Mitochondrial mutation spectrum in Chordates: damage versus replication signatures, causes, and dynamics

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

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

Abstract

To elucidate the primary factors shaping mitochondrial DNA (mtDNA) mutagenesis, we derived a comprehensive 192-component mtDNA mutational spectrum using 86,149 polymorphic synonymous mutations reconstructed from the CytB gene of 967 chordate species. The mtDNA spectrum analysis provided numerous findings on repair and mutation processes, breaking it down into three main signatures: (i) symmetrical, evenly distributed across both strands, mutations, induced by gamma DNA polymerase (about 50% of all mutations); (ii) asymmetrical, heavy-strand-specific, C>T mutations (about 30%); and (iii) asymmetrical, heavy-strand-specific A>G mutations, influenced by metabolic and age-specific factors (about 20%). We propose that both asymmetrical signatures are driven by single-strand specific damage coupled with inefficient base excision repair on the lagging (heavy) strand of mtDNA. Understanding the detailed mechanisms of this damage is crucial for developing strategies to reduce somatic mtDNA mutational load, which is vital for combating age-related diseases.

Introduction

DNA mutations can be a result of either replication error or damage followed by incorrect repair[1], with a range of mutagens responsible for these changes[2]. Reconstruction of a mutational spectrum, with 96 or 192 components, helps to decompose it into distinct mutational signatures, allowing us to trace the effects of various mutagens[3]. In the human nuclear genome, both germline and somatic spectra, including cancerous ones, have been instrumental in reconstructing and deconvoluting these signatures, leading to major breakthroughs in both fundamental[4] and applied[5] 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[6] and ageing[7]. 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 reactive oxygen species (ROS), UV light nor tobacco smoke in associated cancer data[8]. Furthermore, the mechanisms of mtDNA replication and repair have been under ongoing debates, and recent detailed analysis of the mitochondrial mutational spectrum offered important insights into these processes[9]. On an evolutionary scale, understanding mtDNA mutagenesis and implementation of mtDNA spectrum-aware approaches is crucial for accurate phylogenetic inferences and for uncovering variations in selection processes between species[3,10] as well as for understanding the dynamics of deleterious human mtDNA 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 characteristic of the mutational spectrum[11]. 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)[12–14] and 12-component (doubling the 6-component spectrum to account for asymmetrical mutagenesis on complementary strands)[10] spectra in analyses for comparative studies. However, more comprehensive 96-component (expanding the 6-component spectrum with an inclusion of the nucleotide context) 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[8] and SARS-CoV-2[15]. 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 mtDNA polymorphisms into a comprehensive spectrum representative of all *Chordata*. Reconstructing 86,149 synonymous polymorphic mutations from CytB sequences of 967 chordate species, we compiled a 192-component mtDNA spectrum. This extensive spectrum

enables the exploration of critical questions, related to damage and replication-driven mechanisms of mtDNA mutagenesis, their etiologies, and dynamics.

METHODS

In this paper we use heavy strand notation for all mtDNA mutations and spectra.

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 the majority of mtDNA synonymous polymorphic variants within chordate species are effectively neutral. Although recent suggestions indicate that certain synonymous sites in mtDNA may not be entirely neutral[16], we evaluated the effect of removing highly-constrained synonymous sites on our data. The results suggested that omitting these conservative sites did not markedly change the results (see Supplementary Figure 1).

We followed these steps:

1a) Observed mutations.
Only species with at least five unique sequences of CytB gene from GenBank were used in this work. For each of 967 such species (listed in Supplementary Table 1) the observed mutational spectrum was derived through the following steps: (i) obtaining a codon-based multiple alignment using macse v2[17]; (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[18]; (iv) identifying all single-nucleotide synonymous substitutions on each tree branch; (v) categorising these mutations into 192 groups based on the nucleotide context (16 unique contexts of downstream and upstream nucleotides within each of 12 mutation types).

1b) Expected mutations.
To consider differences in nucleotide and trinucleotide composition, for each of 967 species we calculated the expected mutational spectra. Focusing on the reference CytB gene, obtained from GenBank for each analysed chordates species, we executed an *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.

1c) To obtain species-specific mutational spectra we normalised observed mutations to expected ones. For each species, we calculated the mutation rates by dividing the number of observed substitutions by the number of expected substitutions in the corresponding context, resulting in 192 substitution rates. These 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.
We primarily used cosine similarity to estimate the closeness of the compared spectra. Comparison has been performed on two levels: between species and between classes. Firstly, we conducted pairwise comparisons of mutational spectra at the species-specific level within and between classes, considering all possible combinations. Second, we estimated differences between classes using jackknife resampling of species spectra. 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 of either the overall mutational spectrum or its parts (transitions and transversions). 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) Annotation of mitochondrial mutational signatures.
To decompose class-specific mtDNA mutational spectra into COSMIC signatures we used SigProfilerAssignment tool v0.0.30[19]. Since main COSMIC signatures are symmetrical and do not account for strand-specific mutagenesis, we split the asymmetrical mitochondrial 192-component spectra into two ("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. Also, to analyse the asymmetrical component of mutagenesis, we derived “diff” spectra subtracting “low” from “high” spectra in a context-dependent manner (Supplementary Fig. 6). All complementary pairs of transversions rates (for example, A_H>C_H and T_H>G_H) were averaged and equally added to the "low", “diff” and "high" spectra; otherwise, to test the impact of noisy and rare transversions, they were zeroed in these spectra. In addition, SigProfilerAssignment, like any other SigProfiler tool, is capable of processing distinct substitution counts on the reference human nuclear genome. To ensure correspondence of input spectra to the human nuclear genome, we rescaled spectra of analysed classes multiplying them by the trinucleotide frequencies of the human nuclear genome. We applied SigProfilerAssignment 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.

<u>4) Abasic sites patterns in mtDNA</u>.

The distribution of abasic sites (AP) at single-nucleotide and single strand (separately for light and heavy) resolution of in mouse mitochondrial genome (mm10) was obtained from Cai et al.[20] We quantified the AP sites occurrences within all 64 trinucleotide sequences of each strand of mouse mtDNA, excluding the control region. These values were then adjusted by the counts of trinucleotide motifs in a strand-specific manner.
We created trinucleotide logos using the Python library logomaker[21]. These logos used the normalised AP sites count for each trinucleotide as the nucleotide weights. These weights were averaged across all 64 trinucleotides, and final nucleotide counts at each trinucleotide position were normalised and converted to frequencies.

5) Calculation of the mtDNA mutational spectrum asymmetry

The assessment of mtDNA asymmetry involved the transformation of a 192-component mutational 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 mutational spectrum in human cancers

Mutations in the full human mtDNA were derived from comprehensive analysis of the human mitochondrial genome by Yuan et al.[8] The mutational 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 mutational 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

Variations in mtDNA Mutational Spectrum within Chordates: Importance of Damage

The 192-component mitochondrial mutational spectrum (Fig. 1, Supplementary Table 1) was generated by integrating 86,149 synonymous mutations obtained from CytB sequences across 967 chordate species. Following the exclusion of highly-constrained synonymous sites which did not significantly affect the overall results (Methods, Supplementary Fig. 1) we assume that our mutational spectrum can be considered neutral. The resulting spectrum highlights a significant prevalence of transitions, with C_H>T_H and A_H>G_H being the most frequent types (Fig. 2a). This mutational pattern mirrors findings observed in mammalian germline mutations[10], somatic mutations in human cancers[8], and healthy tissues[22] (Supplementary Table 2).

To assess variation in spectra, we calculated pairwise cosine similarities between all species and revealed no clustering of species spectra within classes (Supplementary Fig. 2). The lack of similarities within class-specific spectra points to fast evolutionary changes in the mtDNA mutation spectra[2]. Such changes can be driven either by variable replication-driven component (if POLG's properties vary strongly between species) or variable damage-driven component (if mtDNA damage is affected by the metabolism-related traits which can be highly variable even among close species). Taking into account the high evolutionary conservatism of chordate POLG[23–25] and rather variable levels of aerobic metabolism in different chordates[26] we propose that the damage effect can be the main reason for the observed high variation. To test it we performed class-specific analyses, expecting that aves, a class characterised on average by the highest level of basal metabolic rate (BMR)[26], can show the most divergent patterns of the spectrum. We reconstructed mutational spectra for each of the five chordate classes (Supplementary Fig. 3), revealing notable similarities across them and to the overall mtDNA mutational pattern (Fig. 2b), suggesting a conserved mutational process within chordates. Further analysis, involving median pairwise cosine similarities between different classes, confirmed that birds (aves) possess the most distinct mtDNA mutational spectrum among all classes (Methods; Fig. 3a), with the lowest cosine similarities to all other chordate classes. This is in line with a hypothesis that the mtDNA mutational spectrum is shaped by metabolism-associated chemical damage[27].

To further test the potential damage effect, we analysed somatic mtDNA mutations from human cancers, which are often hypoxic[28] and, thus, may have less oxidative damage[29]. Conducting pairwise comparisons between human cancers[8] (Methods) and five chordate classes we indeed revealed the least cosine similarity between birds and cancers (Fig. 3a), indicating their strong deviations from each other presumably due to the lowest damage in cancers and the highest damage in birds[30,31]. Altogether we propose that due to the differences in the level of aerobic metabolism all chordate species and classes as well as human cancers are affected by mtDNA damage of differential impact.

Conservative and variable patterns of the mtDNA spectrum: symmetrical replication-driven and asymmetrical damage-driven mutations

To differentiate between stable and variable mtDNA mutational patterns across classes, we analysed transitions and transversions separately. Transitions showed consistently high similarities among chordate classes and human cancers, aligning with the full spectrum trends (Fig. 3b). Conversely, transversions presented lower similarities, likely due to their stochastic occurrence and rarity in our dataset (Fig. 3c; Supplementary Fig. 4). Subsequently, we focused on an analysis of each transition type separately.

