Risk of somatic mitochondrial deletions is affected by the secondary structure of the mitochondrial genome

Alina A. Mikhailova[#1], Viktor Shamanskiy[#1], Kristina Ushakova[#1], Alina G. Mikhailova[#1], Sergei Oreshkov[1], Dmitrii Knorre[2,3], Evgenii O. Tretiakov[4], Marianna Zazhytska[5], Wolfram Kunz[6], Alexandre Reymond[7], Ilia Mazunin[8], Georgii A. Bazykin[8-9], Konstantin Gunbin[1,10], Jacques Fellay[11], Masashi Tanaka[12-14], Konstantin Khrapko[15], Konstantin Popadin[1,7,11]*

1. Center for Mitochondrial Functional Genomics, Immanuel Kant Baltic Federal University, Kaliningrad, Russia.
2. The A.N Belozersky Institute Of Physico-Chemical Biology, MSU, Moscow, Russian Federation.
3. Institute of Molecular Medicine Sechenov First Moscow State Medical University Trubetskaya str. 8-2, 119991, Moscow.
4. Department of Molecular Neurosciences, Center for Brain Research, Medical University of Vienna, Vienna, Austria.
5. Department of Biochemistry and Molecular Biophysics, Mortimer B. Zuckerman Mind Brain and Behavior Institute, Columbia University, New York.
6. Division of Neurochemistry, Department of Experimental Epileptology and Cognition Research, University Bonn Medical Center and Life & Brain Center, University Bonn, Bonn, Germany.
7. University of Lausanne, Center for Integrative Genomics, Lausanne, Switzerland.
8. Center of Life Sciences, Skolkovo Institute of Science and Technology, Skolkovo, Russia.
9. Sector of Molecular Evolution, Institute for Information Transmission Problems (Kharkevich Institute) of the Russian Academy of Sciences, Moscow, Russia.
10. Institute of Molecular and Cellular Biology SB RAS, Novosibirsk, Russia.
11. Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland.
12. Department for Health and Longevity Research, National Institutes of Biomedical Innovation, Health and Nutrition, 1-23-1 Toyama, Shinjuku-ku, Tokyo 162-8636, Japan.
13. Department of Neurology, Juntendo University Graduate School of Medicine, 2-1-1 Hongo, Bunkyo-ku Tokyo 113-8421, Japan.
14. Department of Clinical Laboratory, IMS Miyoshi General Hospital, 974-3, Fujikubo, Miyoshi-machi, Iruma, Saitama Prefecture, 354-0041 Japan.
15. Northeastern University, Boston, MA, USA.

# These authors contributed equally to this work
*corresponding author: konstantin.popadin@epfl.ch

Abstract

Aging is associated with accumulation of somatic mutations. This process is especially pronounced in mitochondrial genomes (mtDNA) of postmitotic cells, where accumulation of large-scale somatic mitochondrial deletions is associated with age-related mitochondrial encephalomyopathies. Here we revise risks of somatic mtDNA deletions and uncover that additionally to direct repeats, known to affect deletions, there is a strong impact of the secondary structure of single-stranded mtDNA during replication. The secondary structure is shaped by inverted repeats which can form stems and shorten effective distance between two direct repeats increasing chances of deletions. Our results support recent discovery that the replication slippage can be the main mechanism of origin of mtDNA deletions. Taking into account strong interaction between direct and inverted repeats it is possible to predict the deletion burden of different haplogroups and associate it with age-related phenotypes. For example, the increased longevity of D4a and D5a haplogroups is in line with their expected low deletion burden.

Introduction

Ageing is associated with the accumulation of DNA damage. Mitochondrial DNA (mtDNA) exists in a cell in a large number of copies with continuous turnover and high mutation rate. These conditions assure coexistence of different mtDNA variants within the same cell (heteroplasmy), which might compete with each other. The intracellular mtDNA selection is expected to be especially influential in slowly-dividing tissues where selfish mtDNA mutants (with replicative advantage within a cell but fitness loss to the host cell) have time for clonal expansion. The typical examples of such mutations are mtDNA deletions. For example, in substantia nigra neurons first mtDNA deletions appear approximately at the age of 50 years (Kraytsberg et al. 2006a; Bender et al. 2006). Initially, an intraneuronal fraction of these mutated mitochondrial genomes is negligible (1 among approximately 10,000 of wild type full-length copies of mtDNA inside a cell) and has no phenotypic effect. However, it increases 1-2% of heteroplasmy every year reaching phenotypically essential 50-80% during several dozens of years (Kraytsberg et al. 2006b)). An accumulation of the mtDNA deletions (Kraytsberg et al. 2006b) eventually leads to neurodegeneration. The other tissue, predisposed to the accumulation of mtDNA somatic deletions, is skeletal muscles. Myofibrils are also post-mitotic, and the expansion of somatic mtDNA deletions is associated with sarcopenia, loss of muscle weight and strength with age (A. Herbst et al. 2007; Allen Herbst et al. 2016). Other tissues with slow-dividing cells also suffer from mtDNA deletions such as extraocular muscles (Yu-Wai-Man et al. 2010) and oocytes (Blakely et al. 2004; Barritt et al. 1999; Chan et al. 2005). If mtDNA deletions are expanded in oocyte or during early embryogenesis the mtDNA deletion syndrome can appear and manifest itself in all tissues including the proliferative ones leading to multisystem disorders ((Schon et al. 1989; Goldstein and Falk 2003))

It has not yet been proven unambiguously, but mtDNA deletions are considered to be causative for the host cell degeneration and several corresponding age-related phenotypes. Several lines of evidence vote for this point of view: (i) the proof-reading-deficient version of mtDNA polymerase causes somatic accumulation of point mutations (Trifunovic et al. 2004) in mice, leading to reduced lifespan and premature onset of ageing-specific phenotypes; (ii) colocalization of mtDNA deletions with regions of myofibre breakage ((Hiona et al. 2010)) and with respiratory chain-deficient neurons (Kraytsberg et al 2006) also vote for the causative effects of mtDNA deletions on ageing; (iii) there are no neurons with too high (> 80%) deletion burden suggesting that such cells have been degraded and disappeared from the analyzed tissue (Kraytsberg et al. 2006b). Thus, the mtDNA deletion burden is not a neutral hallmark of aged cells but probably a causative agent. An investigation of therapeutic application targeting expansion of selfish mtDNA may have various outcomes: postpone cell death, account for

healthier aging and possibly increase the human lifespan. Thus, understanding of the molecular mechanisms, affecting the origin of the somatic mtDNA deletions as well as their velocity are primary importance questions (Khrapko 2011; Popadin et al. 2014).

It has been shown that somatic mtDNA deletions are flanked by direct nucleotide repeats (Samuels, Schon, and Chinnery 2004)long imperfect duplexes consisting of short stretches of direct repeats (Guo et al. 2010). Since direct repeats predispose mtDNA to somatic deletions they are considered as "Deleterious In Late Life" alleles (DILL): neutral or slightly-deleterious during reproductive age but deleterious in late life (Cortopassi 2002). The negative correlation between the abundance of direct repeats in mtDNA and species-specific lifespan of mammals (Samuels 2004; Khaidakov, Siegel, and Shmookler Reis 2006) has been interpreted as an evidence of the deleterious effect of repeats in mtDNA of long-lived mammals. Similarly, with these comparative-species results, we proposed that the decreased number of direct repeats in the mitochondrial genome of some human haplogroups can lead to the lower prevalence of somatic mtDNA deletions and thus may make aging more healthy or postpone it (Popadin and Bazykin 2008).

Analysis of associations between human longevity and properties of mtDNA showed that the longest (a 13 base-pairs) and the most severe direct repeat (so called the "common repeat")in the human mitochondrial genome had been disrupted by the point mutation m.8473T>C in the Japanese D4a haplogroup. This haplogroup is famous because of substantial excess of centenarians (persons who live 100 or more years) and supercentenarians (persons who live 110 or more years) (Bilal et al. 2008; Alexe et al. 2007). We hypothesized that the disruption of the common repeat has a beneficial effect on the D4a haplogroup (Konstantin Popadin 2008) since it decreases the probability of the corresponding somatic deletion (which is also called "common" deletion because it appears very often in aged postmitotic tissues) (Figure 1A). Bacterial deletions, which probably have the same mechanism of origin as mtDNA deletions, indeed become an order of magnitude less frequent if a one-nucleotide mismatch is introduced within the direct repeat (Albertini et al. 1982). Using frontal cortex samples of two aged individuals from haplogroup N1b with a disrupted common repeat (m.8472C>T) we analyzed their mtDNA deletion spectra. In line with our hypothesis (Konstantin Popadin 2008) and bacterial data (Albertini et al. 1982), we observed no common deletions in N1b samples at all (Guo et al. 2010). Hence, the disruption of the 13 bp repeat even by a single-nucleotide germline variant can significantly reduce the formation of the somatic common deletion.

Here we analyze factors affecting the rate of origin of somatic mtDNA deletions with the ultimate goal to calculate the risk and spectrum of somatic deletions for different mtDNA sequences.

Results

In order to uncover the main factors behind the formation of mtDNA deletions we focus on key properties of the most mutagenic direct repeat of the human mtDNA - the common repeat.

1. the common repeat is human-specific

The common repeat is the longest (13bp) perfect (with two identical arms) direct repeat located in the major arc of the human mtDNA. Analysis of reference sequences of primate mtDNAs and several available ancient human mtDNAs shows that the common repeat is a human-specific one: it presents in Homo heidelbergensis, Homo neanderthalensis and Homo denisova but absent (i.e. disrupted) in chimpanzee (pan paniscus and pan troglodytes) (Fig 1B). To examine the prevalence of the common repeat in the modern human population we analyzed complete mitochondrial genomes from the HmtDB database (Clima et al. 2017) and reconstructed human mtDNA phylogeny (see Methods). As expected, we observed that more than 98% (42641 out of 43437) of humans have the perfect common repeat, while less than 2% of humans lost it as a result of single-nucleotide substitutions in the proximal arm (8470 - 8482 bp, see table 1).

