Mitochondrial mutational spectrum provides an universal marker of cellular and organismal longevity

A. G. Mikhaylova[1]*, A. A. Mikhailova[1]*, K. Ushakova[1]*, E. Tretiakov[1,2]*, A. Yurchenko[3], M. Zazhytska[4], D.A. Knorre[5,6], E Zdobnov[7], Z. Fleischmann[8], S. Annis[8], M. Franco[8], K. Wasko[8], W.S, Kunz[9], I. Mazunin[1], S. Nikolaev[3], A. Reymond[4], K. Khrapko[8]**, K. Gunbin[1,10]**, K. Popadin[1,4,11]****

[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]*Gustave Roussy Cancer Center, Paris, France;*

[4]*Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland;*

[5]*The A.N. Belozersky Institute Of Physico-Chemical Biology, MSU, Moscow, Russian Federation;*

[6]*Institute of Molecular Medicine Sechenov First Moscow State Medical University Trubetskaya str. 8-2, 119991, Moscow;*

[7]*Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland;*

[8]*Northeastern University, Massachusetts, USA;*

[9]*Institute of Experimental Epileptology and Cognition Research, University Bonn, Sigmund-Freud-Str. 25, 53105 Bonn, Germany;*

[10]*The Institute of Cytology and Genetics of the SB RAS, Novosibirsk, Russian Federation;*

[11]*Swiss Institute of Bioinformatics, Lausanne, Switzerland.*

* *equal contribution*

** *equal contribution*

Abstract

It has been shown recently that mitochondrial (mtDNA) somatic variants are numerous enough to trace cellular lineages in our body. Here we extend this statement and demonstrate that mtDNA variants can be interpreted not only as neutral markers of cell divisions but the relative frequency of different mtDNA substitutions (i.e. mtDNA mutational spectrum) can inform us about important biological properties such as cell longevity. Analysing 7611 somatic mtDNA mutations from 37 types of human cancers and more than 2000 somatic mtDNA mutations from 25 healthy human tissues we observed that mtDNA mutational spectrum is associated with cell turnover rate: the ratio of T>C to G>A is increasing with cell longevity. To extend this logic we considered that, if universal, the discovered mutation bias may drive the differences in mtDNA mutational spectrum between mammalian species with short- ('mice') and long- ('elephants') lived oocytes. Based on presumably neutral polymorphisms in *MT-CYB* we reconstructed mutational spectra for 424 mammalian species and obtained that the fraction of T>C positively correlated with the species-specific generation length, which is a good proxy for oocyte longevity. Next, comparing complete mitochondrial genomes of 650 mammalian species we confirmed that exactly the same process shapes the nucleotide content of the most neutral sites in the whole mitochondrial genomes of short- (high T, low C) versus long- (low T, high C) lived mammals. Altogether analysing mtDNA mutations in time interval from dozens of years (somatic mutations) through the hundreds of thousands of years (within species polymorphisms) to millions of years (between species substitutions) we demonstrated that T>C/G>A positively correlates with cellular and organismal longevity. We hypothesize that the discovered mtDNA signature presents a chemical damage which is associated with the level of oxidative metabolism which, in turn, correlates with cellular and organismal longevity. The described properties of mtDNA mutational spectrum shed light on mtDNA replication, mtDNA evolution of mammals and can be used as a marker of cell longevity in single-cell analyses of heterogeneous samples.

Introduction

It has been recently shown, that because of high mitochondrial DNA (mtDNA) mutation rate and high mtDNA copy number per cell somatic variants in mtDNA are informative to trace different cellular lineages in our body (Ludwig et al. 2019). Here we hypothesize further that mitochondrial variants can be interpreted not just as neutral markers of cellular lineages but the relative frequency of different mtDNA substitutions (i.e. mtDNA mutational spectrum) may also contain a functional signature of cellular properties such as the level of metabolism or turnover rate.

Indeed there are several properties of mtDNA mutagenesis which make it especially predisposed to sense cell-specific metabolism. First, during asynchronous mtDNA replication parental heavy chain is spending significant amount of time in single-stranded state during which (i) spontaneous deamination of adenine to guanine is increasing as a linear function of a time (hereafter we will follow widely accepted in mitochondrial community light chain notation according to which A>G on heavy chain is called T>C on light chain), (ii) spontaneous deamination of cytosine to thymine (G>A, light chain notation) is less sensitive to the time being single-stranded and (iii) all other rare transitions and transversions are not sensitive at all (Faith and Pollock 2003). These different time dynamics of substitutions will lead to unique signatures (ratios of substitutions) corresponding to various time intervals of mtDNA replication. Second, mitochondria is tightly involved in the determination of the cell specific level of oxidative metabolism and thus mutagens, associated with it are expected to affect mtDNA directly and strongly (Ericson et al. 2012). Third, cancer studies have shown that mtDNA is apparently affected by very strong endogenous mutagen because it completely overwhelms effects of other expected exogenous mutagens such as tobacco smoke in lung cancers or ultraviolet light in melanomas (Yuan et al. 2017; Ju et al. 2014). All these properties (time dependent dynamics, tight involvement into the determination of the oxidative metabolism, low sensitivity to exogenous mutagens) make mtDNA mutational spectrum a good potential sensor of cellular metabolic processes.

In our work we analysed several cellular phenotypes, related to the level of metabolism such as the number of mtDNA copies, expression level of mitochondrial genes and the turnover rate of cells from different tissues. Correlating mtDNA mutational spectrum (i.e. probabilities of different substitutions) with all these cellular phenotypes we found the only robust relationship with cell turnover rate. This might show that cell turnover rate is indeed one of the most important cellular phenotypes, affecting mutagenesis (Tomasetti and Vogelstein 2015; Tomasetti et al. 2017). Interestingly, this metric due to its simplicity allowed us to extend our analyses to other mammalian species. Since in all mammals mtDNA is inherited exclusively through oocytes, and the division of oocytes is arrested after the birth, the species-specific generation length well enough approximates the longevity of oocytes from different mammalian species (i.e. time which mtDNA spent in dormant oocyte) and thus might correlate with specific mtDNA signature of longevity.

Altogether, using a collection of 7611 somatic mtDNA mutations from 37 types of human cancers (TCGA and ICGC), more than 2000 somatic mtDNA mutations from 25 human healthy tissues (GTEx), 39112 polymorphic synonymous substitutions in four-fold degenerate *MT-CYB* sites from 424 mammalian species and nucleotide content in whole mitochondrial genomes of 650 mammalian species we observed one universal trend: ratio of T>C to G>A positively correlates with cellular and organismal longevity. Out of several potential explanations of the observed associations we prefer one according to which T>C/G>A is sensitive to the level of oxidative metabolism which is expected to be higher in cancer cells in early versus advanced stages, in slow-dividing versus fast-dividing tissues and in long- versus short-lived species. The discovered mtDNA signature of cellular and organismal longevity helps to understand better the mechanism of mtDNA replication, shed light on evolution of mammalian mtDNA and opens a possibility to mark cell longevity in heterogeneous tissues.

Results and discussion

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

Recent survey of somatic mtDNA mutations in human cancers demonstrated that the majority of substitutions are G>A and T>C transitions irrespectively of the cancer type (Yuan et al. 2017). In order to compare the minor differences in mtDNA mutational spectrum between different cancers and associate them with cellular phenotypes we performed two analyses. First we checked if mtDNA mutational spectrum is changing during the transformation of healthy cells to cancerous ones (this might be driven for example by the increase in cell turnover rate or switch to more glycolytic metabolism). Second we tested if mtDNA mutational spectrum depends on properties of the ancestral tissues (precancerous cells), such as tissue specific turnover rate. To perform both these analyses we worked with 7611 somatic mitochondrial mutations observed in 37 cancer types grouped into 21 ancestral tissues (Yuan et al. 2017).

1.a) mtDNA mutational spectrum is changing during tumorigenesis

MtDNA somatic mutations observed in human cancer cells can represent two populations: (i) early mutations which occurred before the tumorigenesis and expanded later by clonal expansion of cancer cells and (ii) late mutations, originated in cancer cells. To split all somatic mutations into presumably early and presumably late we can use variant allele frequencies (VAF) observed in cancer tissues. To prove this logic we compared VAF in matched cancer and normal tissues and observed positive correlation (Spearman's rho = 0.07, p = 1.226e-09, N = 7611; if we take into account only variants with non-zero VAF in normal tissues, the correlation is still positive: Spearman's rho = 0.09, p = 9.638e-11, N = 5436; see also figure 1A) suggesting that many mtDNA somatic mutations indeed originated before the tumorigenesis and that cancer VAF can approximate the time of origin of mutations: high VAF marks mutations before the tumorigenesis and low VAF - after the tumorigenesis.