Cosine similarities across classes were uniformly high for each of the four transition types (Fig. 4), yet interesting variations were observed. Notably, G_H>A_H exhibited higher similarity between classes than C_H>T_H, despite being complementary equivalents. This indicates that symmetrical C>T mutations on double-stranded DNA (dsDNA) (approximated by G_H>A_H, Fig. 2c) are more conserved compared to the asymmetrical part of C>T mutation in single-stranded DNA (ssDNA) (approximated by the difference between C_H>T_H and G_H>A_H, Fig. 2c). We suggest that the conserved pattern of symmetrical C>T substitutions (equal to G_H>A_H) primarily results from internal replication errors, introduced by POLG. Conversely, the less conserved pattern of asymmetrical C>T mutations (C_H>T_H minus G_H>A_H) on ssDNA, particularly divergent between aves and cancer, may be influenced by chemical damage including spontaneous deamination or oxydation.

Taking into account that asymmetrical C_H>T_H (i.e. C_H>T_H minus G_H>A_H) is most likely associated with chemical damage such as spontaneous deamination or oxidation at ssDNA[9], we explored the profile of the C_H>T_H substitutions in detail. A yeast experiment with ssDNA and ROS (oxygen peroxide and paraquat) demonstrated that cCc>cTc mutations are a primary target of oxidative damage in ssDNA[32]. Notably, completely in line with these experimental results, we consistently identified cCc>cTc as the most common motif for C_H>T_H substitutions in all chordate classes (Supplementary Fig. 5a). Therefore, the cCc>cTc mutation, the most prevalent in mtDNA across all chordates, likely arises not only from spontaneous deamination but also includes contributions from oxidative or other types of damage to ssDNA. Conversely, cancer data again demonstrates a distinct pattern: asymmetrical C_H>T_H substitutions predominantly appear in the nCg context rather than cCc (Supplementary Fig. 5c), suggesting influences from potentially lower levels of oxidative damage[7] or from cytosine methylation in mtDNA[33], paralleling the well-known CpG > TpG substitutions in the nuclear genome.

Analysing the local patterns of A>G substitutions we observed that nAt>nGt and nAg>nGg are the most common motifs across all chordates (Supplementary Fig. 5b). Conversely, it alters in cancers, demonstrating once again diverse environments in somatic cancer cells relative to germ-line tissues (Supplementary Fig. 5d). The A_H>G_H mutation, recently linked to ageing in mammals[10] and body temperature in chordates[34], suggests a signature of damage related to aerobic metabolism[10], yet the exact process driving these substitutions is still unknown.

The potential link to this mutation can be due to the N6-methyldeoxyadenosine (6mA) in mtDNA, being strongly enriched on the heavy strand[35] and associated with stressful hypoxic conditions. Although enrichment in 6mA in mammalian mitochondria is still contradictory[36,37] and a link between 6mA and A_H>G_H is rather suggestive, we observed that the 6mA motifs described as (c/a)At and A(t/g) in previous studies[38,35] are similar to our A_H>G_H motifs nAt>nGt and nAg>nGg (Supplementary Fig. 5b). Future studies are needed to uncover a mechanism behind A_H>G_H substitutions in mtDNA.

Altogether, splitting mtDNA spectrum into conservative and variable parts we propose that the symmetrical part (C_H>T_H equal to G_H>A_H) can be shaped by replication-driven POLG-induced mutations while the asymmetrical part (C_H>T_H minus G_H>A_H and A_H>G_H minus T_H>C_H) is most likely shaped by damage-driven mutations, induced by deamination, oxidation, or methylation.

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

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 (Supplementary Fig. 6; Methods): "high" spectrum for both symmetric and asymmetric transitions (substitutions with the highest rate out of two complementary substitutions: nCn>nTn and nAn>nGn), "low" spectrum for symmetric transitions (substitutions with the lowest rate out of two complementary substitutions: nGn>nAn and nTn>nCn), and "diff" spectrum, typical for the heavy strand ssDNA, calculated by subtraction of "low" spectrum from "high" (nCn>nTn minus nGn>nAn and nAn>nGn minus nTn>nCn). All transversions were complementary averaged and added to each set. Given the rarity and variability of transversions in mtDNA, we conducted analyses with all substitutions and separately with only transitions, which confirmed that excluding transversions does not significantly change our findings. We observed that five signatures, namely SBS30, SBS44, SBS21, SBS5 and SBS12 are predominant in mtDNA mutations with varying contributions to high, low and diff spectra (Fig. 5a). SBS5 shows a consistently uniform signature, more pronounced when including transversions .

SBS30 is associated with deficient base excision repair (BER) in the nuclear genome and predominantly results in C>T mutations[39,40]. In mitochondria, BER is the primary repair pathway for chemically damaged bases, including deaminated, oxidised, and alkylated bases[41]. Our findings indicate that in comparison with the nuclear genome, BER in mtDNA is less efficient, and leads to C_H>T_H mutations, particularly in conditions of ssDNA ("high" and "diff" spectra, see Fig. 5a), when there is no complementary strand to replace removed damaged bases. Another notable observation is the association of SBS30 in the nuclear genome with malfunctioning NTHL1 glycosylase, suggesting an inherent decreased efficiency of this enzyme in mtDNA.

NTHL1 removes oxidised pyrimidine lesions in both nuclear DNA (nDNA)[42] and mtDNA[43]; significantly decreased NTHL1 activity in Friedreich's ataxia leads to the accumulation of mtDNA mutations dramatically affecting mitochondrial functioning[44]. These observations suggest that NTHL1 could serve as a strategic target for mitigating the mitochondrial mutation rate in clinical settings.

Further insights into mtDNA mutagenesis can be obtained by examining abasic sites (AP sites), which emerge either as an intermediate step of the BER process or spontaneous base loss. In cases where BER is ineffective on ssDNA and there is no glycosylase activity, AP site motifs resembling SBS30 patterns (C_H>T_H mutations) are not expected. Spontaneous base loss will show distinct tendency: depurination (A and G loss) outpaces depyrimidination (C and T loss) by 20-fold, guanines are 1.5 times more prone to depurination than adenines, depurination happens faster in ssDNA than in dsDNA[45]. Analysing AP sites in mouse mtDNA, with precise single-nucleotide resolution on both strands[20], we observed that motifs with AP sites correlate well between the light and heavy strands, occurring about twice as often on the heavy strand; with Guanine being the most common AP site, followed by Adenine (Fig. 5b). This pattern is more consistent with spontaneous base loss, not BER. Consequently, due to BER deficiency, we do not detect AP sites corresponding to SBS30 motifs. Rather, we find motifs indicative of spontaneous base loss (G and A loss). This process, upon replication by POLG that incorporates a dA residue opposite the abasic site[46], can result in G>T and A>T transversions. The asymmetry observed for G>T and A>T transversions (Fig. 2a) goes in line with a predominance of AP sites on the heavy strand (Fig. 5b).

SBS44 and SBS21 are two of seven known signatures linked to defective DNA mismatch repair (MMR), crucial for correcting mismatches during DNA replication[47]. Although the presence of MMR in mtDNA has been debated, it is now widely accepted that there is no MMR in mtDNA[41]. 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 symmetric polymerase errors (Fig. 5a). We suggest that MMR-deficiency signatures, i.e. the symmetrical part of C>T (equal to G_H>A_H) (Fig. 5d) is shaped by the gamma DNA polymerase, which is expected to introduce symmetrical mutations[40,46,48].

SBS12, despite its unknown origins in the nuclear genome, can be a hallmark of chemically damaged adenines[22] in mtDNA's single-stranded heavy strand due to several reasons. First, SBS12 is based on A_H>G_H mutations (T>C in COSMIC notation, Fig. 5e), which in mtDNA are sensitive to age and temperature[10,49]. Second, SBS12 is presented in the "high" and "diff" spectra (Fig. 5a), reflecting its impact on the heavy strand. Third, SBS12 shows: an increase with replication timing in the nuclear genome ("Replication timing" section in COSMIC), 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). Fourth, SBS12 has the highest prevalence in birds (Fig. 5a), known for elevated

metabolic rates, which further underscores its association with chemical damage. SBS12 therefore offers great promise for uncovering the mechanism of mitochondrial ssDNA-specific damage.

Our analysis highlights three primary mutational signatures in mtDNA: (i) Symmetrical $C_H>T_H$ (and complementary $G_H>A_H$) mutations, reflecting mtDNA polymerase errors due to the lack of MMR in mtDNA; (ii) Asymmetric $C_H>T_H$ mutations, likely resulting from cytosine modifications coupled with inefficient BER on ssDNA; (iii) Asymmetric $A_H>G_H$ mutations, linked to metabolism-related adenine damage on ssDNA.

Strong asymmetry in mtDNA mutagenesis is shaped by single-strand DNA damage coupled with Base Excision Repair deficiency during asynchronous replication

Throughout our paper we have repeatedly demonstrated the pronounced asymmetry of mtDNA mutagenesis: (i) the most common transitions C>T and A>G occur several times more often on heavy strand; (ii) transversions G>T, A>T and C>G also demonstrate increased frequencies on heavy strand in our (Fig. 2a) and other studies[8]; (iii) both mostly pronounced signatures, BER deficient SBS30 and mito-specific SBS12, are highly asymmetrical with a much stronger impact on a heavy strand (Fig. 5a-b). Estimation of the total level of asymmetry in mtDNA (Methods) shows that approximately 50% of all mtDNA mutations are asymmetrical, i.e. occur exclusively on a heavy stand, while the rest of mutations occur symmetrically on both heavy and light strands. Here, we analyse deeper the phenomenon of asymmetry to identify its primary causes.