2. the common repeat disrupted by synonymous variants is associated with increased longevity

Among the carriers of the disrupted common repeat, the first most frequent variant is m.8473T>C which occurred several times independently leading to 5 big clusters with more than 20 genomes each, and many small clusters (see table 1). The best-studied and the most sequenced haplogroup with this variant is D4a which is characterized by increased longevity (Bilal et al. 2008) and gave a basis for our initial hypothesis (Popadin and Bazykin 2008). The second most common variant of the disrupted common repeat is m.8472C>T which belongs to Ashkenazi specific haplogroup N1b, where a significant reduction of the common deletion in aged neurons has been demonstrated (Guo et al. 2010). The third most common variant is m.8479A>G which belongs to D5a haplogroup, which is also, and independently of D4a, is characterized by increased longevity (Alexe et al. 2007). Recent database (Bolze et al. 2019) containing the maximal up to date (196,554) number of complete mitochondrial human genomes confirms rareness of the disrupted common repeat: the most common disrupted variant is m.8473T>C (1%: 2118 out of 196,554) and the second most common variant is m.8472C>T (0.4%: 833 out of 196,554). It is interesting to mention that both frequent synonymous variants m.8473T>C and m.8479A>G are associated with increased longevity. Synonymous variants might be associated with longevity because their only expected effect is the disruption of the common repeat, without any potentially-deleterious amino-acid substitutions. On the contrary, nonsynonymous N1b specific variant (m.8472C>T) has not been associated with increased longevity, probably, because the potentially beneficial effect of the disruption of the repeat, i.e. decreased deletional burden (Guo et al. 2010), has been

counterbalanced by the deleterious effect of the rare amino-acid substitution.

Table 1. The list of variants disrupting the proximal arm of the common repeat in HmtDB database (43437 mtDNAs).

Sequence of the proximal arm of the direct repeat (8470 - 8482 bp)	Haplogroups	Number of individuals
the perfect direct repeat: ACCTCCCTCACCA	**majority of haplogroups, ancestral state and RefSeq**	**42641**
m.8473T>C; ACCcCCCTCACCA; SYN	D4a (89), R2 (68), U2e (65), H1c (56), U6a (42), and sporadic cases in other haplogroups (79)	399
m.8472C>T; ACtTCCCTCACCA; Pro>Leu	N1b (106), and sporadic cases in other haplogroups (38)	144
m.8479A>G; ACCTCCCTCgCCA; SYN	D5a (57), and sporadic cases in other haplogroups (3)	60
m.8470A>G; gCCTCCCTCACCA; SYN	Dispersed sporadic cases in various haplogroups (32)	32
m.8478C>T; ACCTCCCTtACCA; Ser>Leu	L4b (18), and sporadic cases in other haplogroups (12)	30
m.8477T>C; ACCTCCCcCACCA; Ser>Pro	Dispersed sporadic cases in various haplogroups (26)	26

3. the motif of the common repeat can be associated with specific recombination - related proteins

Next peculiarity of the common repeat is that the nucleotide sequence of the repeat (**aCCTCCCTCACCAx**) strongly resembles degenerate 13 bp motif (CCNCCNTNNCCNC) that is overrepresented in predicted human recombination hotspots in the nuclear genome (Myers et al. 2008) and matches the predicted binding domain for the PRDM9 gene (Baudat et al. 2010; Myers et al. 2008; Brick et al. 2012). Despite the fact that PRDM9 is not officially included into the mitoproteome (Calvo, Clauser, and Mootha 2016), the similarity is striking. Also, the motif of the common repeat is highly C-rich (8 out of 13) which is a hallmark of highly mutagenic repeats (Persson et al. 2019).

4. the most mutagenic region of the major arc - effect of the common repeat or the secondary structure?

Another important feature of the common repeat is that its arms are located exactly in the peaks of the distribution of all deletion breakpoints across mtDNA. Based on this observation it has been hypothesized that the common repeat is important not only locally - for the formation of the common deletion but also globally - for the formation of all other deletions: the common repeat is “appear to be the principal factor behind the formation of most deletions” (Samuels, Schon, and Chinnery 2004). When we analyzed the mtDNA deletion spectrum in the frontal cortex of N1b samples with disrupted common repeat, we aimed to test both the effects: the local (an absence or significant decrease in the frequency of the common deletion) and the global one (shift in the pattern of distribution of all other deletions). Having a very small sample size (2 cases and 2 controls), we observed strong local effect (almost complete absence of the common deletion in N1b samples) and

apparently no global effect at all (Guo et al. 2010). In our study (Guo et al. 2010) we didn't find a piece of evidence to support the global effect of the common repeat on the pattern of other deletions, and thus we rejected the hypothesis that the common repeat is the main factor behind the formation of the majority of deletions ((Samuels, Schon, and Chinnery 2004). Rejection of this hypothesis left the main observation, emphasized by Samuels and coauthors, unexplained - i.e. why the distribution of deletions within the major arc is strongly non-uniform? And thus this nonuniformity requires a special explanation. Drawing parallels with bacterial deletions (Albertini et al. 1982), experiments with mtDNA common repeat (Persson et al. 2019) and nuclear DNA deletions (Canela et al. 2017) where topological proximity of DNA regions increases their chances to interact and to cause a deletion, we hypothesized that increased probability of deletions nearby of the common repeat might be maintained not by the common repeat per se, but by independent factor such as the spatial organization of mtDNA. Below we reconstruct the potential secondary structure of mtDNA and analyze its effect on the distribution of deletions using two different approaches: the holistic one (chapters 5,6,7 and corresponding Fig 2) where we are focused mainly on integral properties of the whole mitochondrial major arc, and reductionist one (chapters 8,9 and corresponding Fig 3) where we aim to analyze the effect of each particular inverted repeat in the shaping of the secondary structure of mtDNA and affecting the deletion spectra.

5. major arc of mtDNA may be folded into the large-scale loops: Hi-C and in silico folding

We reconstructed the 3D structure of mtDNA using the densest contact matrix of the human lymphoblastoid cells with available resolution 1 kb (Rao et al. 2014). Besides the expected diagonal we observed high-density contacts between 0-1 kb and 15-16.5 kb which reflect a circular nature of mtDNA (nucleotides with positions 1 and 16569 are neighbour nucleotides) and confirm that the quality of the 2D reconstruction of mtDNA from the Hi-C data is reasonable (Fig 2A, ovals mark the contacts reflecting the circular nature of mtDNA). The second strongest contact was observed within the major arc (Fig 2A, dotted square marks the potential contact zone). This mtDNA contact suggests that major arc is shaped as a loop, and the whole mtDNA may resemble an "infinity sign" with the potential contact zone between 6-9 kb and 13-16 kb and with the centre of the major arc (11 kb) being the centre of the folding axis (Fig 2 central panel). Despite the fact that this pilot result is promising, different Hi-C experiments show different mtDNA contact matrices making this result not robust. Moreover, the structure of the mtDNA molecule reconstructed by Hi-C experiment represents mainly the 3D structure of the double-stranded DNA, which is relevant to mtDNA deletion formation according to some (Krishnan et al. 2008) but not all the models. According to the most recent model (Persson et al. 2019), mtDNA deletions originate during mtDNA replication when a long stretch of the major arc (specifically - the parental heavy chain) is single-stranded. Thus, using the potential contact of the Hi-C matrix (Fig 2A) exclusively as a suggestive one, in the downstream analyses we focus on the reconstruction of the secondary

structure of the single-stranded mtDNA during replication.

In order to reconstruct the potential shape of the single-stranded parental heavy chain of the major arc, we performed an *in silico* folding in the Mfold program (see Methods). Using the output structure of the molecule, we derived a contact matrix as the number of hydrogen bonds between two DNA regions with 1kb resolution (Fig 2B). Interestingly, we observed an unexpected diagonal from the lower left to the upper right part of the matrix with the strongest contact zone overlapping the similar one from the Hi-C matrix (see the dotted line zones in Fig 2A and 2B). This cross-like contact matrix resembles bacterial Hi-C data (Le et al. 2013), and it might be interpreted as if the single-stranded heavy chain forms a loop with a center close to 10-11 kb and the large-scale stem with correspondingly aligned to each other regions: 9.5 kb in front of 11.5 kb, 8.5 kb in front of 12.5 kb, 7.5 kb in front of 13.5 kb, and the strongest contact - 6.5 kb in front of 14.5 kb. Secondary structures resulting from the reshuffled nucleotide sequence of the major arc didn't show stable and strong cross-like structure. Interestingly, the strongest contacts of the *in silico* folding matrix are in line with the potential contact zone from Hi-C data: see the dotted squares in Fig 2A-B. The secondary structure of the single-stranded DNA is expected to be maintained by the distribution of the inverted repeats, which can hybridize with each other forming stems and loops. To prove it we correlated the density of the inverted repeats in 100 bp windows (see Methods) with the corresponding number of hydrogen bonds between these DNA regions in the *in silico* folding matrix and observed a positive correlation (Spearman's rho = 0.05, p = 0.0017, N= 4005, diagonal elements were removed from the analyses).

6. single-stranded heavy chain of the major arc may have secondary structure: evidence from deletion spectra

In silico folding experiment (Fig 2B) described above suggests that the single-stranded heavy chain of the major arc might be shaped like a loop. If the formation of deletions depends on spatial proximity of single-stranded DNA regions, we expect that the distribution of deletions will follow this structure. First, the major arc is expected to be folded into the loop with center (folding axes) around 11 kb, and thus the majority of deletions should be centered around the same axis. Analyzing the collection of the human mtDNA deletions from the MitoBreak database (Damas et al. 2014), we observed that the median of the centers is indeed close to the expected folding axis - 11 kb (Fig 2C, the median of deletion centers is 11'463, N = 1060). Second, we expect that deletions would be more abundant in the region close to the potential contact zone as compared to the loop. To check it we grouped individual deletions into several clusters (Fig 2D, see Methods) and confirmed that the majority of clusters are indeed located within the potential contact zone (Fig 2D, p-value < 0.001; several biggest clusters of deletions are plotted as colored dotted lines on our infinity sign scheme in the central panel of Fig 2).

