a). Strikingly, Aves displayed the most distinct mtDNA mutational spectrum, with the lowest cosine similarities to all other chordate classes: each median cosine similarity being less than or equal to 0.83. This supports the hypothesis that the mtDNA mutational spectrum (asymmetric transitions), possibly shaped by spontaneous chemical damage [19], reflects the basal metabolic rate, highest in birds among all chordates. The $A_H>G_H$ mutation, most probably resulting in damage via rapid adenine deamination to hypoxanthine [20,21], is notably more prevalent in birds compared to other chordate groups.

To test additionally the role of the damage versus replication in formation of the mtDNA Mutational Spectrum, we hypothesised that somatic mtDNA mutations in human cancers, often occurring in a hypoxic microenvironment, would exhibit rather distinct mutational patterns compared to avian species. Conducting pairwise comparisons between somatic mutations in human cancers [16] (Methods) and germline variants in five chordate classes (all our previous results were based on germline variants) we observed that comparisons between birds and cancers indeed revealed the highest cosine dissimilarity (Fig. 3a), indicating opposing deviations from the core germline mutagenesis observed in chordates. This divergence may be attributed to variations in aerobic metabolism, since birds (specifically, the germline cells of birds — oocytes, the sole carriers of mtDNA) exhibit high oxidative phosphorylation rates due to their elevated basal metabolic rate [22,23], while the majority of cancers typically thrive in a relatively hypoxic state [24].

Conservative Patterns in mtDNA Mutational Spectrum

We aimed to identify the components of the mtDNA mutational spectrum that exhibit conservation in their patterns across different groups. We firstly categorised substitutions into transitions (4*16) and transversions (8*16), finding that transitions displayed relatively high cosine similarities between various chordate classes and human cancers (Fig. 3b), while

transversions showed lower similarities (Fig. 3c), possibly due to class-specific mutagens or data stochasticity arising from transversions' rarity in our sample (Supplementary Fig. 3). Interestingly, transitions displayed a pattern consistent with our previous findings (Fig. 3b): (i) a high degree of similarity between core chordate classes, including Actinopterygii, Amphibia, Lepidosauria, and Mammalia, all with cosine similarities exceeding or equaling 0.84; (ii) decreased similarity of Aves and human cancers with the core chordate spectra, with all cosine similarities falling below or equaling 0.83; (iii) minimal similarity observed between Aves and human cancers. These findings suggest a significant impact of mitochondrial mutagen(s), linked to the level of aerobic metabolism, specifically on transitions.

After categorising transitions into four types, we consistently observed high cosine similarity across all four categories (Fig. 4). Notably, $G_H>A_H$ exhibited higher similarity than $C_H>T_H$, despite being complementary equivalents ($G_H>A_H$ is equivalent to $C_L>T_L$). This suggests that C>T mutagenesis on the double-stranded DNA is conserved and defines $C_H>T_H$ as well as $C_L>T_L$ ($G_H>A_H$). On the other hand C>T mutagenesis on the single-stranded DNA defines the asymmetrical part of $C_H>T_H$. Importantly, this pattern remains consistent across all chordate classes and human cancers (Fig. 4). Given the prevalence of C>T substitutions [25], its symmetry, and consistency across different taxa, we propose that the underlying mutagen responsible for the symmetrical part of C>T substitutions is mainly associated with internal mistakes during replication, caused by the gamma DNA polymerase (Supplementary Fig. 4) . It can occur during the most thermodynamically stable T/G mispairing by pol-γ[26].

Taking into account this observation, we deeply explored the asymmetrical aspect of the $C_H>T_H$ substitutions. These mutations are known to increase in single-stranded DNA which is exposed to reactive oxygen species, as demonstrated in a yeast experiment involving oxygen peroxide and paraquat. In this experiment, the trinucleotide motifs cCc>cTc emerged as the hallmark of oxidative damage in single-stranded nuclear DNA[27]. Utilising the 192-component spectrum in our analyses, we consistently identified cCc>cTc as the most common motif for $C_H>T_H$ substitutions in the mtDNA of all chordate classes, perfectly aligning with previous experimental findings (Fig. 5a). This suggests that cCc>cTc, being the predominant mutation in mtDNA across all chordates, likely results from oxidative or other damage to a single-stranded DNA. Interestingly, cancer data reveals a distinct pattern: asymmetrical $C_H>T_H$ substitutions are most frequent in the context of nCg (Supplementary Fig. 5). This suggests two potential explanations: at first, there can be reduced oxidative damage affecting somatic cancer mtDNA mutations[28] compared to germ-line mtDNA mutations, leading to a decrease in cCc>cTc. Secondly, the potentially stronger role of cytosine methylation in somatic mtDNA mutagenesis can result in an increase in CpG > TpG. While being less explored in the context of mtDNA, this phenomenon may be more prominent in somatic tissues echoing the well-known association of CpG > TpG substitutions with ageing and molecular evolution in the nuclear genome[29].

Deep exploration also revealed the second most common substitution A_H>G_H. This mutation is rather rare in the nuclear genome. We observed that the nAt>nGt and nAg>nGg motifs are predominant in A_H>G_H substitutions across all chordates (Fig. 5b). While this pattern remains similar across all chordate classes, it alters in cancers, demonstrating once again diverse environments in cancer cells relative to normal germ-line tissues.

A_H>G_H substitution recently has been identified as a mutation associated with ageing in mammals and correlated with body temperature in all chordates, suggesting that it may serve as a hallmark of a damage, which is a normal by-product of aerobic metabolism.[5] Additionally, A>G mutations in mtDNA are known as highly asymmetric[3]. However, the mechanism behind A_H>G_H substitutions remains unclear. The potential link to this mutation can be due to the N6-methyldeoxyadenosine (6mA) in mtDNA. This modification, concentrated in the mitochondrial genome and strongly enriched on the heavy strand, is associated with stressful hypoxic conditions and reduces the expression level of mtDNA, it is subject to regulation by the methyltransferase METTL4 and the demethylase ALKBH1. Given the density of 6mA is substantially higher in mtDNA than in nuclear DNA, it's plausible that dynamic transitions such as A > 6mA > A could facilitate mutagenesis through deamination mechanisms, leading to A>G mutations. This could occur either via reactive oxygen species (ROS) acting on adenine or through increased deamination susceptibility of 6mA compared to adenine. Conversely, in study on 6mA detection it has shown that unmethylated adenine is prone to transform to guanine during PCR, increasing A>G transitions in unmethylated sites. Additionally, in some cases 6mA is believed to protect genomes from several forms of transversions[30].

Although a link between 6mA and A_H>G_H is rather suggestive, as well as the enrichment in 6mA in mammalian mitochondria is still contradictory[30,31], we observed that the motifs for A_H>G_H substitutions, described as (c/a)At and A(t/g) in previous studies[32,33] are similar to our findings of nAt>nGt and nAg>nGg motifs. There is a suggestion that ROS damaged forms of adenine (i.e. 2-OH-Ade) may lead to strand-specific A:T>G:C transitions in both mammalian cells and *E. coli.* [34]. Future studies are needed to uncover a mechanism of A_H>G_H substitutions in mtDNA.

mtDNA Mutations through the Lens of COSMIC signatures: BER Deficiency and C_H>T_H*, MMR absence and* G_L>A_L*, and SBS12-Driven* A_H>G_H *Alterations*

In our previous analyses (Fig. 5), we examined local substitution patterns and here to deconvolute the overall mutational spectrum of mtDNA into its underlying mutational signatures, we utilised the COSMIC SBS database (https://cancer.sanger.ac.uk/signatures/sbs/). Since the COSMIC SBS database is built upon 96 component signatures, we divided our 192-component mtDNA spectrum into three sets: "high" spectrum for asymmetric mutations on the heavy strand, "low" spectrum for symmetric mutations on both strands, and "diff" spectrum typical for the

single-stranded heavy strand of mtDNA (Methods). We observed that five signatures, namely SBS30, SBS44, SBS21, SBS5 and SBS12 are predominant in mtDNA mutations (Fig. 6a). SBS5 shows a consistently uniform signature, more pronounced when including transversions (Fig. 6a). Due to the rarity and noisy nature of transversions in mtDNA, separate analyses were conducted using all mutations (transitions and transversions) and transitions alone, assuring that eliminating transversions from the analysed spectra did not significantly impact our results (Fig. 6a).

SBS30, linked to base excision repair (BER) deficiencies involving NTHL1, predominantly features C>T mutations[26,35]. NTHL1 is known to excise oxidised pyrimidine lesions[36]. In mitochondria, BER is the primary repair pathway for chemically damaged bases, such as deaminated, oxidised, and alkylated bases[37]. Our results show increased BER deficiency in the "high" and "diff" spectra (Fig. 6a), which can be explained by the extended time of the heavy strand being single-stranded (ss) during asynchronous mtDNA replication[38]. Indeed ssDNA is especially vulnerable to BER deficiency because this repair process requires a complementary strand to replace removed damaged bases, which ssDNA lacks. If BER is deficient on a heavy strand of mtDNA we expect that some fraction of the most common $C_H>T_H$ mutations can be driven by this deficiency. Additionally, $C_H>T_H$ gradient in mtDNA connected with TSSS[5,38] probably can be the result of lagging NTHL1 activity. Interestingly, several NTHL1 variants demonstrate defective ability for the repair of the 5-hydroxyuracil, the most abundant oxidation product of cytosines[39]. Given high concentration of ROS in the mitochondrial matrix it can explain partly high $C_H>T_H$. BER, which includes damaged DNA base recognition and removal, results in an abasic site, followed by cleavage, end processing, gap filling, and ligation[40]. Given mtDNA's glycosylases that recognize and remove damaged bases, abasic sites from these enzymes can shed light on BER's strand-specificity and motifs.

Abasic sites are created through two primary mechanisms: natural base loss and a key phase in the Base Excision Repair (BER) process where glycosylases enzymatically excise damaged nucleotides. According to the first source[41], the following observations were noted: (i) depurination, which involves the loss of adenine (A) and guanine (G), occurs 20 times more frequently than depyrimidination (loss of cytosine [C] and thymine [T]); (ii) Guanines are 1.5 times more prone to depurination than adenines; (iii) depurination occurs over four times faster in single-stranded DNA (ssDNA) compared to double-stranded DNA (dsDNA). In our study, we analysed abasic sites within the mouse mitochondrial genome, with precise single-nucleotide annotation, in both the heavy and light strand coding sequences (excluding the control region). Our observations (Fig. 6) reveal that while trinucleotides at abasic sites are consistent across both strands, the heavy strand exhibits double the damage, particularly in G and A nucleotides (as detailed in Fig. 6). The distribution of abasic sites in mitochondrial DNA (mtDNA) (Fig. 6) aligns well with the expected patterns of spontaneous depurination. The dissimilarity between the distribution of these abasic sites and the mtDNA mutational spectrum, including SBS30, suggests that most abasic sites are either repaired through BER[41] or result in the elimination of the damaged mtDNA[42].