In the nuclear genome, two types of mutational asymmetry: 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) have been described and a tight correlation between both of them has been shown[4]. We calculated mito-asymmetry from our 192-component spectra (Methods) and observed a significant correlation between mito-asymmetry and both R- and T-asymmetries of the nDNA (Fig. 6a). Our findings show that mutations prevalent in the mtDNA heavy strand are also common in the lagging and non-transcribed strands of the nDNA. Further analysis revealed that the correlations (Fig. 6a) are primarily driven by 6 base-specific asymmetries (Fig. 6c), not by their context (Supplementary Fig. 7-8). This indicates, that for example C>T mutations, regardless of context, occur more frequently than G>A across all three areas: (i) mtDNA heavy strand, (ii) nDNA lagging strand, and (iii) nDNA non-transcribed strand. Categorising the six mutation pairs by the degree of asymmetry, from highest to lowest, with the first mutation in each pair occurring more frequently, we got the next ranking: T>C, C>T, C>A, T>A, C>G, T>G (Fig. 6c). Interestingly two the most asymmetrical substitutions, C>T and A>G, are also the most common in the integral mtDNA mutational spectrum (Fig. 2a)

What is the plausible mechanism for the asymmetry's origin? The shared characteristic of the three areas described above is their single-stranded nature, suggesting a uniform mutational

process influenced by ssDNA damage. If single-stranded specific damage[9,50,51] is a viable hypothesis, we anticipate asymmetry to grow with (i) increased time spent single-stranded (TSSS) and (ii) increased total damage level. Assuming that TSSS is linearly increasing during asynchronous replication of mtDNA along the major arc[9], we analysed human cancer mtDNA data[8]. Dividing the major arc into low and high TSSS zones (Methods) we revealed that asymmetry significantly increases in high TSSS areas, with the high to low TSSS asymmetry ratio exceeding one (Fig. 6b, right panel). Testing our hypothesis of different sources of mDNA damage, we leverage the convention that mitochondrial damage is linked to aerobic metabolic rates[7,52]: thus, mito-asymmetry in warm-blooded species should surpass that in cold-blooded ones. Comparing mito-asymmetry within the same gene (CytB) between the warmest (birds) and the coldest[34] (fishes) chordata species in our dataset, we indeed observed an asymmetry ratio of warm to cold species greater than one (Fig. 6b, left panel).

While an alternative explanation for asymmetry, proposed for nDNA, involves low-fidelity translesion DNA synthesis (TLS) polymerases capable of error-prone bypassing of DNA lesions[4], we find it less likely for mito-asymmetry. First, PrimPol, an error-prone TLS polymerase observed in mammalian mitochondria[53], is rarely recruited and has a strong preference for generating base insertions and deletions[54], which are rarely observed in mtDNA[8]. Second, the deamination of C and A on the heavy chain of mtDNA, resulting in the most common and asymmetrical mutations, C>T and A>G, are not expected to be helix-distorting changes that stall replication forks and necessitate PrimPol recruitment (see Zheng et al. 2006 [55]). In fact, C>T substitutions are the most common among POLG-mediated errors in *in vitro* experiment[55], suggesting that POLG can make C>T transitions via cytosine deamination without issues (Supplementary Fig. 9). In summary, we suggest that ssDNA damage is a key factor contributing to mutational asymmetry in mtDNA and potentially has some influence on nDNA as well.

DISCUSSION

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

The first component, driven by POLG replication errors, comprises about 50% of all *de novo* mutations and includes C>T symmetrical mutations (SBS44 and SBS21 signatures, indicative of MMR deficiency in nDNA), A>G symmetrical mutations, and a symmetrical part of the majority of transversions (SBS5-like signature), shown in Fig. 7 in grey colour. The most prevalent type within this component is the symmetrical C>T substitutions, which also shows the most conservative pattern across chordate classes (see G_H>A_H in Fig. 4c) and is known as the most common mutation for all polymerases including POLG due to the mispairing of T opposite to G, leading to C>T mutations[40,48]. If the frequency of symmetrical C>T (G_H>A_H) mutations and most transversions increases with each mtDNA replication round, we expect a positive correlation among mutation types in the component. Interestingly, this correlation has been confirmed recently by a principal component analysis in our comparative-species study, which shows collinearity between the majority of transversions and G_H>A_H substitutions (Fig. 2c in Mihailova et al. 2022[10]).

The second component, representing roughly 30% of mutations and depicted in red in Fig. 7, comprises asymmetrical C>T mutations due to damage (SBS30 signature, indicating BER deficiency in nDNA). It is likely shaped by ssDNA damage caused by spontaneous deamination and oxidative stress, aggravated by deficient BER on the single-stranded heavy strand (Fig. 5a,c). The association of these mutations with dysfunctional NTHL1 in nDNA (Fig. 5), a glycosylase addressing oxidised pyrimidine lesions, and the prominent cCc>cTc signature (Supplementary Fig. 5a,c), experimentally confirmed as indicative of oxidative damage on ssDNA, suggest a significant oxidised component of these C>T mutations. Despite their high prevalence in mtDNA, these mutations maintain a rather constant rate across species, showing no sensitivity to metabolic or life-history changes[10,34]. This component's insensitivity to metabolic or life-history traits and replication, highlighted by its distinct position relative to both the first component (G_H>A_H mutations and the majority of transversions) and the third component (A_H>G_H) in our comparative species PCA plot (Fig. 2c in Mihailova et al. 2022[10]), suggests its potential as a molecular clock in mtDNA, similar to SBS1 in nDNA, which is characterised by C>T mutations in a CpG context due to deamination of 5-methylcytosine (SBS1). Although the methylation of cytosine in mtDNA remains uncertain and probably low[33], somatic mtDNA mutations C>T in cancers clearly show a CpG context, indicating that formation of 5-methylcytosine and subsequent deamination may occur (Supplementary Fig. 5c, see Yuan et al. 2020[8]).

The third component, representing around 20% of the mutation spectrum, consists of asymmetrical A_H>G_H mutations (depicted in red in Fig. 7), associated with ssDNA damage, primarily attributed to adenosine deamination. This mutation type, in contrast to the second

component, exhibits notable correlations with eco-physiological traits in mammals[10] and across chordates[49]. Its distinctiveness may be linked to its status as the most asymmetrical mutation in mtDNA, indicative of ssDNA damage. Moreover, chemical studies show that adenine is the most susceptible to deamination nucleotide[56], and the process of deamination quickens at higher temperatures[57] and in alkaline conditions[56]. This temperature dependency could account for the increased $A_H>G_H$ mutations and $A_H>G_H$ asymmetry, estimated as $A_H>G_H/T_H>C_H$ ratio, in warm-blooded compared to cold-blooded chordates[49], while pH dependency might explain an association of $A_H>G_H$ with longevity and ageing in mammals[10]. All these findings, alongside the similarity of the third component to SBS12 and the potential implications of adenosine methylation, highlight important avenues for further research.

The categorisation of the mtDNA mutational spectrum into replication-driven, ssDNA damage-driven molecular clock-like and ssDNA damage-driven metabolism-associated components can enrich our understanding of species' molecular evolution. It may reveal new aspects of species' origins, history, and potential genetic factors influencing mutation rates, as it was shown in comparative studies on nDNA[2,58–61]. Additionally, these components offer insights into the dynamics of somatic mtDNA mutation patterns in different cancer and healthy tissues[62].

Understanding the primary mutagens in mtDNA mutagenesis can guide strategies to reduce mtDNA mutation load, causative for many human diseases and ageing. For instance, evidence that damage contributes to up to 50% of de novo mutations suggests approaches to mitigate damage-induced mutations, either by reducing damage (for example by inducing hypoxia[29]) or by enhancing repair processes, like boosting mitochondrial BER via upregulation of NTHL1[44]. Further research is required to elucidate the damage processes impacting single-stranded mtDNA.

Acknowledgements

This work was supported by the Federal Academic Leadership Program Priority 2030 at the Immanuel Kant Baltic Federal University (to D.I. and K.P.). A.G.M. and B.E. 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. I.M. is supported by the Russian Science Foundation grant No. 21-75-10081. V.S. is supported by the Ministry of Science and Higher Education of the Russian Federation (agreement no. 075-15-2021-1084)
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., P.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.

Conflict of interest statement. None declared.

Figure 1. Pipeline overview. For 967 chordate species two groups of mutations were obtained from CytB gene: observed synonymous mutations (from our polymorphic database with total number of substitutions 86,149; Species 1, left column) and expected synonymous substitutions (based on NCBI RefSeq database; Species 1, right column). Both groups of mutations were normalised and integrated into a comprehensive 192-component spectrum for all chordates (Methods).

Figure 2. Integral mtDNA mutational spectrum of Chordata. (a) 12-component mutational spectrum (n=967); (b) 192-component mutational spectrum (n=967). 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.
(c) A scheme, visualising symmetrical and asymmetrical parts of a given mutation.

Figure 3. Cosine similarities in pairwise comparisons of somatic and germline variants across five chordate classes and human cancer. Each box presents three values: Q1, Q2 (median), 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. Comparative cosine similarity analysis of transition types across chordate classes and human cancer. Building on the conceptual framework of Figure 3, this figure delineates the cosine similarities for four transition types - (a) $C_H>T_H$, (b) $A_H>G_H$, (c) $G_H>A_H$, and (d) $T_H>C_H$ - across five chordate classes and human cancer mutations. Each transition type is analysed with all possible nucleotide contexts (n=16). The results indicate a high degree of similarity for all transition types examined. Notably, the mutational base $G_H>A_H$ exhibits more conservation compared to the $C_H>T_H$ substitution.

Figure 5. The 192-component mutational spectra deconvoluted into COSMIC SBS signatures. (a) Five signatures SBS30, SBS44, SBS21, SBS5, and SBS12 predominate in spectra within three distinct mutation sets: high, low and diff, as detailed in the main text and Methods section.
(b) Analysis of the trinucleotide pattern of AP 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[63].
(d) Pattern of SBS44 and SBS21 signatures: MMR deficiency mutations[63].

(e) Pattern of SBS12 signature: ssDNA-specific mutations[63].

Figure 6. Comparison of the mitochondrial asymmetry based on the global mitochondrial mutational spectrum with the T and R asymmetries in nDNA.
(a) The analysis of the mitochondrial asymmetry 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 mitochondrial asymmetries between warm- and cold-blooded species shows an increased strength of the asymmetry in warm-blooded (the ratio is higher than one) (Wilcoxon test, p = 1.56e-06, N=61). Similar trend is observed for high versus low TSSS regions (Wilcoxon test, p = 0.06, N=34).
(c) Among the six primary substitution types, a robust association is observed between mito-asymmetry and T (left panel) and R (right panel) asymmetries, superseding the influence of nucleotide content.