Classically, the distribution of deletions is explained by direct repeats within the human mtDNA (Samuels, Schon, and

Chinnery 2004; Guo et al. 2010) and bacterial genomes (Albertini et al. 1982). To prove the importance of the secondary DNA structure as an additional factor affecting the formation of the deletions we take into account both direct repeats and secondary structure information. First of all, we compared the distribution of the perfect direct repeats (see Methods) and deletions from MitoBreak. We observed more or less uniform distribution of the direct repeats within the major arc versus the strongly biased distribution of the deletions (data not shown). This result is in line with the previous observation of Samuels et al (Samuels, Schon, and Chinnery 2004) and confirms that direct repeats alone poorly explain the distribution of the deletions. There is also a possibility that deletions are induced by the special motifs - for example, it has been shown that repeats inducing deletions are more C-rich (Persson et al. 2019). Thus, the shifted distribution of deletions might be explained by the shifted distribution of such motifs. To test it we used our database of degraded repeats of the human mtDNA (Shamanskiy et al. 2019) and grouped all degraded direct repeats into clusters with similar motifs. Next, we annotated all possible pairs of repeats within the same cluster as realized repeats (if there is a MitoBreak deletion flanked by these arms) or non-realized repeats (if there is no corresponding deletion in MitoBreak). The comparison of the positions of realized and non-realized repeats showed strict difference: realized repeats tend to be located closer to the potential contact zone, while non-realized repeats are distributed more uniformly in the major arc (Fig 2E). Cluster-specific analyses confirmed that within the same cluster (the same motif) non-realized repeats tend to start 700 bp later and end 1300 bp earlier leading to 2000 bp shorter distance between non-realized repeats versus realized (average interval between arms is 5643 for realized repeats and 3450 for non-realized ones, average start is 8678 for realized and 9394 for non-realized repeats, average end is 14321 for realized and 13039 for non-realized repeats; all p-values $< 2.2e-16$, paired Mann-Whitney U-test; Number of clusters = 618, Fig 2E). Thus, we conclude that neither position of repeats nor repeat motifs (Fig 2E) can explain the shifted distribution of deletions. Because 80% of realized repeats start in the interval from 6465 to 10954 and end at the interval from 13286 to 15863 we suggest that this biased distribution can be explained by the potential contact zone of the single-stranded DNA, which we annotated arbitrary (see Fig 2B) as frequent contacts between 6-9 kb and 13-16 kb. Indeed, the odds ratio of realized repeats within the potential contact zone versus non-contact zone is 7.5 (Fig 2F, mosaic-plot; Fisher odds ratio = 7.5, $p < 2.2e-16$). Thus, the most parsimonious explanation of the non-uniform distribution of deletions (Fig 2CDEF) is that the regions with the high density of deletions (or realized repeats) are located in spatial proximity to each other during mtDNA replication (Fig 2B).

7. probability of deletions as a function of both DNA microhomology and the proximity to contact point

We have shown that the distribution of the deletions within the major arc (Fig 2CDEF) may be affected by the potential secondary structure of the single-stranded mtDNA (Fig 2B). Also, it has been shown before that local positions of

deletion breakpoints is well described by direct repeats (Samuels, Schon, and Chinnery 2004) or more generally - by long imperfect duplexes (Guo et al. 2010). Thus, as the next step we will incorporate both these variables into one model. To derive a metric of similarity between two regions of DNA which will integrate short stretches of direct repeats we estimated all pairwise microhomology scores between all 100 bp windows inside the major arc (Fig 2G, see Methods). As expected, the microhomology score correlated positively with the density of direct repeats in the corresponding windows (see Methods, Spearman's Rho = 0.07, P = 1.698e-06, N = 4950 regions 100 bp x 100 bp windows). Next, to analyze an association between deletions and the microhomology scores we performed logistic regression where for each cell (100x100bp) of our matrix we derived presence (coded as 1, N = 484) or absence (coded as 0, N = 4466) of deletions from the MitoBreak database (Damas et al. 2014). We estimated the logit function of deletions as a function of the microhomology score:

log(p/(1-p)) = -2.25264 + 0.27442 * microhomology score, with all p values (intercept, coefficient) less or equal 5.13e-09, N = 4005. This result confirms our previous findings (Guo et al. 2010) and demonstrates that a high microhomology score is associated with a higher probability of deletion. Next, we added to our model the second independent variable - "the potential contact zone" coded as one within the zone (6-9kb and 13-16kb) and zero - outside:

log(p/(1-p)) = -2.38296 + 0.32592*microhomology score + 0.90579* potential contact zone, with all p values less or equal 3.8e-10, N = 4005. We can see that the variable "potential contact zone" significantly and positively affects the probability of deletions. Because we used scaled variables (see Methods) we can conclude that the potential contact zone variable affects the odds of probabilities 3 times stronger as compared to the microhomology score. Our potential contact zone: 6-9kb vs 13-16kb was derived arbitrarily from the two structural analyses (Fig 2AB), and despite the apparent congruency with distribution of deletions (Fig2CDEF) it is advantageous to perform the data-driven derivation of the contact zone. In order to do it, we substituted the potential contact zone by one contact point and estimated the Euclidean distance from the contact point to any cell of our contact matrix. Assuming that there is one contact point which together with microhomology score describes the distribution of deletions in the best way, we decided to run 4005 logistic regressions and test all possible contact points as centers of all 4005 cells in our matrix. We observed that the strongest contact point (from the model with the minimal Akaike information criterion, AIC) has coordinates: 5' = 7550, 3' = 15150. Plotting the heatmap with AIC for each contact point, we can see that the data-driven contact zone is similar to our visually-derived 6-9kb vs 13-16kb potential contact zone (fig 2H). If instead of AIC as a metric for model selection we use residual deviance, or P-value or the coefficient of the second independent variable - results are qualitatively similar (data not shown). Altogether, using three orthogonal sources of information: *in silico* folding (Fig 2B), distribution of deletions (Fig 2CDEFG) and microhomology scores (Fig2HG), we propose that the

distribution of the human mtDNA deletions is a function of both global secondary structure of the single-stranded major arc and local similarity between DNA regions, i.e. microhomologies. Below we shift to a more reductionist approach considering individual repeats.

8. deletions are more frequent if direct repeats are nested with inverted ones: DI...I.D combinations and their properties

The process of replication slippage, which is also called a copy-choice recombination, has been shown to affect deletion formation in bacteria (Viguera, Canceill, and Ehrlich 2001; Albertini et al. 1982) and in experimental templates simulating mtDNA in vitro (Persson et al. 2019). It is based on the (i) arrest of replication due to the secondary structure formed by the inverted repeats on single-stranded DNA (ii), dissociation of the newly synthesized (nascent) proximal arm of the direct repeat and (iii) realigning of the nascent arm to the distal parental one (Fig 3A). The suggested role of inverted repeats in this process is to suspend the replication fork and to decrease the effective distance between two arms of the direct repeat which increases the chances of successful realigning.

To test the involvement of the replication slippage in the formation of the mtDNA deletions *in vivo* we calculated all Direct-Inverted-Inverted-Direct combinations of repeats located within the major arc. Using our database (Shamanskiy et al. 2019) of all degraded repeats of the human mtDNA (2957 direct and 764 inverted repeats within the major arc, see Methods), we searched the nested combinations of repeats. Following the experimental studies which confirmed the replication slippage mechanism (Persson et al. 2019; Albertini et al. 1982), we assumed that the proximal direct and inverted repeats should be located close to each other (maximal distance is 10 bp) while the distal inverted and direct repeats can be located up to 1000 bp from each other. To emphasize this condition hereafter we use a notation DI...I.D. Within the major arc we observed 260 DI...I.D combinations and 2697 ordinary direct repeats (DD) without the nested inverted ones. Next, we asked if the mutagenic potential of the DI...I.D combinations is higher than the ordinary DD. Comparing the fraction of DI...I.D combinations associated with deletions (7 were associated and 253 were not associated) and fractions of DD combinations associated with deletions (20 were associated and 2677 were not associated), we observed that mutagenic potential of DI...I.D combination is indeed higher (Fisher odds ratio 3.7, p-value = 0.0072). The increased mutagenic potential of DI...I.D is even more pronounced within the potential contact zone (DI...I.D: 6/34, DD: 12/438; Fisher odds ratio = 6.4; p-value = 0.0018).

The increased mutagenicity of DI...I.D combinations within the potential contact zone suggests that both levels of the mtDNA secondary structure might be important: the global folding integrated for the whole major arc (Fig 2B-H) and the local repeat-specific DI...I.D combinations (Fig 3A). Next, we tested if the potential contact zone can be explained by the increased density of the DI...I.D combinations. Taking into account that the area of the potential contact zone (3kb*3kb) is about 18% of

the major arc (10kb*10kb/2 = 50) we observed the proportional fraction of DI...I.D - 15% (40 DI...I.D combinations in the potential contact zone out of total 260 in the major arc). This assumes that existence of the potential contact zone can not be explained by the increased density of DI..I.D combinations. On the contrary, increased mutagenicity of DI...I.D within the potential contact zone assumes that there is an additional, probably structural factor maintaining increased mutagenicity of the potential contact zone of single-stranded DNA (Fig 2B-H).

To validate additionally the hypothesis that the majority of deletions occurred during the replication of the lagging strand (Persson et al. 2019) we compared the mutagenic potential of DI...I.D combinations (short distance, max 10bp, between proximal direct and inverted repeats and potentially high, up to 1000 bp, between distal direct and inverted repeats) and D.I...ID combinations (potentially long distance, up to 1000 bp, between proximal direct and inverted repeats and short distance, max 10 bp, between distal inverted and direct repeats). The logic behind this analysis is that D.I...ID pattern on the lagging strand is equivalent to the DI...I.D pattern on the leading strand and if both strands are equally mutagenic we expect that both combinations will be equally good associated with deletions (Fig 3C). Analyzing the associations of D.I...ID with deletions, we observed Fisher's odds ratio 2.61 (D.I...ID: 5/234, DD: 22/2696; Fisher Odd's ratio = 2.61; p=0.0615) which is less than 3.7 observed for DI...I.D (see above). D.I...ID within the potential contact zone is not associated with deletions at all (D.I...ID: 3/44, DD: 15/428; Fisher Odd's ratio = 1.9421, p-value = 0.4) while DI...I.D demonstrated strong signal (see above). Our results are in line with the original finding that the majority of deletions occur during the replication on the lagging strand (Persson et al. 2019).

If DI...I.D combinations are strongly mutagenic, we expect that purifying selection would eliminate them maintaining the number of observed DI...I.D combinations less than randomly expected. Permuting 2957 direct and 764 inverted repeats within the major arc, we obtained median value 365 DI...I.D combinations which is 29% higher than 260 observed ones (p <0.0001, 10000 permutations). Within the potential contact zone results were similar: permutations showed the median value of 54 DI...I.D combinations which is 25% higher than 40 observed (p=0.0243, 10000 permutations). The decreased number of the observed versus expected DI...I.D combinations can be used as a hallmark of the purifying selection eliminating these mutagenic combinations.