Using cancer VAF as an approximation for the time of origin of the variant we wanted to observe how mutational spectra is changing during tumorigenesis. First, we have split all variants by the median value of VAF (4.5%) and estimated Ts/Tv for variants with low and high VAFs as 10.3 and 14.9 correspondingly (Fisher Odds Ratio = 0.69, p = 2.524e-05, N = 7611, figure 1B left panel). This result is robust to an elimination of potentially low-quality variants: when we removed 25% of variants with the lowest p-values (variants with -log10(p-value) <= 59) we saw nearly identical result (median of VAF = 9.4%, Fisher Odds Ratio = 0.67, p = 0.0002, N = 5709). Both common transitions T>C and G>A contribute significantly to this trend: T>C/Tv: Fisher Odds Ratio = 0.68, p = 5.26e-05; G>A/Tv: Fisher Odds Ratio = 0.72, p = 0.0002.

Second, to rule out the potential effect of different cancer types, we tested if there is an increase in Ts/Tv with VAF for each cancer type individually (Figure 1B middle panel). Only for Pediatric Brain Cancer (PBCA) we observed significant increase in Ts/Tv with VAF (median of VAF = 10.1%, Fisher Odds Ratio = 0.06, p = 1.14e-06, N = 186). However, when we merged the trends observed in all cancer types together, we observed the increase in Ts/Tv with VAFs (p = 0.0006, paired Mann-Whitney U-test, N = 36 cancer types, figure 1B middle panel). This trend is still significant if we remove PBCA from the dataset (p = 0.001, paired Mann-Whitney U-test, N = 35 cancer types, figure 1B middle panel). The same trend is supported by both common transitions separately: T>C/Tv (p = 0.0004, paired Mann-Whitney U-test, N = 36) and G>A/Tv (p = 0.0006, paired Mann-Whitney U-test, N = 36).

Third, to rule out additional confounders such as different purity of samples, we tested if there is an increase in Ts/Tv with VAF within each sample. To do this we used a subset of samples with at least one transition and at least one transversion (N = 419) and estimated mean VAF for Ts and Tv in each sample. Comparing them within each sample we observed higher VAF for Ts than for Tv (p = 5.109e-15, Mann-Whitney Paired test, N = 419) and correspondingly log ratios of VAF(Ts)/VAF(Tv) was higher than expected value of zero (median of the log ratios of Ts/Tv is 1.1665; p-value < 2.2e-16, Wilcoxon test with mu = 0; figure 1B right panel). Both common transitions contribute significantly to this trend: T>C/Tv (p =1.501e-06, Mann-Whitney Paired test, N = 286) and G>A/Tv (p = 2.651e-09, Mann-Whitney Paired test, N = 346). In order to analyze contribution of T>C versus G>A we compared VAF(T>C) and VAF(G>A) within each sample and found out that T>C is characterized by slightly higher VAF as

compared to G>A (median of the differences between VAF (T>C) and VAF(G>A) is 0.1%, p =0.0097, Mann-Whitney paired test, N = 983), meaning that T>C on average happens more early on than G>A.

Putting together all these analyses: the integral one (Figure 1B left), the cancer-type-specific one (Figure 1B middle) and sample-specific one (Figure 1B right), we conclude that mutational spectrum is changing during tumorigenesis - becoming less transition rich and especially less T>C rich.

The changes in mutational spectra during the tumorigenesis might be associated with changes in cell turnover rate or with a switch from aerobic to non-oxidative (glycolytic) type of metabolism. It has been shown, for example, that 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 (G>A and T>C), and since they are expected to be the results of the reactive oxygen species-mediated mtDNA damage authors hypothesized that the changes in the mutational rate were associated with the shift from oxidative phosphorylation to anaerobic glycolysis. Authors proved this hypothesis by analyzing the relative expression of protein markers for glycolysis and oxidative phosphorylation. Taking into account that literally all cancer types are switching to glycolysis (Rosario et al. 2018) we can interpret VAF as the level of aerobic conditions at a given moment in cancer development: from the highly aerobic precancerous stages (high VAF) to more glycolytic advanced cancers (low VAF). Our results, based on 37 cancer types, confirm and extend the conclusion of this paper (Ericson et al. 2012) and suggest that T>C might be a better mark for the level of oxidative metabolism as compared to G>A. If chemical damage, associated with cell turn over rate is the driver of the observed changes in mutational spectra, we can expect that comparison of various cancer types, derived from tissues with different cell division rate may reveal the same trend.

1.b) mtDNA mutational spectrum differs between cancer types

Using an approximate turnover rate of stem cell divisions in each of 21 normal tissue samples (Supplementary materials), ancestral to our analyzed cancer types we split all cancer tissues 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, CNS, ovary). We estimated Ts/Tv for each group as 10.0, 12.6 and 14.5 correspondingly (the observed difference is robust according to a thousand 50% jackknife re-samplings, Figure 1C left panel). Both common transitions T>C and G>A contribute to this trend (T>C/Tv: 2.8, 3.8, 4.6; G>A/Tv: 5.0, 6.6, 7.5 for fast, intermediate and slow-replicating tissues, figure 1C left panel). Since transversions are very low frequency variants and thus some of them might be sequencing errors (Chen et al. 2017) we wanted to see whether this trend would still be visible when analyze only transitions. We estimated the fraction of T>C to all other transitions (C>T, G>A, A>G) as well as fraction of G>A to all other transitions (C>T, G>A, A>G) and observed a corresponding increase in the fraction of T>C (T>C/Ts: 0.39, 0.44, 0.46) but not in the fraction of G>A (G>A/Ts: 1.00,1.12, 1.08) (Figure 1C middle panel). Finally, we compared the effects of these common transitions with each other and observed an increase in T>C fraction over G>A in slow- versus fast-dividing tissues (T>C/G>A: 0.56, 0.58, 0.61 for fast-, intermediate and slow-replicating tissues, Figure 1C right panel). We conclude that the fraction of T>C is increasing from fast- to slow-replicating tissues.

Putting together the two observations described above: the increase in Ts/Tv with VAF mainly because of T>C (Figure 1B) and increase in T>C in slow dividing cancers (Fig 1C), we can expect that a probability of a T>C transition is a function of both cancer type and VAF. To test it we performed multiple logistic regression and observed, that probability of a T>C transition is indeed increasing with VAF and turnover rate of the tissue (Figure 1D). The results are quantitatively similar and sometimes are even stronger if we use dummy variables, coding for fast-replicating cancers, instead of turnover days; eliminate 25% of the most rare variants (with VAF less or equal 1.74%); repeat this analysis without transversions; add to the model patient age and the number of mitochondrial copies in cancer samples (Supplementary materials). Altogether we demonstrated that somatic T>C substitutions are associated with slow-replicating cells, which are expected to be more frequent (i) before tumorigenesis, and (ii) in cancers derived from slow-replicating tissues.

Interestingly, there is a unique breast cancer sample with the maximal number of somatic mtDNA mutations: 33 mutations, 29 of which are T>C (Yuan et al. 2017). The vast majority of these mutations is concentrated near the

origin of replication of the light chain and they have very similar VAFs, meaning that all of them have been originated at the same mitochondrial genome as one complex mutation. We suggest, that this event was associated with replication fork stalling, which as a result of long time exposition of the single stranded heavy chain to a mutagen lead to massive spontaneous deamination of adenine to guanine (T>C on light chain). Interestingly, if we analyze top 3 mutated cancers in all these dataset of Yuan et al we see an excess of T>C substitutions (82%, 38% and 72% as compared to the expected average of 34.4%) which is against the common logic that the most abundant type of transitions is G>A.