SBS44 and SBS21 are two of seven known signatures linked to defective DNA mismatch repair (MMR), crucial for correcting mismatches during DNA replication[43]. Although the presence of MMR in mtDNA has been debated, it is widely accepted that there is no MMR in mtDNA[37]. Our analysis confirms it showing, that even if MMR partially exists in mtDNA, it is highly deficient. Notably, a pronounced MMR deficiency signature appears in the "low" spectra, associated with symmetrical mutations consistent with polymerase errors. We suggest that MMR-deficiency signatures, i.e. the symmetrical part of C>T (G_L>A_L) (Fig. 6d) are shaped by the gamma DNA polymerase, which is expected to introduce symmetrical mutations among which C>T (and G>A) are the most common[25,26](Fig. 6d, Supplementary Fig. 4).

SBS12, despite its unknown origins, is a potential hallmark of chemical damage in mtDNA's single-stranded heavy strand. This is apparent from its high A_H>G_H mutation rate (T>C in COSMIC notation), indicative of mtDNA specific age- and temperature- related stress[5,18], and its dominant presence in the "high" and "diff" spectra (Fig 6a), reflecting its impact on the heavy strand. SBS12 also shows an increase with replication timing in the nuclear genome (Fig 6e), transcriptional strand asymmetry with more A>G mutations on the non-transcribed strand (more T>C mutations on transcribed strand in COSMIC notation); and replication strand asymmetry with A>G mutations on the lagging strand (T>C on leading strand in COSMIC notation). Additionally, its highest prevalence in birds (Fig 6a), known for elevated metabolic rates, further underscores its association with chemical damage. SBS12 therefore offers great promise for uncovering the mechanism of mitochondrial ssDNA-specific damage.

In summary, our study identifies three key mutational signatures in mtDNA: (i) Asymmetric C_H>T_H mutations, resulting from BER deficiency in single-stranded DNA, highlighting its repair vulnerabilities; (ii) Symmetrical C_H>T_H (complementary G_L>A_L) mutations, due to the absence of MMR in mtDNA, mirroring errors made by mtDNA polymerase; (iii) Asymmetric A_H>G_H mutations, associated with damage, predominantly affecting single-stranded mtDNA. These findings can elucidate the primary mutagens shaping the mtDNA mutational landscape.

Strong Asymmetry in mtDNA Mutagenesis is shaped by Single-Stranded DNA Damage

mtDNA mutagenesis is characterised by a rather strong asymmetry: (i) the most common transitions C_H>T_H and A_H>G_H are considered to occur mainly on a heavy strand during replication of mtDNA (C_H>T_H >> G_H>A_H, A_H>G_H >> T_H>C_H) as a result of deamination of cytosine and adenine[38] , besides level of deamination is proportional to single strandedness [44]; (ii) Transversions G>T, A>T, C>G, and A>C also demonstrate increased frequencies on heavy versus light strands[16]; (iii) both the most pronounced signatures - BER deficiency and mito-specific SBS12 are highly asymmetrical with a much stronger impact on a heavy strand (Fig. 6). Estimation of the total level of asymmetry in mtDNA (Methods) shows that more than

50% of all mtDNA mutations are asymmetrical. This high asymmetry observed for almost all 96 types of substitutions suggests that either mutagenesis or the reparation process is asymmetrical.

In order to comprehend the underlying characteristics of asymmetric mutations, we conducted a comparative analysis, drawing insights from a recent study by Seplyarskiy et al.[3]. This study identified two distinct types of asymmetry within the human nuclear genome: T-asymmetry, transcription asymmetry, which originates from mutations on the non-transcribed strand, and R-asymmetry, replication asymmetry, which predominantly occurs on the lagging strand. Notably, a strong correlation was observed between these two asymmetry types, leading the authors to propose that the maintenance of such asymmetries is influenced by error-prone polymerases involved in DNA repair processes. Considering that mtDNA has no error-prone polymerases involved in the preparation process, we anticipate the involvement of other factors in contributing to asymmetry within mtDNA compared to the nuclear genome.

It is important to note, that the heavy strand of mtDNA can be under both pressures: T asymmetry (the majority of genes are coded on the light strand and it is more intensively transcribed than the heavy strand) and R asymmetry (the heavy strand and especially CYTB, located at the end of the major arc, spent a significant amount of time being single-stranded) and thus it is difficult to deconvolute which damage mechanism is responsible for the mito-asymmetry.

To compare the asymmetries, we calculated mito asymmetry from our 192-component mutational spectra, (Methods). Remarkably, we found a significant correlation between mito asymmetry and both R and T asymmetry (Fig. 7a, including relevant statistical information).

The high similarity of the nuclear T and R asymmetries with the mitochondrial ones means that mutation types more common on a heavy strand are also more predominant on the lagging strand in the nuclear human genome and are more common on the non-transcribed strand of the nuclear human genome. This similarity means that either (i) all these substitutions are due to the same reparation deficiency, such as error-prone polymerases as was proposed by Seplyarskiy et al. for nuclear genome or (ii) reflect the same mutational process, for example a damage of the single-stranded DNA or (iii) different mechanisms lead to very similar correlation due to various internal chemical fragility of various substitutions.. In mtDNA there are no described additional polymerases and thus the first explanation is not supported.

(ii) The single-strand specific damage can be an interesting explanation which is rather universal for both genomes (but see Seplyarskiy et al. for some counterarguments). Taking into account that single-strand specific mutations in mtDNA are well known[38,45,46] and have been proposed for decades we tested the potential effect of the ssDNA on the strength of the asymmetry with two datasets: (i) whether the asymmetry is stronger in mtDNA regions, characterised by high TSSS (time being single stranded) versus low TSSS (Fig. 7b) and (ii) whether the asymmetry is stronger in warm versus cold-blooded chordates, due to increased chemical damage in formers. Comparing the strength of asymmetry of the same substitutions in high TSSS and low TSSS regions from human cancers we observed, as expected, that high TSSS

is characterised by stronger asymmetry (Fig. 7b). Next, comparing mutational spectra asymmetries of warm versus cold-blooded animals we observed quantitatively the same results - the majority of substitutions' ratios are more than 1. (Fig. 7b).

(iii) Finally, even if mechanisms responsible for this correlation are different we can assume similar fragility of nucleotides. Interestingly, the associations between T-, R- and mito-asymmetries also work on the level of six main substitution types. For example, C>T in different contexts occurs more often than G>A on the heavy strand of mtDNA, the lagging strand of the human nuclear genome, and the non-transcribed strand in the human nuclear genome. Polarising the substitutions into complementary pairs where the first one has a higher substitution rate on the strand of interest (heavy strand of mtDNA, lagging strand of the human nuclear genome, non-transcribed strand of the human nuclear genome) than the second, we got the following list: C>T >> G>A, A>G >> T>C, G>T >> C>A, A>T >> T>A, C>G >> G>C, A>C ~ T>G (Fig 2b). Moreover, additional analysis showed that the main correlation is maintained by the association of the 6 main types of substitutions (Fig. 7c) rather than by the content (Supplementary Fig. 7).

Altogether, we conclude that the high similarity of MitoAsymmetry with T and R nuclear asymmetry (Fig. 7a), maintained mainly by base-specific rates (Fig 7c) and more pronounced in high versus low TSSS regions as well as in warm-versus cold-blooded species (Fig. 7b) support that a hypothesis that increased fragility of ssDNA to chemical damage can lead to the asymmetry of mutagenesis in mtDNA and partially in nucDNA.

Altogether all these asymmetries support that DNA damage substantially contributes to human mutations in both mitochondrial and nuclear genomes.

DISCUSSION

In this study, by integrating species-specific mtDNA mutational spectra from various chordates, we have reconstructed a comprehensive 192-component mutational spectrum. Our analysis deconvolutes this spectrum into three main fundamental sources: replication-driven mutations and two damage-driven categories of mutations characterised by distinct etiologies and dynamics.

The first component is the symmetrical, replication-driven component, which is composed of C>T symmetrical mutations and SBS5-like component shaped by numerous transversions and altogether introduces 38% of all de novo mutations (25% for symmetrical part C>T and 13% for all transversions). C>T component (Fig. 8, both grey components of $C_H>T_H$ and $G_H>A_H$), is primarily a result of pol-γ's internal replication errors (Supplementary Fig. 4)[25]. This component exhibits a high degree of conservation across chordate classes (Fig. 4c), indicative of POLG's inherent properties. Its mutational pattern resembles MMR deficiency in the nuclear genome (Fig. 6a), further underscoring the absence of MMR in mtDNA. The SBS5-like component is most likely also driven by replication[47]. It is characterised by a uniform pattern encompassing all mutation types, notably including symmetrical transversions (Supplementary Fig. 4). If the symmetrical part of C>T, as well as transversions, are replication-driven we expect a positive correlation between these types of mutations. Indeed, in our recent comparative-species analysis we saw collinearity of the majority of transversions with Gh>Ah substitutions (Fig 2C). The asymmetrical C>T component (Fig. 8, red component of $C_H>T_H$), representing about 30% of mutations, is associated with single-stranded DNA damage, most likely caused by spontaneous deamination and potentially oxidative damage (Fig. 5a). This type of damage can be traced by intermediate steps as abasic sites on the heavy strand (Fig. 6b) and, coupling with deficient BER on a single-stranded heavy strand (Fig. 6a,c) culminate in the formation of C>T mutations. (iii) The asymmetrical $A_H>G_H$ component (Fig. 8, red component of $A_H>G_H$), accounting for approximately 19% of the spectrum, are linked to single-stranded DNA damage, likely due to adenosine deamination. As the most asymmetrical mutation in mtDNA and, according to chemical studies, the nucleotide most susceptible to deamination[20], this mutation type presents a range of intriguing correlations with eco-physiological traits in mammals[5] and across all chordates[18]. These findings, alongside their similarity to SBS12 and the potential implications of adenosine methylation, highlight important avenues for further research.

The classification of mutation types and the deconvolution of the mtDNA spectrum into distinct signatures present an effective method for determining the roles of replication and damage-driven mutations in various human tissues [48] and across different species.

Investigating how chemical damage in mtDNA is affected by eco-physiological factors can help us understand variations in mtDNA spectra across species[5,18]. For example, adenine's hydrolytic deamination to hypoxanthine happens slower than cytosine's to uracil[49] and is highly mutagenic [50]; this process quickens at higher temperatures[21] (more significant at 25°C than at 8°C) and in alkaline conditions[20]. This temperature dependency could account for the increased $A_H>G_H$ mutations in warm-blooded compared to cold-blooded chordates [18]. The pH dependency, assuming the higher pH in the mitochondrial matrix of aged tissues, might explain why $A_H>G_H$ mutations are more common in long-lived mammals[5]. Considering that mitochondrial environments tend to be more alkaline and warmer compared to the nucleus, these physical and chemical factors might significantly contribute to the increased incidence of $A_H>G_H$ substitutions in mitochondrial DNA compared to nuclear DNA. For instance, the highest divergence of the 192-component mutational spectra in birds and human cancers (Fig 3,4) from the core spectrum of all chordate groups aligns with the idea that bird-specific mutations are influenced by increased damage due to a higher basal metabolic rate, while cancer-specific mutations are shaped by heightened replication-driven processes. This is further supported by the higher proportion of $A_H>G_H$ substitutions in the 12-component spectrum in birds compared to cancers.

