Mitochondrial mutational spectrum in mammals is sensitive to cellular and organismal longevity by means of A>G transitions.

A. G. Mikhaylova (1)*, A. A. Mikhailova (1, 17)*, K. Ushakova (1)*, E.O. Tretiakov (1,2)*, A. Yurchenko (3), M. Zazhytska (4), E Zdobnov (5,6), V. Makeev (7), Valerian Yurov (1), M. Tanaka (8), Irina Gostimskaya (18), Z Fleischmann (9), S. Annis (9), M. Franco (9), K. Wasko (9), W.S, Kunz (10), D.A. Knorre (11,12), I. Mazunin (13, 16) S. Nikolaev (3), J. Fellay (15), A. Reymond (4)**, K. Khrapko (9)**, K. Gunbin (1,14)**, K. Popadin (1,4,6,15)**

1. Center for Mitochondrial Functional Genomics, Immanuel Kant Baltic Federal University, Kaliningrad, Russian Federation.
2. Department of Molecular Neurosciences, Center for Brain Research, Medical University of Vienna, Vienna, Austria.
3. INSERM U981, Gustave Roussy Cancer Campus, Université Paris Saclay, Villejuif, France.
4. Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland.
5. Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland.
6. Swiss Institute of Bioinformatics, Lausanne, Switzerland.
7. Vavilov Institute of General Genetics, Moscow, Russia.
8. Department for Health and Longevity Research, National Institutes of Biomedical Innovation, Health and Nutrition, Shinjuku-ku, Tokyo 162-8636, Japan.
9. Northeastern University, Massachusetts, USA.
10. Institute of Experimental Epileptology and Cognition Research, University Bonn, Sigmund-Freud-Str. 25, 53105 Bonn, Germany.
11. The A.N. Belozersky Institute Of Physico-Chemical Biology, MSU, Moscow, Russian Federation.
12. Institute of Molecular Medicine Sechenov First Moscow State Medical University Trubetskaya str. 8-2, 119991, Moscow.
13. Center of Life Sciences, Skolkovo Institute of Science and Technology, Skolkovo, Russia.
14. The Institute of Cytology and Genetics of the SB RAS, Novosibirsk, Russian Federation.
15. School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.
16. Fomin Women's Health Clinic, Moscow, Russia
17. Institute for Evolution and Biodiversity, University of Münster, Münster, Germany.
18. The University of Manchester, Manchester Institute of Biotechnology, United Kingdom

** equal contribution*

*** equal contribution*

Abstract 150 max:

Mutational spectrum of the mitochondrial genome (mtDNA) does not resemble any of the known mutational signatures of the nuclear genome and variation in mtDNA mutational spectra between different tissues and organisms is still incomprehensible. Since mitochondria is tightly involved in energy production, we expect that mtDNA mutational spectra can reflect the level of cellular aerobic metabolism, which varies in different tissues. Analyzing a collection of somatic mtDNA mutations from human cancers, de novo mtDNA germline mutations from the human mother-offspring pairs, as well as mtDNA substitutions in hundreds of mammalian species, we observed that the frequency of A_H>G_H (heavy strand notation) substitutions is positively correlated with cellular and organismal longevity. For example, epithelium, oocytes of young mothers and mice have decreased A_H>G_H frequencies. We propose that A_H>G_H is a marker of cellular and organismal age, which is driven by oxidative damage of the single-stranded mtDNA during replication.

Introduction

Various mutagens are characterized by their specific spectrum of de novo mutations - mutational signatures. Investigation of these mutational signatures in the human nuclear genome helped to understand main cancer-specific mutagens (Alexandrov et al. 2020). Unlike the nuclear genome, mutagenesis of mitochondrial genome (mtDNA) is very poorly known despite its 10-times increased mutation rate and numerous associated diseases (Wallace 2005).

Pan-cancer studies have shown that mitochondrial DNA (mtDNA) has a unique mutational signature which differs from all known nuclear signatures but remains very consistent across various cancer types (Yuan et al. 2017; Ju et al. 2014). As opposed to the situation with well known strong exogenous mutagens (such as tobacco smoke in lung cancers or ultraviolet light in melanomas) which haven't been shown to trigger significant damage in mtDNA, the observed mitochondrial mutational spectrum has to be attributed to a potent endogenous mutagen rather than an exogenous factor (Yuan et al. 2017; Ju et al. 2014). However, it has not been shown yet, if the proposed mitochondrial endogenous mutagen is sensitive to internal metabolic cellular properties which are expected to vary among tissues. Interestingly, a similar gap of knowledge exists also on a comparative species level; although variation in mtDNA mutational signature between species has been shown in several works (Belle et al. 2005; Montooth and Rand 2008), no common explanation has been proposed yet.

It has been recently shown that, because of high mtDNA mutation rate and high mtDNA copy number per cell, somatic variants in mtDNA can provide crucial information allowing to trace different cellular lineages in our body (Ludwig et al. 2019). Here, taking into account potential association of the mtDNA endogenous mutagens with cellular properties, we hypothesize further that the relative frequency of different mtDNA substitutions (i.e. mtDNA mutational spectrum) may contain a functional signature of cellular properties related to the level of metabolism. Indeed, mitochondria are the organelles determining the cell-specific level of oxidative metabolism and, as a consequence, associated mutagens are expected to directly and strongly affect mtDNA (Ericson et al. 2012).

Focusing on cellular longevity as a metric related to cellular metabolism (Tomasetti and Vogelstein 2015; Tomasetti et al. 2017) we observed differences in mtDNA mutational spectrum between tissues with short- and long-lived cells. This metric, due to its simplicity, allowed us to extend our analyses to human mother-offspring pairs and other mammalian species. As the mtDNA of all mammals is inherited exclusively through the oocytes, and oocyte division is stopped after birth, the species-specific generation length is close enough to the longevity of oocytes from different mammalian species (i.e. the time the mtDNA spent in the dormant oocyte) and could therefore be correlated with the specific mtDNA signature of longevity. Altogether, using a collection of somatic mtDNA mutations from 37 types of human cancers (TCGA and ICGC), de novo germline mutations from the mother-offspring pairs, polymorphic synonymous substitutions in most neutral sites from hundreds of mammalian species, as well as nucleotide content in whole mitochondrial genomes of mammals, we observed one universal trend: A_H>G_H substitution positively correlates with cellular and organismal longevity. We propose that A_H>G_H is driven by oxidative damage, which is known to occur in cells with high levels of oxidative metabolism. The discovered mtDNA mutational signature is a marker of cellular and organismal longevity.

Results

(1) mtDNA mutational spectrum in human cancers changes during tumorigenesis and is associated with cell turnover rate in ancestral tissues

Results of recent comprehensive survey of somatic mtDNA mutations in human cancers demonstrated that mtDNA mutational spectra are conserved across different cancer types with rare transversions and clear prevalence of $C_H>T_H$ and $A_H>G_H$ transitions (Yuan et al. 2020) (hereafter heavy strand notation, see Methods). However, there are minor variations in mutational spectra between cancer types, which might be indicative of cancer- or tissue-specific metabolic rates. To test whether mtDNA mutational spectrum changes during cancerogenesis we used a collection of 7611 somatic mitochondrial mutations from 37 cancer types grouped into 21 ancestral tissues (Yuan et al. 2020). To rank all somatic mutations from early to late, we evaluated the variant allele frequency (VAF) of each variant: high VAFs being a signature of early mutations (originated early during cancerogenesis and thus presented in the majority of cancer cells) and low VAFs being a signature of late mutations (originated at the last stages of cancerogenesis and thus presented in a subclone of the cancer cells) (Figure S1A). Splitting all variants by the median VAF (4.5%), we observed a decrease in transition/transversion ratio (hereafter Ts/Tv) among late variants (from 14.9 to 10.3, Figure S1B). This decrease in Ts/Tv among late variants is observed not only in the total set of 7611 somatic mutations but also within the majority of cancer subtypes and even within individual samples (Figure S1B). Transversions are more common among late variants and, despite numerous quality control steps (Yuan et al. 2020), might be enriched in false-positive variants, partially driving our observation. To avoid the effect of potential false-positive transversions on our calculations, we only focused on two of the most common transitions: $C_H>T_H$ and $A_H>G_H$ and compared their relative time of origin in individual samples. We observed that $A_H>G_H$ occur slightly earlier as compared to $C_H>T_H$ (Supplementary Materials 1.1), providing an independent and time-resolved way of measuring changes associated with tumorigenesis. In conclusion, we demonstrated that the mutational spectrum is changing during tumorigenesis: both Ts/Tv and $A_H>G_H$ / $C_H>T_H$ are decreased among late variants suggesting that mitochondrial microenvironment, responsible for mutagenesis, is dynamic throughout tumorigenesis.