Figure 7. Graphical visualisation of the main mutational signatures in mtDNA: (i) symmetrical mutations, predominantly C>T, A>G and rare transversions (grey component), linked to POLG'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. Chintalapati, M. & Moorjani, P. Evolution of the mutation rate across primates. *Curr. Opin. Genet. Dev.* **62**, 58–64 (2020).
3. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
4. 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).
5. 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).
6. Taylor, R. W. & Turnbull, D. M. MITOCHONDRIAL DNA MUTATIONS IN HUMAN DISEASE. *Nat. Rev. Genet.* **6**, 389 (2005).
7. Ericson, N. G. *et al.* Decreased mitochondrial DNA mutagenesis in human colorectal cancer. *PLoS Genet.* **8**, e1002689 (2012).
8. Yuan, Y. *et al.* Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nat. Genet.* **52**, 342–352 (2020).
9. 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).
10. 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).

11. 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).
12. 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).
13. Saclier, N. *et al.* Bedrock radioactivity influences the rate and spectrum of mutation. *Elife* **9**, (2020).
14. 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).
15. Yi, K. *et al.* Mutational spectrum of SARS-CoV-2 during the global pandemic. *Exp. Mol. Med.* **53**, 1229–1237 (2021).
16. Caleb A. Lareau *et al.* Codon affinity in mitochondrial DNA shapes evolutionary and somatic fitness. *bioRxiv* 2023.04.23.537997 (2023).
17. 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).
18. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
19. Islam, S. M. A. *et al.* Uncovering novel mutational signatures by extraction with SigProfilerExtractor. *Cell Genom* **2**, None (2022).

20. 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).

21. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).

22. Sanchez-Contreras, M. *et al.* The multi-tissue landscape of somatic mtDNA mutations indicates tissue-specific accumulation and removal in aging. *Elife* **12**, (2023).

23. Czernecki, D., Nourisson, A., Legrand, P. & Delarue, M. Reclassification of family A DNA polymerases reveals novel functional subfamilies and distinctive structural features. *Nucleic Acids Res.* **51**, 4488–4507 (2023).

24. Oliveira, M. T., Haukka, J. & Kaguni, L. S. Evolution of the Metazoan Mitochondrial Replicase. *Genome Biol. Evol.* **7**, 943–959 (2015).

25. Khan, Y. A. *et al.* Evidence for a novel overlapping coding sequence in POLG initiated at a CUG start codon. *BMC Genet.* **21**, 1–16 (2020).

26. Gavrilov, V. M., Golubeva, T. B. & Bushuev, A. V. Evolution of metabolic scaling among the tetrapod: effect of phylogeny, the geologic time of class formation, and uniformity of species within a class. *Integr. Zool.* **17**, 904–917 (2022).

27. 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).

28. Bhandari, V. *et al.* Molecular landmarks of tumor hypoxia across cancer types. *Nat. Genet.* **51**, 308–318 (2019).

29. Ericsson, A. C. *et al.* Differential susceptibility to colorectal cancer due to naturally

occurring gut microbiota. *Oncotarget* **6**, 33689–33704 (2015).

30. Sugimura, S. *et al.* Oxidative phosphorylation-linked respiration in individual bovine oocytes. *J. Reprod. Dev.* **58**, 636–641 (2012).
31. 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).
32. 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).
33. Stoccoro, A. & Coppedè, F. Mitochondrial DNA Methylation and Human Diseases. *Int. J. Mol. Sci.* **22**, (2021).
34. Mikhailova, A. G. *et al.* A>G substitutions on a heavy chain of mitochondrial genome marks an increased level of aerobic metabolism in warm versus cold vertebrates. *bioRxiv* 2020.07.25.221184 (2023) doi:10.1101/2020.07.25.221184.
35. 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).
36. Kong, Y. *et al.* Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution. *Science* **375**, 515–522 (2022).
37. Li, X. *et al.* The exploration of N6-deoxyadenosine methylation in mammalian genomes. *Protein Cell* **12**, 756–768 (2021).
38. Hao, Z. *et al.* N6-Deoxyadenosine Methylation in Mammalian Mitochondrial DNA. *Mol. Cell* **78**, 382–395.e8 (2020).

39. 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).

40. 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).

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

42. Das, L., Quintana, V. G. & Sweasy, J. B. NTHL1 in genomic integrity, aging and cancer. *DNA Repair* **93**, 102920 (2020).

43. Karahalil, B., de Souza-Pinto, N. C., Parsons, J. L., Elder, R. H. & Bohr, V. A. Compromised incision of oxidized pyrimidines in liver mitochondria of mice deficient in NTH1 and OGG1 glycosylases. *J. Biol. Chem.* **278**, 33701–33707 (2003).

44. Bhalla, A. D., Khodadadi-Jamayran, A., Li, Y., Lynch, D. R. & Napierala, M. Deep sequencing of mitochondrial genomes reveals increased mutation load in Friedreich's ataxia. *Ann Clin Transl Neurol* **3**, 523–536 (2016).

45. Thompson, P. S. & Cortez, D. New insights into abasic site repair and tolerance. *DNA Repair* **90**, 102866 (2020).

46. Pinz, K. G., Shibutani, S. & Bogenhagen, D. F. Action of mitochondrial DNA polymerase gamma at sites of base loss or oxidative damage. *J. Biol. Chem.* **270**, 9202–9206 (1995).

47. Li, G.-M. Mechanisms and functions of DNA mismatch repair. *Cell Res.* **18**, 85–98 (2008).

48. Lee, H. R. & Johnson, K. A. Fidelity of the human mitochondrial DNA polymerase. *J. Biol. Chem.* **281**, (2006).

49. 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.

50. Tanaka, M. & Ozawa, T. Strand asymmetry in human mitochondrial DNA mutations. *Genomics* **22**, 327–335 (1994).
51. Faith, J. J. & Pollock, D. D. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* **165**, 735–745 (2003).
52. Mikhailova, A. G. *et al.* A>G substitutions on a heavy chain of mitochondrial genome marks an increased level of aerobic metabolism in warm versus cold vertebrates. *bioRxiv* 2020.07.25.221184 (2023) doi:10.1101/2020.07.25.221184.
53. Rudd, S. G., Bianchi, J. & Doherty, A. J. PrimPol-A new polymerase on the block. *Mol Cell Oncol* **1**, e960754 (2014).
54. Guilliam, T. A. *et al.* Human PrimPol is a highly error-prone polymerase regulated by single-stranded DNA binding proteins. *Nucleic Acids Res.* **43**, 1056–1068 (2015).
55. 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).
56. 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).
57. 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).
58. Harris, K. & Pritchard, J. K. Rapid evolution of the human mutation spectrum. *Elife* **6**,

(2017).

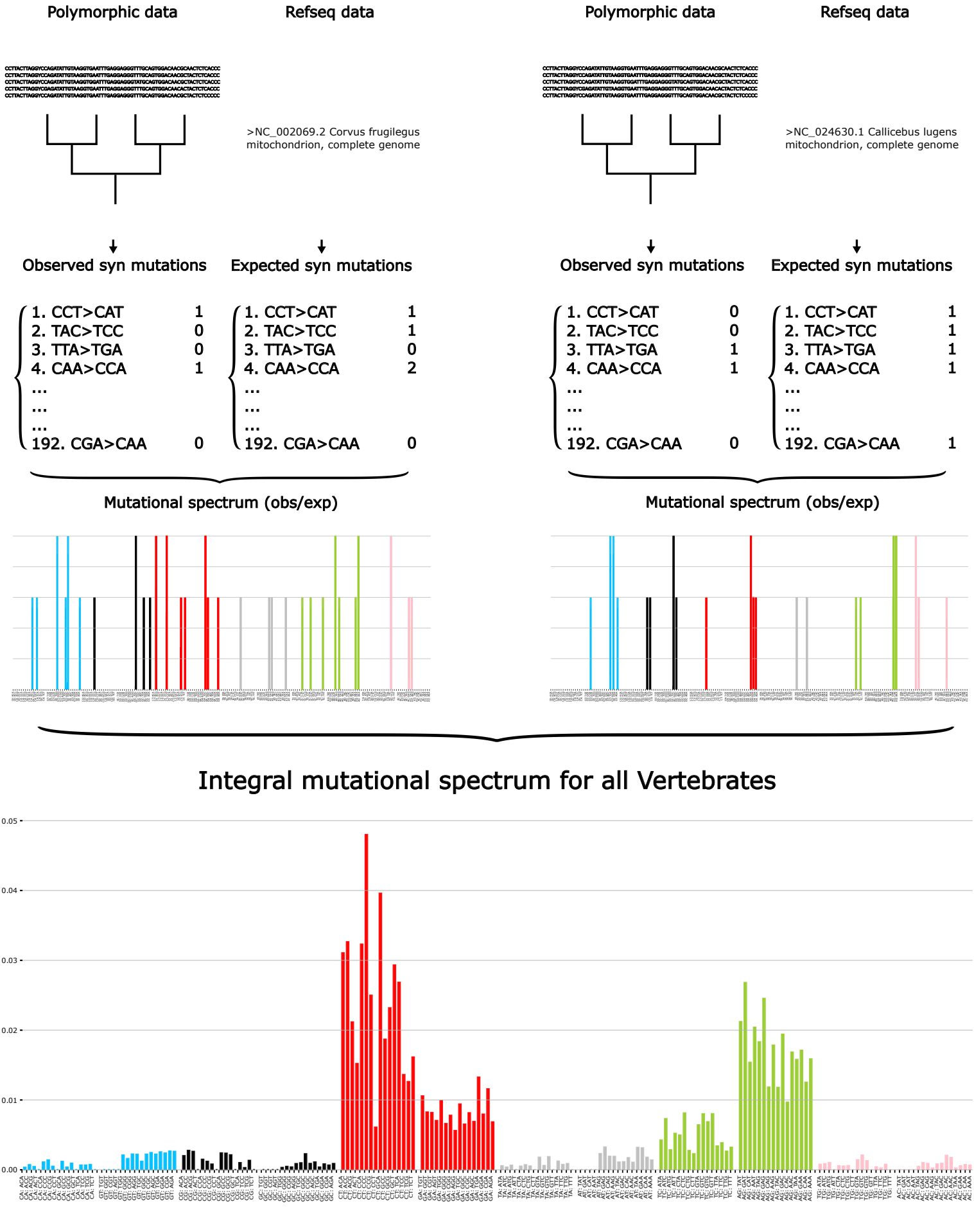
59. Sasani, T. A. *et al.* A natural mutator allele shapes mutation spectrum variation in mice. *Nature* **605**, 497–502 (2022).