In exploring the intricate dynamics of mtDNA, it becomes evident that the heavy strand is subject to dual pressures—T asymmetry and R asymmetry. The prevalence of genes on the heavy strand and its more intensive transcription, coupled with its single-stranded exposure, particularly at the CYTB locus on the major arc, presents a complex landscape. This intricacy poses a challenge in discerning the specific damage mechanisms responsible for the observed mito-asymmetry, adding a layer of complexity to our understanding of mitochondrial genomic dynamics.

In future research, particular focus should be given to genomes that display asymmetrical spectra. For these cases, employing a comprehensive 192-component analysis is recommended, moving beyond the standard 96-component framework typically utilised in cosmic and signal spectra. This approach, as exemplified in our study, is crucial for accurately assessing complex mutational patterns. Additionally, the development and availability of normalised 192-component mutational spectra across diverse species and taxa will be instrumental in advancing our understanding of species/taxa-specific mutagens.

Main source of pathogenic mutations in human mtDNA (if we remove normalisation) - would it be replication-driven or damage-driven?

Acknowledgements

This work was supported by the Federal Academic Leadership Program Priority 2030 at the Immanuel Kant Baltic Federal University (to D.I, K.P. and B.E). A.G.M. is supported by the Russian Science Foundation grant No. 21-75-20143. K.G. is supported by the Russian Science Foundation grant No. 21-75-20145.
We thank the high-performance computing platform at the Immanuel Kant Baltic Federal University.

Contributions

The design of the study developed by K.P. Data mining and processing performed by D.I., B.E. and K.G. Manuscript prepared by K.P., D.I., B.E. and A.G.M. All authors (D.I., B.E., A.G.M., V.S., M.K.S., I.M., D.K., W.S.K., S.D., K.K., J.F., K.G. and K.P.) discussed in depth the manuscript and the rationale behind the project. All authors read and approved the final manuscript. The authors express their appreciation to Vladimir Seplyarskiy for his valuable contributions and insightful discussions regarding asymmetry and repair mechanisms, and Philipp Kapranov for engaging discussions related to the analysis of abasic sites.

Conflict of interest statement. None declared.

Figures

Figure 1. **Overview of the developed pipeline.** From the mitochondrial genomes of 1040 chordata species two groups of mutations were obtained: observed 129'100 synonymous mutations from our polymorphic database (Species 1, left column) and expected synonymous substitutions from NCBI RefSeq database (Species 2, left column). Both obtained groups of substitutions were normalised and transformed into the integral species-specific 192-component spectra for all chordates (Methods).

Figure 2. Pattern of the integral mutational spectrum reflects mitochondrial mutagenesis in chordates: there is an excess of $C_H>T_H$ and $A_H>G_H$ transitions in 192-component (b) and 12-component (a) mutational spectra (n=1040).
The order of substitutions is based on reverse complemented mutations, where, for example, the third bin for C>A substitutions is represented as ACG, and the third bin for G>T substitutions is CGT. Missing (zero) bins are explained by the absence of observed synonymous substitutions, the absence of expected substitutions or both.

Figure 3. Median cosine similarities in pairwise comparisons of somatic and germline variants across five chordate classes and human cancer. Each box presents three values: Q1, Q2, and Q3, which were derived from 1000 cosine similarity comparisons between two classes. (a) High median cosine similarities (0.65 or higher) were consistently observed across five classes of chordates and human cancer when comparing the whole 192-component mutational spectrum (n=192). (b) Similarly, high median cosine similarities (0.83 or higher) were observed specifically for transitions alone (n = 4x16). (c) In contrast, low median cosine similarities (less than 0.6) were observed exclusively for transversions (n = 8x16).

Figure 4. Cosine similarities among chordate groups for four types of substitutions. High cosine similarities were observed when comparing four types of substitutions among different chordate groups. Notably, the similarities between two complementary substitutions, $G_H>A_H$ and $C_H>T_H$, demonstrated higher values for $G_H>A_H$. This higher similarity in $G_H>A_H$ substitutions can be interpreted as a potential marker of asymmetrical mutagenesis occurring on different mtDNA strands.

Figure 5. Deep exploration of the observed common motifs.(a) The motif cCc>cTc is consistently the most frequently observed motif among all $C_H>T_H$ substitutions in mtDNA across all chordate species. This pattern was previously reported as the hallmark of oxidative damage in single-stranded DNA. (b) nAt>nGt and nAg>nGg are the second two of the most common motifs of $A_H>G_H$ substitutions in all chordates. It has been reported that such motifs can be linked to the 6mA ROS induced mutations in mtDNA.

Figure 6. The 196-component mutational spectra demonstrates COSMIC SBS signatures. (a) Five signatures SBS30, SBS44, SBS21, SBS5, and SBS12 predominate in spectra within three distinct mutation sets, as detailed in the Methods section.
(b) Analysis of the trinucleotide pattern of abasic sites within coding sequences of both the heavy and light strands reveals that trinucleotides on the heavy strand exhibit higher levels of damage, with G being the most frequently damaged nucleotide.
(c) Pattern of SBS30 signature: BER deficiency mutations.
(d) Pattern of SBS44 and SBS21 signatures: MMR deficiency mutations.
(e) Pattern of SBS12 signature: ssDNA-specific mutations.

Figure 7. Comparison of the mitochondrial asymmetry based on the global mitochondrial mutational spectrum with the nuclear asymmetries in T and R states.
(a) The analysis of 192-component mitochondrial mutational spectra reveals a notable positive correlation with both T (upper panel, Spearman's Rho =0.34, p = 0.0007, N = 96) and R (bottom panel, Spearman's Rho = 0.29, p = 0.004 , N = 96) nuclear asymmetries (each dot represents substitution type with a context). The elimination of zero-rated substitutions (rare transversions, never observed in chordates) from the mito-asymmetry significantly improved the positive associations with both T- (Spearman's Rho = 0.64, p = 4.6*e-09, N = 68) and R-asymmetry (Spearman's Rho = 0.62, p = 1.3*e-08 , N = 68)
(b) Direct comparison of 192-component mitochondrial mutational spectra asymmetries demonstrates that the mutational spectrum of warm-blooded species is more asymmetrical when contrasted with their cold-blooded counterparts, probably due to increased metabolic damage (Wilcoxon test, p = 1.56e-06, N=61). Additionally, we can see a slight trend that high TSSS shows asymmetry as well (Wilcoxon test, p = 0.06, N=34).
(c) Among the six primary substitution types, a robust association is observed with T (left panel) and R (right panel) asymmetries, superseding the influence of nucleotide content.

Figure 8. Main mutagens found in mtDNA. (i) symmetrical mutations, predominantly C>T and rare transversions (grey component), linked to pol-γ's replication errors; similar to SBS21, SBS44 and SBS5; (ii) asymmetrical C>T mutations (red component), indicative of single-stranded DNA damage; similar to SBS30; (iii) asymmetrical A>G mutations (red component), also resulting from single-stranded DNA damage but particularly influenced by metabolic and age-specific mitochondrial environment; similar to SBS12.

REFERENCES

1. Chatterjee, N. & Walker, G. C. Mechanisms of DNA damage, repair, and mutagenesis. *Environ. Mol. Mutagen.* **58**, 235–263 (2017).
2. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
3. Seplyarskiy, V. B. *et al.* Error-prone bypass of DNA lesions during lagging-strand replication is a common source of germline and cancer mutations. *Nat. Genet.* **51**, 36–41 (2019).
4. Koh, G., Degasperi, A., Zou, X., Momen, S. & Nik-Zainal, S. Mutational signatures: emerging concepts, caveats and clinical applications. *Nat. Rev. Cancer* **21**, 619–637 (2021).
5. Mikhailova, A. G. *et al.* A mitochondria-specific mutational signature of aging: increased rate of A > G substitutions on the heavy strand. *Nucleic Acids Research* vol. 50 10264–10277 Preprint at https://doi.org/10.1093/nar/gkac779 (2022).
6. Belle, E. M. S., Piganeau, G., Gardner, M. & Eyre-Walker, A. An investigation of the variation in the transition bias among various animal mitochondrial DNA. *Gene* **355**, 58–66 (2005).
7. Chu, X.-L. *et al.* Temperature responses of mutation rate and mutational spectrum in an Escherichia coli strain and the correlation with metabolic rate. *BMC Evol. Biol.* **18**, 126 (2018).
8. Saclier, N. *et al.* Bedrock radioactivity influences the rate and spectrum of mutation. *Elife* **9**, (2020).
9. Dillon, M. M., Sung, W., Sebra, R., Lynch, M. & Cooper, V. S. Genome-Wide Biases in the Rate and Molecular Spectrum of Spontaneous Mutations in Vibrio cholerae and Vibrio

fischeri. *Mol. Biol. Evol.* **34**, 93–109 (2017).

10. Website. https://doi.org/10.1101/2023.04.23.537997 doi:10.1101/2023.04.23.537997.
11. Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N. & Delsuc, F. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Mol. Biol. Evol.* **35**, 2582–2584 (2018).
12. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
13. Islam, S. M. A. *et al.* Uncovering novel mutational signatures by extraction with SigProfilerExtractor. *Cell Genom* **2**, None (2022).
14. Cai, Y., Cao, H., Wang, F., Zhang, Y. & Kapranov, P. Complex genomic patterns of abasic sites in mammalian DNA revealed by a high-resolution SSiNGLe-AP method. *Nat. Commun.* **13**, 5868 (2022).
15. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).
16. Yuan, Y. *et al.* Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nat. Genet.* **52**, 342–352 (2020).
17. Sanchez-Contreras, M. *et al.* The multi-tissue landscape of somatic mtDNA mutations indicates tissue-specific accumulation and removal in aging. *Elife* **12**, (2023).
18. Mikhailova, A. G. *et al.* A mitochondrial mutational signature of temperature in ectothermic and endothermic vertebrates. *bioRxiv* 2020.07.25.221184 (2021) doi:10.1101/2020.07.25.221184.
19. Almatarneh, M. H., Flinn, C. G., Poirier, R. A. & Sokalski, W. A. Computational study of the deamination reaction of cytosine with H2O and OH-. *J. Phys. Chem. A* **110**, 8227–8234

(2006).