(2) mtDNA mutational spectrum derived from somatic mutations in healthy human tissues

Next we expanded our analysis considering somatic mtDNA mutations from healthy human tissues. Using a collection of somatic mtDNA mutations from GTEx tissues (Ludwig et al. 2019) we derived the mutational spectrum for 25 tissues with ten or more somatic mutations and tested for an association between mutation spectrum and expression level of mitochondrial genes, number of mtDNA copies and tissue-specific cell longevity (Supplementary materials). The only significant correlation was observed between G>A and cell longevity (Spearman's rho = -0.532615, p-value = 0.006125, N = 25, Figure 2A). This negative correlation was robust to the effect of VAFs and expression level of mitochondrial genes. If we estimate ratio of T>C/G>A the effect is reverting to positive as expected (Spearman's rho = 0.5791803, p = 0.002415, N = 25).

It has been shown recently that the number of de novo mtDNA heteroplasmies in children is increasing with maternal age (Rebolledo-Jaramillo et al. 2014), that might be attributed to oocyte aging (oocyte longevity). In order to test which type of nucleotide substitutions drives this correlation we reanalyzed all de novo germline mutations from this study and found that T>C is characterized by the highest age of fertilization (average age of fertilization is 34.1, N = 4) as compared to G>A (average age of fertilization is 29.9, N = 4) or all other substitutions (average age of fertilization is = 32.1, N = 12). Due to the small sample size (16 mutations in total) this trend was not significant (p = 0.19 if we compare T>C with all other substitutions and p = 0.0956 if we compare T>C with G>A; one-sided Mann-Whitney U test) and future mtDNA analyses of human mother-offspring duos are needed.

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

Till now we compared different cells and tissues from the human body and observed that the fraction of mtDNA transitions T>C positively correlates with cell longevity (Figure 1) while the fraction of G>A - negatively (Figure 2). If cell longevity is the main driver of the observed association we may extend our logic and analyse oocytes from different mammalian species. First, 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). Second, mammalian oocytes are arrested from birth to puberty, which takes months (mouse) or decades (humans) (Von Stetina and Orr-Weaver 2011) and allows us to use species - specific generation length as a good proxy for oocyte longevity in different mammalian species. Thus in this chapter we switch from somatic mtDNA mutations to germline mtDNA mutations and analyse intraspecies mtDNA polymorphisms in different mammalian species. Correspondingly in our analyses we change the time scale from several dozens of years (human somatic mutations) to hundreds of thousands of years which is the average time of segregation of neutral mtDNA polymorphisms (Atkinson, Gray, and Drummond 2008).

Mitochondrial polymorphisms in mammals are especially advantageous to reconstruct species-specific mtDNA spectrum because they are extensively used in ecological, evolution and population genetics studies, for example *MT-CO1* was selected for the DNA barcoding project (Hebert et al. 2003). Several studies revealed that there is variation in mtDNA mutational spectra between different species (Montooth and Rand 2008; Belle et al. 2005), however, no overarching explanation to explain these differences has been suggested so far. For 611 mammalian species we collected all available intra-species protein-coding mtDNA sequences, reconstructed within-species phylogeny using sequences from sister species as outgroups, reconstructed sequences at each internal node, derived a list of polarized single-nucleotide substitutions between the nodes, and, focusing on the most neutral synonymous fourfold degenerate sites, counted the observed frequencies of twelve types of nucleotide substitutions (Figure 3A left panel). We observed strong excess of transitions over transversions which is typical for animal mtDNA (Belle et al. 2005). These fractions of observed substitutions depend on the frequencies of ancestral nucleotides in the

synonymous fourfold degenerate sites, which are highly non-uniform (Figure 3A middle panel). After normalizing the observed nucleotide substitutions (Figure 3A left panel) by the ancestral nucleotide frequency (Figure 3A middle panel) we obtained a mutational spectrum as a probability of a given nucleotide to mutate to any other nucleotide (Figure 3A right panel) irrespective of its frequency. Hereafter in all downstream analyses of this chapter we used these species-specific mutational spectra, presented as vectors of twelve probabilities with the total sum equal to one.

To uncover the nature of mitochondrial mutagenesis shaping the species-specific variation in mutational spectra we focused on *MT-CYB*, which represent 56% of all analyzed mutations (39,112 out of 70,053 used to draw Fig 3A) 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 first and simplest approximation of the mutational spectrum we used transition transversion ratio (Ts/Tv). For 424 mammalian species with known Ts/Tv and generation length we observed a positive correlation between Ts/Tv and generation length (Figure 3B left panel; spearman's rho = 0.23, p-value = 2.021e-06; N = 424). To test the robustness of this correlation we performed several additional analysis and showed that the results were not affected by phylogenetic inertia, varying number of mutations between different species and different algorithms of reconstruction of ancestral sequences.

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 T>C positively correlated with generation length (Spearman's rho = 0.252, p value = 1.188e-07), while several rare transversions showed negative correlation (A>T, A>C, G>T and C>A: all Spearman's rhos < -0.17, all p-values are < 0.0003). We included these five types of substitutions into a multiple linear model and determined by removing in a stepwise manner the most non-significant variable that three types of substitutions were independently associated with generation length (*GL*):

log2(GL) ~ 10.39 + 0.26(T>C) - 0.18*(A>T) - 0.16*(G>T), N = 424, total p-value = 1.249e-10; p-values of transversions are < 0.002, p-value of T>C = 6.42e-06, R^2=0.104, presented coefficients are scaled.* *equation (I)*

Comparing the scaled regression coefficients we observed that the strongest (the highest effect size), as well as the most significant (the lowest p value) effect was associated with T>C transitions (Figure 3B right panel). Positive correlation of generation length with the frequency of a T>C transition and negative correlation with the frequencies of G>T and A>T transversions as expected lead to the increased Ts/Tv in long-lived species described above (Figure 3A). Of note the inclusion in the linear model of the number of mutations used to estimate the species-specific mutational spectrum does not affect these results quantitatively (Supplementary materials).

In order to derive 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 G>A substitutions, whereas the second is driven mainly by T>C substitutions (Figure 3C 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 3C right panel). Interestingly, the correlation of the second principal component with generation length was significantly higher (rho = -0.39, p < 2.2e-16, N = 424) as compared to the sole effect of T>C (Spearman's rho = 0.252, p value = 1.188e-07, N = 424) and the sole effect of Ts/Tv (Spearman's rho = 0.23, p-value = 2.021e-06; N = 424), suggesting that this second component could reflect a complex signature of a specific mutagen associated with generation length. Indeed, both transversions A>T and G>T, negatively associated with generation length (equation I), have strong effects on the second principal component and, as expected, point to opposite direction compared to T>C (Figure 3C left panel). Again inclusion in the analysis of the number of mutations used to estimate the species-specific mutational spectrum did not affect the results quantitatively, nor did it changed the separate analysis performed on species having more than 60 mutations (Supplementary materials). The first principal component, loaded mainly by the most common G>A transition, appears to represent an unknown source of variation, which we were unable to associate with any life-history traits such as generation length, body temperature, body mass, metabolic rate or the number of mutations used to derive the mutational spectrum.

Since some low frequency genetic variants might be due to DNA damage (Chen et al. 2017, 2018; Stewart et al. 2018) we replicated our results taking into account only two the most common transitions: T>C and G>A. Deriving T>C as a fraction of G>A plus T>C, we still observed the positive correlation with generation length (Spearman's rho = 0.21, p = 8.441e-06, n = 424).

The fraction of T>C in each species-specific mutational spectra depends on both the number of observed T>C substitutions within a species and the number of T nucleotides in fourfold degenerate synonymous position of *MT-CYB* of a given species, used for normalization (Figure 3A). To demonstrate that the observed substitution pattern (Fig 3B) was not solely due to the variation in the frequency of ancestral nucleotides, we recalculated T>C fraction for each species using only substitutions from T: T>C / (T>C + T>G + T>A). Whereas this means that we did not take into account species-specific nucleotide content , we still observed a positive correlation between T>C and the mammalian generation length (Spearman's rho = 0.164, p = 0.0007, N = 424).

(4) The long-term effect of the mutational bias: neutral nucleotide content of the whole mitochondrial genome

Then we assessed if the whole-genome nucleotide content is affected by the species-specific mutational spectra. Mutational bias may affect long-term changes in nucleotide content in the absence of selection. To test it we used 650 complete mitochondrial genomes of mammalian species and for each of them we estimated nucleotide content of the most neutral sites: synonymous fourfold degenerate positions in all 13 protein-coding genes without overlapped regions between neighbor genes. We expect, that discovered above mutational bias - increased T>C in long-lived mammals - will shape the nucleotide content in these presumably neutral sites.