9. the common direct repeat has its nested inverted repeat (!)

There are 8.8% of direct repeats involved into DI...I.D combinations in the major arc (260/2957). We observed the common repeat among 2.3% of DI...I.D combinations which are additionally associated with deletion (7/260) (Fig 3B). The 10 bp degraded inverted repeat, associated with the common repeat (8484-8493: AGCcCATAAA; 13354-13363: TTTATGtGCT)

has a potential to form a stem structure which can arrest the replication fork and decrease an effective distance between the direct repeat arms from approximately 5kb (13447-8482=4965) to less than 100 base pairs (Fig3B). Both these roles of the nested inverted repeat are crucial in the explanation of the high mutagenic potential of the common repeat (Fig 3B). Different haplogroups have mismatches in the common repeat or in the inverted one. Despite the fact that these substitutions are rare enough, sometimes both types of mismatches exist in the same mitochondrial genome. For example, some representatives of haplogroup R22 (GenBank IDs: KP346026.1, KP346018.1), distributed mainly in Nicobar archipelago (Trivedi et al. 2006) and found also in south-central Indonesia and Vietnam (Peng et al. 2010), contain one synonymous substitution disrupting the common repeat (m.8473T>C) and another synonymous substitution disrupting the inverted repeat (m.13359G>A). We expect an epistatic interaction between these two mismatches: destabilized stem will unlikely arrest the replication fork while degraded common repeat will unlikely, with one order of magnitude decreased probability (Albertini et al. 1982), allow the realigning of the newly-synthesized proximal arm to the distal arm on the parental chain (Fig 3B). Altogether both these mismatches are expected to lead to the vanishing of the common deletion. Investigation of the deletion burden in the age-matched tissues of these and other interesting haplogroups will help to establish the prediction of the genetic risks of somatic mtDNA deletions.

Discussion

We demonstrated that the mtDNA deletion spectrum is associated not only with mtDNA microhomology, but also with the secondary structure of a single stranded heavy chain of mtDNA (Fig 2B-H, Fig 3). This means that direct repeats, which are shaping the mtDNA microhomology, interact with inverted repeats, which in turn are shaping the secondary structure of the single-stranded heavy chain of the major arc. More profound understanding of this interaction will open a possibility to predict mtDNA deletion spectra (distribution of different types of deletions) and deletion burden (total fraction of mutated versus wild-type mtDNA) based on mtDNA sequence and thus will aid in the uncovering of haplogroup-specific disease risks (mitochondrial polygenic risk score of somatic deletions). For example, in the haplogroups with the common repeat disrupted by the synonymous variants (D4a and D5a), due to decreased somatic mtDNA deletion burden (Guo et al. 2010), we expect to observe not only postponed aging (Bilal et al. 2008; Alexe et al. 2007), but also less pronounced age-related phenotypes associated with mtDNA deletions such as: the rate of neurodegeneration in a healthy cohort (Kraytsberg et al. 2006b) and among Parkinson patients (Bender et al. 2006), skeletal muscle myopathies (A. Herbst et al. 2007; Allen Herbst et al. 2016), extraocular muscle weakness (Yu-Wai-Man et al. 2010) and the rate of mitochondrial aging of Hiv patients (Payne et al. 2011). It is important to note also that beside the above mentioned very common late-onset mtDNA phenotypes, mtDNA deletions are associated with several rare early-onset "mtDNA deletion syndromes" classically consisting of Kearns-Sayre syndrome (KSS), Pearson syndrome and progressive external ophthalmoplegia (PEO) (Schon et al. 1989; Goldstein and Falk 2003) with (rarely manifested) Leigh syndrome. These mtDNA deletion syndromes are typically caused by mtDNA deletions being inherited through the oocyte or occuring in early embryonic development, as a consequence they are usually present in all tissues of the affected individuals, including proliferative tissues (however still more pronounced in postmitotic tissues) hence why these syndromes typically are multisystem disorders. We expect the prevalence of these mtDNA deletion syndromes to be decreased among the haplogroups with the disrupted DI...I.D combinations including, but not limited to the common repeat (Fig 3B).

Our results are compatible with the replication slippage mechanism (Persson et al. 2019), (Albertini et al. 1982) where the nested pattern of direct and inverted repeats (DI...I.D) leads to the formation of deletions (Fig 3A-B). We have shown that in the human mtDNA the DI...I.D combinations are indeed more mutagenic as compared to the ordinary DD patterns and DI...I.D combinations potentially evolve under the pressure of purifying selection. Similar effects of selection against DI...I.D is expected to be evident on evolutionary and interspecies scale. Initially, it has been shown, that the mammalian lifespan negatively correlates with an abundance of direct repeats in mtDNA (Samuels 2004; Khaidakov, Siegel, and Shmookler Reis

2006), while later it was demonstrated that inverted repeats show an even stronger negative correlation with mammalian lifespan (Yang, Seluanov, and Gorbunova 2013). Recently, it has been shown that both types of repeats: direct and inverted ones correlate with each other well enough (Shamanskiy et al. 2019). Thus, the analysis of interactions between direct and inverted repeats and their association with mammalian longevity is the next anticipated step of the research.

A possibility to predict mtDNA deletion burden and spectrum based on mtDNA sequence (Fig 3B) might be an important step forward for mitochondrial medicine. Haplogroups with low deletion burden can provide a donor mtDNA in (i) mitochondrial donation and (ii) mitochondrial transplantation, while known spectrum of deletions can help to develop (iii) targeted systems eliminating expected deletions in neurons and muscle cells. (i) Mitochondrial donation has been introduced to practical medicine to prevent the inheritance of mutant mtDNA (Kang et al. 2016; Hyslop et al. 2016; Gorman et al. 2018). Currently the technology is based on removal of the nuclear genome from an oocyte or zygote that carries a pathogenic mtDNA mutation, followed by the transfer to an enucleated donor oocyte or zygote with wild-type mtDNA. We propose that it is worth thinking about the donor mtDNA haplogroup during the mitochondrial donation procedure, considering not only any random mtDNA sequences without deleterious variants, but also mtDNA sequences, such as D4a, carrying a variant that potentially protects against age-related encephalomyopathies. (ii) Mitochondrial transplantation is another therapeutic approach when exogenous healthy mitochondria are delivered to the tissue in an injured/diseased state (e.g. the central nervous system (Gollihue, Patel, and Rabchevsky 2018) or cardiomyocytes (Ali Pour, Kenney, and Kheradvar 2020)) and the source of donor mitochondria and mtDNA sequence can potentially impact the success rate of the treatment. (iii) The last strategy for mitochondrial medicine is based on the heteroplasmic shift within the patient (Bacman et al. 2013; Gammage et al. 2014). It is induced by targeted double-strand breaks in the mtDNA structure which leads to its elimination by replicative machinery (Peeva et al. 2018). Heteroplasmic shift introduced on mito-TALEn and /or ZFN may become an important clinical target in the near future (Bacman et al. 2018; Gammage et al. 2018). The main problem with treatment of somatic mtDNA variants, is that TALEN / ZFN technology requires an exact nucleotide target to induce double strand breaks and delete damaged DNA. We hope that our results open a way towards more precise prediction of the haplogroup-specific spectra of deletions and thus - towards finding the most optimal cocktail of mito-TALEn / ZFN to decrease heteroplasmy level of the most expected deletions.

If our considerations above are correct and the disruption of the common repeat and other repeats from DI...I.D combinations are beneficial to human post-reproductive health status and aging, it can still be evolutionary neutral, since the increased longevity is not necessarily associated with prolonged reproductive period and/or increased number of offspring.

Thus according to the NULL hypothesis: synonymous variants affecting repeats from DI...I.D combinations are neutral. However, the possibility that disruption of DI...I.D is associated with some fitness advantages is highly intriguing and worth further investigations. For example, the disruption of DI...I.D may increase fitness (i) directly if the carriers are healthier at reproductive age and thus have higher fertility, or (ii) indirectly if increased longevity of parents and grandparents is advantageous to offspring due to the grandmother effect ((Chapman et al. 2019; Engelhardt et al. 2019). Currently the potential beneficial loss of DI...I.D combinations is supported by the (i) deficit of observed DI...I.D combinations in the human mtDNA as compared to the expected (chapter 8) and (ii) by the negative correlations between the abundance of repeats and mammalian longevity (Samuels 2004; Khaidakov, Siegel, and Shmookler Reis 2006; Yang, Seluanov, and Gorbunova 2013). Further tests are needed to test an evolutionary advantage of the loss of DI...I.D combination in humans and other long-lived mammals.

Methods

Hi-C mtDNA contact matrix: *The mtDNA matrix was used from the Juicebox.*

http://aidenlab.org/juicebox/?juiceboxURL=http://bit.ly/2Rmz4wy.

The corresponding paper describing the data is Rao et al (Rao et al. 2014).

Human phylogenetic mtDNA tree: We used multiple alignment from HmtDB containing 43437 mtDNA genome sequences (Clima et al. 2017). For each mtDNA we predict a haplogroup by the HaploGrep v. 2.1.20 software (Weissensteiner et al. 2016). Based on the whole multiple alignment of complete mtDNA genomes of healthy cases we reconstructed general phylogenetic tree by the IQ-TREE v. 1.6.1 software (Nguyen et al. 2015) using general time reversible (GTR) model of base substitutions. FreeRate site variation model (Soubrier et al. 2012) allowing for a proportion of invariable sites (option: -m GTR+F+I+R6).

In silico folding: We used the reference human mtDNA sequence (NC_012920.1) in heavy chain notation as it spends the most time being single-stranded according to the asymmetric model of mtDNA replication (Persson et al. 2019). Using mfold (Zuker 2003) with parameters set for DNA folding and a circular sequence we constrained everything but the major arc from forming base pairs. The obtained global (genome-wide) secondary structure,we then translated into the number of hydrogen bonds connecting our regions of interest (1kb windows for visualization and 100 bp windows for the majority of the analyses).

The density of inverted/direct repeats: For each pair of 100 bp window we estimated the number of nucleotides involved in at least one inverted/direct degraded repeat. The corresponding repeat should have one arm located in the first window and another arm located in the second window. All degraded (with the maximal level of imperfection 80%) repeats in the human mtDNA were called using our algorithm described here (Shamanskiy et al. 2019).