Many somatic mitochondrial mutations, annotated in human cancers, would have occurred before the tumorigenesis (Figure S1A) and thus can be indicative of healthy tissue properties. One of the most prominent tissue-specific properties, linked to both mutagenesis (Tomasetti and Vogelstein 2015) and mitochondrial metabolism, is the turnover rate of stem cells in a given tissue. Based on extensive literature search we derived an approximate turnover rate of stem cells in each of 21 normal tissue samples, ancestral to the analyzed cancer types (Table S1). Correspondingly, we split all these tissues, which are representative of different cancers, into three categories: fast-replicating - with stem cell turnover rate less than a month (uterus, colon/rectum, stomach, cervix, esophagus, head/neck, lymphoid and myeloid tissues); intermediate - with stem cell turnover rate higher than a month and less than ten years (breast, prostate, skin, bladder, biliary, pancreas, liver and kidney) and slow-replicating - with stem cell turnover rate higher than ten years (thyroid, lung, bone/soft tissue, central nervous system (CNS) and ovary). Estimating the mutational spectrum for each category we observed an increase in Ts/Tv ratio from 10.0 in a fastly-dividing category, to 12.6 in intermediate, and to 14.5 in the slow-dividing one (Figure S1C, left panel). The observed difference is robust to jackknife resampling and is maintained mainly by a common transition $A_H>G_H$ (Figure S1C, middle panel). Comparing the frequencies of both these transitions with each other we observed an increase in the ratio $A_H>G_H$ / $C_H>T_H$ from fast- to slow-replicating tissues ($A_H>G_H$ / $C_H>T_H$: 0.56, 0.58, 0.61 for fast-, intermediate and slow-replicating tissues; Figure S1C, right panel). Thus, we conclude that the Ts/Tv is increasing from fast- to slow-replicating tissues and this effect is associated with an increased fraction of $A_H>G_H$.

Putting together the two observations described above: the increased $A_H>G_H$ in early stages of cancer (Supplementary Materials 1.1) and increased $A_H>G_H$ in slow dividing cancers (Supplementary Materials 1.1, Figure S1C), we can expect that a probability of an $A_H>G_H$ transition is a function of both relative time of origin (approximated as 1 minus VAF) and tissue-specific turnover rate. To test it, we performed multiple logistic regression and observed that probability of a $A_H>G_H$ transition indeed decreases with both time of cancerogenesis and turnover rate of the tissue (Figure 1, see Table S2 and Supplementary Materials 1.3 for details). The results are quantitatively similar if we use dummy variables, coding for fast-replicating cancers, eliminate 25% of the rarest variants, rerun this analysis only with transitions and add to the model patient age and the number of mitochondrial

copies in cancer samples (Table S2). Altogether we demonstrated a robust trend that somatic A_H>G_H substitutions are more frequent (i) at earlier stages of tumorigenesis and (ii) in cancers derived from slow-replicating tissues.

The changes in mutational spectra during the tumorigenesis as well as between slowly and fastly dividing tissues might be associated with different levels of aerobic metabolism, ranging from hypoxia to normoxia. It has been shown, for example, that mtDNA in colon cancer exhibited threefold decreased mutation rate as compared to adjacent non-tumor tissue (Ericson et al. 2012). This effect was mainly driven by the decrease in both common transitions (C_H>T_H and A_H>G_H), which are expected to be associated with oxidative damage and thus become rarer in hypoxic cancer conditions. Taking into account that practically all cancer types are becoming more hypoxic with time (Rosario et al. 2018) we interpret the fraction of A_H>G_H as a marker of oxidative damage at a given moment in cancer development: from the highly aerobic, normoxic, early stages (high fraction of A_H>G_H) to late, hypoxic stages (low fraction of A_H>G_H). Using hypoxia scores derived for many individual cancer samples in the framework of the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium (Bhandari et al. 2020) we were able to test directly an association between hypoxia and fraction of A_H>G_H. Analysis of the whole dataset showed only suggestive evidence of correlation (Spearman's rho = -0.046, p-value = 0.1869, N = 828), however, this relationship became more evident when we compared the upper decile of the most hypoxic cancers (frequency of A_H>G_H is 0.20, N = 94 with Buffa score higher or equal to 32) versus all other, less hypoxic cancers (frequency of A_H>G_H is 0.28, N = 734, Buffa score lower than 32) (p = 0.0047, Mann-Whitney U-test, Figure 1B right panel). Altogether, we suggest that mtDNA mutational spectrum and especially A_H>G_H is sensitive to the level of the metabolic rate through the level of oxidative damage, which is higher in slow-dividing tissues and tissues in early stages of cancer.

(2) mtDNA mutational spectrum changes with female reproductive age

Analyzing cancer data we observed that mtDNA mutational spectrum, and particularly the high A_H>G_H rate, marks tissues with slow turn-over rate (Figure 1A). We assume that high A_H>G_H can be a result of increased chemical (oxidative) damage in slow-dividing long-lived cells. If so, we can expand our logic and switch from somatic tissues to human oocytes, which demonstrate a high degree of variation in longevity (time of fertilization) and may have an altered mtDNA mutational spectrum.

By analyzing mitochondrial mutations in offsprings and mothers, we can reconstruct the de novo mitochondrial mutations that occurred in oocytes. It has been shown that the number of de novo mtDNA mutations in children is increasing with maternal age (Rebolledo-Jaramillo et al. 2014; Wei et al. 2019; Zaidi et al. 2019), that might be attributed to oocyte aging (oocyte longevity). In order to test potential switch in mtDNA mutational spectra that occurs when oocyte age, we reanalyzed all of the de novo germline mutations from three recent studies of mother-offspring pairs (Rebolledo-Jaramillo et al. 2014; Wei et al. 2019; Zaidi et al. 2019). All the analyses of de novo mutations in mother-offspring pairs were only suggestive due to low available sample sizes but consistently showed a trend of A_H>G_H rate increasing with oocyte longevity. Having access to the information about mothers' age at the moment of giving birth from studies of Rebolledo-Jaramillo and Zaidi (Rebolledo-Jaramillo et al. 2014; Wei et al. 2019; Zaidi et al. 2019) we observed that A_H>G_H de novo mutations were characterized by the highest age of reproduction (Supplementary Materials 2). Having access to the VAF values of de novo mtDNA mutations from the study of Wei et al 2020 (Wei et al. 2019) we observed the later onset of A_H>G_H as compared to other mutations (Supplementary Materials 2). These results are in line with our hypothesis despite the low sample size.

(3) Mutational spectrum, derived from the intra-species mtDNA polymorphisms of mammals, is shaped by generation time

Until now we have been comparing different cells and tissues from the human body and observed that the fraction of mtDNA A_H>G_H transitions out of all possible single-nucleotide substitutions positively correlates with cell longevity (Figure 1). However, if cell longevity is the main driver of the observed association, we may extend our logic further and analyze oocytes from different mammalian species. Oocytes are the only lineage through which mtDNA is transmitted (with rare exceptions of paternal inheritance) from generation to generation in mammals (Sato and Sato 2017). Mammalian oocytes are arrested from birth to puberty, which takes weeks (mouse) or decades (humans) (Von Stetina and Orr-Weaver 2011) and allows us to use the species-specific generation length as a good proxy for oocyte longevity in different mammalian species. Mitochondrial polymorphisms in mammals are especially advantageous for the reconstruction of species-specific mtDNA mutational spectrum because they are extensively used in ecological, evolutionary, and population genetics studies. As an example, the *MT-CO1* gene was selected for

the DNA barcoding project (Hebert et al. 2003). Several studies revealed that there is some variation in mtDNA mutational spectra between different species (Montooth and Rand 2008; Belle et al. 2005), however, no overarching explanation has been suggested so far.

For 611 mammalian species, we collected all available mtDNA polymorphisms and reconstructed normalized 12-component mutational spectra as a probability of each nucleotide to mutate to any other nucleotide (see Methods, Figure S2A). The average mutational spectrum for all mammalian species (Figure 2A) demonstrates an excess of C_H>T_H and A_H>G_H substitutions which is similar to previous studies (Ju et al. 2014; Yuan et al. 2020).

To uncover the nature of mitochondrial mutagenesis shaping the species-specific variation in mutational spectra we focused on *MT-CYB* gene, which contained 56% of all analyzed mutations (39,112 out of 70,053 used to draw Figure 2A) and compared the *MT-CYB*-derived spectrum between species with variable life-history traits. We used the well-characterized mammalian generation length as a metric (Pacifici et al. 2013; Tacutu et al. 2013) which is associated with oocyte longevity and also with numerous ecological (body mass, litter size, effective population size) and physiological (basal metabolic rate) parameters of mammalian species (Ollason 1987; Damuth 1987). As the first and simplest approximation of the mutational spectrum, we used Ts/Tv.