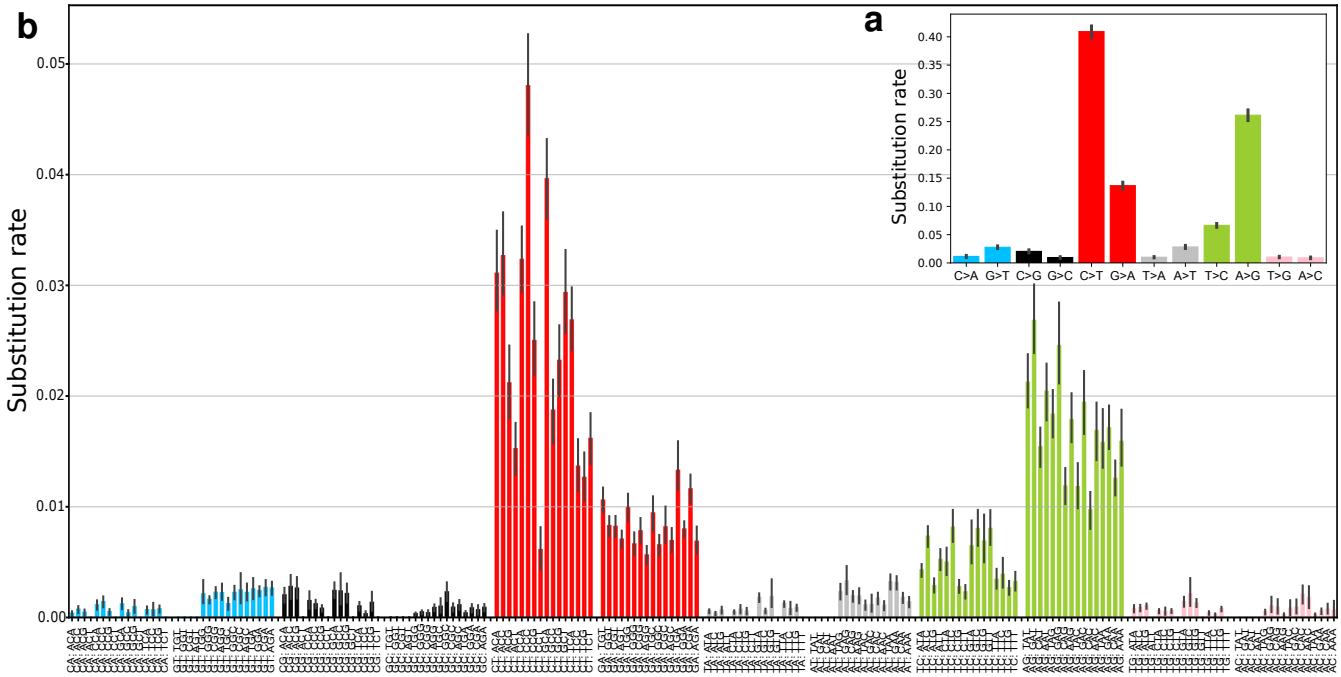
60. Jayakodi, M. *et al.* The giant diploid faba genome unlocks variation in a global protein crop. *Nature* **615**, 652–659 (2023).

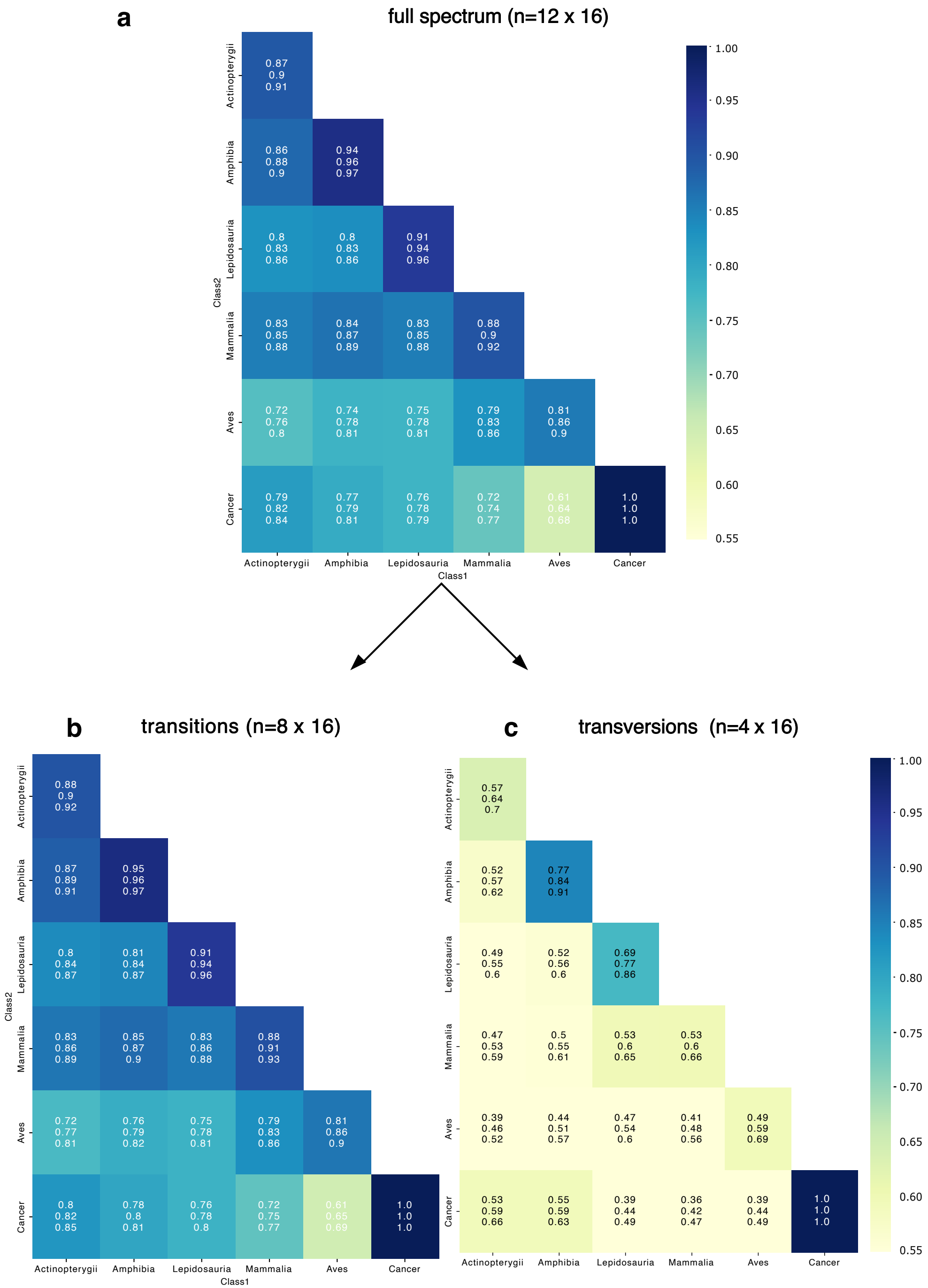
61. Gao, Z., Zhang, Y., Cramer, N., Przeworski, M. & Moorjani, P. Limited role of generation time changes in driving the evolution of the mutation spectrum in humans. (2023) doi:10.7554/eLife.81188.

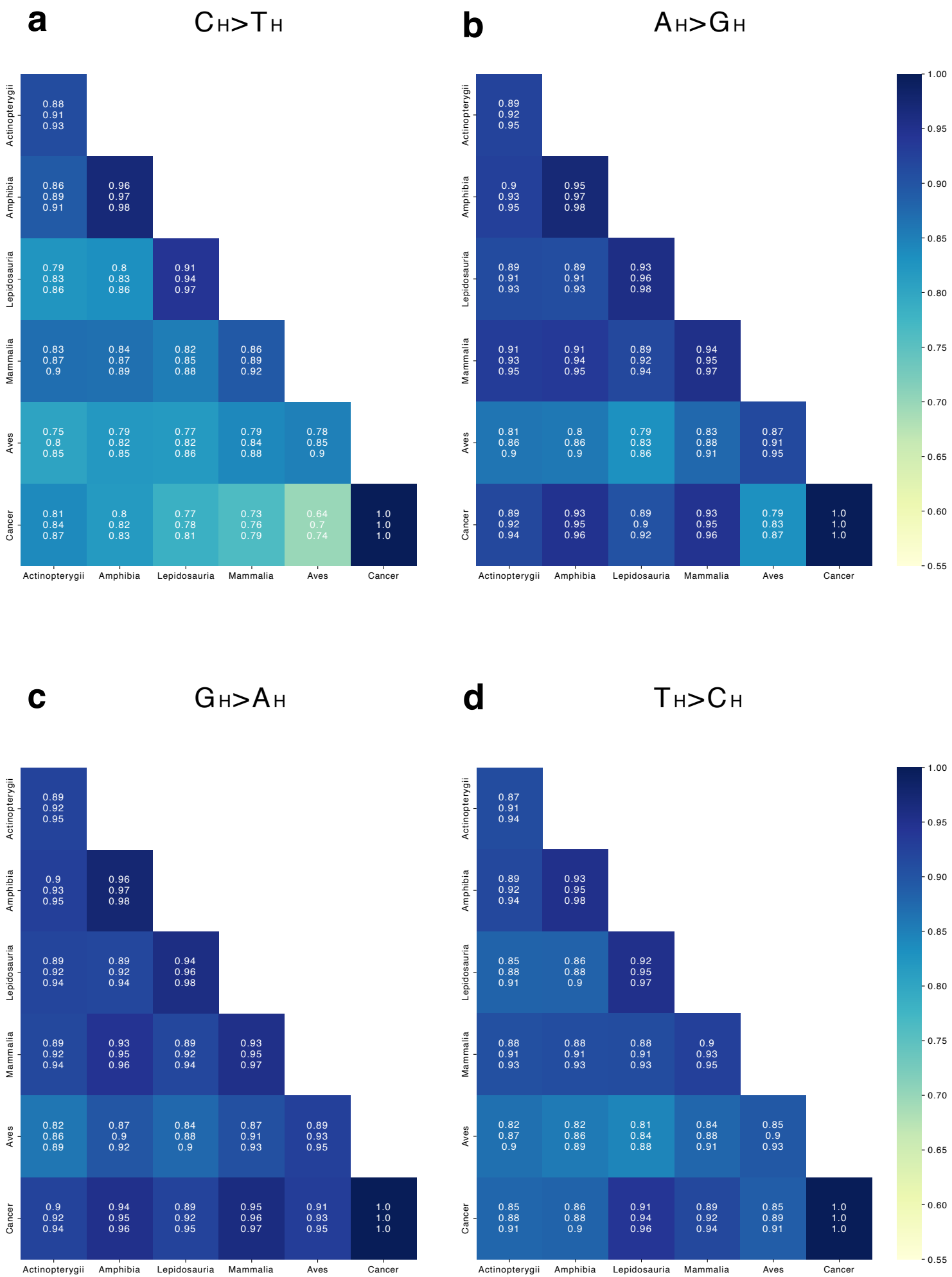
62. A. G. Mikhaylova, A. A. Mikhailova, K. Ushakova, E.O. Tretiakov, V. Shamansky, A. Yurchenko, M. Zazhytska, E. Zdobnov, V. Makeev, V. Yurov, M. Tanaka, I. Gostimskaya, Z. Fleischmann, S. Annis, M. Franco, K. Wasko, W.S Kunz, D.A. Knorre, I. Mazunin, S. Nikolaev, J. Fellay, A. Reymond, K. Khrapko, K. Gunbin, K. Popadin. Mammalian mitochondrial mutational spectrum as a hallmark of cellular and organismal aging. *bioRxiv* doi:10.1101/589168.