Clusterization of deletions: For clusterization, we used all MitoBreak (Damas et al. 2014) deletions from the major arc. We used 5' and 3'coordinates as input for a hierarchical density based clustering algorithm (python hdbscan v0.8.24). DBSCAN is a well-known algorithm for probability density based clusterization, which detects clusters as regions with more densely located sample data as well as outlier samples. Advantage of this method is soft clustering. We variated cluster density parameters in order to ensure clusters stability and found that cluster formations stay relatively stable for a wide range of parameters.Thus, DBSCAN produces a robust set of clusters, producing additional evidence for regions with elevated deletion rates. We also performed affinity propagation clustering (Frey and Dueck 2007) as a data exploration experiment, which also yields to robust clustering.

Perfect direct repeats of the human mtDNA: the list of the perfect direct repeats with length 10 or more base pairs was used

from our algorithm described in Guo et al (Guo et al. 2010).

Realized and non-realized direct degraded repeats: We used our database of degraded mtDNA repeats (Shamanskiy et al. 2019) with length 10bp or more and similarity 80% or more. We took into account only direct repeats with both arms located in the major arc. We grouped repeats with similar motifs into clusters so that each considered cluster should contain at least three arms of the repeat and at least one deletion should be associated with two of them. We additionally restricted our subset of clusters considering only non-realized repeats as pairs of arms, where at least one of them (the first or the second) is the same as in realized repeat. Visually on Fig 2E it means that within each cluster we compare realized repeats (red dot) with non-realized ones (grey dot) located on the same horizontal (the same Y coordinate) or vertical (the same X coordinate) axis. We got 618 clusters like this.

Pairwise alignments for microhomology matrix: A measure for the degree of similarity between segments of the major arc was obtained by aligning small windows of mitochondrial major arc sequence with each other. We sliced the mitochondrial major arc sequence into 100 nucleotide pieces and aligned them against each other using EMBOSS Needle (Needleman and Wunsch 1970) with default parameters (match +5, gap open - 10, gap extend - 0.5), parsed out the alignment scores, thus obtaining data for the matrix of microhomology.

Colocalization of degraded direct and inverted repeats: Our database with all human mtDNA degraded repeats (with the maximal level of imperfection 80%) with length 10 bp and more has been used for all analyses (Shamanskiy et al. 2019).

Code and data availability: All data-sets and scripts used in the manuscript are available on Github https://github.com/Aragret/ComparativeGenomics/tree/PhilTransactions, with the key scripts being MitoBreakDeletionsAndOrlovRepeats.R, MitoBreakDeletionsDistribution.R, RealizedVsNonrealizedDeletions.R, SlipAndJump.R and MitoBreakDeletionsAndInteractionOfDirectAndInvertedVictorRepeats.R.

Acknowledgments

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; No. 19-29-04101 to I.M] and Russian Science Foundation [No. 17- 75-20015 to I.M.]. E.O.T. is supported by a Ph.D. fellowship from the Austrian Science Fund FWF (DOC 33-B27). We acknowledge Filipe Pereira and Joana Damas for discussion of the

MitoBreak database, Maria Falkenberg for discussion of the potential structure of mtDNA and Nariman Battulin for discussion of mtDNA Hi-C data.

Author contributions

KP designed the study, KG reconstructed and analyzed the phylogenetic tree, A.A.M, A.G.M, V.SH and K.P performed main statistical analyses, K.U derived in silico folding and microhomologies, all authors wrote the manuscript.

Figures

Figure 1 **A.** A hypothesis: disrupted common repeat impedes the origin of somatic mtDNA deletions maintaining postmitotic cells (neurons and skeletal muscular cells) in more healthy conditions (green circles – wild type mtDNA; red small circles – mtDNA with common deletion) thus postponing age-related phenotypes. **B.** The phylogenetic scheme of the common repeat. Perfect common repeat is human specific.

Figure_2. The secondary structure of the mtDNA. The dotted square marks a potential contact zone within the major arc. **A.** Hi-C contact matrix; **B.** *In silico* global folding. Contact matrix derived from the *in silico* folding of the single-stranded heavy chain of the major arc; **C.** Centers of deletions. The median of the distribution is close to the expected center, 11 kb, of the loop of the major arc; **D.** Clusters of deletions within the major arc. Majority of clusters are located within the potential contact zone; **E.** Realized (red) versus non-realized (gray) repeats tend to be enriched in the potential contact zone; **F.** Mosaic-plot of repeats (realized versus non-realized) within and outside the potential contact zone; **G.** Heatmap of the microhomologies between 100 bp windows within the major arc. Microhomology alone poorly explains the distribution of the deletions within the major arc; **H.** Heatmap of the data-driven contact zone, based on AIC of the compared models.

Figure 3**.** The replication slippage can explain mtDNA deletion formation, where nested inverted (green) and direct (red) repeats play a crucial role during the replication of the lagging (parental heavy chain) strand. **A.** Main steps of the deletion formation. First: according to the asymmetric replication of mtDNA the lagging strand (solid black line) spent a significant amount of time in the single-stranded condition. Second: single-stranded parental heavy chain forms a secondary structure shaped by inverted repeats. This stem can arrest the replication fork. Third: replication arrest leads to partial dissociation of the newly-synthesized proximal arm of the direct repeat (blue arrow). Fourth: dissociated chain realigns to the closest distal arm of the direct repeat. Fifth: replication is resembled but the newly-synthesized light chain (dotted gray line) has a deletion - both

inverted repeats and DNA between them as well as the proximal arm of the direct repeat are removed; **B.** The common repeat (red) has nested inverted repeat (green) which can explain high mutability of the common repeat. Color codes are the same as in Fig 3A. First: synthesis of the nascent light chain (blue) and arrest of the replication near the stem, formed by the inverted repeat. Second: partial dissociation of the nascent proximal arm of the common repeat. Third: realigning of the newly-synthesized proximal arm of the common repeat with the distal arm of the common repeat and continuation of the replication. The most frequent polymorphisms (with 100 or more mtDNAs from HmtDB) disrupting these repeats are marked by rectangles. These substitutions can substantially slow down the process of the deletion formation. The common repeat here is extended from classical perfect 13bp-long repeat (ACCTCCCTCACCA) to 15bp-long degraded repeat with one mismatch (CtACCTCCCTCACCA); **C.** Colocalization of inverted and direct repeats within the major arc demonstrates asymmetry in the deletion formations: DI...I.D are more mutagenic than D.I...ID. This suggests that the majority of deletions originate during the replication of the lagging (parental heavy) strand (see text).

Literature:

Albertini, A. M., M. Hofer, M. P. Calos, and J. H. Miller. 1982. "On the Formation of Spontaneous Deletions: The Importance of Short Sequence Homologies in the Generation of Large Deletions." *Cell* 29 (2): 319–28.

Alexe, Gabriela, Noriyuki Fuku, Erhan Bilal, Hitomi Ueno, Yutaka Nishigaki, Yasunori Fujita, Masafumi Ito, et al. 2007. "Enrichment of Longevity Phenotype in mtDNA Haplogroups D4b2b, D4a, and D5 in the Japanese Population." *Human Genetics* 121 (3-4): 347–56.

Ali Pour, Paria, M. Cristina Kenney, and Arash Kheradvar. 2020. "Bioenergetics Consequences of Mitochondrial Transplantation in Cardiomyocytes." *Journal of the American Heart Association* 9 (7): e014501.

Bacman, Sandra R., Johanna H. K. Kauppila, Claudia V. Pereira, Nadee Nissanka, Maria Miranda, Milena Pinto, Sion L. Williams, Nils-Göran Larsson, James B. Stewart, and Carlos T. Moraes. 2018. "MitoTALEN Reduces Mutant mtDNA Load and Restores tRNA Levels in a Mouse Model of Heteroplasmic mtDNA Mutation." *Nature Medicine* 24 (11): 1696–1700.

Bacman, Sandra R., Siôn L. Williams, Milena Pinto, Susana Peralta, and Carlos T. Moraes. 2013. "Specific Elimination of Mutant Mitochondrial Genomes in Patient-Derived Cells by mitoTALENs." *Nature Medicine* 19 (9): 1111–13.

Barritt, J. A., C. A. Brenner, J. Cohen, and D. W. Matt. 1999. "Mitochondrial DNA Rearrangements in Human Oocytes and Embryos." *Molecular Human Reproduction* 5 (10): 927–33.

Baudat, F., J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B. de Massy. 2010. "PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice." *Science* 327 (5967): 836–40.

Bender, Andreas, Kim J. Krishnan, Christopher M. Morris, Geoffrey A. Taylor, Amy K. Reeve, Robert H. Perry, Evelyn Jaros, et al. 2006. "High Levels of Mitochondrial DNA Deletions in Substantia Nigra Neurons in Aging and Parkinson Disease." *Nature Genetics* 38 (5): 515–17.

Bilal, Erhan, Raul Rabadan, Gabriela Alexe, Noriyuki Fuku, Hitomi Ueno, Yutaka Nishigaki, Yasunori Fujita, et al. 2008. "Mitochondrial DNA Haplogroup D4a Is a Marker for Extreme Longevity in Japan." *PloS One* 3 (6): e2421.

Blakely, E. L., L. He, R. W. Taylor, P. F. Chinnery, R. N. Lightowlers, A. M. Schaefer, and D. M. Turnbull. 2004. "Mitochondrial DNA Deletion in 'Identical' Twin Brothers." *Journal of Medical Genetics* 41 (2): e19.

Bolze, Alexandre, Fernando Mendez, Simon White, Francisco Tanudjaja, Magnus Isaksson, Misha Rashkin, Johnathan Bowes, et al. 2019. "Selective Constraints and Pathogenicity of Mitochondrial DNA Variants Inferred from a Novel Database of 196,554 Unrelated Individuals." *bioRxiv*. https://doi.org/10.1101/798264.

Brick, Kevin, Fatima Smagulova, Pavel Khil, R. Daniel Camerini-Otero, and Galina V. Petukhova. 2012. "Genetic Recombination Is Directed Away from Functional Genomic Elements in Mice." *Nature* 485 (7400): 642–45.

Calvo, Sarah E., Karl R. Clauser, and Vamsi K. Mootha. 2016. "MitoCarta2.0: An Updated Inventory of Mammalian Mitochondrial Proteins." *Nucleic Acids Research* 44 (D1): D1251–57.