20. Wang, S. & Hu, A. Comparative study of spontaneous deamination of adenine and cytosine in unbuffered aqueous solution at room temperature. *Chem. Phys. Lett.* **653**, 207–211 (2016).
21. Karran, P. & Lindahl, T. Hypoxanthine in deoxyribonucleic acid: generation by heat-induced hydrolysis of adenine residues and release in free form by a deoxyribonucleic acid glycosylase from calf thymus. *Biochemistry* **19**, 6005–6011 (1980).
22. Sugimura, S. *et al.* Oxidative phosphorylation-linked respiration in individual bovine oocytes. *J. Reprod. Dev.* **58**, 636–641 (2012).
23. Trimarchi, J. R., Liu, L., Porterfield, D. M., Smith, P. J. & Keefe, D. L. Oxidative phosphorylation-dependent and -independent oxygen consumption by individual preimplantation mouse embryos. *Biol. Reprod.* **62**, 1866–1874 (2000).
24. Bhandari, V. *et al.* Molecular landmarks of tumor hypoxia across cancer types. *Nat. Genet.* **51**, 308–318 (2019).
25. Zheng, W., Khrapko, K., Coller, H. A., Thilly, W. G. & Copeland, W. C. Origins of human mitochondrial point mutations as DNA polymerase gamma-mediated errors. *Mutat. Res.* **599**, 11–20 (2006).
26. Zou, X. *et al.* A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nat Cancer* **2**, 643–657 (2021).
27. Degtyareva, N. P. *et al.* Mutational signatures of redox stress in yeast single-strand DNA and of aging in human mitochondrial DNA share a common feature. *PLoS Biol.* **17**, e3000263 (2019).

28. Ericson, N. G. *et al.* Decreased mitochondrial DNA mutagenesis in human colorectal cancer. *PLoS Genet.* **8**, e1002689 (2012).

29. Cosmic. COSMIC. (2020) doi:10.1093/nar/gkw1121.

30. Li, X. *et al.* The exploration of N6-deoxyadenosine methylation in mammalian genomes. *Protein Cell* **12**, 756–768 (2021).

31. Kong, Y. *et al.* Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution. *Science* **375**, 515–522 (2022).

32. Hao, Z. *et al.* N6-Deoxyadenosine Methylation in Mammalian Mitochondrial DNA. *Mol. Cell* **78**, 382–395.e8 (2020).

33. Koh, C. W. Q. *et al.* Single-nucleotide-resolution sequencing of human N6-methyldeoxyadenosine reveals strand-asymmetric clusters associated with SSBP1 on the mitochondrial genome. *Nucleic Acids Res.* **46**, 11659–11670 (2018).

34. Kamiya, H. & Kasai, H. Mutations induced by 2-hydroxyadenine on a shuttle vector during leading and lagging strand syntheses in mammalian cells. *Biochemistry* **36**, 11125–11130 (1997).

35. Drost, J. *et al.* Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* **358**, 234–238 (2017).

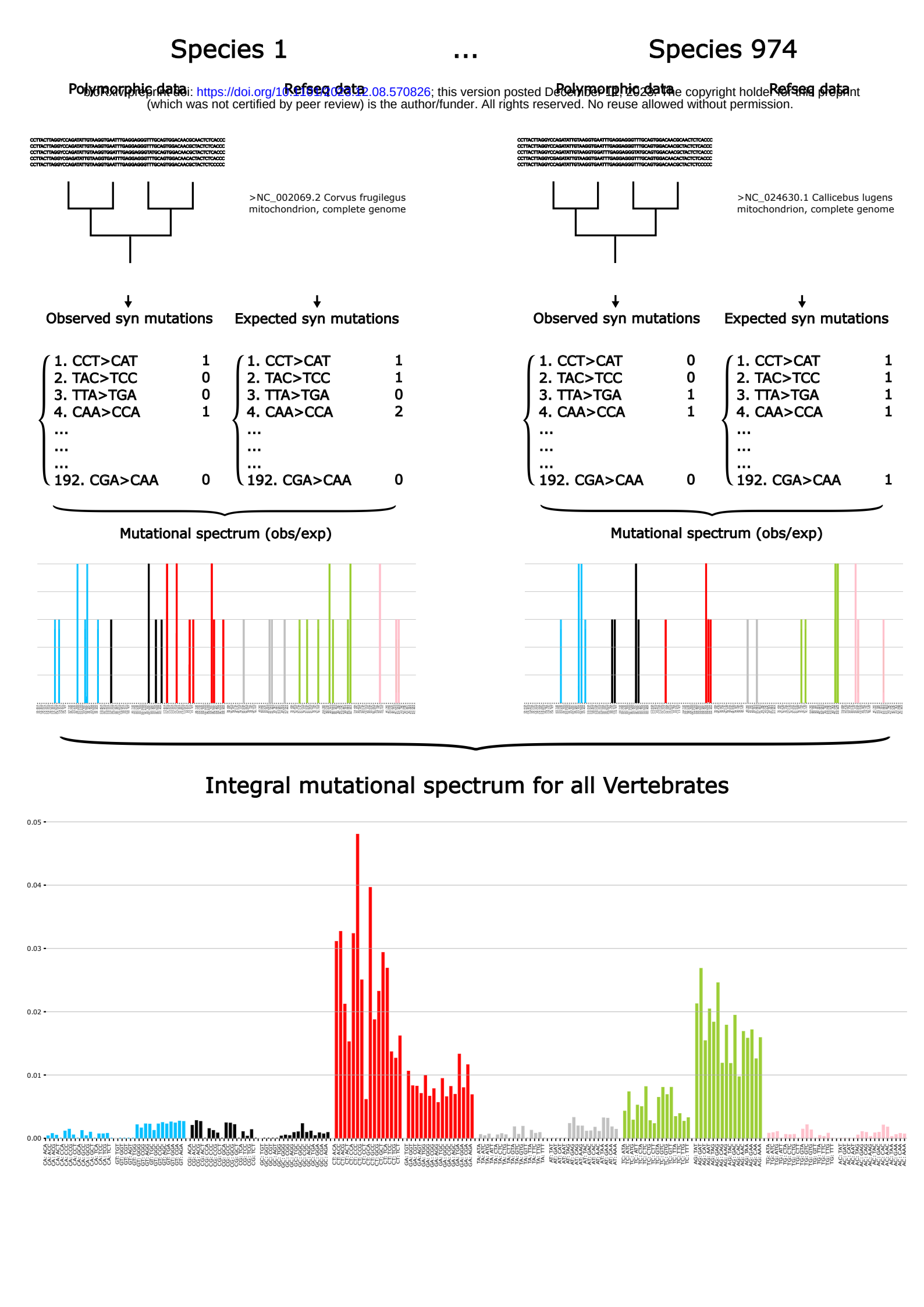
36. NTHL1 in genomic integrity, aging and cancer. *DNA Repair* **93**, 102920 (2020).

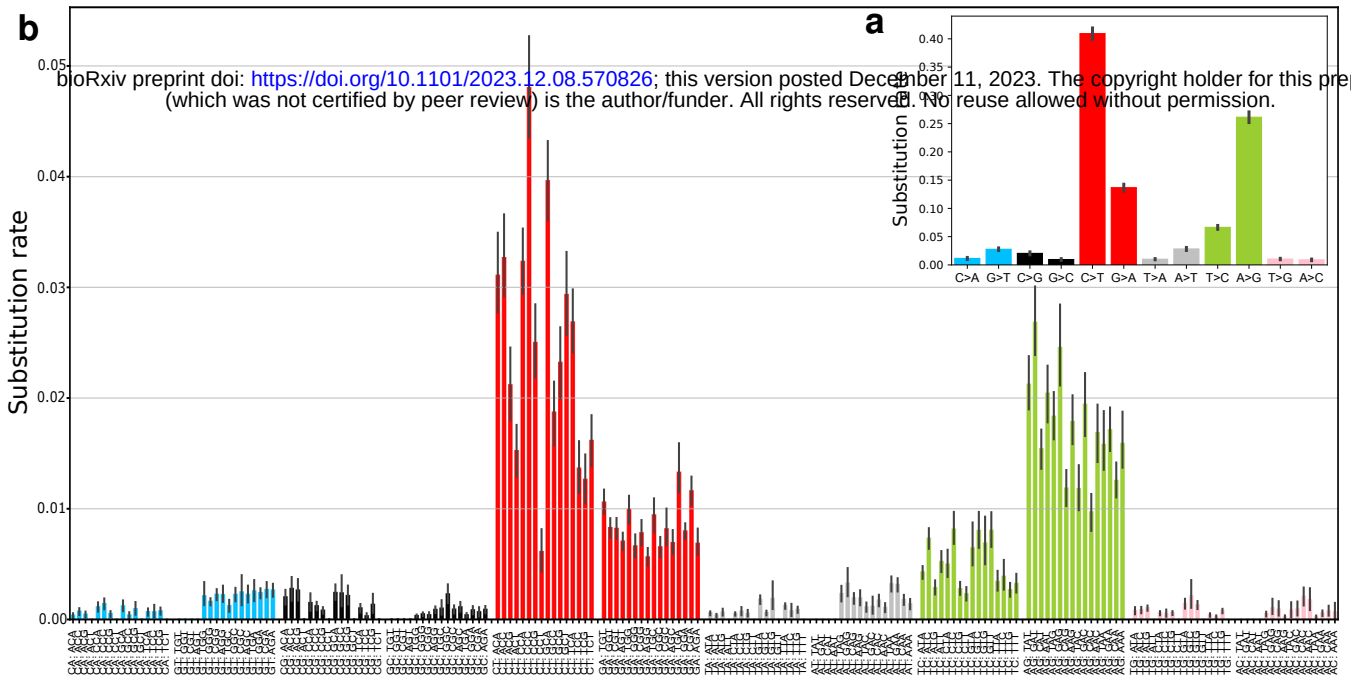
37. Rong, Z. *et al.* The Mitochondrial Response to DNA Damage. *Front Cell Dev Biol* **9**, 669379 (2021).

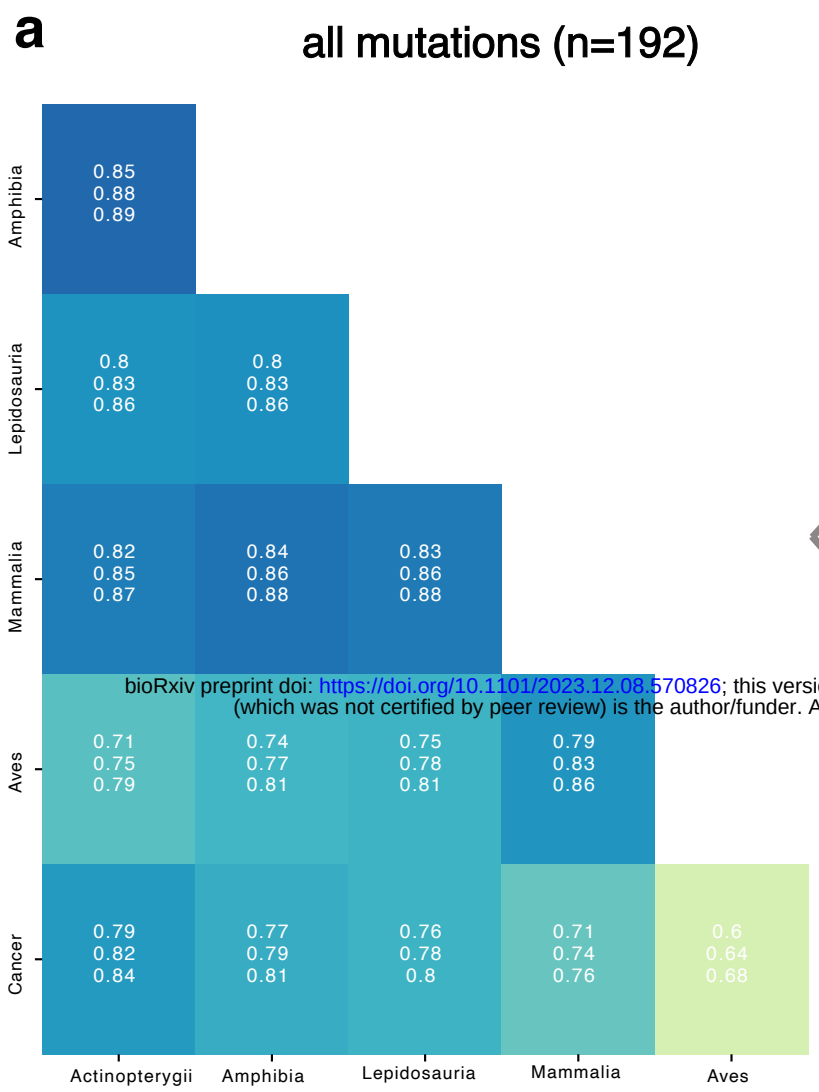
38. Sanchez-Contreras, M. *et al.* A replication-linked mutational gradient drives somatic mutation accumulation and influences germline polymorphisms and genome composition in mitochondrial DNA. *Nucleic Acids Res.* **49**, 11103–11118 (2021).

