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

Alina A. Mikhailova[#1], Viktor Shamanskiy[#1], Kristina Ushakova[#1], Alina G. Mikhailova[#1,2], Sergei Oreshkov[1], Dmitry Knorre[3,4], Evgenii O. Tretiakov[5], Marianna Zazhytska[6], Samuel W. Lukowski[7], Chia-Wei Liou[8], Tsu-Kung Lin[8], Wolfram S. Kunz[9,10], Alexandre Reymond[11], Ilia Mazunin[12], Georgii A. Bazykin[12,13], Konstantin Gunbin[1,14], Jacques Fellay[15], Masashi Tanaka[16-18], Konstantin Khrapko[19], Konstantin Popadin[1,11,15]*

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

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

Abstract

Ageing is associated with accumulation of somatic mutations. This process is especially pronounced in mitochondrial genomes (mtDNA) of postmitotic cells, where the accumulation of somatic mitochondrial deletions is associated with healthy ageing and mitochondrial encephalomyopathies. Deletions are often flanked by direct nucleotide repeats, however, they do not provide an exhaustive explanation of deletion distribution. We hypothesized that in parallel with the role of direct repeats there is also a global secondary structure of mtDNA, shaping deletion formation. Analyzing the folding energies of the heavy chain, which stays single-stranded during mtDNA replication, we observed a potential contact zone between 6-9kb and 13-16kb of the major arc of mtDNA. Describing the distribution of deletions in the human mtDNA we demonstrated that the contact zone is 3-times more mutagenic under all else equal. The proposed topological model improves our understanding of the mechanisms of deletion formation in the human mitochondrial genome.

Introduction

Ageing is associated with accumulation of DNA damage. The mitochondrial genome (mtDNA) is more pronounced to accumulate damage since it exists within a cell in a large number of copies with continuous turnover (Poovathingal et al. 2012) and high mutation rate (Rebolledo-Jaramillo et al. 2014). These conditions assure the coexistence of different mtDNA variants within the same cell (heteroplasmy) (Wilton et al. 2018), which might compete with each other. The intracellular mtDNA selection is thought to be especially influential in slowly-dividing tissues, where selfish mtDNA mutants, such as mtDNA deletions, have time for clonal expansion. For example, in substantia nigra neurons, the first mtDNA deletions appear at around 50 years of age (Kraytsberg et al. 2006a; Bender et al. 2006). Each year, this fraction of heteroplasmy increases by 1-2% until, after several decades, a phenotypically essential threshold of 50-80% is reached (Kraytsberg et al. 2006b) leading to neurodegeneration. A second tissue that is predisposed to the accumulation of mtDNA somatic deletions is skeletal muscle. Myofibrils are 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 that are also affected by mtDNA deletions include extraocular muscles (Yu-Wai-Man et al. 2010) and oocytes (Blakely et al. 2004; Barritt et al. 1999; Chan et al. 2005). In the case of oocytes, the expansion of mtDNA deletions could potentially manifest itself across all tissues, including proliferative ones, leading to multisystem disorders (Schon et al. 1989; Goldstein and Falk 2003).

mtDNA deletions are considered to be causative for the host cell degeneration and several corresponding age-related phenotypes. This is supported by several lines of evidence: (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 myofiber breakage (Hiona et al. 2010) and with respiratory chain-deficient neurons (Kraytsberg et al 2006) also supports the causative effects of mtDNA deletions on ageing; (iii) there is a deficit of neurons carrying an extremely high (> 80%) deletion burden, suggesting that such cells degraded and were no longer present in the analyzed tissue (Kraytsberg et al. 2006b). Thus, the high mtDNA deletion burden is not a neutral hallmark of aged cells but more likely a causative agent. Thus, understanding the molecular mechanisms underlying the origin of somatic mtDNA deletions, as well as their velocity, is of primary importance (Khrapko 2011; Popadin et al. 2014).

It has been shown that the majority of somatic mtDNA deletions are flanked by direct nucleotide repeats (Samuels, Schon, and Chinnery 2004) or by 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 "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 evidence of the deleterious effect of repeats in mtDNA of long-lived mammals. Similarly, with these comparative-species results, we previously proposed that the decreased number of direct repeats in the mitochondrial genome of some human haplogroups can lead to a lower prevalence of somatic mtDNA deletions, thereby enabling healthier ageing or even postponing the ageing process (Popadin and Bazykin 2008).

Analysis of associations between human longevity and properties of mtDNA showed that the longest (13 base-pairs) and the most severe direct repeat (so-called "common repeat") in the human mitochondrial genome had been disrupted by the point mutation m.8473T>C in the Japanese D4a haplogroup (Popadin and Bazykin 2008). This haplogroup is well known due to the substantial relative 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 by one-nucleotide mismatch has a beneficial effect on the D4a haplogroup (Popadin and Bazykin 2008) since it decreases the probability of the occurrence of the corresponding somatic deletion. This is also called the "common" deletion because it is frequently observed in aged postmitotic tissues (Figure 1A). Interestingly, bacterial deletions, which may share 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 examined their mtDNA deletion spectra. In line with our hypothesis (Popadin and Bazykin 2008) we observed no common deletions in N1b samples (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 in order to uncover the main factors behind the formation of mtDNA deletions we focused on key properties of the most mutagenic direct repeat of the human mtDNA - the common repeat.