4.a) Neutral positions of mtDNA in long-lived species are depleted in T and enriched in C

T>C substitutions are more common in long-lived species (Figure 3B, 3C). To test this effect on the whole mitochondrial genome nucleotide content we correlated the nucleotide content in neutral sites of different species with their generation time. We were able to analyze 650 mammalian species in terms of both genomic (mtDNA nucleotide content) and ecological (generation length) data. Testing all pairwise correlations between the generation length and presumably neutral nucleotide content we observed two the strongest correlations: positive with C and negative with T (Figure 4A left panel). Stepwise backward multiple linear model confirmed the importance of these two nucleotides: in the final model, after the removing of all non-significant variables, generation length depends negatively on T and positively on C:

log2(GL) = 11.06 - 0.11(fraction of T) + 0.46*(fraction of C), p-value of the fraction of T is 0.023, p-value of the fraction of C is <2e-16, R^2 = 0.229, presented coefficients are scaled.* equation (ii)

Consideration of phylogenetic inertia by means of phylogenetically independent contrasts demonstrate the same trends, however less significant (fraction of T: spearman's rho = - 0.09, p = 0.025; fraction of C: spearman's rho = 0.09, p = 0.021). Analysis of individual signals from each of 13 genes demonstrated quantitatively similar results.

4.b) AT and GC skews are more pronounced in long-lived species

All animals have positive AT skew (excess of A as compared to T) and negative GC skew (deficit of G as compared to C) on the light chain of mtDNA. Two most common transitions G>A and T>C are expected to drive it since each such substitution increases AT and decreases GC skew. Generation-time sensitivity of T>C should lead to the more pronounced skews (more positive AT skew and more negative GC skew) in long-lived versus short-lived species. Estimating AT and GC skew for each species based on all neutral sites of thirteen genes we observed indeed that long-lived species have strongly increased AT skew (Figure 4A right panel; spearman's rho = 0.12, p = 0.0025, N = 650) and a slight trend towards the decrease in GC skew (spearman's rho = -0.06, p = 0.14, N = 650). Interestingly when we focus on long-lived mammals only (with generation length higher than the median log2(GL) > 11.1 the negative correlation of the GC skew with generation length became much more evident: spearman's rho -0.31, p = 8.205e-09, N = 323). The more pronounced AT and GC skews in long-lived mammals is supported individually by each of the 13 protein-coding genes.

4.c) Nucleotide gradient along mtDNA is more pronounced in long-lived species

Till now we have been analyzing the average properties of the whole mtDNA such as average nucleotide content and average skews. It has been proposed, that the majority of mtDNA mutations occurs in a single-stranded state during replication of the heavy chain of mtDNA and thus the mutation rate and spectrum may depend on a particular region of mtDNA forming a gradient along the genome (Faith and Pollock 2003). We are interested if these changes (the strength of the gradient) differ between short- and long- lived species as a result of the mutation bias. We plotted nucleotide content at neutral sites of each of 13 genes and correlated it with gene location, ranked from the shortest (MT-CO1) to the longest time (MT-CYB) of being single-stranded (Figure 4B). We observed that the frequency of C is increasing along the gradient of time of being single stranded (Spearman's Rho >= 0.62, p-values <= 0.028). Given the fact that this analyses is limited by the low number of genes (13), we performed more sensitive analysis splitting the genome of each species into 50 windows containing 25 neutral (fourfold degenerate synonymous) nucleotides and using only genes located on the same strand of the major arc (all genes except ND1, ND2 and ND6). For each window we estimated the frequencies of four nucleotides as well as AT GC skew and for each species we calculated the Spearman Rho coefficient and p-value, estimating the changes in a given metric along the genome (Figure 4C). We can see, that the frequency of C indeed has the strongest bias towards positive rho values, and additionally we can see also, that the second strongest bias belongs to T, which is decreasing along the genome (Figure 4C). In line with the increase in C frequency GC skew is becoming more negative along the genome, and in line with the decrease in T frequency AT skew is becoming more positive along the genome (Figure 4C). To test if some of these gradients are more pronounced in long versus short lived mammals, we estimated the fraction of long-lived mammals (with generation length more than median, 1101.1 day) among significant (with p-values <=0.01) rho values. Only for C nucleotides and GC skew we observed significant excess of long-lived mammals among the significant ones (Figure 4C, odds ratio = 1.61, p-value = 0.003521 for C nucleotide, Fisher test; odds ratio = 1.75; p-value = 0.003534, Fisher test; all other p-values > 0.1). This means, that the rate of increase in the frequency of C as well as the rate of decrease in GC skew is faster in long- versus short- lived mammals.

Altogether, our results demonstrate, that the nucleotide content in synonymous fourfold degenerate positions reflects the mutational bias from T to C: long-lived mammals are more T poor and C rich, have stronger AT and GC skew (Figure 4A), have stronger gradient (decrease in T and increase in C) along the genome (Figure 4B, 4C). All these correlations support our hypothesis, that the majority of these sites are effectively neutral and that the mtDNA mutational bias is stable over long period of time. Similar conclusions about neutrality of mtDNA codon usage has been derived from the comparison of somatic cancer-derived mutational spectrum of human mtDNA with corresponding mtDNA codon usage in our genome (Ju et al. 2014) as well as from the analysis of mtDNA codon usage in vertebrate species (Uddin and Chakraborty 2017; Faith and Pollock 2003).

(5) The model of mtDNA mutagenesis

We demonstrated that mtDNA mutational spectrum is changing during tumorigenesis, between tissues and species: T>C/G>A is increasing with longevity on different time scales - from several years to millions of years. We propose that the same background mechanism explains all our findings. Taking into account extremely low VAFs of analyzed heteroplasmic mtDNA mutations in both human datasets (with median VAF < 5%) and presumed absence of selection on synonymous four-fold degenerate sites in mtDNA of mammals (Faith and Pollock 2003; Uddin and Chakraborty 2017) we don't consider potential effect of selection and propose, that the observed changes in mutational spectrum are driven solely by mutagenesis.

To reveal a potential mechanistic explanation of all our findings we considered the current model of mtDNA mutagenesis (Faith and Pollock 2003). During asymmetric replication of mtDNA the heavy chain spends a significant amount of time in single-stranded condition where the majority of these transitions occur: G>A (C>T on heavy chain) and T>C (A>G on heavy chain). G>A is characterized by the highest frequency but it is just slightly sensitive to the time spent single-stranded; oppositely the probability of T>C is modest, but it linearly increases with time spent single-stranded (see Fig 4 in (Faith and Pollock 2003)). Taking into account weak increase in the probability of G>A and strong increase in the probability of T>C, the fraction of T>C is increasing while the fraction of G>A is decreasing with the time spent single-stranded, correspondingly leading to an increase in the T>C to G>A ratio (Figure 5). Parsimoniously, all our results might be explained if duration of mtDNA replication differs between cell types and species: if it is shorter in short-lived cells / species and longer in long-lived cells / species. In

this case on average short-lived cells will accumulate less T>C (correspondingly low T>C/G>A) while long-lived cells will accumulate more T>C (correspondingly high T>C/G>A). However, the process of mtDNA replication is most likely highly constrained and currently there is no information related to the variation in replication timing between cells and species. Thus, some additional, but not mutually exclusive scenarios might explain why accumulation of T>C is increased in long-lived cells. For example, there is a possibility that mtDNA replication modes (asynchronous versus strand-coupled: (Herbers et al. 2019; Cluett et al. 2018) alternate in each species / cell line and the relative proportion of these modes may affect the mutational signature. Asymmetrical mode of replication is more common and it is expected to be marked by the linear increase in T>C with time spent single-stranded. If so, increased T>C will mark the preponderance of asymmetrical over strand-coupled modes of replication in long-lived cells. Currently there are no well established links between the balance of the modes of replication in different cell lines, however it is evident that the switch between modes may depend on accumulation of damaged mtDNA, energy consumption, as well as the availability of mitochondrial transcripts (mitochondria transcription activity) ((Herbers et al. 2019; Cluett et al. 2018)). Apart from the replication time and switch between replication modes our results might be explained if the intercept of T>C regression line is higher in long-lived cells / species because of specific chemical mutagens. We suppose that the level of oxidative metabolism (or, oppositely - frequency of glycolysis) might be one of the universal causes. Long-lived species, slow-dividing tissues and cancer cells in early stages might be characterized by proportionally higher fraction of mtDNA replications under high oxidative conditions as compared to short-lived mammals, fast-dividing tissues and cancer cells in the more advanced stages of cancer.