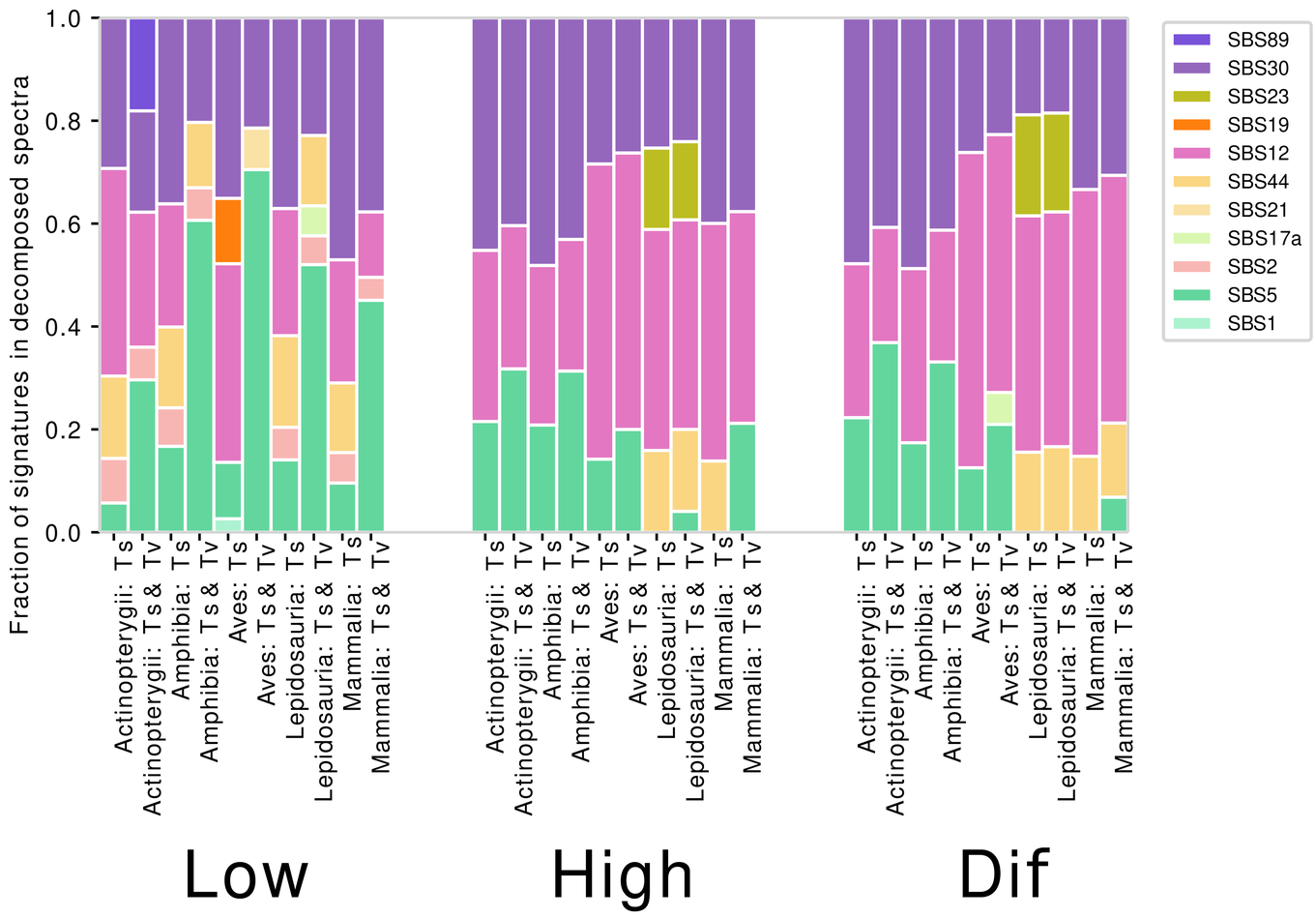
63. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).

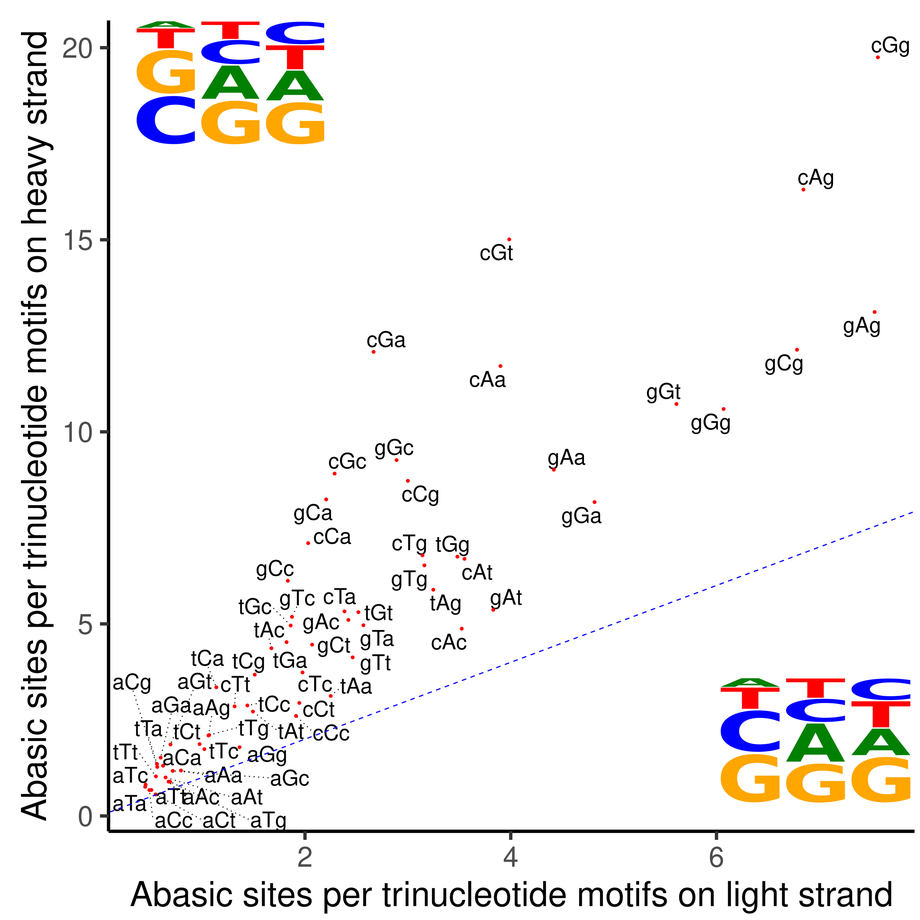

Species 1
...
Species 974
Polymorphic data
Refseq data
Polymorphic data
Refseq data
>NC_002069.2 Corvus frugilegus mitochondrion, complete genome
>NC_024630.1 Callicebus lugens mitochondrion, complete genome
Observed syn mutations
Expected syn mutations
Observed syn mutations
Expected syn mutations
1. CCT>CAT 1
2. TAC>TCC 0
3. TTA>TGA 0
4. CAA>CCA 1
...
192. CGA>CAA 0
1. CCT>CAT 1
2. TAC>TCC 1
3. TTA>TGA 0
4. CAA>CCA 2
...
192. CGA>CAA 0
1. CCT>CAT 0
2. TAC>TCC 0
3. TTA>TGA 1
4. CAA>CCA 1
...
192. CGA>CAA 0
1. CCT>CAT 1
2. TAC>TCC 1
3. TTA>TGA 1
4. CAA>CCA 1
...
192. CGA>CAA 1
Mutational spectrum (obs/exp)
Mutational spectrum (obs/exp)
Integral mutational spectrum for all Vertebrates
0.05
0.04
0.03
0.02
0.01
0.00