For 424 mammalian species with known Ts/Tv and generation length, we observed a positive correlation between those two parameters (Figure S2B). To test the robustness of this correlation we performed several additional analyses and showed that the results were not affected by phylogenetic inertia, the total number of polymorphisms observed in different species and algorithms used for reconstruction of ancestral sequences (data not shown).

To understand which substitution types shaped the observed correlation between Ts/Tv and generation length we performed twelve pairwise rank correlation analyses. We observed that only A_H>G_H frequency positively correlated with the generation length (Spearman's rho = 0.252, p value = 1.188e-07) (Figure 2B), while several rare transversions showed a negative correlation (T_H>A_H, T_H>G_H, C_H>A_H and G_H>T_H: all Spearman's rhos < -0.17, all p-values are < 0.0003). Inclusion of all five types of substitutions into a multiple linear model showed A_H>G_H being the strongest component associated with generation length (Supplementary Materials 3.3, 3.4, equation I).

In order to derive mtDNA mutational signatures in an unsupervised way we performed principal component analysis (PCA) of mutational spectra of 424 mammalian species. We observed that the first component is mainly driven by C_H>T_H substitutions, whereas the second is driven mainly by A_H>G_H substitutions (Figure 2C left panel). We assessed if the first ten principal components were correlated with generation length and observed that only the second component was significantly correlated (Figure 2C right panel).

Interestingly, the correlation of the second principal component with generation length was higher as compared to the sole effect of A_H>G_H frequency, suggesting that the second component could reflect a complex nature of longevity-associated mutational signature (Supplementary Materials 3.5). Additional analyses which take into account the effects of the total number of mutations, potential sequencing errors and nucleotide content confirmed the robustness of our results (Supplementary Materials 3.5).

Altogether, our analyses of within-species neutral polymorphisms in hundreds of mammalian species demonstrated a clear sensitivity of mtDNA mutational spectrum (A_H>G_H) to the lifespan.

(4) The long-term mutational bias affects nucleotide composition of complete mammalian mitochondrial genomes

Mutational bias, if stronger than selection, is expected to change genome-wide nucleotide content. We expect that increased A_H>G_H (Figure 2) would decrease frequencies of A_H and increase frequencies of G_H in effectively neutral positions of long-lived species. Using 650 complete mitochondrial genomes of mammalian species we estimated their nucleotide content in the most neutral - synonymous fourfold degenerate - positions: in 13 protein-coding genes without overlapped regions. Testing all four pairwise correlations between the species-specific generation length and the nucleotide content we observed two strongest correlations: negative with A_H and positive with G_H (Figure 3A; Table S3). Inclusion of all four types of nucleotide frequencies into the multiple model confirmed the importance of A_H and G_H only, the effect of which was also robust to the phylogenetic inertia (Supplementary Materials 4.1). Thus we concluded that mtDNA of long- versus short-lived mammals is more A_H poor and G_H rich (Figure 3A), which is in line with the more intensive A_H>G_H mutagenesis in the former (Figure 2).

An excess of G_H and deficit of A_H in long-lived species determines the G_HA_H nucleotide skew which approximates the level of asymmetry in the distribution of these two nucleotides and is calculated as $(G_H-A_H)/(G_H+A_H)$. As expected we observed positive correlation between G_HA_H skew and generation length of mammalian species (phylogenetic independent contrasts: Spearman's rho = 0.09, p-value = 0.0164; phylogenetic generalized least squares: coefficient = 0.13, p-value = $2.9*10^{-4}$; see also Figure 3C). To visualize a contrast in G_HA_H skew between the shortest- and the longest-lived species we plotted A_H and G_H fractions along the major arc of mtDNA for opossum (generation length 341 days) and whale (generation length 18980 days) (Figure 3B). An excess of G_H and deficit of A_H were more pronounced in whale versus opossum.

Moreover, the frequencies of G_H and A_H are changing along the major arc (Figure 3B, 3C), reflecting the change in G_HA_H skew with gene position. It has been shown that the frequency of $A_H>G_H$ substitutions depends on Time which heavy strand Spent in Single-Stranded (TSSS) condition during asynchronous mtDNA replication (Faith and Pollock 2003). Genes, located close to the origin of light strand replication (O_L), such as *COX1*, spend minimal time being single-stranded and demonstrate the low frequency of $A_H>G_H$, while genes located far away from O_L spend more time being single-stranded and demonstrate correspondingly higher frequencies of $A_H>G_H$. Thus, we expect that G_HA_H skew is a function of both: species-specific generation length and gene-specific TSSS. Plotting G_HA_H skew of short- and long-lived mammals for each gene we observed that indeed G_HA_H skew increases with both species-specific generation length and gene-specific TSSS (Figure 3C).

Increase in G_HA_H skew with TSSS is supported by corresponding decrease in A_H and an increase in G_H in all species (Figure 3A, Supplementary Materials 4.2). Interestingly, comparing the rate of changes in the nucleotide frequencies with TSSS we observed that long-lived mammals demonstrate faster decrease in A_H and increase in G_H along the genome (Supplementary Materials 4.2, the same effect is visible on Figure 3B). Fast changes (stronger gradients) in A_H and G_H in long-lived mammals can be interpreted as an interaction between TSSS and longevity probably mediated by an increased oxidative damage in genes with high TSSS in long-lived mammals.

Altogether, our results demonstrate that the neutral nucleotide content reflects the mutational bias from A_H to G_H: long-lived mammals are more A_H poor and G_H rich (Figure 3A), have stronger G_HA_H skew (Figure 3B, 3C) and faster changes in nucleotide content (decrease in A_H and increase in G_H) along the genome (Supplementary Materials 4.2).

Discussion

We observed that the fraction of A_H>G_H substitutions increases with cellular and organismal longevity on different time scales - from several years in somatic tissues (chapter 1, Figure 1) and dozens of years in the human de novo germline substitutions (chapter 2) to hundreds of thousands years (the average time of segregation of neutral mtDNA polymorphisms (Atkinson, Gray, and Drummond 2008) in within-species polymorphisms (chapter 3, Figure 2) and millions of years as in the case of a comparative species scale (chapter 4, Figure 3).

We propose that a process of mtDNA mutagenesis, rather than selection, is primarily responsible for all our findings. Due to little or no evidence of selection on mtDNA somatic mutations in human cancers (Yuan et al. 2020; Ju et al. 2014), on low-heteroplasmy de novo mutations in the germline (Arbeithuber et al. 2020) and on synonymous four-fold degenerate sites in mtDNA of mammals (Faith and Pollock 2003; Uddin and Chakraborty 2017) we consider the vast majority of variants analysed in this study as effectively neutral and thus driven solely by mutagenesis. The propagation of the same trend - an excess of A_H>G_H in long-lived cells - from somatic cancer data to germline mutations on population and comparative-species levels, additionally support the mutagenic explanation of this effect as the most parsimonious.

We propose that mtDNA mutagenesis is sensitive to the mitochondrial microenvironment which in turn depends on cell- and organism-specific metabolic rate. We assume that oxidative damage - a main by-product of the aerobic metabolism - through the deamination of adenosine induces A_H>G_H substitution on a single-stranded mtDNA heavy strand.

It has been shown for example that colon cancer, in the course of transformation to a more hypoxic stage, decreases its mtDNA mutation rate due to slowing down both common transitions: C_H>T_H and A_H>G_H (Ericson et al. 2012). In our study, we extended the hypothesis of Ericson to all cancer types and all healthy tissues including oocytes of humans and other mammals, claiming that hypoxic conditions, more often observed in fast-dividing cells, slow down mtDNA mutation rate and especially fraction of A_H>G_H transitions.

Recent deep sequencing of de novo mtDNA mutations in aged versus young mice oocytes confirmed that an increased fraction of A_H>G_H substitutions is the best hallmark of oocyte aging (Arbeithuber et al. 2020), which is in line with our trend observed in human mother-offspring duos (chapter 2). This discovery also suggests that continuous mtDNA turnover within the dormant oocytes can be a universal trait of all mammalian species (Arbeithuber et al. 2020) - and thus variation in mtDNA mutational spectra between different species can reflect the level of oxidative damage in oocytes of different species. Extending the results of Arbeithuber et al (Arbeithuber et al. 2020) from young and old mice oocytes to oocytes of short- and long-lived mammalian species we observed the same trend with A_H>G_H marking more long-lived oocytes (Figure 2). These results suggest increased oxidative damage in mtDNA of oocytes of long-lived mammals.

Extending the trend of the species-specific mtDNA mutational spectra (Figure 2) to nucleotide content of mitochondrial genomes (Figure 3) confirms long-term effects of the described tendencies and goes in line with an excess of G_H in mtDNA of long-lived mammals which had been previously shown (Lehmann et al. 2006, Lehmann et al. 2008). Although Lehmann et al proposed a selective explanation of this observation assuming increased stability of the G_H-rich genomes which may confer an advantage to long-lived mammals, our findings demonstrate that an excess of G_H in long-lived mammals may be a result of mutagenesis, not selection.