Evolution is a function of mutagenesis and selection. To understand it we have to split observed DNA changes into the ones, driven by the mutation process and the others, driven by selection. For example, long-lived mammals have a greater fraction of GC in mitochondrial genomes than short-lived mammals (Lehmann et al. 2008), but without additional investigation it is impossible to disentangle whether this is the result of positive selection (for example GC rich mtDNA could provide a selective advantage to long-lived mammals due to its higher stability in somatic cells) or a neutral consequence of mutational bias (more mutations towards GC in long-lived mammals). According to our results the increased mutational bias from T to C in long-lived species will lead to higher fraction of C in mtDNA and thus, the positive correlation between GC and longevity might be at least partially interpreted as a neutral consequences of mutagenesis. Further studies are needed to reveal the mtDNA component of mammalian longevity.

The described properties of mtDNA mutational spectrum may shed light on the basic mechanisms of replication and mtDNA mutagenesis, improve our understanding of evolution of mtDNA in mammals and can be used as a marker of cell longevity in heterogeneous tissues.

Figure captions

FIGURE 1: mtDNA mutational spectra is changing during the cancerogenesis and differs between cancer types

1A. Variant Allele Frequency (VAF) in cancer approximates the time of origin of the variant: it is high for early variants (observed also in normal tissues) and it is low for late variants (not observed in normal tissues).

1B. Ts/Tv is higher for early (with high VAF) versus late (with low VAF) variants. Integral analysis includes all somatic variants (left panel). Cancer-type specific analysis (middle panel) takes into account the trend specific for each cancer type; each cancer type is marked by a segment, PBCA is marked by red segment. Sample-specific analysis demonstrates increased VAF(Ts) versus VAF(Tv) within each sample.

1C. Ts/Tv, TC/Tv and GA/Tv are higher for cancer types, descendant from slow versus fast-replicating tissues (left panel). When we normalize by transitions only T>C, but not G>A demonstrates the concordant behavior (middle panel). T>C/G>A is also increasing from fast to slow-replicating tissues (right panel).

1D. The probability of T>C is increasing with VAF and cell longevity (Turnover Rate). The size of the circles is proportional to a probability obtained from the logistic regression model 1 and normalized to 0-1 range. Each circle reflects one mutation, red circles mark T>C, grey circles - all other substitutions.

FIGURE 2: the frequency of G>A substitutions is higher in tissues with short-lived cells

FIGURE 3: variation in mammalian mutational spectrum, derived from within species polymorphisms, is driven by generation length

3A. Derivation of mtDNA mutational spectrum for mammalian species (N = 611). Left panel: observed frequencies of twelve types of nucleotide substitutions in four fold degenerate synonymous sites of all available mtDNA protein-coding genes. Middle panel: nucleotide content in four fold degenerate synonymous sites of all available mtDNA protein-coding genes. Right panel: mutational spectrum calculated as a probability of each nucleotide to mutate to each other.

3B. Mutational spectra vary with species-specific generation length (N = 424). Left panel: positive correlation between generation length and Ts/Tv in mammals. Right panel: frequency of T>C is the best type of substitutions, correlated with generation length.

3C. The principal component analysis (N = 424). Left panel: the biplot of the principal component analyses (first and the second components explains 16% and 12% of variation correspondingly). G>A has the highest loading on the first principal component while T>C has the highest loading on the second principal component. Note that we reverted PC2 to make it positively correlated with generation length (in this case it is concordant throughout the paper that T>C, generation length and -PC2 positively correlate with each other). 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_4: the long-term effect of the mutational bias: neutral nucleotide content in mammalian species

4A. Left panel: nucleotide frequencies in neutral sites of all 13 protein-coding genes as a function of generation time - fraction of T is decreasing while fraction of C in increasing (N = 650). Right panel: AT and GC skews are more pronounced in long-lived species (N = 650).

4B. Changes in nucleotide content along mammalian mtDNA. (N = 650). All genes located in major arc are ranked according to the time spent single stranded: from COX1 to CYTB. Nucleotide content used for ND6 is reversed to L chain notation. ND2 gene spent more time than ND1 in the single-strand state but we do not compare directly these two genes with all others from the major arc.

4C. The gradient of nucleotide changes with time being single stranded (increase in C and decrease in GC skew) is more pronounced in long-lived mammals (N = 650). Histograms of spearman Rho values demonstrate that T is decreasing with the time being single stranded while C is increasing. Correspondingly, AT skew is increasing and GC skew is decreasing with the time being single stranded. Long-lived mammals are in excess among ones with

significant Rho for C and GC skew (see the mosaic-plots which are based on corresponding histograms of Rho values) meaning that these changes in nucleotide content are more step in mtDNA of long-lived mammals.

FIGURE 5: the model of mtDNA mutagenesis as a function of a time of parental heavy strand being single stranded during mtDNA replication

5A. The current model of mtDNA mutagenesis based on figure 4 from (Faith and Pollock 2003) demonstrates that T>C (green line) is increasing fast and G>A (red line) - slowly with the time being single stranded. Other rare transitions and all transversions (all other colors) are expected to be non-sensitive to the time being single stranded.

5B. Different rate of T>C and G>A leads to the fact that fraction of T>C is increasing while the fraction of G>A is decreasing with the time being single stranded. Correspondingly ratio of T>C to G>A is increasing with the time being single stranded.

5C. Replication rate hypothesis assumes that mtDNA replication time is shorter in short-lived cells / organisms that leads to decreased T>C/G>A in them.

5D. Chemical damage hypothesis assumes that the intercept of the regression line T>C is higher in long lived cells and organisms due to increased oxidative damage. This hypothesis is our preferred working hypothesis, which is marked by red box.

5E. Replication mode switch hypothesis assumes that in long-lived cells / organisms the predominant mode of the mtDNA replication is the asynchronous one, while in short-lived cells / organisms the alternative strand-coupled mode can be more common. Strand-coupled mode doesn't assume the long time of existence of parental heavy strand in single-stranded condition and thus the slope of T>C regression line is zero.

5F. Summary of all our results: T>C/G>A is increasing with cellular / organismal longevity.

Methods

Analyses of somatic mtDNA mutations

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 (see Supplementary materials for details). Somatic mtDNA mutations from healthy tissues we obtained from Ludwig et al. (Ludwig et al. 2019) and treated them in the same way as cancer data.

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. 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.

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, T, G and C nucleotides in synonymous fourfold degenerate positions. For each protein-coding gene of each reference mitochondrial genome of mammalia species we estimated AT skew as (A-T)/(A+T) and GC skew as (G-C)/(G+C) using only synonymous fourfold degenerate sites.

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. This work was also supported by Russian Foundation of Basic Research [No. 18-29-13055 & 18-04-01143 to K.P] and Russian Science Foundation [No. 17- 75-20015 to I.M.]

Literature

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. https://doi.org/10.1016/j.gene.2005.05.019.

Chen, Lixin, Pingfang Liu, Thomas C. Evans Jr, and Laurence M. Ettwiller. 2017. "DNA Damage Is a Pervasive Cause of Sequencing Errors, Directly Confounding Variant Identification." *Science* 355 (6326): 752–56. https://doi.org/10.1126/science.aai8690.

———. 2018. "Response to Comment on 'DNA Damage Is a Pervasive Cause of Sequencing Errors, Directly Confounding Variant Identification.'" *Science* 361 (6409). https://doi.org/10.1126/science.aat0958.

Cluett, Tricia J., Gokhan Akman, Aurelio Reyes, Lawrence Kazak, Alice Mitchell, Stuart R. Wood, Antonella Spinazzola, Johannes N. Spelbrink, and Ian J. Holt. 2018. "Transcript Availability Dictates the Balance between Strand-Asynchronous and Strand-Coupled Mitochondrial DNA Replication." *Nucleic Acids Research* 46 (20): 10771–81. https://doi.org/10.1093/nar/gky852.

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.

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. https://doi.org/10.1371/journal.pgen.1002689.

Faith, Jeremiah J., and David D. Pollock. 2003. "Likelihood Analysis of Asymmetrical Mutation Bias Gradients in Vertebrate Mitochondrial Genomes." *Genetics* 165 (2): 735–45. https://www.ncbi.nlm.nih.gov/pubmed/14573484.