Results

1. the common repeat is human-specific

Analysis of reference sequences of primate mtDNAs and several available ancient human mtDNAs shows that the common repeat is human-specific. It presents in Homo heidelbergensis, Homo neanderthalensis and Homo denisova but is absent (i.e. has a mismatch) 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). We observed that more than 98% (42641 out of 43437) of sequenced humans have the perfect common repeat, while less than 2% of humans lost the repeat as a result of different single-nucleotide substitutions in the proximal arm (8470 - 8482 bp, see table 1).

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

Among the carriers of the disrupted common repeat, the most frequent variant is m.8473T>C which occurred independently several times, 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). This provided 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 the 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 the D5a haplogroup, which is also characterized by increased longevity independently of D4a (Alexe et al. 2007). A recent analysis of a human mitochondrial genome database (Bolze et al. 2019) containing 196,554 complete entries confirms the rarity of the disrupted common repeat. In this data, the most common disrupted variant was m.8473T>C (1%: 2118 out of 196,554) and the second most common variant was m.8472C>T (0.4%: 833 out of 196,554). It is interesting to note that both frequent synonymous variants m.8473T>C and m.8479A>G are associated with increased longevity (Alexe et al. 2007; Bilal et al. 2008). Synonymous variants may 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, the nonsynonymous N1b specific variant (m.8472C>T) has not been associated with increased longevity, which may be due to the potentially beneficial effect of the disruption of the repeat, i.e. decreased deletional burden (Guo et al. 2010), has been counterbalanced by the potentially 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).

The sequences 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 most mutagenic region of the major arc - an effect of the common repeat or the global 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 "appears to be the principal factor behind the formation of most deletions" (Samuels, Schon, and Chinnery 2004). It means that the common repeat may be important not only for the formation of the common deletion but also for the formation of all other deletions. When we analyzed the mtDNA deletion spectrum in the frontal cortex of N1b samples with disrupted common repeat, we aimed to test both these effects: (i) the decrease in the frequency of the common deletion and (ii) a potential shift in the pattern of distribution of all other deletions. Having a very small sample size (2 cases and 2 controls), we observed an almost complete absence of the common deletion in N1b samples, but we didn't find any changes in other deletions (Guo et al. 2010). 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 is the distribution of deletions within the major arc 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 an increased probability of deletions near the common repeat might be maintained not by the common repeat per se, but by an independent

factor such as the spatial organization of mtDNA.

Below we reconstruct the potential secondary structure of the whole mitochondrial major arc and analyze - how it may affect the deletion spectra.

4. 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 (Supplementary figure 1, ovals mark the contacts reflecting the circular nature of mtDNA). The second strongest contact was observed within the major arc (Supplementary figure 1, dotted ovals mark 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. The Hi-C contact matrix may be predictive for a folding of double-stranded mtDNA, however, according to the most recent model (Persson et al. 2019), mtDNA deletions originate during mtDNA replication when a long stretch of the parental heavy chain stays single-stranded. Thus, in our downstream analyses we focused on the reconstruction of the secondary structure of the single-stranded major arc during replication.

In order to reconstruct the potential secondary structure of the single-stranded parental heavy chain of the whole major arc, we performed an *in silico* folding using Mfold (see Methods). Using the output structure of the whole molecule, we derived a contact matrix as the number of hydrogen bonds between two DNA regions with 1kb resolution (Fig 2B, bottom-left triangle). Interestingly, we observed an unexpected diagonal from the lower left to the upper right part of the matrix with the potential contact zone, which also overlapped the one from the Hi-C matrix (Supplementary file 1). 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 hairpin with a center close to 10-11 kb and the large-scale stem correspondingly aligned to each of the 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 found at 6.5 kb in front of 14.5 kb (see several alternative schemes on Figure 2A3-5). The global secondary structure of the single-stranded DNA is expected to be maintained by the inverted repeats, which can hybridize with each other to form stems and loops. To test this, 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 weak positive correlation (Spearman's rho =

0.05, p = 0.0017, N= 4005, diagonal elements were removed from the analyses).

In-silico folding of a very long (~10kb) single-stranded DNA molecule (Fig2B, bottom-left triangle) may have it's own computational limitations and artificially force the origin of the hairpin-like structures (Figure 2A3-5). To avoid this potential problem we split the major arc into short (100 bp) windows, folded them in all pairwise combinations (see Methods), estimated Gibbs Energies for each pair and finally reconstructed the contact matrix (Fig 2B, upper-right triangle). The contact matrix shows several stripes corresponding to the strongest contacts (three horizontal lines with ordinate equals 6100, 6900 and 7900 and one vertical with abscissa equals 15000) and intersection of these lines gives a potential contact zone.