Altogether, we propose that A_H>G_H substitutions are a marker of the metabolic rate. It is important to mention that A_H>G_H transitions are the substitution type which is the most sensitive to the time being single-stranded during asymmetric replication of mammalian mtDNA (Faith and Pollock 2003; Tanaka and Ozawa 1994). This mutational signature may be driven by the increased oxidative damage of single-stranded DNA. For example, it has been shown on mouse cells (Shin and Turker 2002) and bacteria (Shewaramani et al. 2017) that oxidative damage predominantly increases adenine deamination especially on single-stranded DNA (Shewaramani et al. 2017). Also it has been shown recently, that the fraction of A_H>G_H positively correlates with the ambient temperature of Actinopterygii species and it is higher in non-hibernating versus hibernating mammals (Mikhailova et al. 2020). Because higher temperature is associated with an increased level of aerobic metabolism (Martin and Palumbi 1993) we assume that A_H>G_H can be a marker of the oxidative damage.

The results of our study can significantly expand the usability of mtDNA mutational spectra. Low-heteroplasmy somatic mtDNA mutations from a merely neutral marker used to trace cellular lineages (Ludwig et al. 2019) can be transformed to a metric, associated with cell-specific metabolic rate - a metric which can be especially important in highly heterogeneous tissues such as cancers. Low-heteroplasmy de novo germ-line mtDNA mutations (Arbeithuber et al. 2020) can predict the biological age of human oocytes, which can be used in *in vitro* fertilization techniques. MtDNA mutational spectra, reconstructed for non-model vertebrate species, may help to approximate their metabolic rate - a trait for which experimental estimation is too laborious or even impossible.

Direct experimental investigation of the effect of different mutagens (Kucab et al. 2019) and especially an effect of oxidative damage (Degtyareva et al. 2019) and chemical modifications of nucleotides (Koh et al. 2018, Hao et al. 2020) on single-stranded mtDNA will significantly improve our understanding of mtDNA mutagenesis.

Figures

FIGURE 1: mtDNA mutational spectra is changing between tissues and during carcinogenesis.

1A. The probability of A_H>G_H is higher in early stages of cancer (when cells are more normoxic and slow dividing) and in tissues with low turnover rate. X axis reflects the stages of cancerogenesis approximated by 1 - VAF. Y axis marks the stem cell turnover rate. Each circle reflects one mutation. The size of the circles is proportional to a probability that a given mutation is A_H>G_H, where the probability is obtained from the logistic regression model 1 and normalized to 0-1 range. Red circles mark A_H>G_H, grey circles - all other substitutions. It is seen that circles (probability of A_H>G_H) are decreasing from left to right reflecting increased hypoxia in more advanced stages of cancers and from top to bottom, reflecting a gradient of tissues with low to high turnover rate.

1B. The probability of A_H>G_H is decreased in hypoxic samples. The scheme on the left demonstrates expected changes in normoxia level among different tissues (for example slow dividing oocytes and fastly dividing colon epithelium) and stages of cancer (all cancers slowly become more hypoxic with time). A fraction of A_H>G_H (red asterisks) reflects oxidative damage at a given moment in cancer development: from the highly aerobic, normoxic (bright-blue) early stages to late, hypoxic stages (pale-blue and white).

The barplots on the right demonstrate differences in the frequencies of A_H>G_H between normoxic samples (with Buffa score lower than 32, N = 734) and hypoxic samples (with Buffa score higher or equal to 32, N = 94).

FIGURE 2: variation in mammalian mutational spectrum is driven by generation length

2A. Average mtDNA mutational spectrum of mammalian species (N = 611). Mutational spectrum as a probability of each nucleotide to mutate to each other based on the observed and normalized frequencies of twelve types of nucleotide substitutions in four fold degenerate synonymous sites of all available within-species polymorphisms of mtDNA protein-coding genes.

2B. Mutational spectra vary with species-specific generation length (N = 424). Frequency of A_H>G_H is the best type of substitutions, correlated with generation length.

2C. The principal component analysis of mtDNA mutational spectra of mammalian species (N = 424). Left panel: the biplot of the principal component analyses (first and the second components explains 16% and 12% of variation correspondingly). C_H>T_H has the highest loading on the first principal component while A_H>G_H has the highest loading on the second principal component. Note that we reverted PC2 to make it positively correlated with generation length. Right panel: The second principal component correlates with generation length in mammals. Generation length is color-coded from dark green (the shortest generation length) to light green (the longest generation length).

FIGURE_3: the long-term effect of the mutational bias: neutral nucleotide content in mammalian species

3A. Nucleotide frequencies in neutral sites of all 13 protein-coding genes as a function of generation length - fraction of A_H is decreasing while fraction of G_H is increasing (N = 650).

3B. Frequencies of A_H (red) and G_H (grey) nucleotides along the major arc of mtDNA of the most short-lived and most long-lived species from our dataset. Each bar represents the nucleotide frequency in a 20 nucleotide window. In both mammals A_H is decreasing and G_H is increasing along the mtDNA: from bottom left (origin of replication of light strand) to top right (origin of replication of heavy strand). Also, mtDNA of a long-lived mammal has a deficit of A_H and excess of G_H genome-wide - a signature of increased longevity.

3C. Changes in nucleotide content along mtDNA of short- and long-lived mammals. (N = 650). All genes (except for ND6) located in the major arc are ranked according to the time spent single stranded: from COX1 to CYTB. Pairs of boxplots for each gene represent G_HA_H skew for short- and long-lived mammals splitted by the median generation length.

Methods

Analyses of somatic mtDNA mutations and mother-offspring pairs

We used 7611 somatic single-nucleotide mtDNA substitutions observed in human cancer samples (Yuan et al. 2017). Taking into account presumably very weak selection on somatic mtDNA mutations in cancers and their low VAFs we used all substitutions (not only synonymous fourfold degenerate) as neutral to derive the mutational spectrum. Correspondingly, to normalize the relative probabilities of mutations we used all nucleotides of the human genome (A, T, G, C) not just the ones in synonymous fourfold degenerate positions. Tissue-specific number of divisions of each stem cell per lifetime we sampled from the supplementary table 1 of (Tomasetti and Vogelstein 2015) and other sources (Supplementary Materials).

Data for the analyses of mother-offspring pairs were taken from three studies (Rebolledo-Jaramillo et al. 2014; Wei et al. 2019; Zaidi et al. 2019). In each dataset we filtered out all inherited mutations and took into account only de novo mtDNA mutations in offspring.

Reconstruction of the species-specific mutational spectrum for mammalian species

Using all available intraspecies sequences (at April 2016) of mitochondrial protein-coding genes we derived the mutational spectrum for each species. Briefly, we collected all available mtDNA sequences of any protein-coding genes for any chordate species, reconstructed the intraspecies phylogeny using an outgroup sequence (closest species for analyzed one), reconstructed ancestral states spectra in all positions at all inner tree nodes and finally got the list of single-nucleotide substitutions for each gene of each species. The pipeline is described in more details in Supplementary Materials 3. Using species with at least 15 single-nucleotide synonymous mutations at four-fold degenerate sites we estimated mutational spectrum as a probability of each nucleotide to mutate into any other nucleotide (vector of 12 types of substitutions with sum equals one) for more than a thousand of chordata species. We normalized observed frequencies by nucleotide content in the third position of four-fold degenerative synonymous sites of a given gene. To eliminate the effect of nonuniform sampling of analyzed genes between different species the vast majority of our analyses were performed with *MT-CYB* gene - the most common gene in our dataset.

Generation length in days (as the average age of parents of the current cohort, reflecting the turnover rate of breeding individuals in a population) for mammals was downloaded from Pacifici et al. (Pacifici et al. 2013).

Analyses of complete mitochondrial genomes

We downloaded whole mitochondrial genomes from GenBank using the following search query: “Chordata”[Organism] AND (complete genome[All Fields] AND mitochondrion [All Fields] AND mitochondrion[filter]. We extracted non-overlapping regions of protein-coding genes, calculated codon usage and extracted fractions of A_H, T_H, G_H and C_H nucleotides in synonymous fourfold degenerate positions. For each protein-coding gene of each reference mitochondrial genome of mammalian species we estimated G_HA_H skew as $(G_H-A_H)/(G_H+A_H)$ using only synonymous fourfold degenerate sites.

All statistical analyses were performed in R.

Acknowledgments

We thank Han Liang, Peter Campbell and Yuan Yuan for the sharing of somatic mtDNA mutations from cancer samples, Athanasios Kousathanas for statistical comments, the whole laboratory of Alexandre Reymond, Mikhail Gelfand and Vladimir Katanaev for discussions and valuable suggestions. K.P. was supported by the 5 Top 100 Russian Academic Excellence Project at the Immanuel Kant Baltic Federal University. E.O.T. is supported by a scholarship from the Austrian Science Fund (FWF, DOC 33-B27). This work was also supported by Russian Foundation of Basic Research [No. 18-29-13055 & 18-04-01143 to K.P; No. 19-29-04101 to I.M].