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. https://doi.org/10.1098/rspb.2002.2218.

Herbers, Elena, Nina J. Kekäläinen, Anu Hangas, Jaakko L. Pohjoismäki, and Steffi Goffart. 2019. "Tissue Specific Differences in Mitochondrial DNA Maintenance and Expression." *Mitochondrion* 44 (January): 85–92. https://doi.org/10.1016/j.mito.2018.01.004.

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.

Lehmann, Gilad, Elena Segal, Khachik K. Muradian, and Vadim E. Fraifeld. 2008. "Do Mitochondrial DNA and Metabolic Rate Complement Each Other in Determination of the Mammalian Maximum Longevity?" *Rejuvenation Research* 11 (2): 409–17. https://doi.org/10.1089/rej.2008.0676.

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. https://doi.org/10.1016/j.cell.2019.01.022.

Montooth, Kristi L., and David M. Rand. 2008. "The Spectrum of Mitochondrial Mutation Differs across Species." *PLoS Biology* 6 (8): e213. https://doi.org/10.1371/journal.pbio.0060213.

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. https://www.researchgate.net/profile/Luca_Santini/publication/258439428_Generation_length_for_mammals/links/0deec5283b58e083cb000000.pdf.

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. https://doi.org/10.1073/pnas.1409328111.

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. https://doi.org/10.1038/s41467-018-07232-8.

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. https://doi.org/10.1093/jb/mvx052.

Stewart, Chip, Ignaty Leshchiner, Julian Hess, and Gad Getz. 2018. "Comment on 'DNA Damage Is a Pervasive Cause of Sequencing Errors, Directly Confounding Variant Identification.'" *Science* 361 (6409). https://doi.org/10.1126/science.aas9824.

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. https://doi.org/10.1093/nar/gks1155.

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. https://doi.org/10.1038/nature23302.

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. https://doi.org/10.1126/science.1260825.

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. https://doi.org/10.3109/19401736.2015.1115842.

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. https://doi.org/10.1101/cshperspect.a005553.

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.