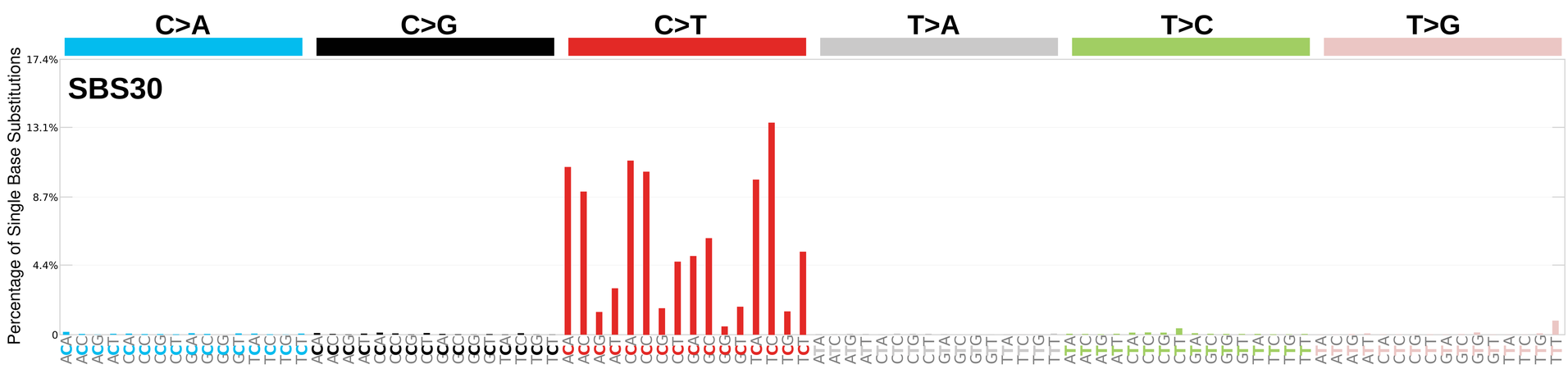

b
a
Substitution rate
0.00
0.01
0.02
0.03
0.04
0.05
Substitution rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
C>A
G>T
C>G
G>C
C>T
G>A
T>A
A>T
T>C
A>G
T>G
A>C

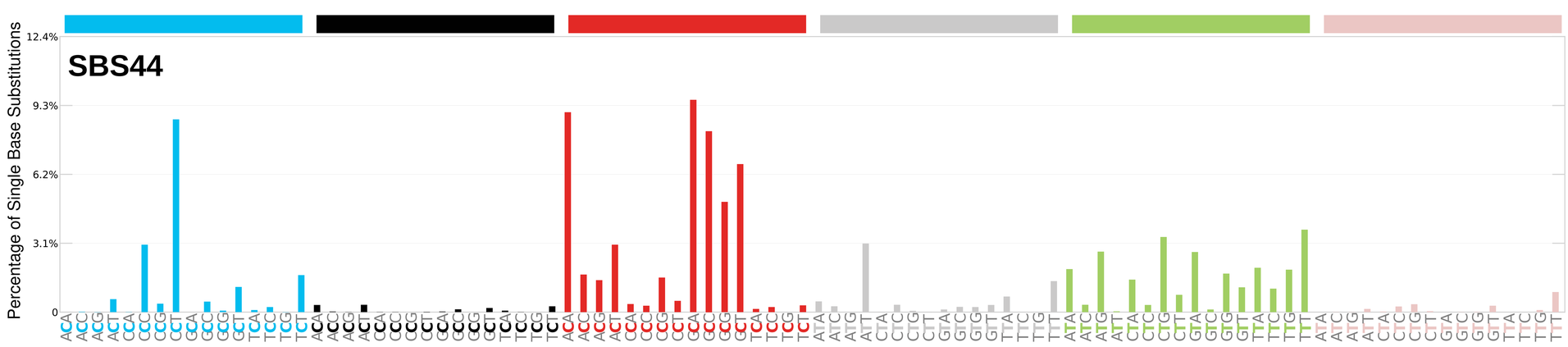

a
full spectrum (n=12 x 16)
Actinopterygii
Amphibia
Lepidosauria
Mammalia
Aves
Cancer
Class2
Class1
0.87 0.9 0.91
0.86 0.88 0.9
0.94 0.96 0.97
0.8 0.83 0.86
0.8 0.83 0.86
0.91 0.94 0.96
0.83 0.85 0.88
0.84 0.87 0.89
0.83 0.85 0.88
0.88 0.9 0.92
0.72 0.76 0.8
0.74 0.78 0.81
0.75 0.78 0.81
0.79 0.83 0.86
0.81 0.86 0.9
0.79 0.82 0.84
0.77 0.79 0.81
0.76 0.78 0.79
0.72 0.74 0.77
0.61 0.64 0.68
1.0 1.0 1.0
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
b
transitions (n=8 x 16)
0.88 0.9 0.92
0.87 0.89 0.91
0.95 0.96 0.97
0.8 0.84 0.87
0.81 0.84 0.87
0.91 0.94 0.96
0.83 0.86 0.89
0.85 0.87 0.9
0.83 0.86 0.88
0.88 0.91 0.93
0.72 0.77 0.81
0.76 0.79 0.82
0.75 0.78 0.81
0.79 0.83 0.86
0.81 0.86 0.9
0.8 0.82 0.85
0.78 0.8 0.81
0.76 0.78 0.8
0.72 0.75 0.77
0.61 0.65 0.69
1.0 1.0 1.0
c
transversions (n=4 x 16)
0.57 0.64 0.7
0.52 0.57 0.62
0.77 0.84 0.91
0.49 0.55 0.6
0.52 0.56 0.6
0.69 0.77 0.86
0.47 0.53 0.59
0.5 0.55 0.61
0.53 0.6 0.65
0.53 0.6 0.66
0.39 0.46 0.52
0.44 0.51 0.57
0.47 0.54 0.6
0.41 0.48 0.56
0.49 0.59 0.69
0.53 0.59 0.66
0.55 0.59 0.63
0.39 0.44 0.49
0.36 0.42 0.47
0.39 0.44 0.49
1.0 1.0 1.0

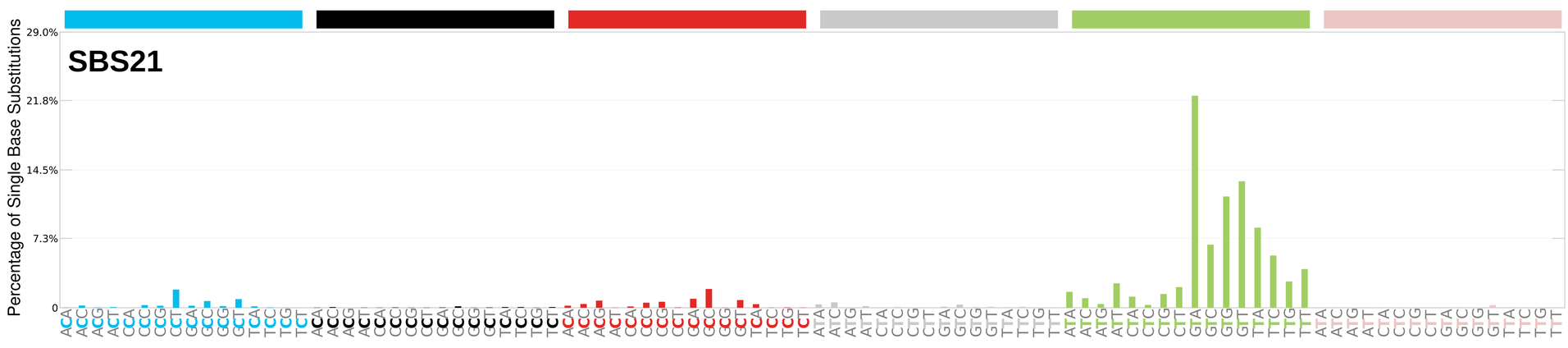

a
CH>TH
Actinopterygii
Amphibia
Lepidosauria
Mammalia
Aves
Cancer
0.88 0.91 0.93
0.86 0.89 0.91
0.96 0.97 0.98
0.79 0.83 0.86
0.8 0.83 0.86
0.91 0.94 0.97
0.83 0.87 0.9
0.84 0.87 0.89
0.82 0.85 0.88
0.86 0.89 0.92
0.75 0.8 0.85
0.79 0.82 0.85
0.77 0.82 0.86
0.79 0.84 0.88
0.78 0.85 0.9
0.81 0.84 0.87
0.8 0.82 0.83
0.77 0.78 0.81
0.73 0.76 0.79
0.64 0.7 0.74
1.0 1.0 1.0
b
AH>GH
0.89 0.92 0.95
0.9 0.93 0.95
0.95 0.97 0.98
0.89 0.91 0.93
0.89 0.91 0.93
0.93 0.96 0.98
0.91 0.93 0.95
0.91 0.94 0.95
0.89 0.92 0.94
0.94 0.95 0.97
0.81 0.86 0.9
0.8 0.86 0.9
0.79 0.83 0.86
0.83 0.88 0.91
0.87 0.91 0.95
0.89 0.92 0.94
0.93 0.95 0.96
0.89 0.9 0.92
0.93 0.95 0.96
0.79 0.83 0.87
1.0 1.0 1.0
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
c
GH>AH
0.89 0.92 0.95
0.9 0.93 0.95
0.96 0.97 0.98
0.89 0.92 0.94
0.89 0.92 0.94
0.94 0.96 0.98
0.89 0.92 0.94
0.93 0.95 0.96
0.89 0.92 0.94
0.93 0.95 0.97
0.82 0.86 0.89
0.87 0.9 0.92
0.84 0.88 0.9
0.87 0.91 0.93
0.89 0.93 0.95
0.9 0.92 0.94
0.94 0.95 0.96
0.89 0.92 0.95
0.95 0.96 0.97
0.91 0.93 0.95
1.0 1.0 1.0
d
TH>CH
0.87 0.91 0.94
0.89 0.92 0.94
0.93 0.95 0.98
0.85 0.88 0.91
0.86 0.88 0.9
0.92 0.95 0.97
0.88 0.91 0.93
0.88 0.91 0.93
0.88 0.91 0.93
0.9 0.93 0.95
0.82 0.87 0.9
0.82 0.86 0.89
0.81 0.84 0.88
0.84 0.88 0.91
0.85 0.9 0.93
0.85 0.88 0.91
0.86 0.88 0.9
0.91 0.94 0.96
0.89 0.92 0.94
0.85 0.89 0.91
1.0 1.0 1.0