Literature

Alexandrov, L. B., J. Kim, N. J. Haradhvala, M. N. Huang, Tian Ng Aw, Y. Wu, A. Boot, et al. 2020. "The Repertoire of Mutational Signatures in Human Cancer." *Nature* 578 (7793). https://doi.org/10.1038/s41586-020-1943-3.

Arbeithuber, Barbara, James Hester, Marzia A. Cremona, Nicholas Stoler, Arslan Zaidi, Bonnie Higgins, Kate Anthony, Francesca Chiaromonte, Francisco J. Diaz, and Kateryna D. Makova. 2020. "Age-Related Accumulation of de Novo Mitochondrial Mutations in Mammalian Oocytes and Somatic Tissues." *PLoS Biology* 18 (7): e3000745.

Atkinson, Q. D., R. D. Gray, and A. J. Drummond. 2008. "mtDNA Variation Predicts Population Size in Humans and Reveals a Major Southern Asian Chapter in Human Prehistory." *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msm277.

Belle, Elise M. S., Gwenael Piganeau, Mike Gardner, and Adam Eyre-Walker. 2005. "An Investigation of the Variation in the Transition Bias among Various Animal Mitochondrial DNA." *Gene* 355 (August): 58–66.

Bhandari, Vinayak, Constance H. Li, Robert G. Bristow, Paul C. Boutros, and PCAWG Consortium. 2020. "Divergent Mutational Processes Distinguish Hypoxic and Normoxic Tumours." *Nature Communications* 11 (1): 737.

Damuth, John. 1987. "Interspecific Allometry of Population Density in Mammals and Other Animals: The Independence of Body Mass and Population Energy-Use." *Biological Journal of the Linnean Society*. https://doi.org/10.1111/j.1095-8312.1987.tb01990.x.

Degtyareva, Natalya P., Natalie Saini, Joan F. Sterling, Victoria C. Placentra, Leszek J. Klimczak, Dmitry A. Gordenin, and Paul W. Doetsch. 2019. "Mutational Signatures of Redox Stress in Yeast Single-Strand DNA and of Aging in Human Mitochondrial DNA Share a Common Feature." *PLOS Biology*. https://doi.org/10.1371/journal.pbio.3000263.

Ericson, Nolan G., Mariola Kulawiec, Marc Vermulst, Kieran Sheahan, Jacintha O'Sullivan, Jesse J. Salk, and Jason H. Bielas. 2012. "Decreased Mitochondrial DNA Mutagenesis in Human Colorectal Cancer." *PLoS Genetics* 8 (6): e1002689.

Faith, Jeremiah J., and David D. Pollock. 2003. "Likelihood Analysis of Asymmetrical Mutation Bias Gradients in Vertebrate Mitochondrial Genomes." *Genetics* 165 (2): 735–45.

Hao, Ziyang, Tong Wu, Xiaolong Cui, Pingping Zhu, Caiping Tan, Xiaoyang Dou, Kai-Wen Hsu, et al. 2020. "N-Deoxyadenosine Methylation in Mammalian Mitochondrial DNA." *Molecular Cell* 78 (3): 382–95.e8.

Hebert, Paul D. N., Alina Cywinska, Shelley L. Ball, and Jeremy R. deWaard. 2003. "Biological Identifications through DNA Barcodes." *Proceedings. Biological Sciences / The Royal Society* 270 (1512): 313–21.

Ju, Young Seok, Ludmil B. Alexandrov, Moritz Gerstung, Inigo Martincorena, Serena Nik-Zainal, Manasa Ramakrishna, Helen R. Davies, et al. 2014. "Origins and Functional Consequences of Somatic Mitochondrial DNA Mutations in Human Cancer." *eLife* 3 (October). https://doi.org/10.7554/eLife.02935.

Koh, Casslynn W. Q., Yeek Teck Goh, Joel D. W. Toh, Suat Peng Neo, Sarah B. Ng, Jayantha Gunaratne, Yong-Gui Gao, Stephen R. Quake, William F. Burkholder, and Wee Siong S. Goh. 2018. "Single-Nucleotide-Resolution Sequencing of Human N6-Methyldeoxyadenosine Reveals Strand-Asymmetric Clusters Associated with SSBP1 on the Mitochondrial Genome." *Nucleic Acids Research* 46 (22): 11659–70.

Kucab, Jill E., Xueqing Zou, Sandro Morganella, Madeleine Joel, A. Scott Nanda, Eszter Nagy, Celine Gomez, et al. 2019. "A Compendium of Mutational Signatures of Environmental Agents." *Cell* 177 (4): 821–36.e16.

Lehmann, Gilad, Arie Budovsky, Khachik K. Muradian, and Vadim E. Fraifeld. 2006. "Mitochondrial Genome Anatomy and Species-Specific Lifespan." *Rejuvenation Research* 9 (2): 223–26.

Ludwig, Leif S., Caleb A. Lareau, Jacob C. Ulirsch, Elena Christian, Christoph Muus, Lauren H. Li, Karin Pelka, et al. 2019. "Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics." *Cell* 176 (6): 1325–39.e22.

Martin, A. P., and S. R. Palumbi. 1993. "Body Size, Metabolic Rate, Generation Time, and the Molecular Clock." *Proceedings of the National Academy of Sciences of the United States of America* 90 (9): 4087–91.

Mikhailova, Alina G., Victor Shamansky, Alina A. Mikhailova, Kristina Ushakova, Evgenii Tretyakov, Sergey Oreshkov, Dmitry Knorre, et al. 2020. "A Mitochondrial Mutational Signature of Temperature and Longevity in Ectothermic and Endothermic Vertebrates." *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.07.25.221184.

Montooth, Kristi L., and David M. Rand. 2008. "The Spectrum of Mitochondrial Mutation Differs across Species." *PLoS Biology* 6 (8): e213.

Ollason, J. G. 1987. "R. H. Peters 1986. The Ecological Implications of Body Size. Cambridge University Press, Cambridge. 329 Pages. ISBN 0-521-2886-X. Price: £12.50, US$16.95 (paperback)." *Journal of Tropical Ecology*. https://doi.org/10.1017/s0266467400002224.

Pacifici, Michela, Luca Santini, Moreno Di Marco, Daniele Baisero, Lucilla Francucci, G. Grottolo Marasini, Piero Visconti, and Carlo Rondinini. 2013. "Generation Length for Mammals." *Nature Conservation* 5: 87–94.

Rebolledo-Jaramillo, B., M. S. -W. Su, N. Stoler, J. A. McElhoe, B. Dickins, D. Blankenberg, T. S. Korneliussen, et al. 2014. "Maternal Age Effect and Severe Germ-Line Bottleneck in the Inheritance of Human Mitochondrial