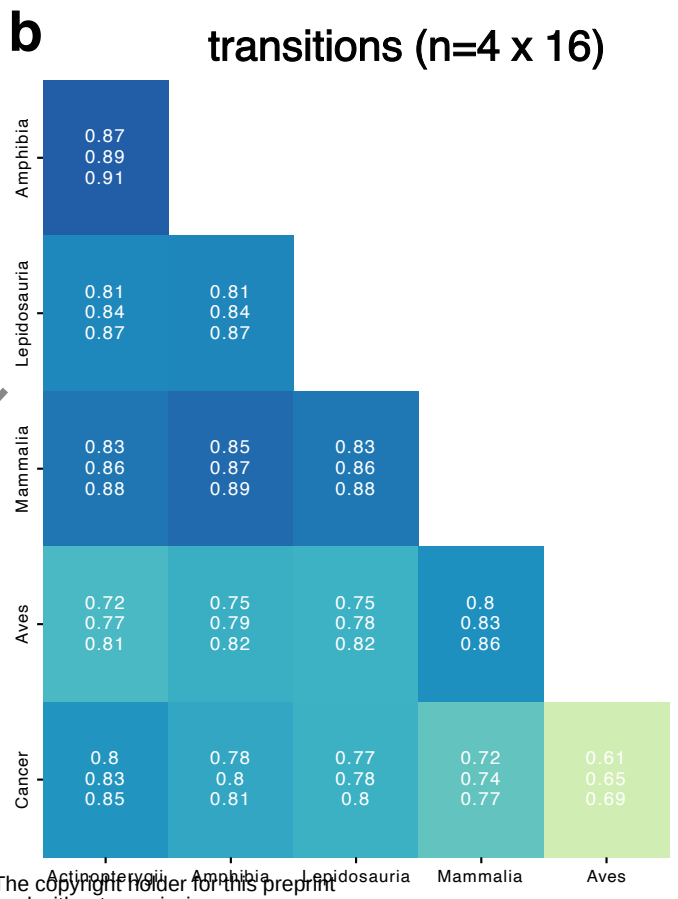
39. Defective repair capacity of variant proteins of the DNA glycosylase NTHL1 for 5-hydroxyuracil, an oxidation product of cytosine. *Free Radical Biology and Medicine* **131**, 264–273 (2019).
40. Fortini, P. *et al.* 8-Oxoguanine DNA damage: at the crossroad of alternative repair pathways. *Mutat. Res.* **531**, 127–139 (2003).
41. Thompson, P. S. & Cortez, D. New insights into abasic site repair and tolerance. *DNA Repair* **90**, 102866 (2020).
42. Khozhukhar, N., Spadafora, D., Rodriguez, Y. & Alexeyev, M. Elimination of Mitochondrial DNA from Mammalian Cells. *Curr. Protoc. Cell Biol.* **78**, 20.11.1–20.11.14 (2018).
43. Li, G.-M. Mechanisms and functions of DNA mismatch repair. *Cell Res.* **18**, 85–98 (2008).
44. Deamination gradients within codons after 12 position swap predict amino acid hydrophobicity and parallel β-sheet conformational preference. *Biosystems.* **191-192**, 104116 (2020).
45. Tanaka, M. & Ozawa, T. Strand asymmetry in human mitochondrial DNA mutations. *Genomics* **22**, 327–335 (1994).
46. Faith, J. J. & Pollock, D. D. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* **165**, 735–745 (2003).
47. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
48. Website. https://doi.org/10.1101/589168 doi:10.1101/589168.
49. Shapiro, R. & Klein, R. S. The deamination of cytidine and cytosine by acidic buffer solutions. Mutagenic implications. *Biochemistry* **5**, 2358–2362 (1966).

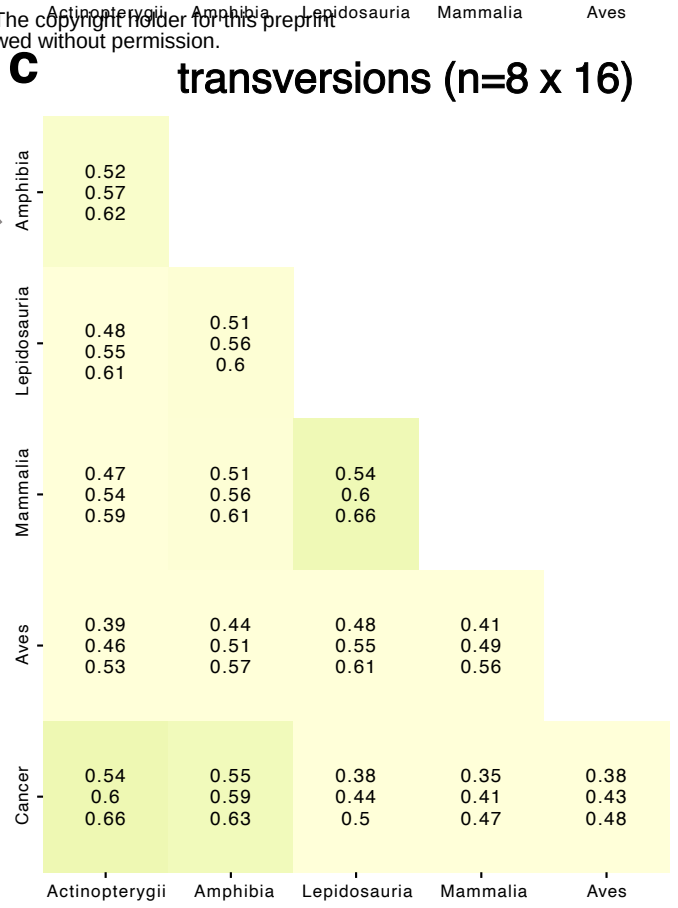
50. Budke, B. & Kuzminov, A. Hypoxanthine Incorporation Is Nonmutagenic in Escherichia coli. *J. Bacteriol.* **188**, 6553 (2006).