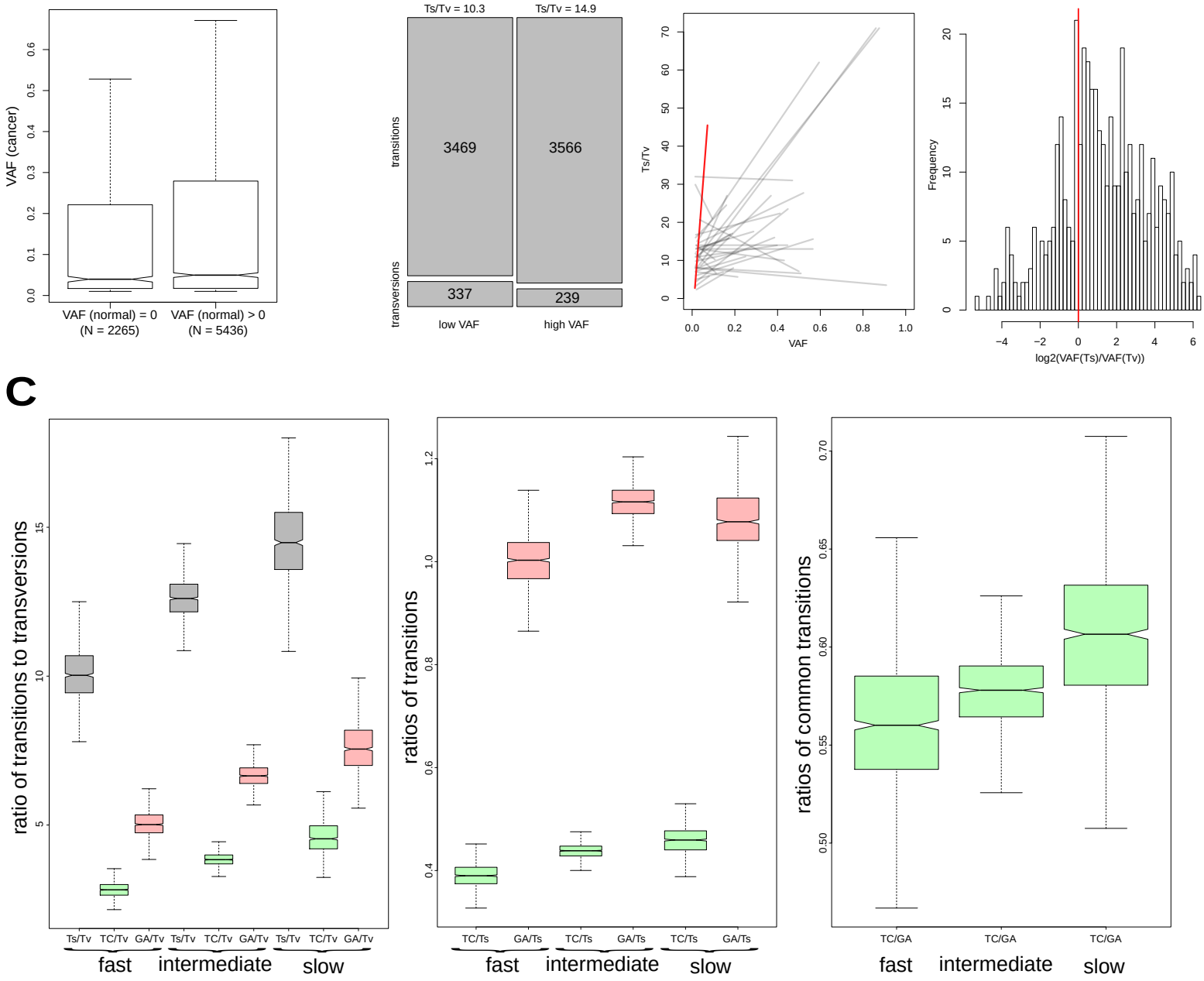


Figure 1

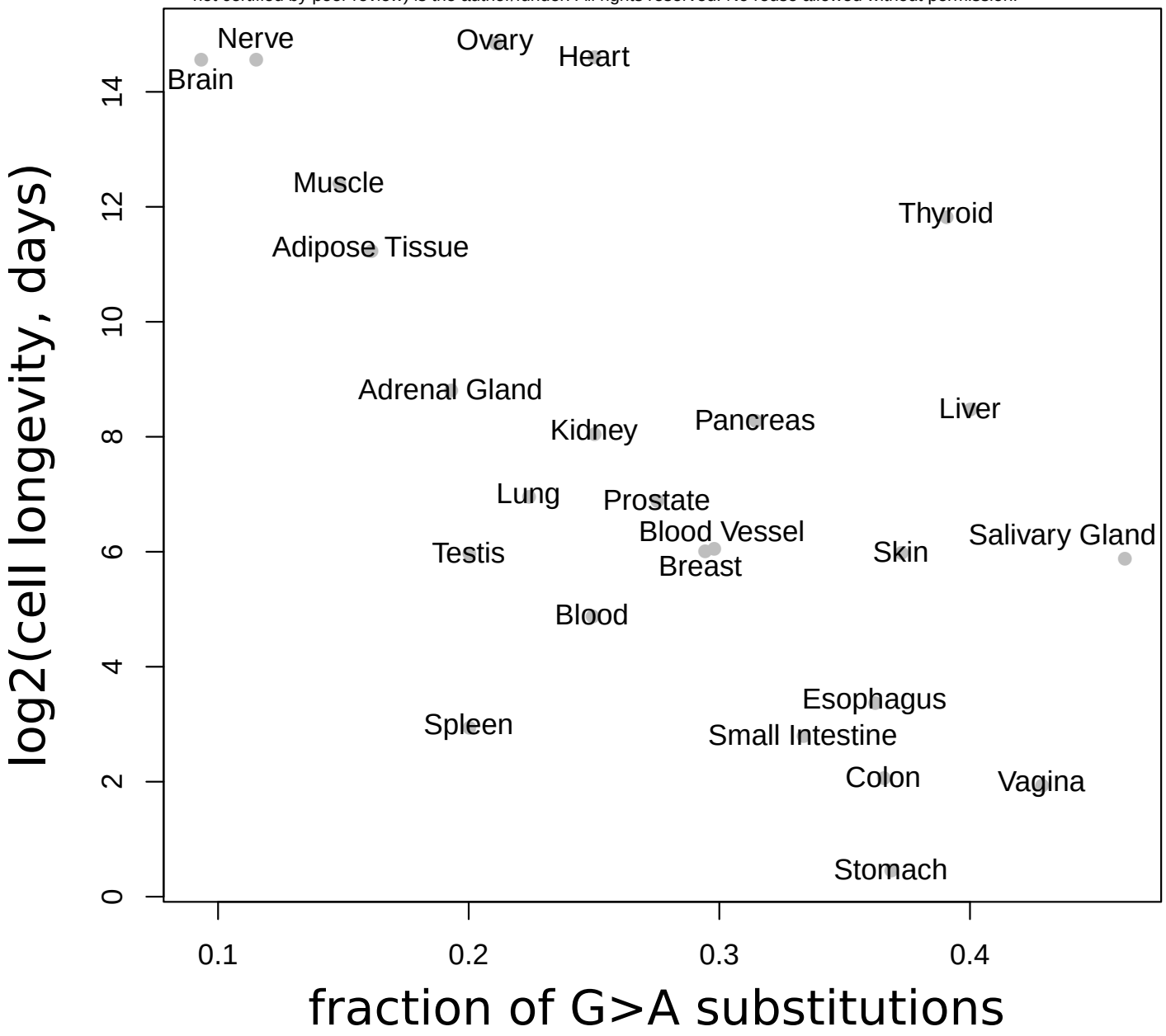
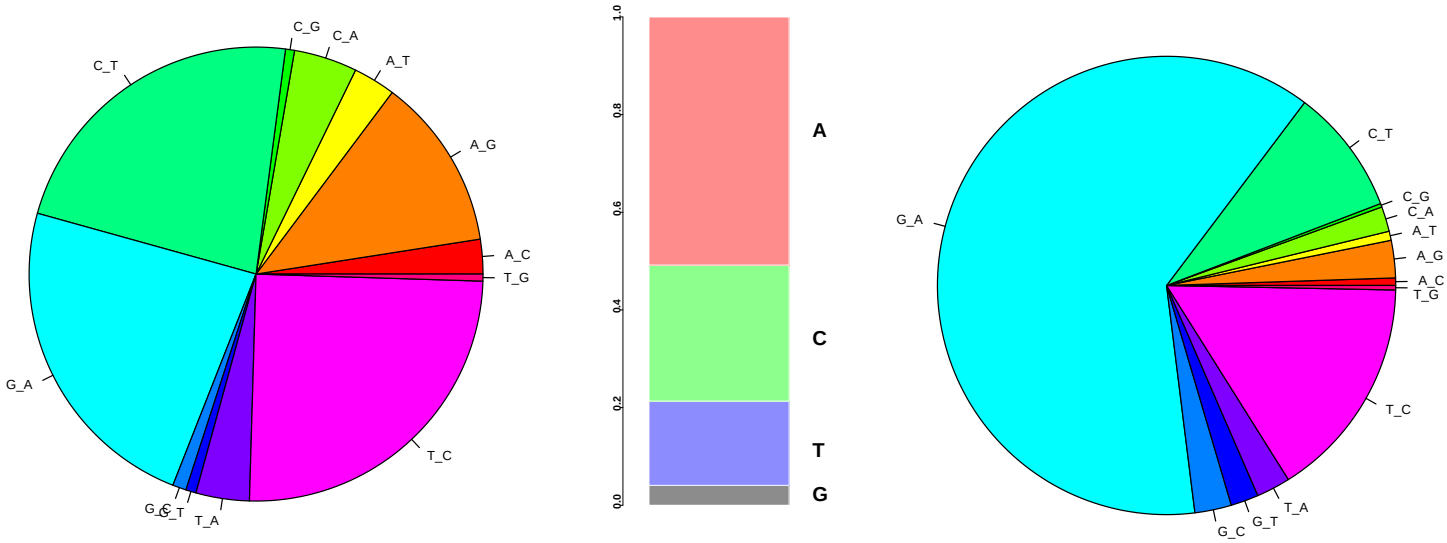
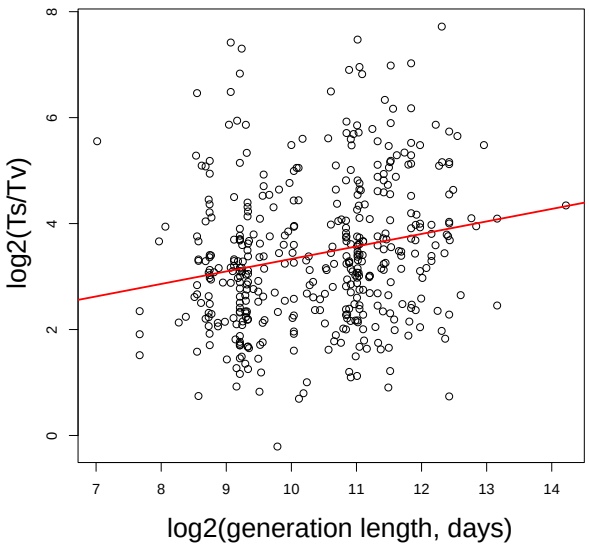