DNA.” *Proceedings of the National Academy of Sciences* 111 (43): 15474–79.
Rosario, S. R., M. D. Long, H. C. Affronti, A. M. Rowsam, K. H. Eng, and D. J. Smiraglia. 2018. “Pan-Cancer Analysis of Transcriptional Metabolic Dysregulation Using The Cancer Genome Atlas.” *Nature Communications* 9 (1): 5330.
Sato, Ken, and Miyuki Sato. 2017. “Multiple Ways to Prevent Transmission of Paternal Mitochondrial DNA for Maternal Inheritance in Animals.” *Journal of Biochemistry* 162 (4): 247–53.
Shewaramani, Sonal, Thomas J. Finn, Sinead C. Leahy, Rees Kassen, Paul B. Rainey, and Christina D. Moon. 2017. “Anaerobically Grown Escherichia Coli Has an Enhanced Mutation Rate and Distinct Mutational Spectra.” *PLoS Genetics* 13 (1): e1006570.
Shin, Chi Y., and Mitchell S. Turker. 2002. “A:T → G:C Base Pair Substitutions Occur at a Higher Rate than Other Substitution Events in Pms2 Deficient Mouse Cells.” *DNA Repair*. https://doi.org/10.1016/s1568-7864(02)00149-0.
Tacutu, Robi, Thomas Craig, Arie Budovsky, Daniel Wuttke, Gilad Lehmann, Dmitri Taranukha, Joana Costa, Vadim E. Fraifeld, and João Pedro de Magalhães. 2013. “Human Ageing Genomic Resources: Integrated Databases and Tools for the Biology and Genetics of Ageing.” *Nucleic Acids Research* 41 (Database issue): D1027–33.
Tanaka, M., and T. Ozawa. 1994. “Strand Asymmetry in Human Mitochondrial DNA Mutations.” *Genomics* 22 (2): 327–35.
Tomasetti, Cristian, Rick Durrett, Marek Kimmel, Amaury Lambert, Giovanni Parmigiani, Ann Zauber, and Bert Vogelstein. 2017. “Role of Stem-Cell Divisions in Cancer Risk.” *Nature* 548 (7666): E13–14.
Tomasetti, Cristian, and Bert Vogelstein. 2015. “Cancer Etiology. Variation in Cancer Risk among Tissues Can Be Explained by the Number of Stem Cell Divisions.” *Science* 347 (6217): 78–81.
Uddin, Arif, and Supriyo Chakraborty. 2017. “Synonymous Codon Usage Pattern in Mitochondrial CYB Gene in Pisces, Aves, and Mammals.” *Mitochondrial DNA. Part A, DNA Mapping, Sequencing, and Analysis* 28 (2): 187–96.
Von Stetina, Jessica R., and Terry L. Orr-Weaver. 2011. “Developmental Control of Oocyte Maturation and Egg Activation in Metazoan Models.” *Cold Spring Harbor Perspectives in Biology* 3 (10): a005553.
Wallace, Douglas C. 2005. “A Mitochondrial Paradigm of Metabolic and Degenerative Diseases, Aging, and Cancer: A Dawn for Evolutionary Medicine.” *Annual Review of Genetics* 39: 359.
Wei, Wei, Salih Tuna, Michael J. Keogh, Katherine R. Smith, Timothy J. Aitman, Phil L. Beales, David L. Bennett, et al. 2019. “Germline Selection Shapes Human Mitochondrial DNA Diversity.” *Science* 364 (6442). https://doi.org/10.1126/science.aau6520.
Yuan, Yuan, Young Seok Ju, Youngwook Kim, Jun Li, Yumeng Wang, Yang Yang, Inigo Martincorena, et al. 2017. “Comprehensive Molecular Characterization of Mitochondrial Genomes in Human Cancers.” *bioRxiv*. https://doi.org/10.1101/161356.
Yuan, Yuan Yuan, PCAWG Consortium, Young Seok Ju, Youngwook Kim, Jun Li, Yumeng Wang, et al. 2020. “Comprehensive Molecular Characterization of Mitochondrial Genomes in Human Cancers.” *Nature Genetics*. https://doi.org/10.1038/s41588-019-0557-x.
Zaidi, Arslan A., Peter R. Wilton, Marcia Shu-Wei Su, Ian M. Paul, Barbara Arbeithuber, Kate Anthony, Anton Nekrutenko, Rasmus Nielsen, and Kateryna D. Makova. 2019. “Bottleneck and Selection in the Germline and Maternal Age Influence Transmission of Mitochondrial DNA in Human Pedigrees.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (50): 25172–78.

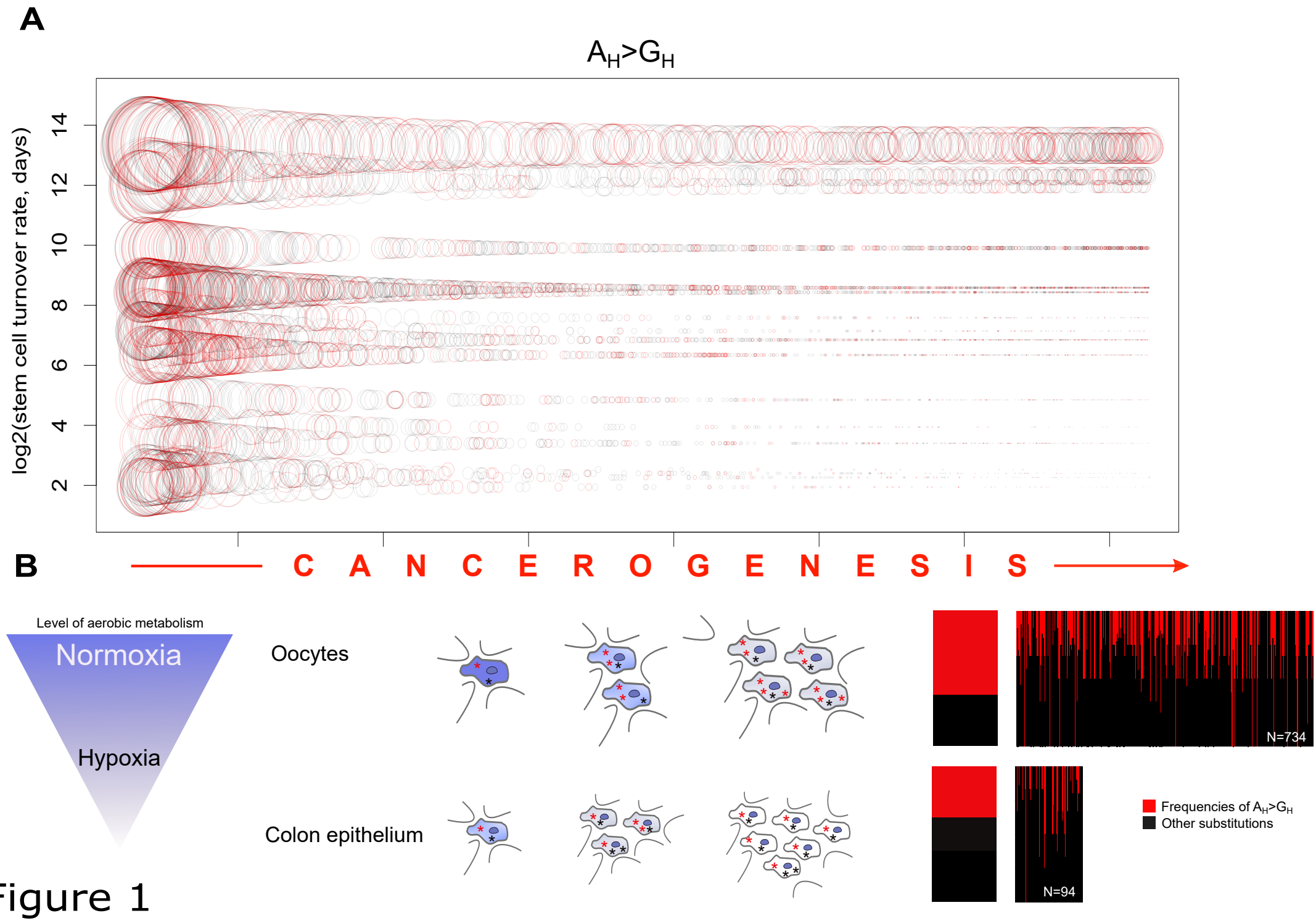
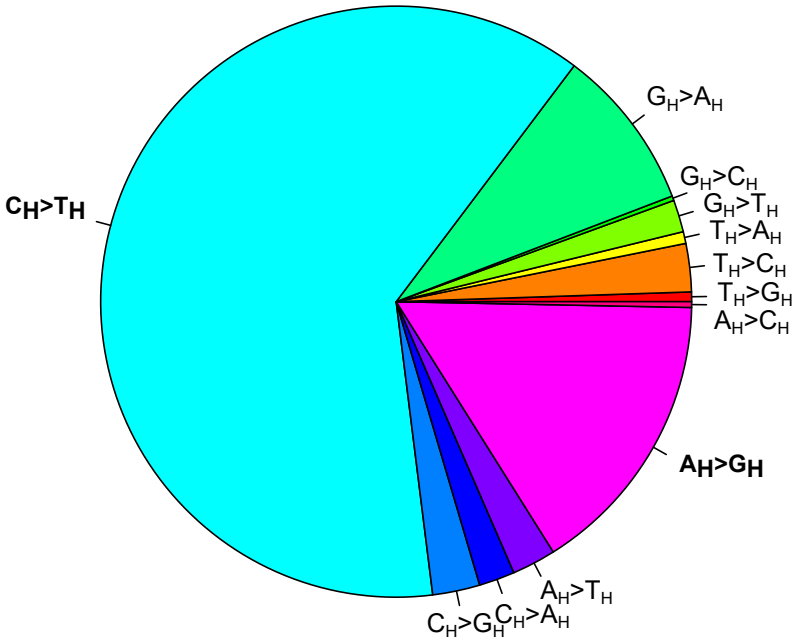
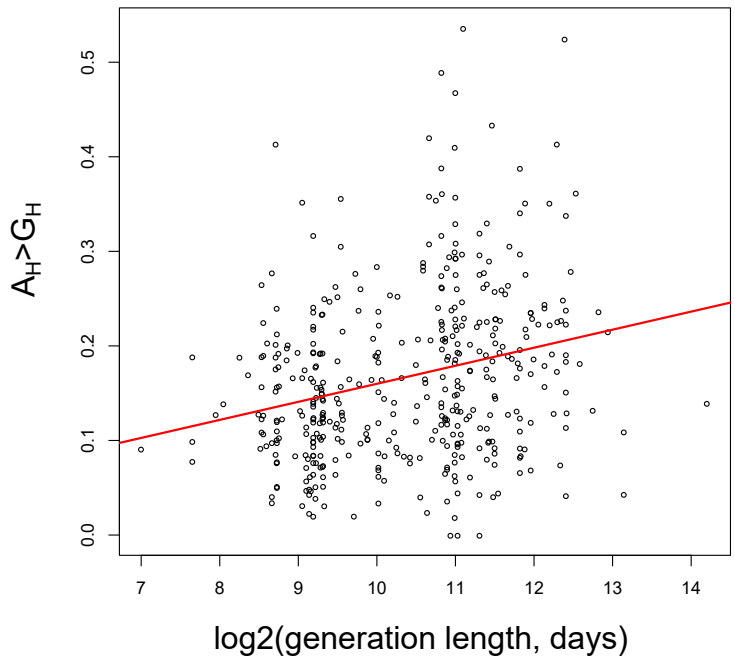