Canela, Andres, Yaakov Maman, Seolkyoung Jung, Nancy Wong, Elsa Callen, Amanda Day, Kyong-Rim Kieffer-Kwon, et al. 2017. "Genome Organization Drives Chromosome Fragility." *Cell* 170 (3): 507–21.e18.

Chan, C. C. W., V. W. S. Liu, E. Y. L. Lau, W. S. B. Yeung, E. H. Y. Ng, and P. C. Ho. 2005. "Mitochondrial DNA Content and 4977 Bp Deletion in Unfertilized Oocytes." *Molecular Human Reproduction* 11 (12): 843–46.

Chapman, Simon N., Jenni E. Pettay, Virpi Lummaa, and Mirkka Lahdenperä. 2019. "Limits to Fitness Benefits of Prolonged Post-Reproductive Lifespan in Women." *Current Biology*. https://doi.org/10.1016/j.cub.2018.12.052.

Clima, Rosanna, Roberto Preste, Claudia Calabrese, Maria Angela Diroma, Mariangela Santorsola, Gaetano Scioscia, Domenico Simone, Lishuang Shen, Giuseppe Gasparre, and Marcella Attimonelli. 2017. "HmtDB 2016: Data Update, a Better Performing Query System and Human Mitochondrial DNA Haplogroup Predictor." *Nucleic Acids Research* 45 (D1): D698–706.

Cortopassi, Gino A. 2002. "A Neutral Theory Predicts Multigenic Aging and Increased Concentrations of Deleterious Mutations on the Mitochondrial and Y Chromosomes." *Free Radical Biology & Medicine* 33 (5): 605–10.

Damas, Joana, João Carneiro, António Amorim, and Filipe Pereira. 2014. "MitoBreak: The Mitochondrial DNA Breakpoints Database." *Nucleic Acids Research* 42 (Database issue): D1261–68.

Engelhardt, Sacha C., Patrick Bergeron, Alain Gagnon, Lisa Dillon, and Fanie Pelletier. 2019. "Using Geographic Distance as a Potential Proxy for Help in the Assessment of the Grandmother Hypothesis." *Current Biology: CB* 29 (4): 651–56.e3.

Frey, Brendan J., and Delbert Dueck. 2007. "Clustering by Passing Messages between Data Points." *Science* 315 (5814): 972–76.

Gammage, Payam A., Joanna Rorbach, Anna I. Vincent, Edward J. Rebar, and Michal Minczuk. 2014. "Mitochondrially Targeted ZFNs for Selective Degradation of Pathogenic Mitochondrial Genomes Bearing Large-Scale Deletions or Point Mutations." *EMBO Molecular Medicine* 6 (4): 458–66.

Gammage, Payam A., Carlo Viscomi, Marie-Lune Simard, Ana S. H. Costa, Edoardo Gaude, Christopher A. Powell, Lindsey

Van Haute, et al. 2018. “Genome Editing in Mitochondria Corrects a Pathogenic mtDNA Mutation in Vivo.” *Nature Medicine* 24 (11): 1691–95.

Goldstein, Amy, and Marni J. Falk. 2003. “Mitochondrial DNA Deletion Syndromes.” In *GeneReviews*, edited by Margaret P. Adam, Holly H. Ardinger, Roberta A. Pagon, Stephanie E. Wallace, Lora J. H. Bean, Karen Stephens, and Anne Amemiya. Seattle (WA): University of Washington, Seattle.

Gollihue, Jenna L., Samir P. Patel, and Alexander G. Rabchevsky. 2018. “Mitochondrial Transplantation Strategies as Potential Therapeutics for Central Nervous System Trauma.” *Neural Regeneration Research* 13 (2): 194–97.

Gorman, Gráinne S., Robert McFarland, Jane Stewart, Catherine Feeney, and Doug M. Turnbull. 2018. “Mitochondrial Donation: From Test Tube to Clinic.” *The Lancet* 392 (10154): 1191–92.

Guo, Xinhong, Konstantin Yu Popadin, Natalya Markuzon, Yuriy L. Orlov, Yevgenya Kraytsberg, Kim J. Krishnan, Gábor Zsurka, Douglas M. Turnbull, Wolfram S. Kunz, and Konstantin Khrapko. 2010. “Repeats, Longevity and the Sources of mtDNA Deletions: Evidence from ‘Deletional Spectra.’” *Trends in Genetics: TIG* 26 (8): 340–43.

Herbst, Allen, Jonathan Wanagat, Nashwa Cheema, Kevin Widjaja, Debbie McKenzie, and Judd M. Aiken. 2016. “Latent Mitochondrial DNA Deletion Mutations Drive Muscle Fiber Loss at Old Age.” *Aging Cell*. https://doi.org/10.1111/acel.12520.

Herbst, A., J. W. Pak, D. McKenzie, E. Bua, M. Bassiouni, and J. M. Aiken. 2007. “Accumulation of Mitochondrial DNA Deletion Mutations in Aged Muscle Fibers: Evidence for a Causal Role in Muscle Fiber Loss.” *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*. https://doi.org/10.1093/gerona/62.3.235.

Hiona, Asimina, Alberto Sanz, Gregory C. Kujoth, Reinald Pamplona, Arnold Y. Seo, Tim Hofer, Shinichi Someya, et al. 2010. “Mitochondrial DNA Mutations Induce Mitochondrial Dysfunction, Apoptosis and Sarcopenia in Skeletal Muscle of Mitochondrial DNA Mutator Mice.” *PloS One* 5 (7): e11468.

Hyslop, Louise A., Paul Blakeley, Lyndsey Craven, Jessica Richardson, Norah M. E. Fogarty, Elpida Fragouli, Mahdi Lamb, et al. 2016. “Towards Clinical Application of Pronuclear Transfer to Prevent Mitochondrial DNA Disease.” *Nature* 534 (7607): 383–86.

Kang, Eunju, Jun Wu, Nuria Marti Gutierrez, Amy Koski, Rebecca Tippner-Hedges, Karen Agaronyan, Aida Platero-Luengo, et al. 2016. “Mitochondrial Replacement in Human Oocytes Carrying Pathogenic Mitochondrial DNA Mutations.” *Nature* 540 (7632): 270–75.

Khaidakov, Magomed, Eric R. Siegel, and Robert J. Shmookler Reis. 2006. “Direct Repeats in Mitochondrial DNA and Mammalian Lifespan.” *Mechanisms of Ageing and Development* 127 (10): 808–12.

Khrapko, Konstantin. 2011. “The Timing of Mitochondrial DNA Mutations in Aging.” *Nature Genetics*.

Konstantin Popadin, Georgii Bazykin. 2008. “Nucleotide Repeats in Mitochondrial Genome Determine Human Lifespan,” no. 713 (October). https://doi.org/10.1038/npre.2008.2399.1.

Kraytsberg, Yevgenya, Elena Kudryavtseva, Ann C. McKee, Changiz Geula, Neil W. Kowall, and Konstantin Khrapko. 2006a. “Mitochondrial DNA Deletions Are Abundant and Cause Functional Impairment in Aged Human Substantia Nigra Neurons.” *Nature Genetics* 38 (5): 518–20.

———. 2006b. “Mitochondrial DNA Deletions Are Abundant and Cause Functional Impairment in Aged Human Substantia Nigra Neurons.” *Nature Genetics* 38 (5): 518–20.

Krishnan, Kim J., Amy K. Reeve, David C. Samuels, Patrick F. Chinnery, John K. Blackwood, Robert W. Taylor, Sjoerd Wanrooij, Johannes N. Spelbrink, Robert N. Lightowlers, and Doug M. Turnbull. 2008. “What Causes Mitochondrial DNA Deletions in Human Cells?” *Nature Genetics*. https://doi.org/10.1038/ng.f.94.

Le, Tung B. K., Maxim V. Imakaev, Leonid A. Mirny, and Michael T. Laub. 2013. “High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome.” *Science* 342 (6159): 731–34.

Myers, Simon, Colin Freeman, Adam Auton, Peter Donnelly, and Gil McVean. 2008. “A Common Sequence Motif Associated with Recombination Hot Spots and Genome Instability in Humans.” *Nature Genetics* 40 (9): 1124–29.

Needleman, Saul B., and Christian D. Wunsch. 1970. “A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins.” *Journal of Molecular Biology* 48 (3): 443–53.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. “IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.” *Molecular Biology and Evolution* 32 (1): 268–74.

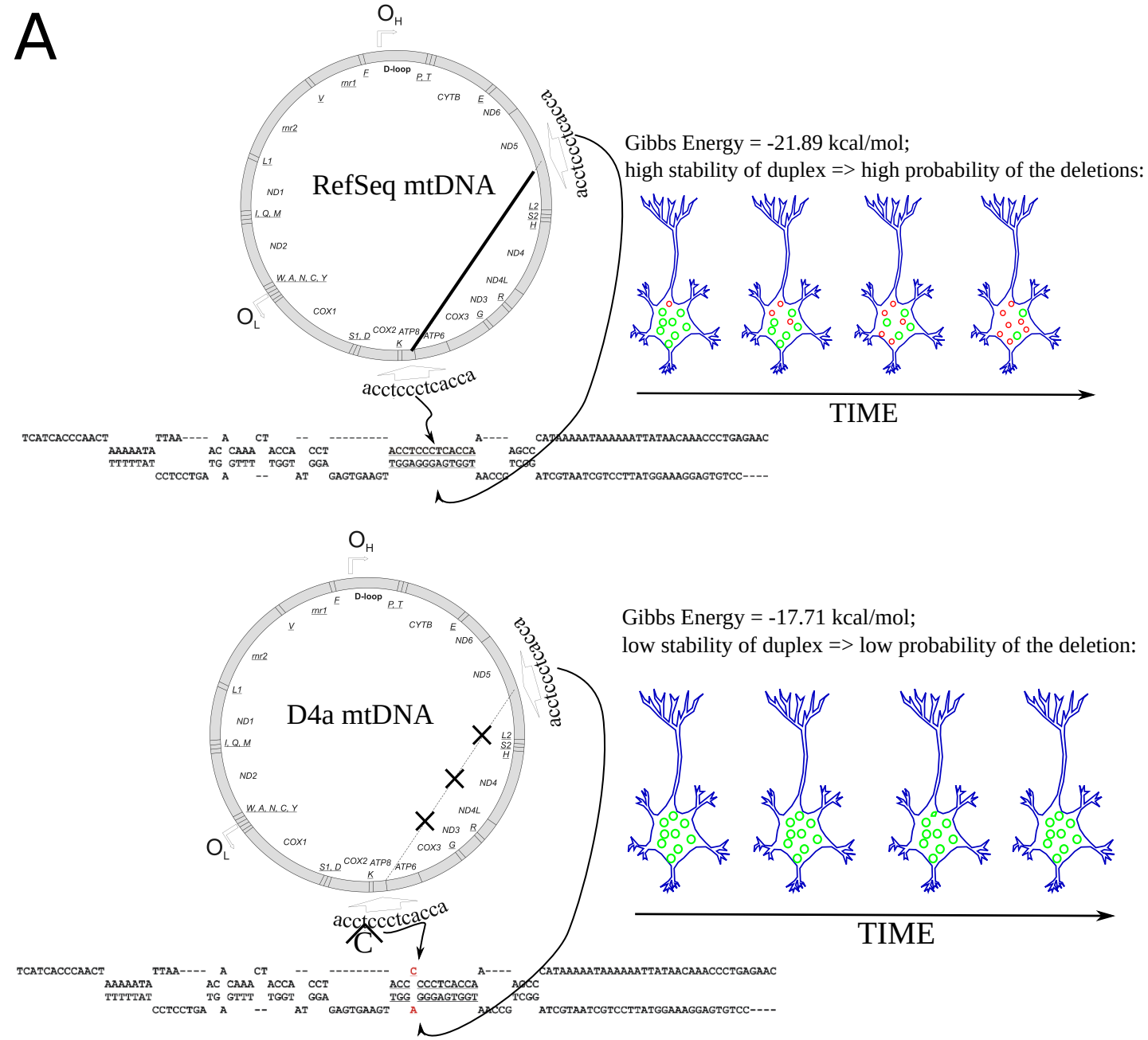
Payne, Brendan A. I., Ian J. Wilson, Charlotte A. Hateley, Rita Horvath, Mauro Santibanez-Koref, David C. Samuels, D. Ashley Price, and Patrick F. Chinnery. 2011. “Mitochondrial Aging Is Accelerated by Anti-Retroviral Therapy through the Clonal Expansion of mtDNA Mutations.” *Nature Genetics* 43 (June): 806.