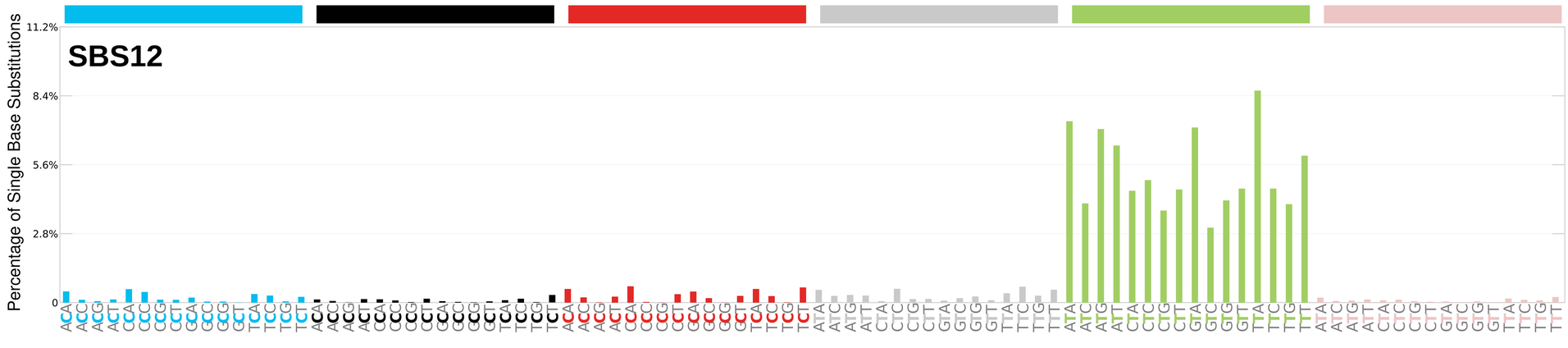

a
Fraction of signatures in decomposed spectra
SBS89
SBS30
SBS23
SBS19
SBS12
SBS44
SBS21
SBS17a
SBS2
SBS5
SBS1
Actinopterygii: Ts
Actinopterygii: Ts & Tv
Amphibia: Ts
Amphibia: Ts & Tv
Aves: Ts
Aves: Ts & Tv
Lepidosauria: Ts
Lepidosauria: Ts & Tv
Mammalia: Ts
Mammalia: Ts & Tv
Low
High
Dif

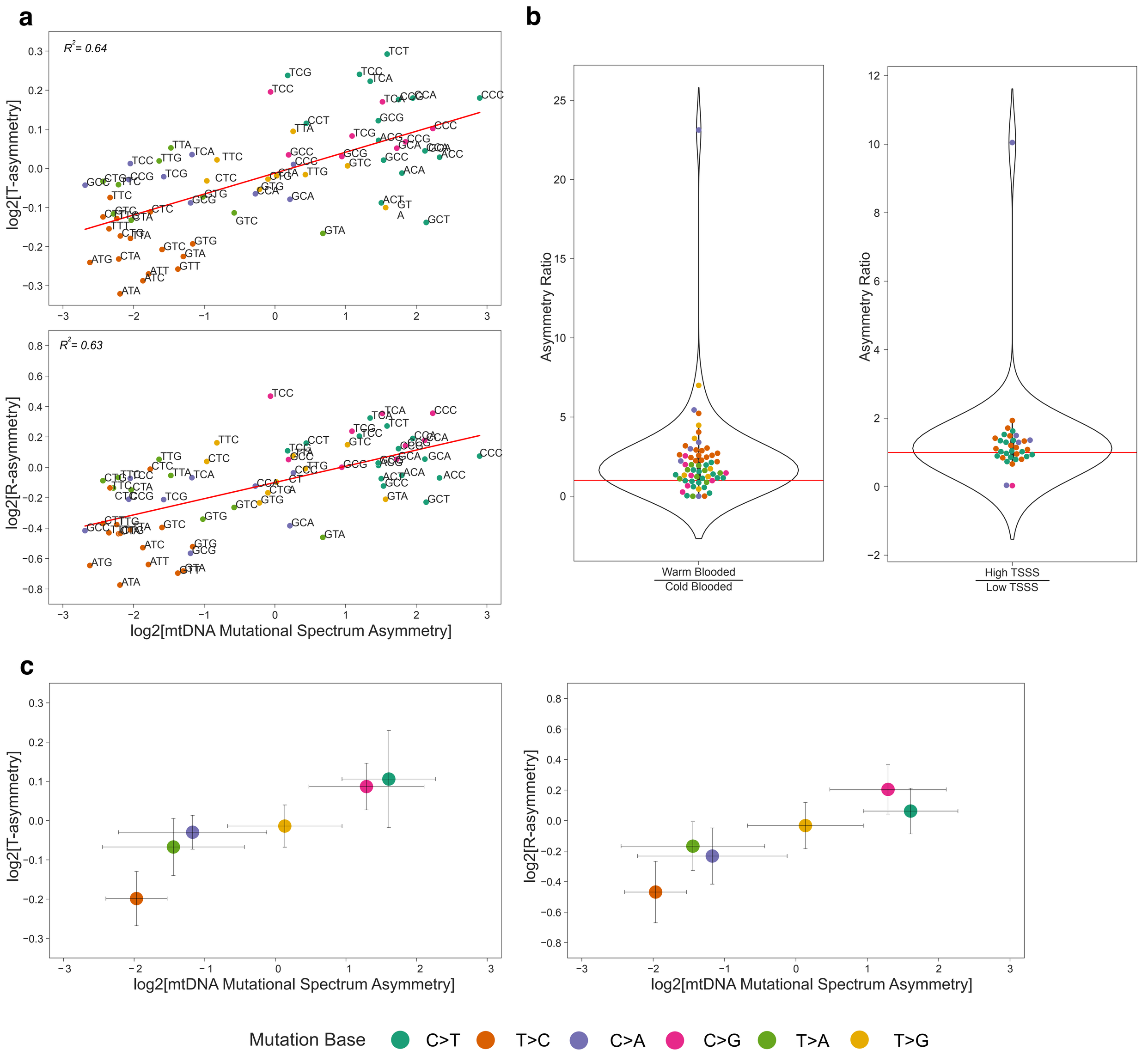

b
Abasic sites per trinucleotide motifs on heavy strand
Abasic sites per trinucleotide motifs on light strand

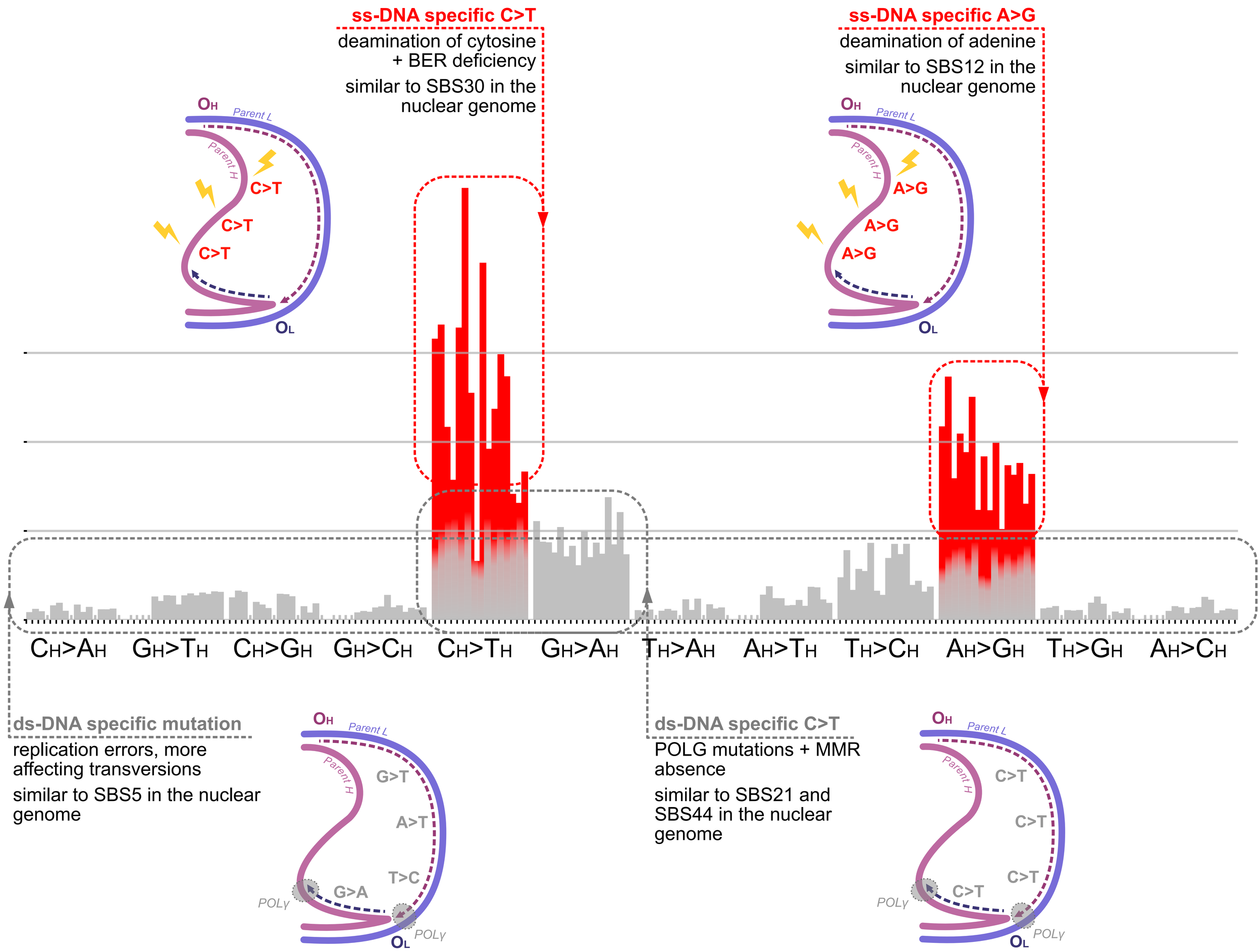

c
C>A
C>G
C>T
T>A
T>C
T>G
SBS30
Percentage of Single Base Substitutions


d
SBS44
Percentage of Single Base Substitutions


e
SBS21
Percentage of Single Base Substitutions


SBS12
Percentage of Single Base Substitutions


a
R^2= 0.64
log2[T-asymmetry]
R^2= 0.63
log2[R-asymmetry]
log2[mtDNA Mutational Spectrum Asymmetry]
b
Asymmetry Ratio
Warm Blooded
Cold Blooded
High TSSS
Low TSSS
c
log2[T-asymmetry]
log2[R-asymmetry]
log2[mtDNA Mutational Spectrum Asymmetry]
Mutation Base
C>T
T>C
C>A
C>G
T>A
T>G



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