A
A_H>G_H
log2(stem cell turnover rate, days)
2
4
6
8
10
12
14
B
C A N C E R O G E N E S I S
Level of aerobic metabolism
Normoxia
Hypoxia
Oocytes
Colon epithelium
N=734
N=94
Frequencies of A_H>G_H
Other substitutions

Figure 1

A


$G_H>A_H$
$G_H>C_H$
$G_H>T_H$
$T_H>A_H$
$T_H>C_H$
$T_H>G_H$
$A_H>C_H$
$C_H>T_H$
$A_H>G_H$
$A_H>T_H$
$C_H>A_H$
$C_H>G_H$

B


$A_H>G_H$
0.0
0.1
0.2
0.3
0.4
0.5
7
8
9
10
11
12
13
14
log2(generation length, days)

C

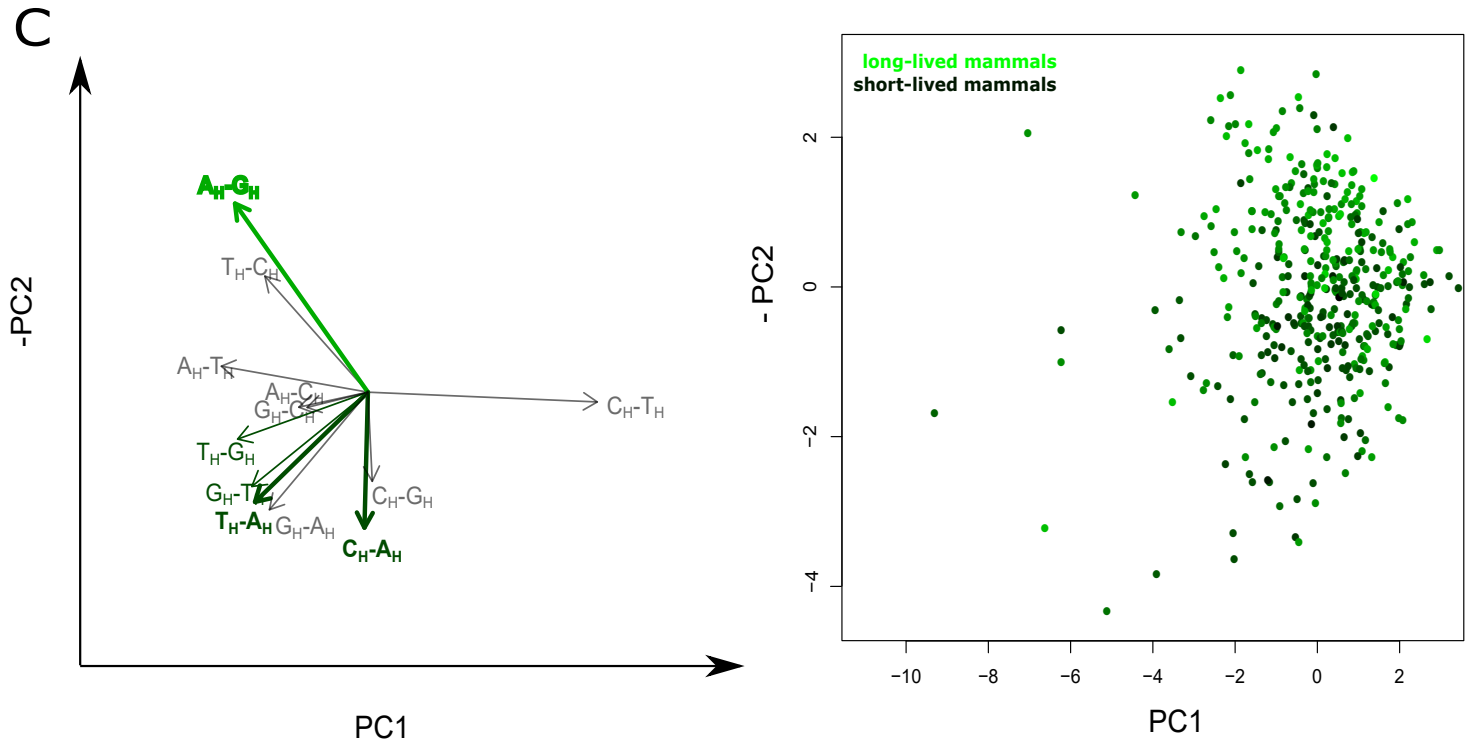
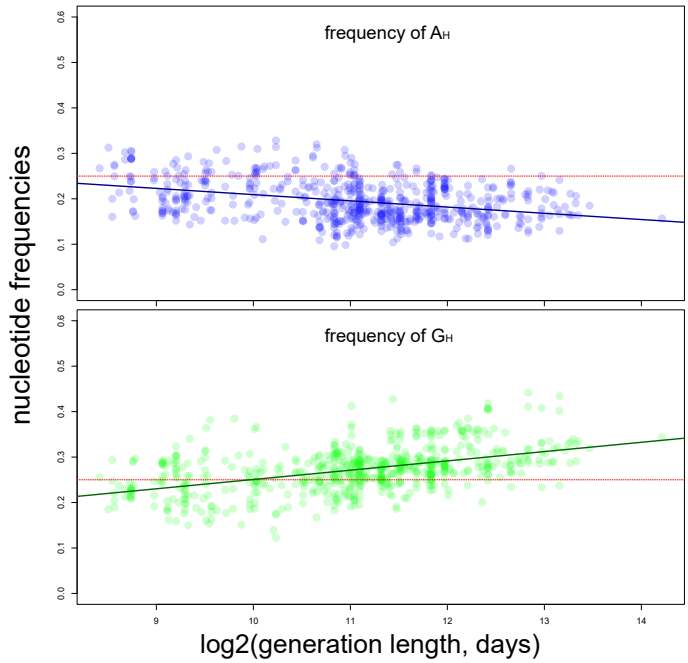
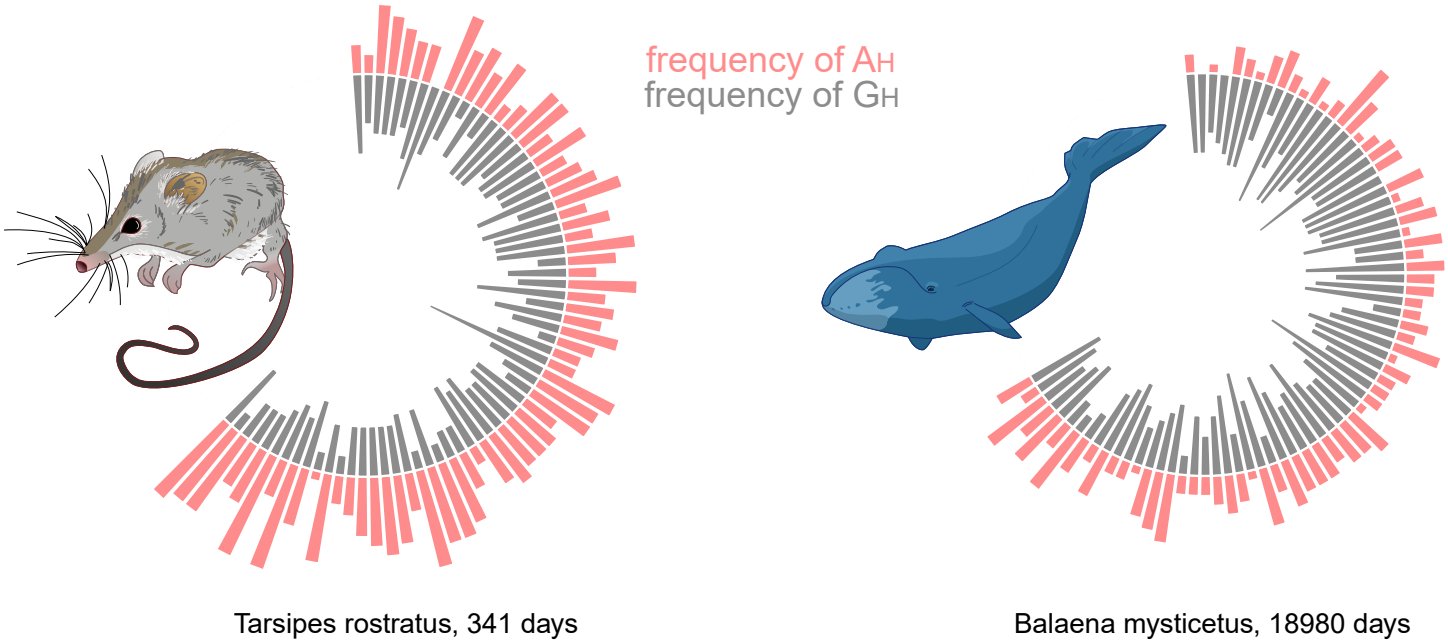