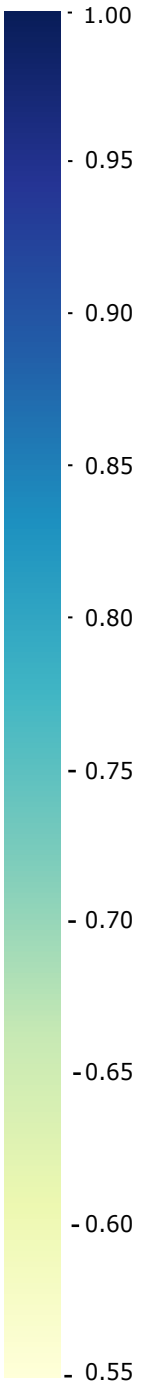


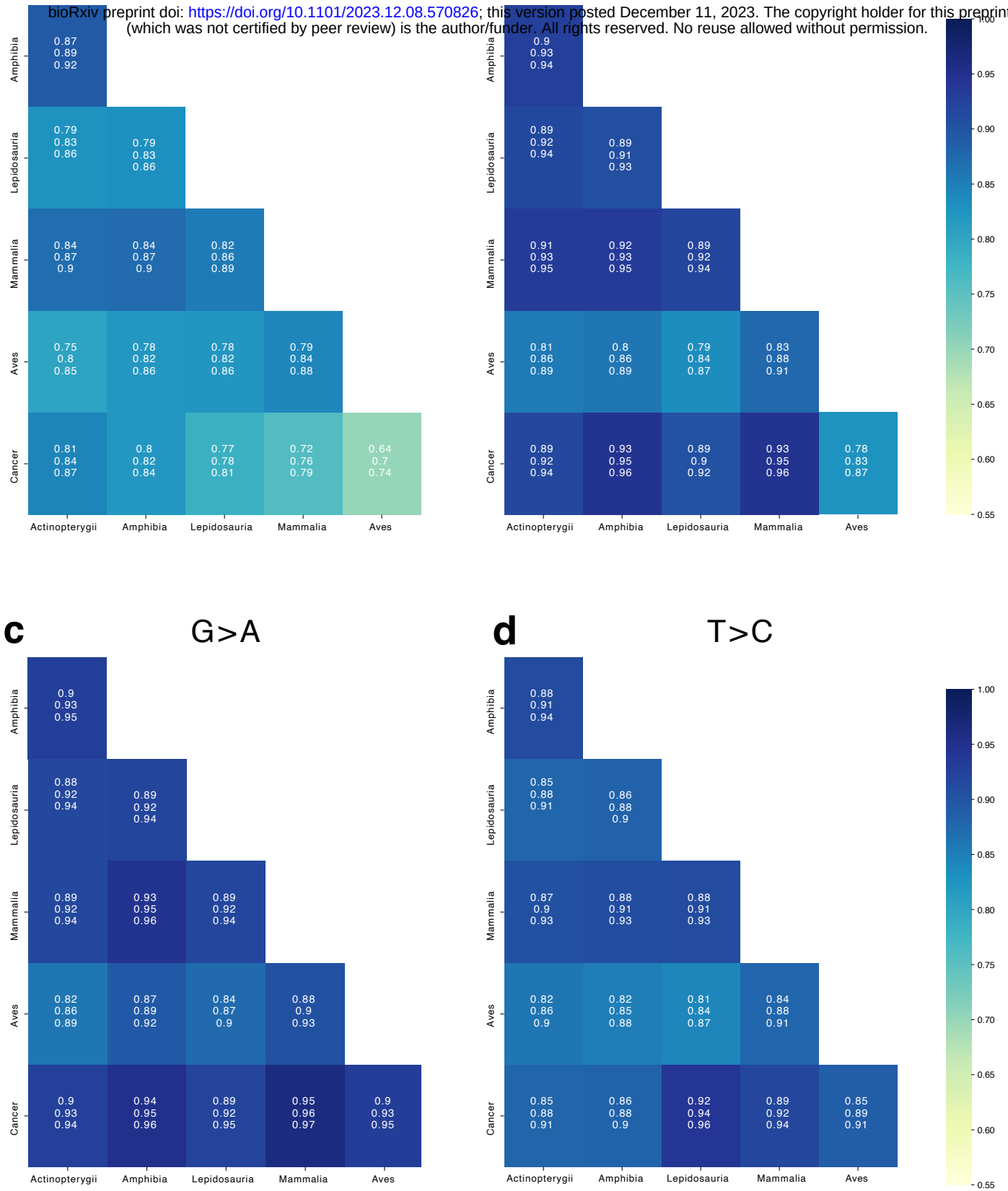


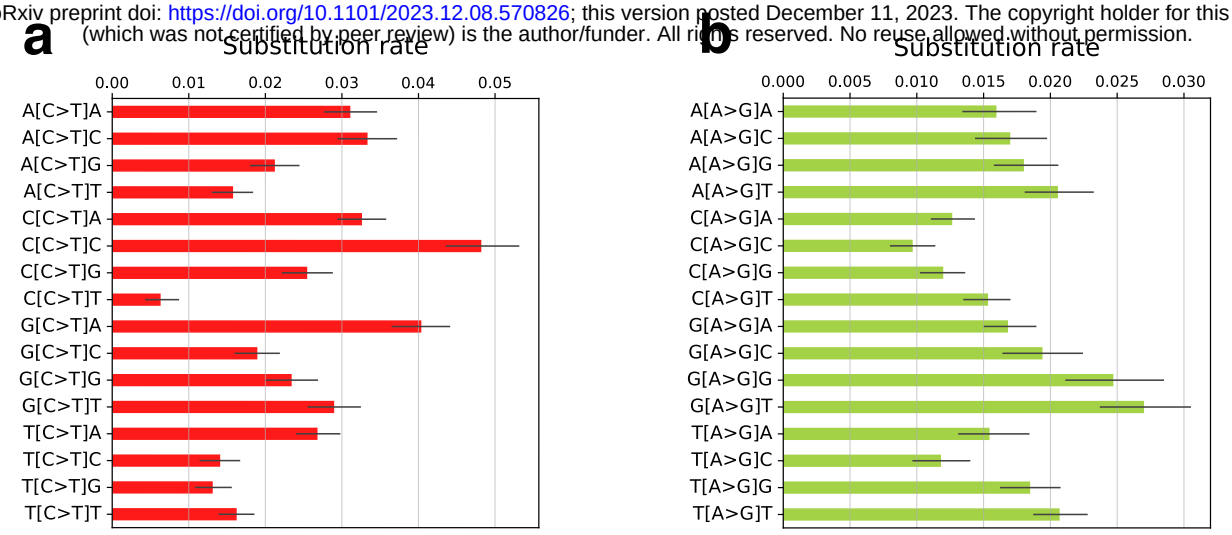




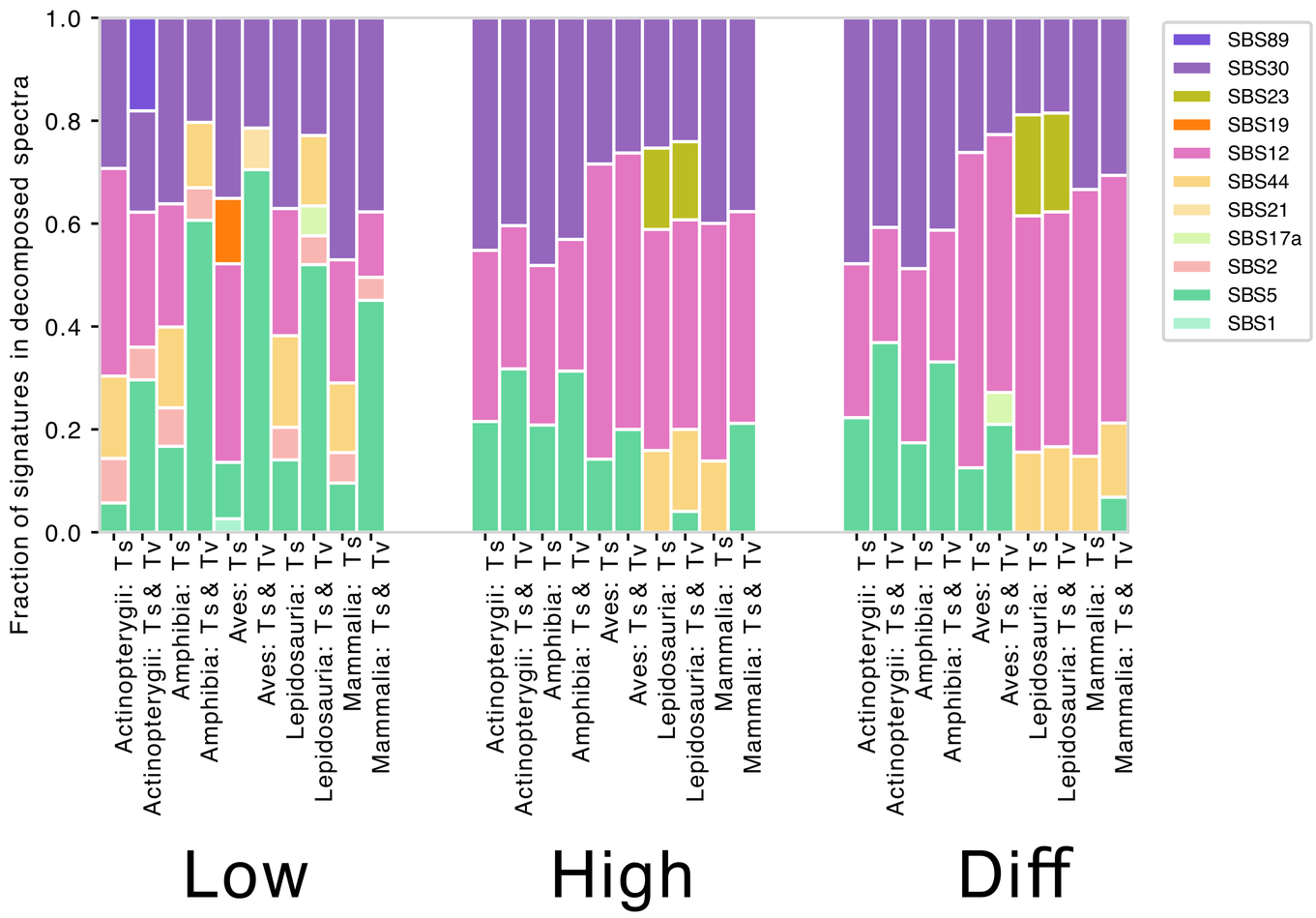




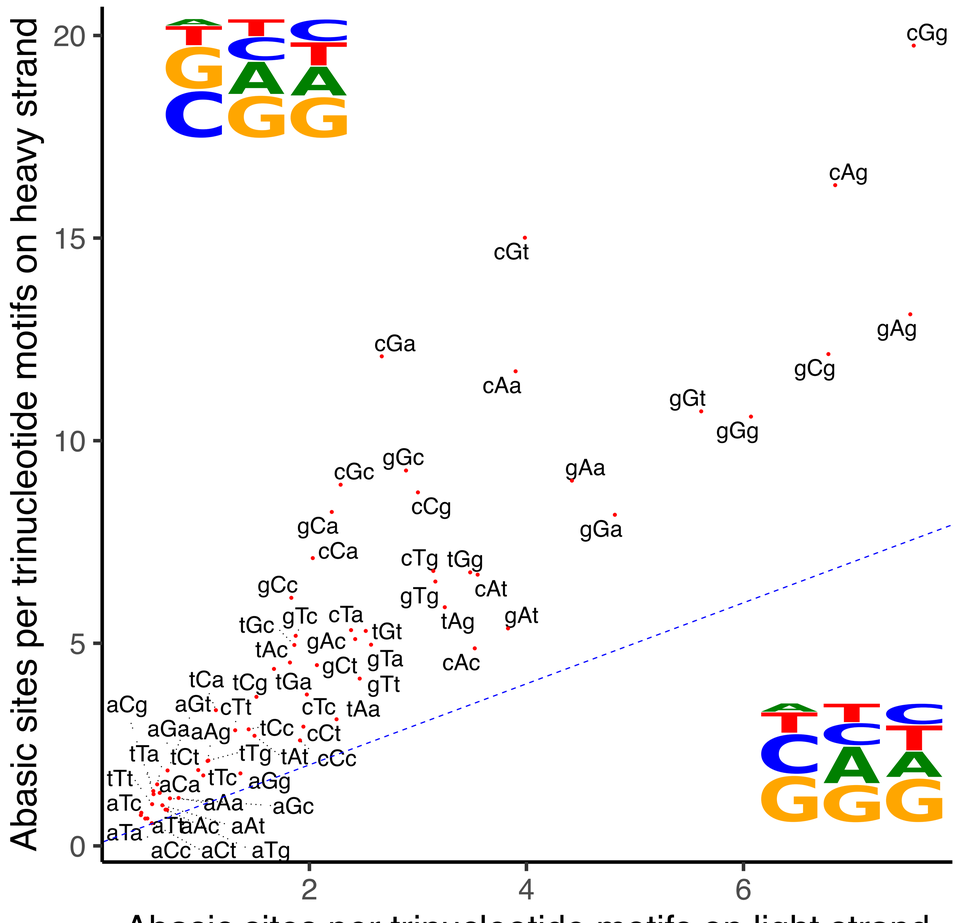




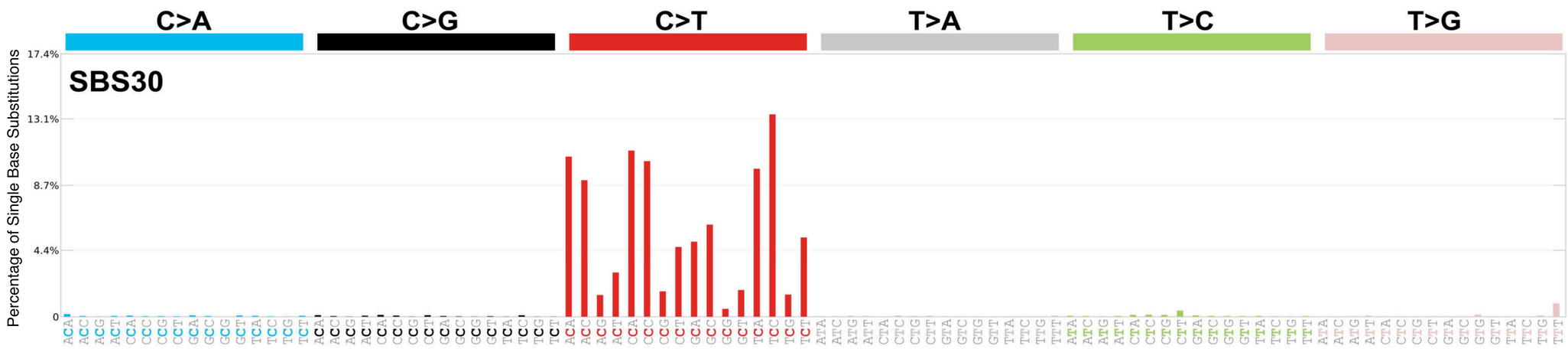
a



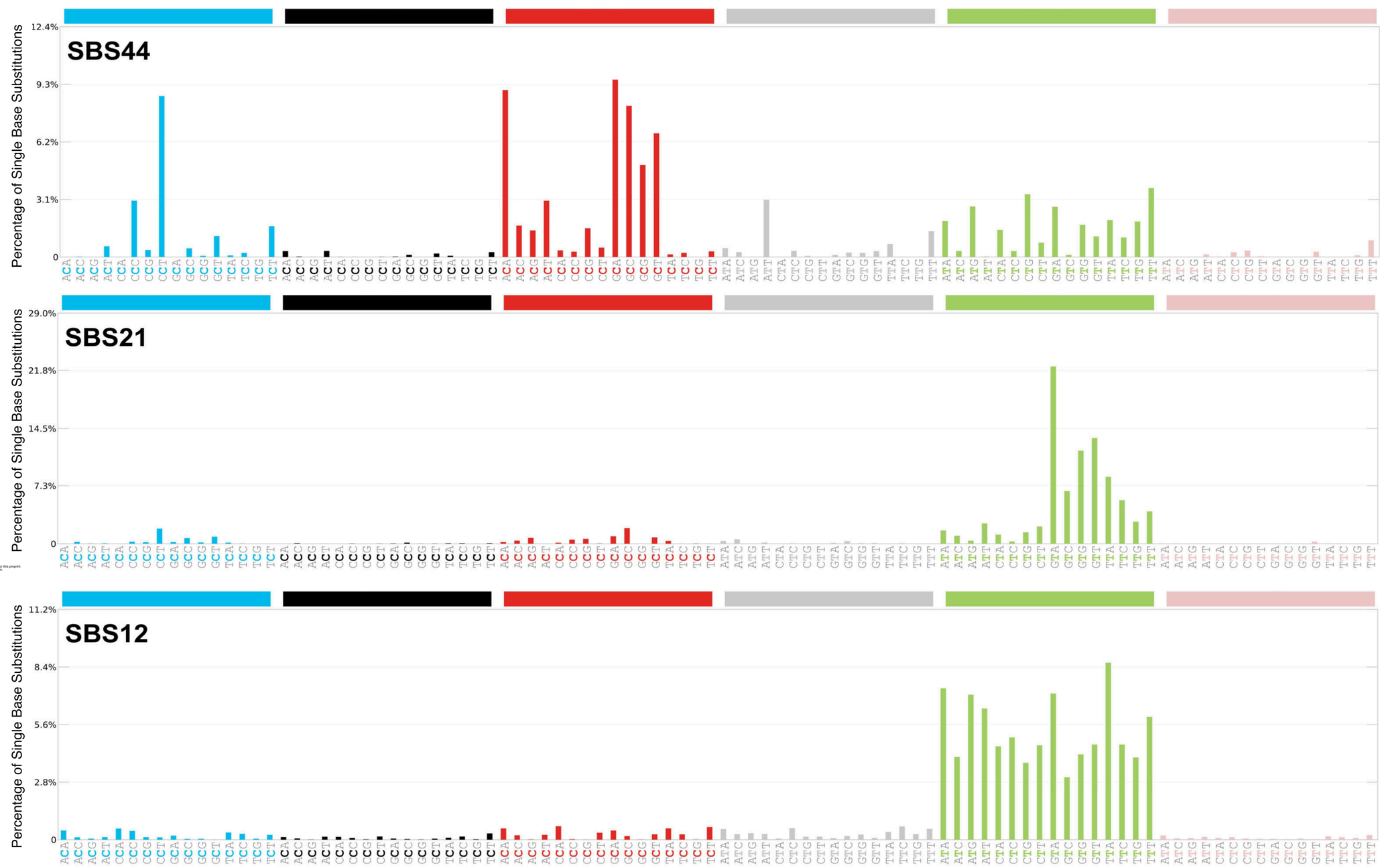
b



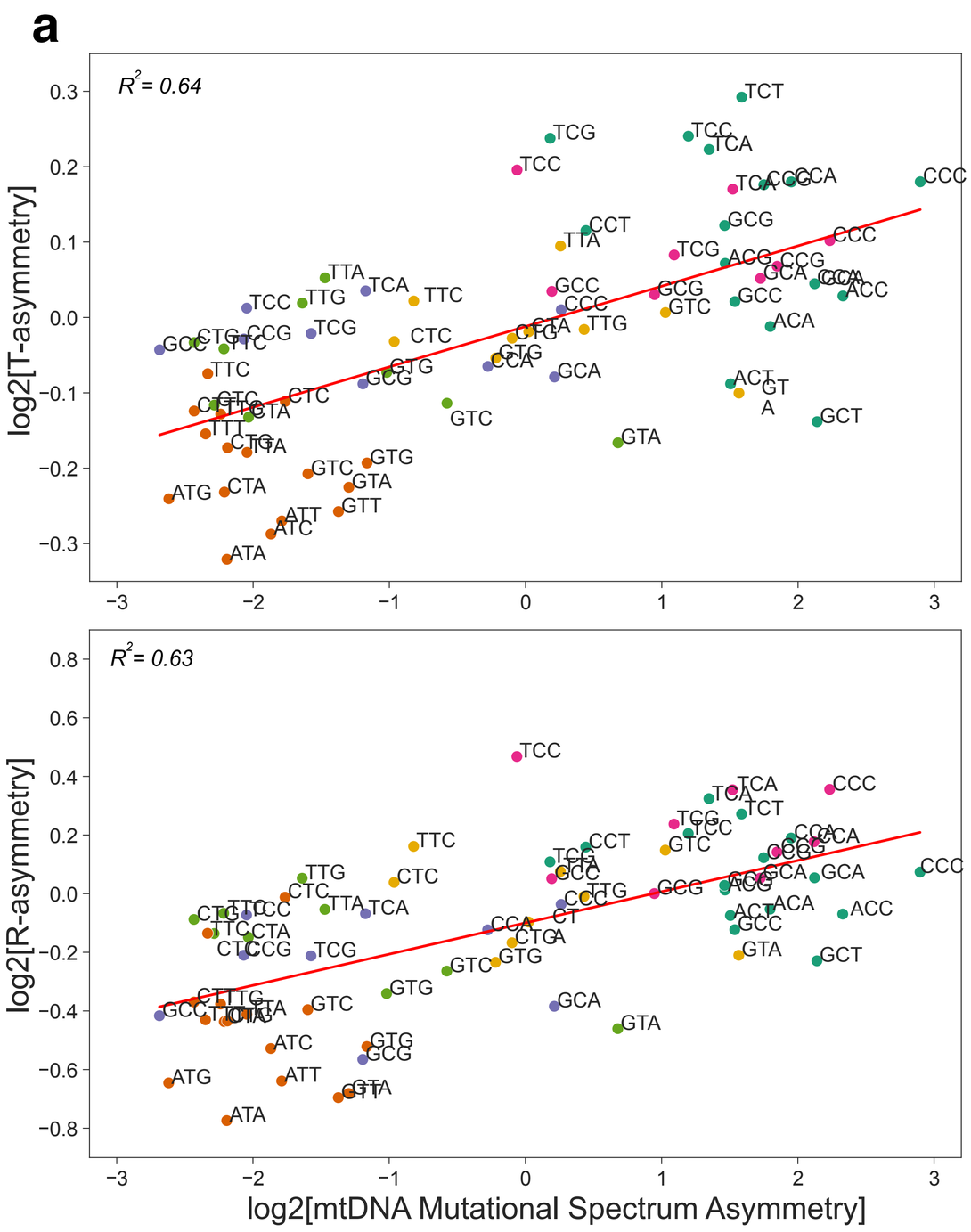
c