Peeva, Viktoriya, Daniel Blei, Genevieve Trombly, Sarah Corsi, Maciej J. Szukszto, Pedro Rebelo-Guiomar, Payam A. Gammage, et al. 2018. “Linear Mitochondrial DNA Is Rapidly Degraded by Components of the Replication Machinery.” *Nature Communications* 9 (1): 1727.

Peng, Min-Sheng, Huy Ho Quang, Khoa Pham Dang, An Vu Trieu, Hua-Wei Wang, Yong-Gang Yao, Qing-Peng Kong, and Ya-Ping Zhang. 2010. “Tracing the Austronesian Footprint in Mainland Southeast Asia: A Perspective from

Mitochondrial DNA." *Molecular Biology and Evolution* 27 (10): 2417–30.
Persson, Örjan, Yazh Muthukumar, Swaraj Basu, Louise Jenninger, Jay P. Uhler, Anna-Karin Berglund, Robert McFarland, et al. 2019. "Copy-Choice Recombination during Mitochondrial L-Strand Synthesis Causes DNA Deletions." *Nature Communications* 10 (1): 1–10.
Popadin, Konstantin, and Georgii Bazykin. 2008. "Nucleotide Repeats in Mitochondrial Genome Determine Human Lifespan." *Nature Precedings*. https://doi.org/10.1038/npre.2008.2399.1.
Popadin, Konstantin, Adeel Safdar, Yevgenya Kraytsberg, and Konstantin Khrapko. 2014. "When Man Got His mtDNA Deletions?" *Aging Cell*.
Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." *Cell* 159 (7): 1665–80.
Samuels, David C. 2004. "Mitochondrial DNA Repeats Constrain the Life Span of Mammals." *Trends in Genetics: TIG* 20 (5): 226–29.
Samuels, David C., Eric A. Schon, and Patrick F. Chinnery. 2004. "Two Direct Repeats Cause Most Human mtDNA Deletions." *Trends in Genetics: TIG* 20 (9): 393–98.
Schon, E., R. Rizzuto, C. Moraes, H. Nakase, M. Zeviani, and S. DiMauro. 1989. "A Direct Repeat Is a Hotspot for Large-Scale Deletion of Human Mitochondrial DNA." *Science*. https://doi.org/10.1126/science.2711184.
Shamanskiy, Viktor A., Valeria N. Timonina, Konstantin Yu Popadin, and Konstantin V. Gunbin. 2019. "ImtRDB: A Database and Software for Mitochondrial Imperfect Interspersed Repeats Annotation." *BMC Genomics* 20 (Suppl 3): 295.
Soubrier, Julien, Mike Steel, Michael S. Y. Lee, Clio Der Sarkissian, Stéphane Guindon, Simon Y. W. Ho, and Alan Cooper. 2012. "The Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates." *Molecular Biology and Evolution* 29 (11): 3345–58.
Trifunovic, Aleksandra, Anna Wredenberg, Maria Falkenberg, Johannes N. Spelbrink, Anja T. Rovio, Carl E. Bruder, Mohammad Bohlooly-Y, et al. 2004. "Premature Ageing in Mice Expressing Defective Mitochondrial DNA Polymerase." *Nature* 429 (6990): 417–23.
Trivedi, Rajni, T. Sitalaximi, Jheelam Banerjee, Anamika Singh, P. K. Sircar, and V. K. Kashyap. 2006. "Molecular Insights into the Origins of the Shompen, a Declining Population of the Nicobar Archipelago." *Journal of Human Genetics* 51 (3): 217–26.
Viguera, E., D. Canceill, and S. D. Ehrlich. 2001. "Replication Slippage Involves DNA Polymerase Pausing and Dissociation." *The EMBO Journal* 20 (10): 2587–95.
Weissensteiner, Hansi, Dominic Pacher, Anita Kloss-Brandstätter, Lukas Forer, Günther Specht, Hans-Jürgen Bandelt, Florian Kronenberg, Antonio Salas, and Sebastian Schönherr. 2016. "HaploGrep 2: Mitochondrial Haplogroup Classification in the Era of High-Throughput Sequencing." *Nucleic Acids Research* 44 (W1): W58–63.
Yang, Jiang-Nan, Andrei Seluanov, and Vera Gorbunova. 2013. "Mitochondrial Inverted Repeats Strongly Correlate with Lifespan: mtDNA Inversions and Aging." *PloS One* 8 (9): e73318.
Yu-Wai-Man, Patrick, Joey Lai-Cheong, Gillian M. Borthwick, Langping He, Geoffrey A. Taylor, Laura C. Greaves, Robert W. Taylor, Philip G. Griffiths, and Douglass M. Turnbull. 2010. "Somatic Mitochondrial DNA Deletions Accumulate to High Levels in Aging Human Extraocular Muscles." *Investigative Ophthalmology & Visual Science* 51 (7): 3347–53.
Zuker, Michael. 2003. "Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction." *Nucleic Acids Research* 31 (13): 3406–15.

A


O_H
D-loop
F
P T
mr1
V
CYTB
E
ND6
mr2
ND5
L1
ND1
I, Q, M
L2
S2
H
ND2
ND4
W, A, N, C, Y
ND4L
ND3
R
COX1
COX3
G
O_L
S1, D
COX2
ATP8
ATP6
K
RefSeq mtDNA
acctccctcacca
Gibbs Energy = -21.89 kcal/mol;
high stability of duplex => high probability of the deletions:
TIME
TCATCACCCAACT TTAA---- A CT -- --------- A---- CATAAAAATAAAAAATTATAACAAACCCTGAGAAC
AAAAATA AC CAAA ACCA CCT ACCTCCCTCACCA AGCC
TTTTTAT TG GTTT TGGT GGA TGGAGGGAGTGGT TCGG
CCTCCTGA A -- AT GAGTGAAGT AACCG ATCGTAATCGTCCTTATGGAAAGGAGTGTCC----
D4a mtDNA
acctccctcacca
C
Gibbs Energy = -17.71 kcal/mol;
low stability of duplex => low probability of the deletion:
TIME
TCATCACCCAACT TTAA---- A CT -- --------- C A---- CATAAAAATAAAAAATTATAACAAACCCTGAGAAC
AAAAATA AC CAAA ACCA CCT ACC CCCTCACCA AGCC
TTTTTAT TG GTTT TGGT GGA TGG GGGAGTGGT TCGG
CCTCCTGA A -- AT GAGTGAAGT A AACCG ATCGTAATCGTCCTTATGGAAAGGAGTGTCC----