-PC2
PC1
A_H-G_H
T_H-C_H
A_H-T_H
A_H-C_H
G_H-C_H
T_H-G_H
G_H-T_H
T_H-A_H
G_H-A_H
C_H-A_H
C_H-G_H
C_H-T_H
long-lived mammals
short-lived mammals
- PC2
PC1
2
0
−2
−4
−10
−8
−6
−4
−2
0
2

Figure 2

A


frequency of AH
frequency of GH
nucleotide frequencies
log2(generation length, days)

B


frequency of AH
frequency of GH
Tarsipes rostratus, 341 days
Balaena mysticetus, 18980 days

C

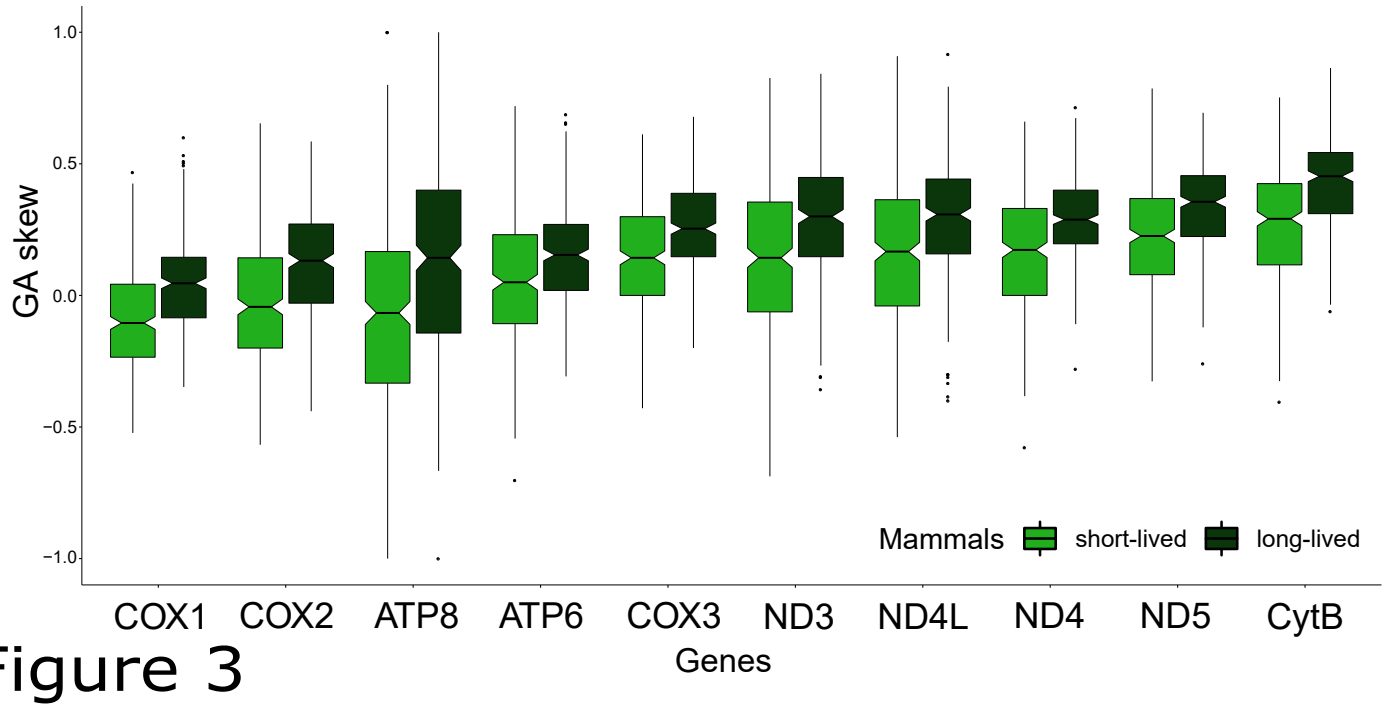
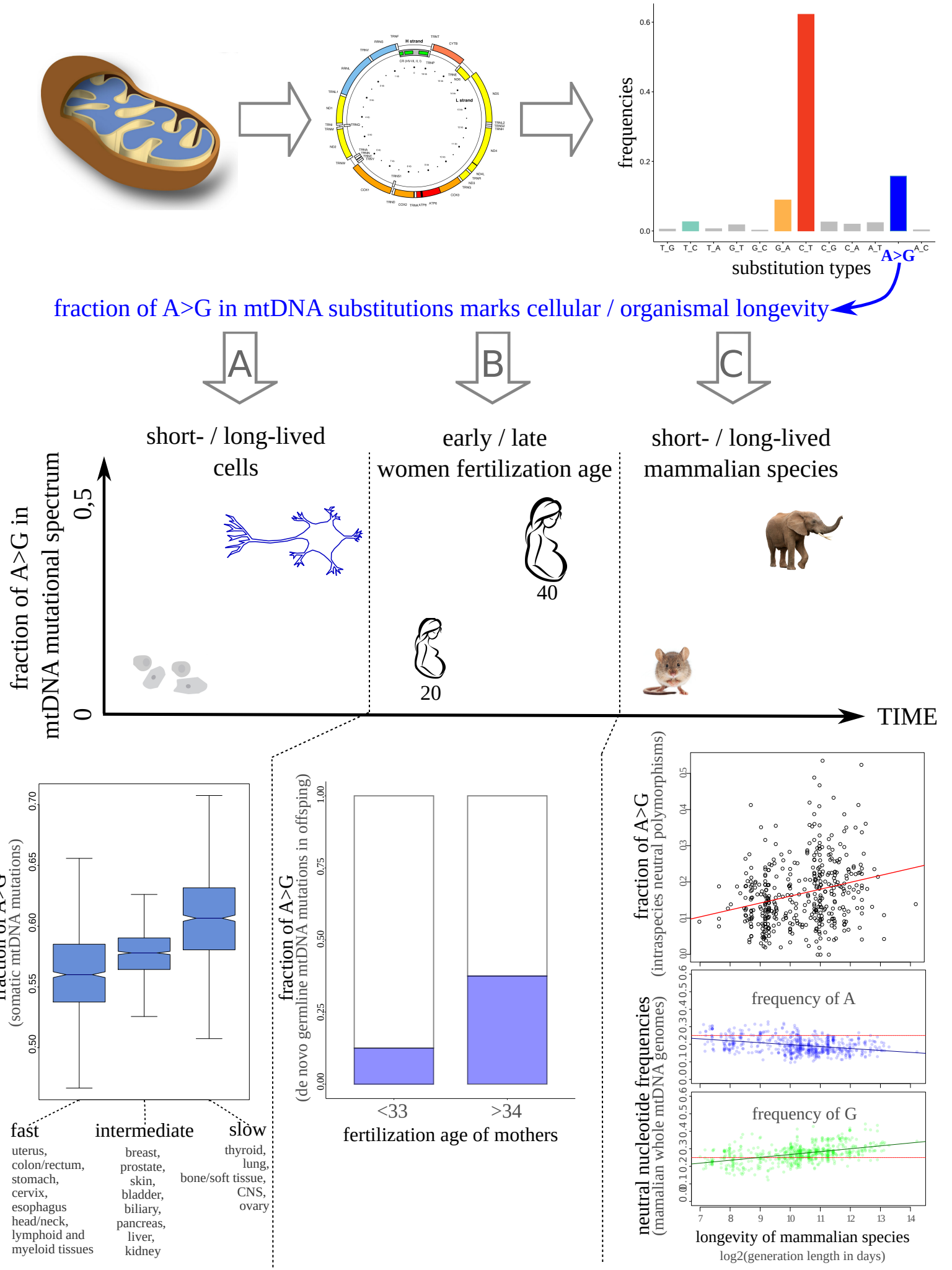

GA skew
1.0
0.5
0.0
−0.5
−1.0
Mammals
short-lived
long-lived
COX1
COX2
ATP8
ATP6
COX3
ND3
ND4L
ND4
ND5
CytB
Genes

Figure 3

Graphical abstract: why melanoma is similar to a mouse and glioblastome resembles an elephant?