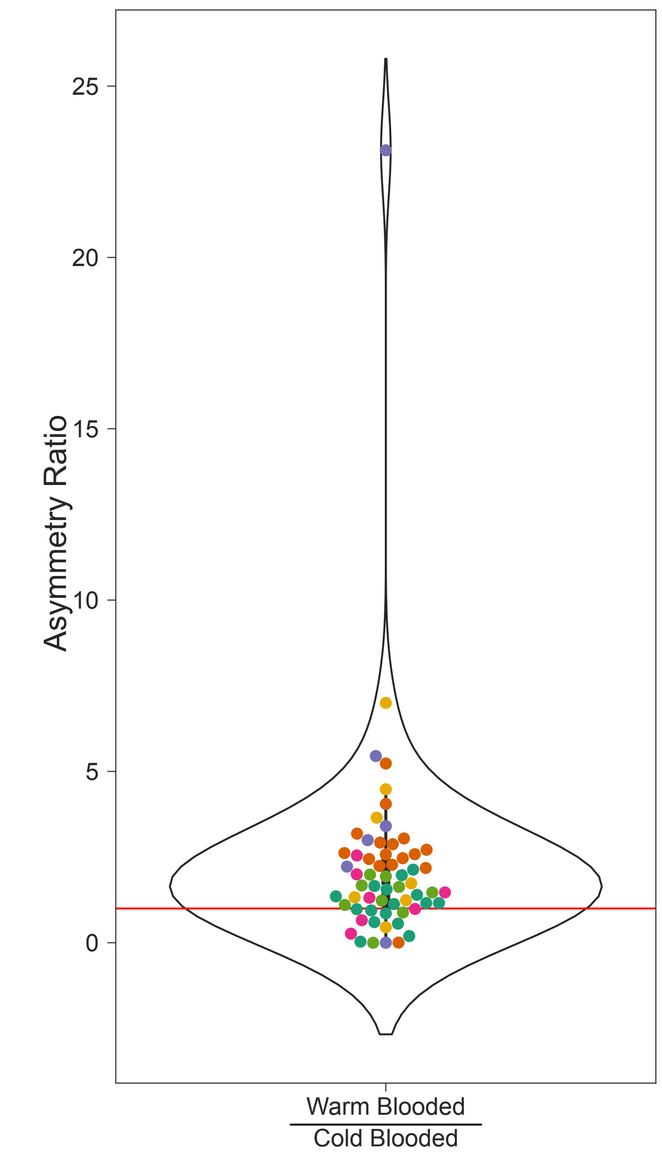


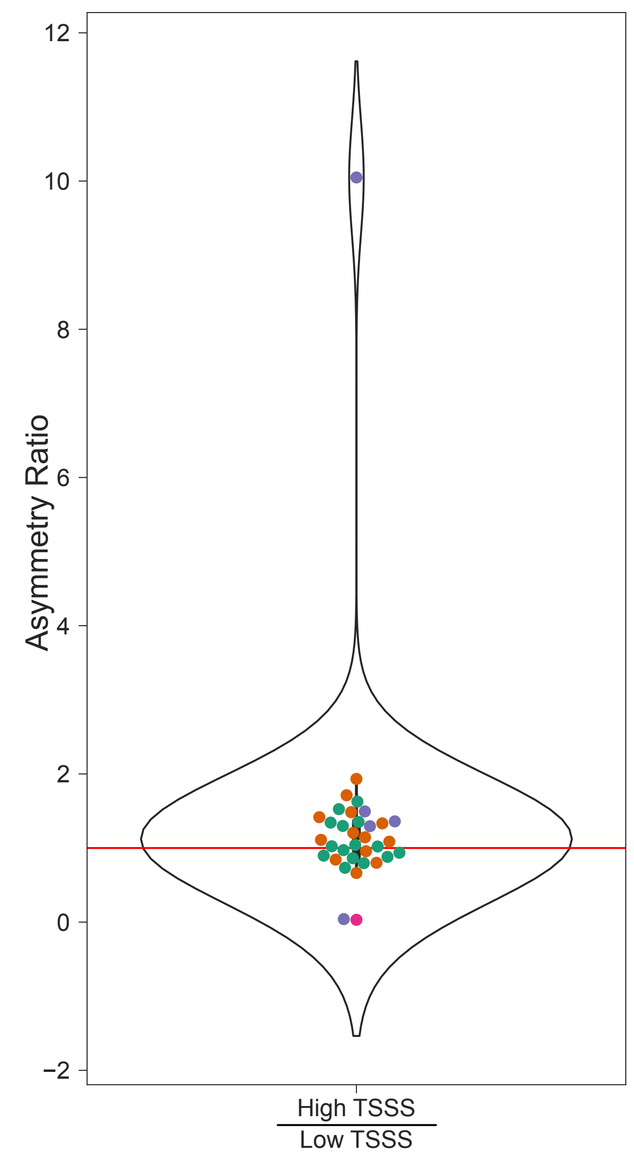
d

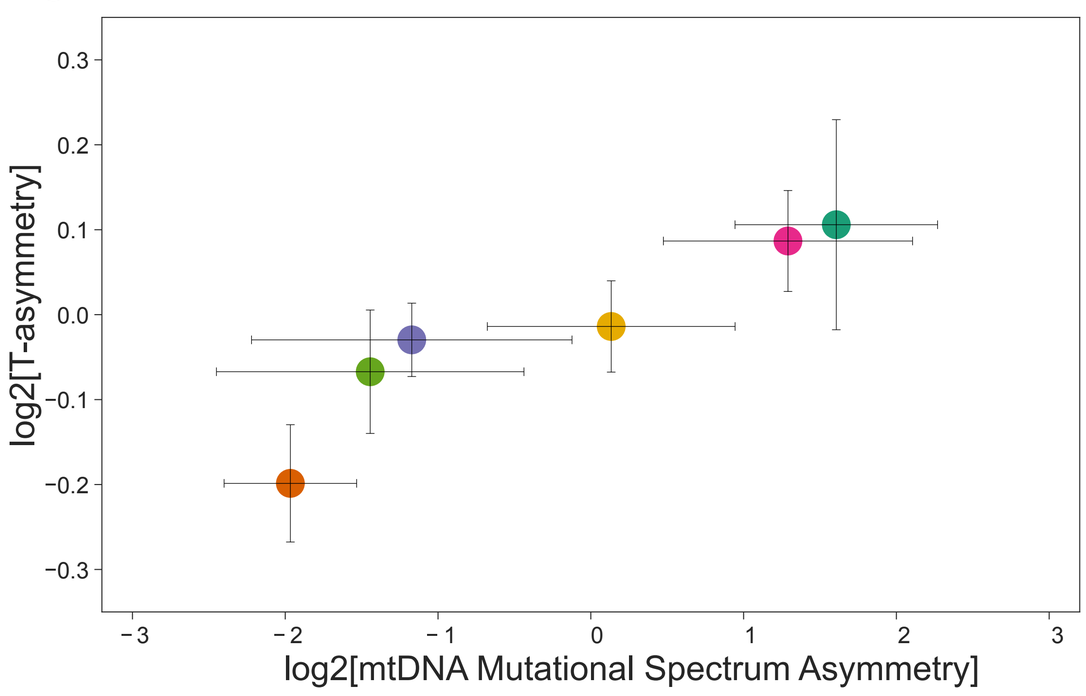


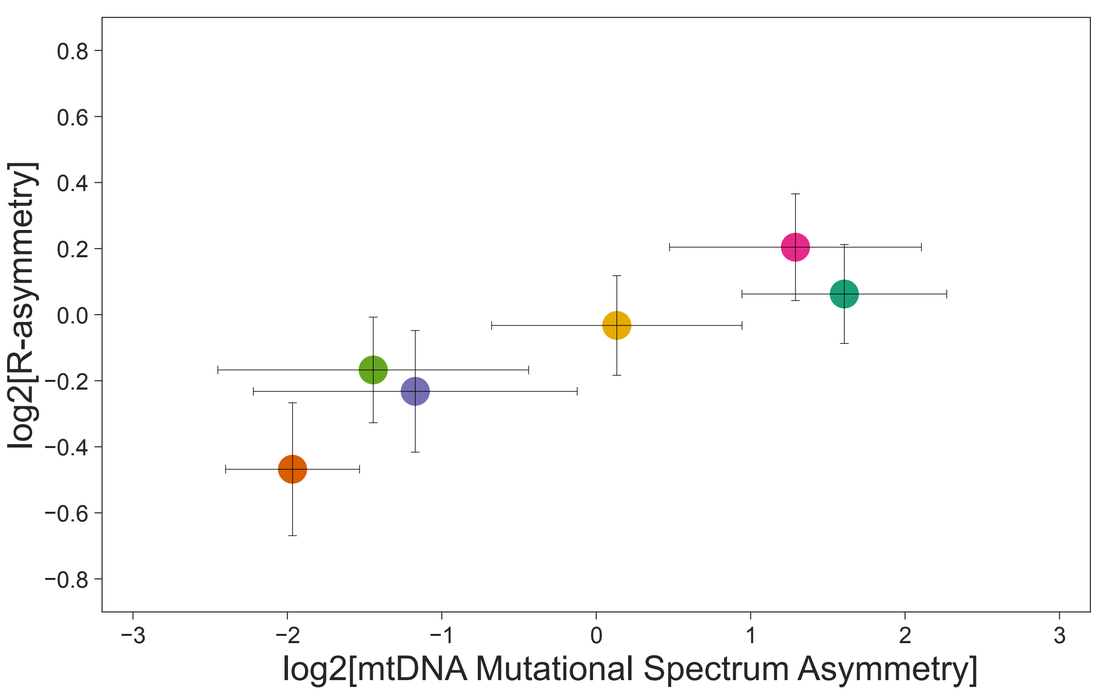
e







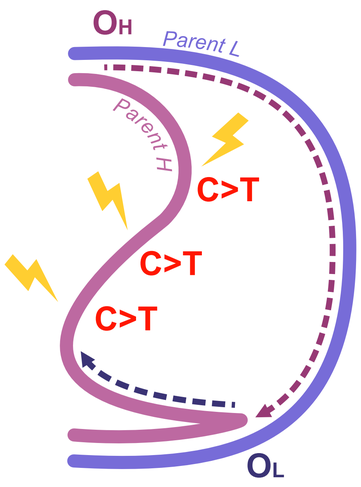




ss-DNA specific C>T

deamination of cytosine + BER deficiency

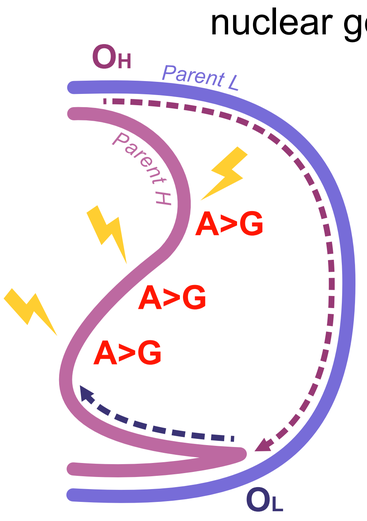
similar to SBS30 in the nuclear genome