B

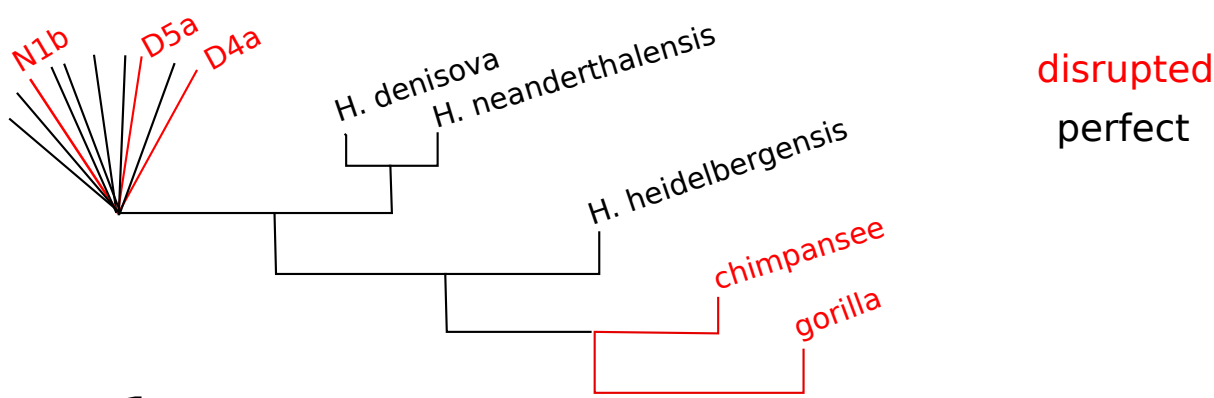
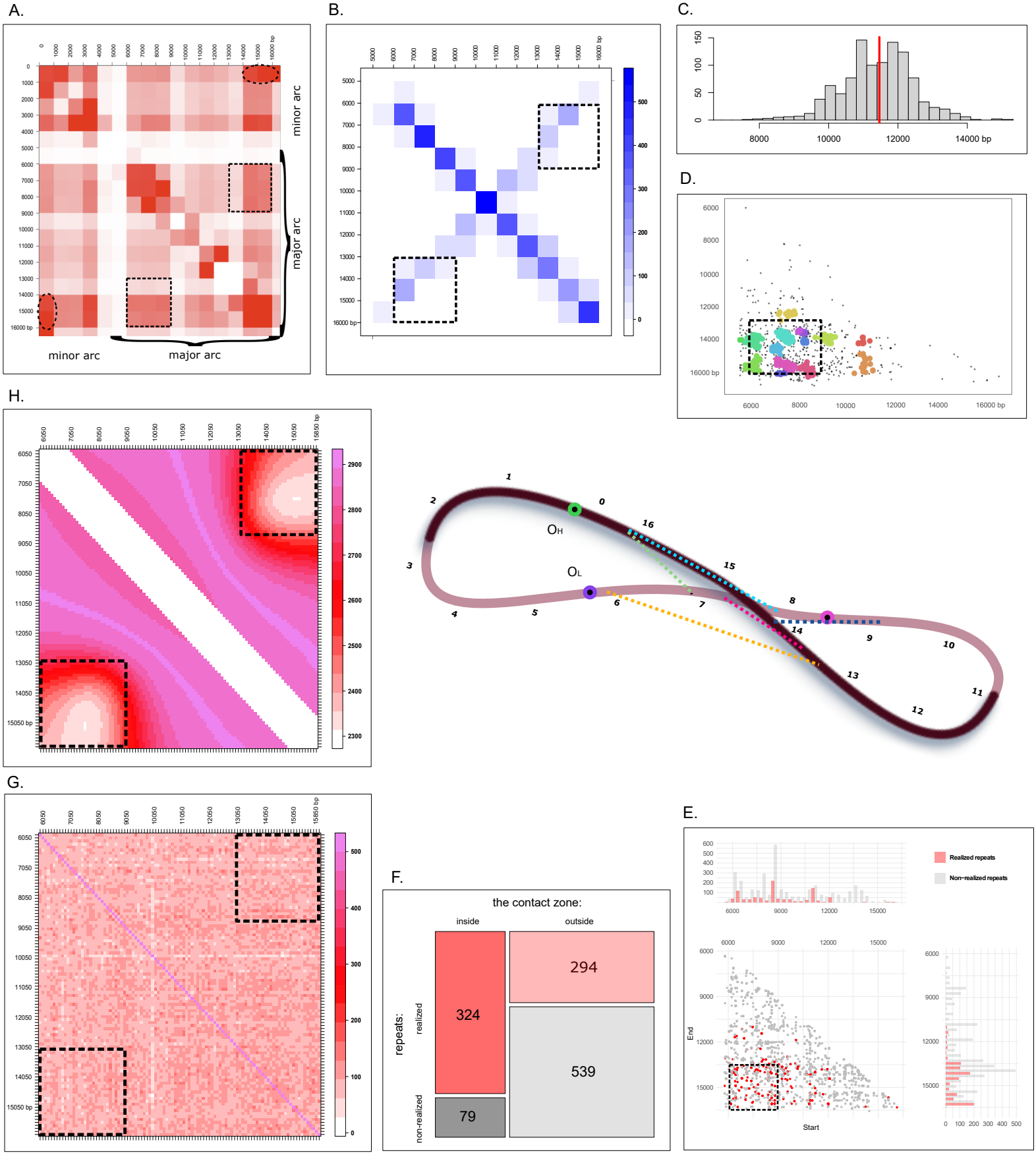
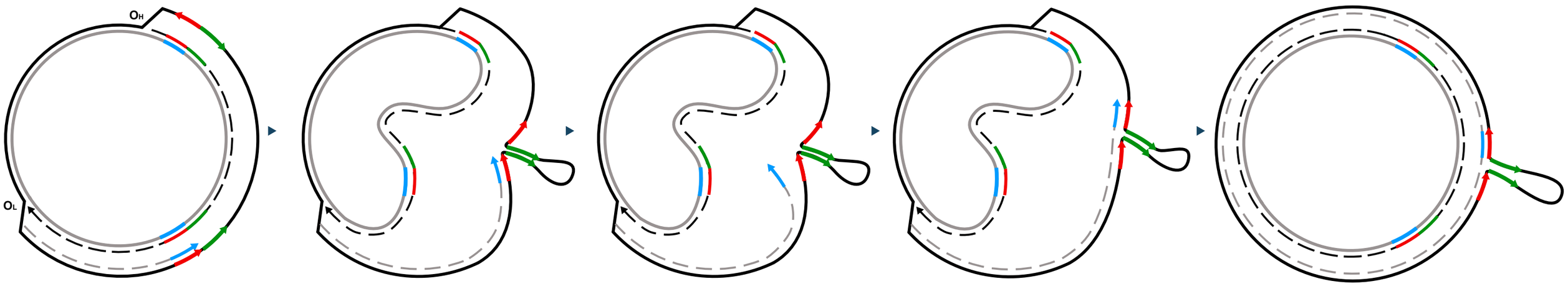

N1b
D5a
D4a
H. denisova
H. neanderthalensis
H. heidelbergensis
chimpansee
gorilla
disrupted
perfect

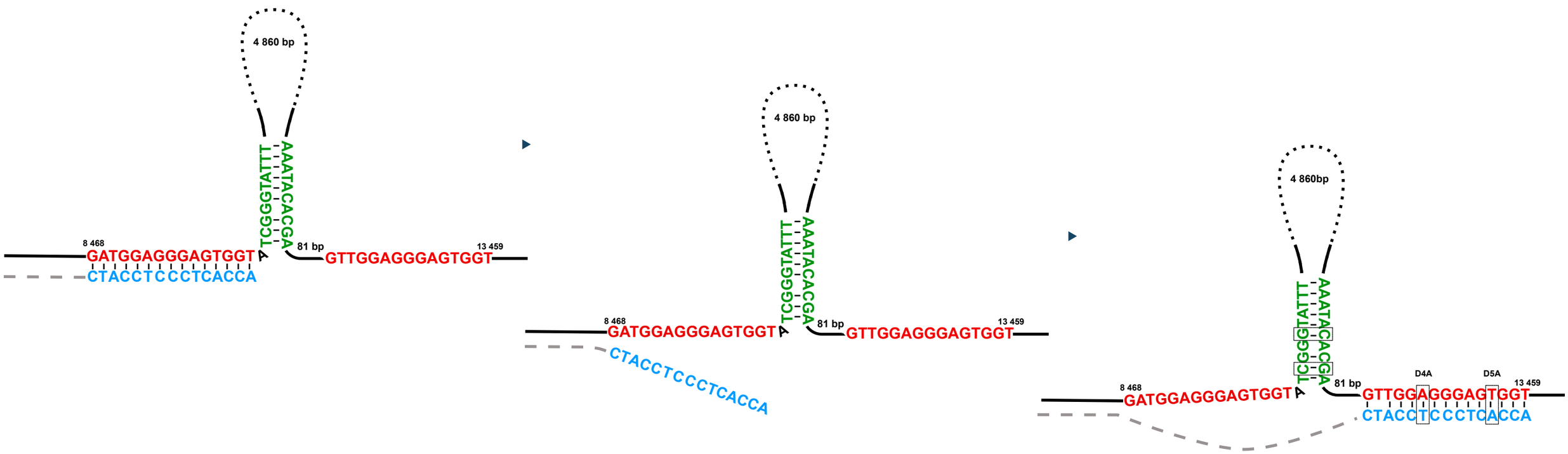
figure 1


A.
minor arc
major arc
minor arc
major arc
B.
C.
8000
10000
12000
14000 bp
D.
6000
8000
10000
12000
14000
16000 bp
H.
G.
F.
the contact zone:
inside
outside
repeats:
realized
non-realized
324
294
539
79
E.
Realized repeats
Non-realized repeats
End
Start

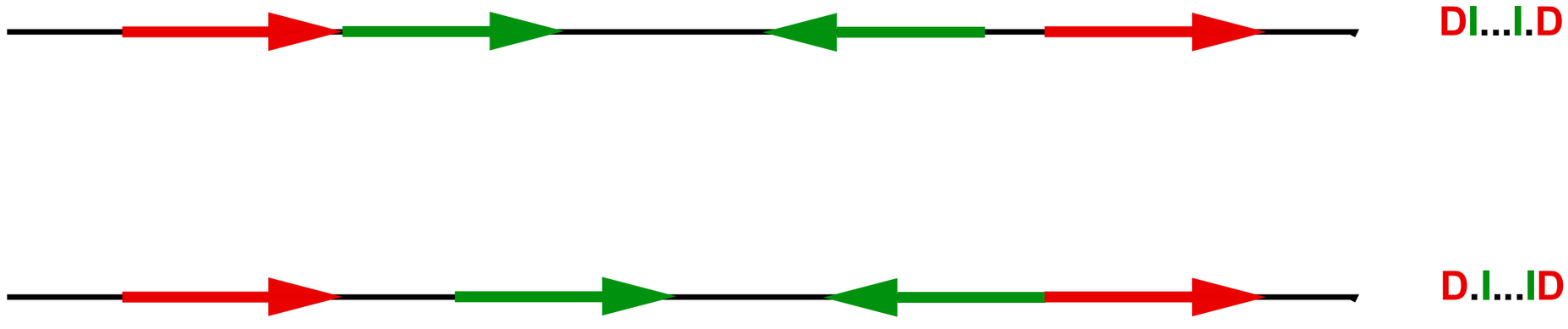
A.


OH
OL

B.


4 860 bp
TCGGGTATTT
AAATACACGA
8 468
GATGGAGGGAGTGGT
CTACCTCCCTCACCA
81 bp
GTTGGAGGGAGTGGT
13 459
4 860 bp
TCGGGTATTT
AAATACACGA
8 468
GATGGAGGGAGTGGT
CTACCTCCCTCACCA
81 bp
GTTGGAGGGAGTGGT
13 459
4 860bp
TCGGGTATTT
AAATACACGA
8 468
GATGGAGGGAGTGGT
81 bp
D4A
D5A
GTTGGAGGGAGTGGT
13 459
CTACCTCCCTCACCA

C.


DI...I.D
D.I...ID