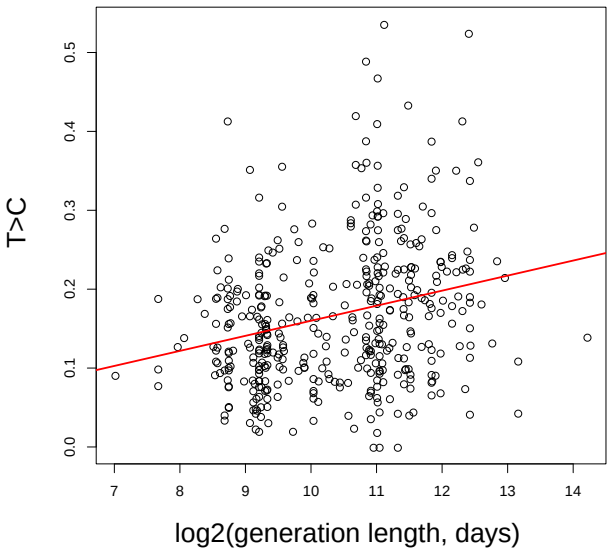
Figure 2

A

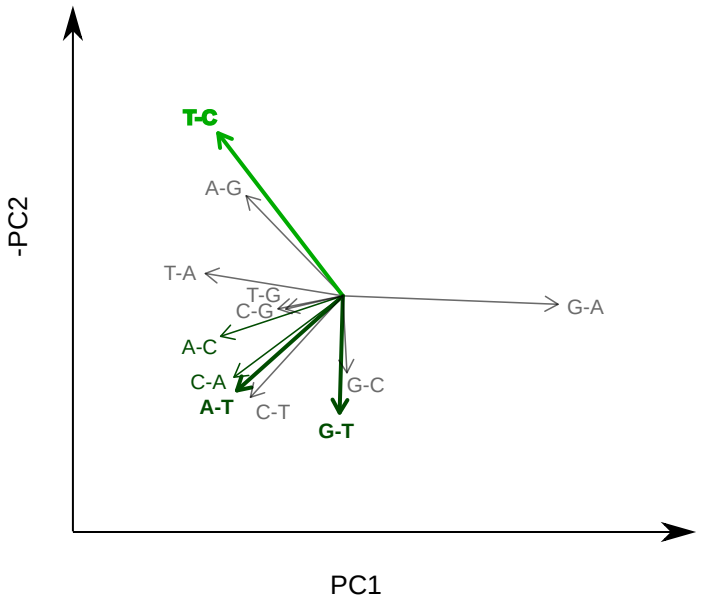


B





C



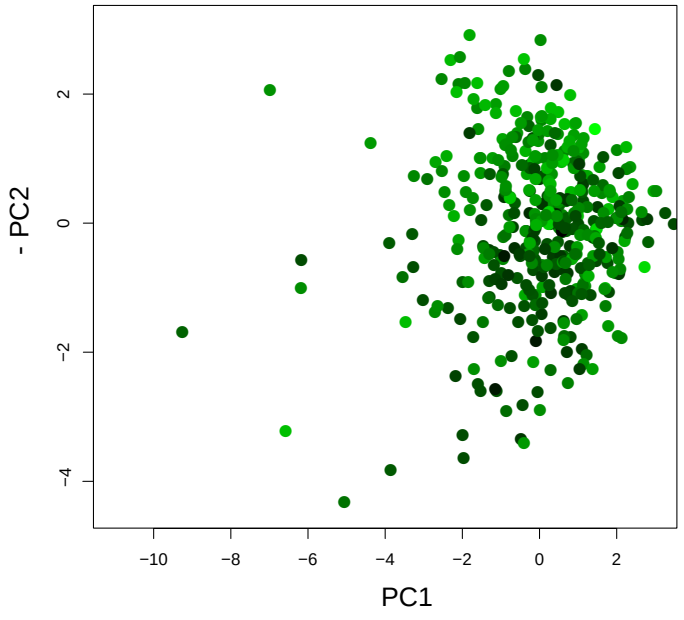


Figure 3

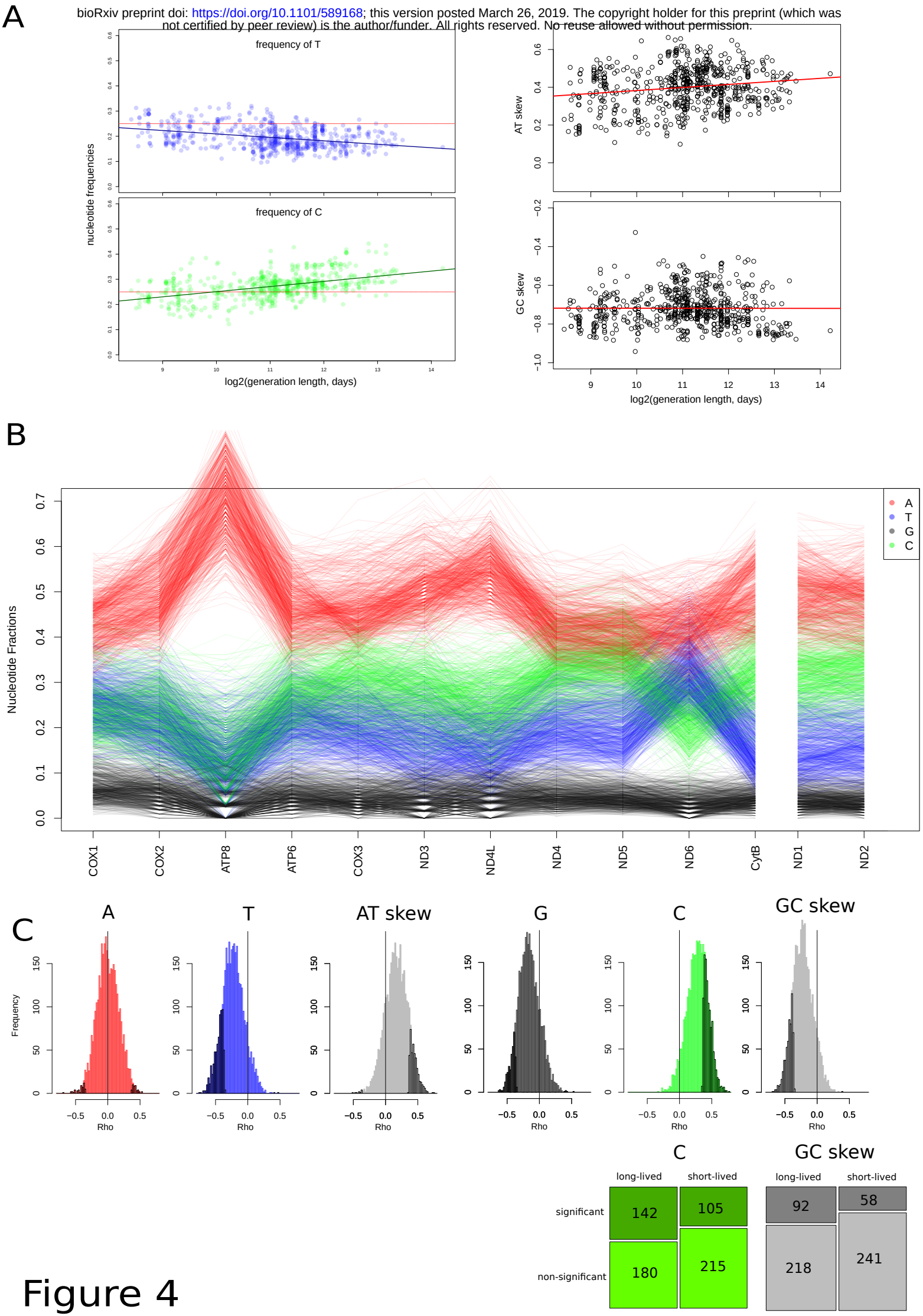


Figure 4

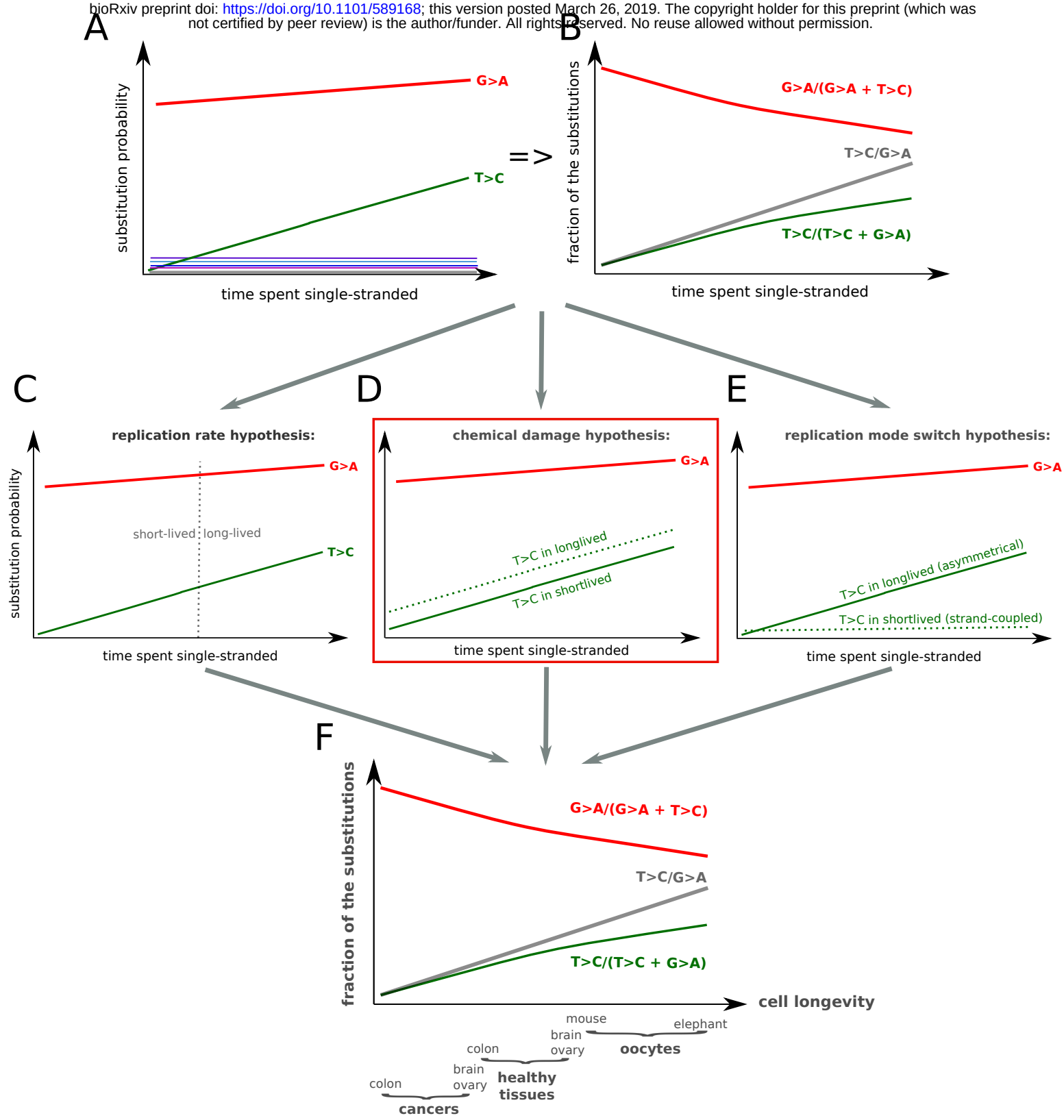


Figure 5