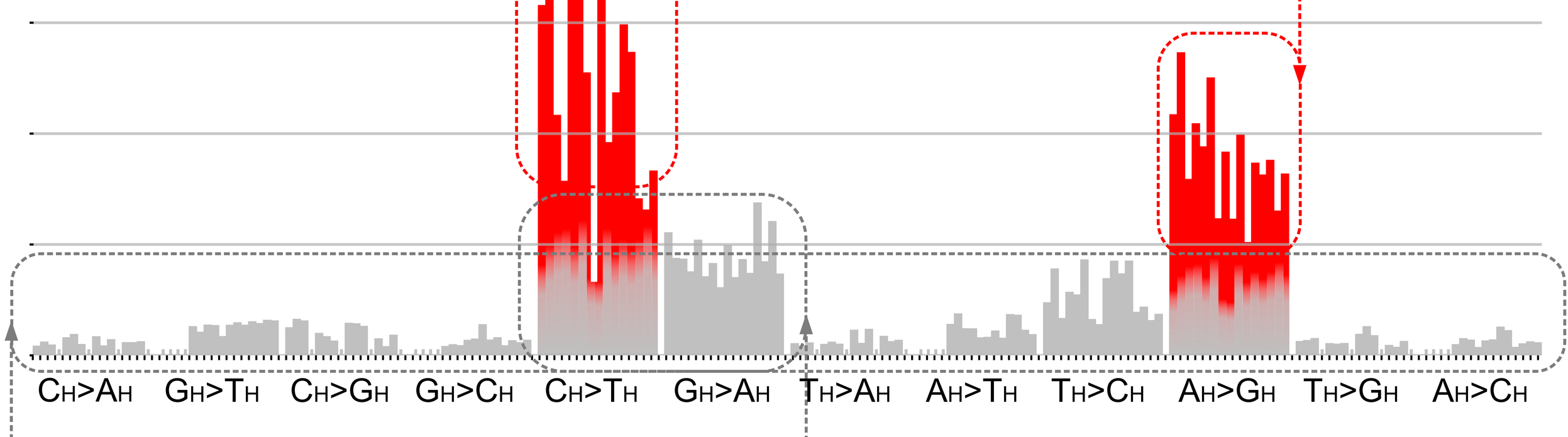


ss-DNA specific A>G

deamination of adenine

similar to SBS12 in the nuclear genome

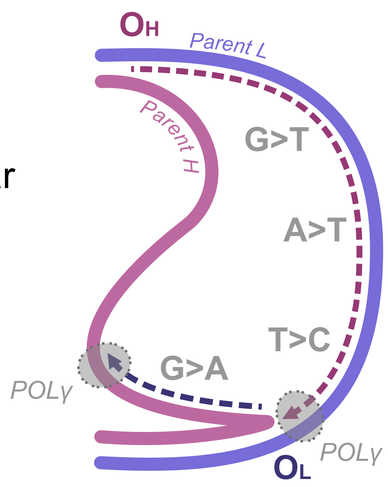




ds-DNA specific mutation

replication errors, more affecting transversions

similar to SBS5 in the nuclear genome



ds-DNA specific C>T

POLG mutations + MMR absence

similar to SBS21 and SBS44 in the nuclear genome