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

In silico folding experiments (Fig 2B) described above suggested that the single-stranded heavy chain of the major arc might be shaped like a hairpin (Fig 2A). 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 hairpin 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 2D, 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 test this, we grouped individual deletions into several clusters (Fig 2C clusters are presented by colored dots, see Methods) and confirmed that the majority of clusters are indeed located within the potential contact zone (several biggest clusters of deletions are plotted as colored dotted lines, connecting regions of the contact zone - see upper sub-panel on Fig2C).

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 test the importance of the global secondary DNA structure as an additional factor affecting the formation of the deletions, we have to take into account direct repeats. Firstly, we compared the distribution of the perfect direct repeats (see Methods) and deletions from MitoBreak. We observed approximately uniform distribution of the direct repeats within the major arc versus the strongly biased distribution of the deletions (Fig 2F). 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 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 this, 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 a distinct difference: realized repeats tended to be located closer to the potential contact zone, while non-realized repeats were distributed more uniformly in the major arc (Fig 2F). 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 a 2000 bp shorter distance between non-realized repeats versus realized. The average interval between arms was 5643 bp for realized repeats and 3450 bp for non-realized ones, the average start was 8678 for realized and 9394 for non-realized repeats, and the average end was 14321 bp for realized and 13039 bp for non-realized repeats (all p-values $< 2.2e-16$, paired Mann-Whitney U-test; Number of clusters = 618, Fig 2E). Together, this suggests that neither position of repeats nor repeat motifs (Fig 2F) can explain the shifted distribution of deletions. Because 80% of realized repeats start in the interval from 6465-10954 bp and end at the interval from 13286-15863 bp, we suggest that this biased distribution can be explained by the potential contact zone of the single-stranded DNA, which we annotated arbitrarily using contact matrixes of our in silico folding experiments (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 was 7.5 (Fig 2F1, mosaic-plot; Fisher odds ratio = 7.5, $p < 2.2e-16$). Thus, the most parsimonious explanation of the non-uniform distribution of deletions (Fig 2C,D,F) 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, Fig2A).

6. 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 is poorly explained by the distribution of direct repeats (Fig 2F) and the potential secondary structure of the single-stranded mtDNA can be an extra factor affecting the deletions (Fig 2A,2B). Here, to prove the deficiency of the direct-repeat-only explanation we extend it to an integral metric of similarity between two regions of DNA estimating all pairwise microhomology scores between all 100 bp windows inside the major arc (Fig 2E bottom - left triangle, 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. (equation 1)

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. (equation 2)

We can see that the variable "potential contact zone" significantly and positively affects the probability of deletions. Because we used standardised variables, we could conclude that the potential contact zone variable affects the odds of probabilities 3 times stronger as compared to the microhomology score.

The potential contact zone between 6-9kb and 13-16kb was derived arbitrarily from two in-silico folding experiments (Fig 2B) and an unexplained excess of deletions in this region (Fig2C,2D,2F, 2E lower-left triangle) supported the potential importance of this region. Here, in order to derive more exact data-driven location of the potential contact zone we performed additional analyses. In order to do this, 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 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 the coordinates 7550 bp (5') and 15150 bp (3'). Plotting the heatmap with AIC for each contact point, we observed the data-driven contact zone was similar to our visually-derived 6-9kb vs 13-16kb potential contact zone (fig 2E upper right triangle). 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 - the results are qualitatively similar (data not shown). Altogether, using three orthogonal sources of information: *in silico* foldings (Fig 2B), empirical distribution of deletions (Fig 2CDF) and microhomology scores (Fig 2E bottom left), we propose that the distribution of the human mtDNA deletions is a function of both global secondary structure of the single-stranded major arc (Fig 2A) and local similarity between DNA regions, i.e. microhomologies.

Discussion

We proposed an existence of the global secondary structure of the human single-stranded heavy chain of mtDNA (Fig 2A) and demonstrated that the mtDNA deletion spectrum is associated not only with DNA similarity but also with the potential contact zone between 6-9 and 13-16 kb. This means that formation of deletions depends on both direct repeats, which shape the mtDNA microhomology, and inverted repeats, which shape the secondary structure of the single-stranded heavy chain of the major arc. These results are best compatible with the replication slippage mechanism (Persson et al. 2019; Albertini et al. 1982) where the nested pattern of direct and inverted repeats leads to the formation of deletions. All our analyses were performed *in silico* and additional experimental evidence is needed to test this hypothesis.

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 mtDNA disease risks. 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), it is expect to observe not only postponed ageing (Bilal et al. 2008; Alexe et al. 2007), but also decreased rate of neurodegeneration (Kraytsberg et al. 2006b), frequency of Parkinson diseases (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), rate of mitochondrial ageing in HIV patients (Payne et al. 2011) and rate of 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). Haplogroup-specific mtDNA secondary structures, which can be obtained experimentally or computationally, can add an additional factor explaining the mtDNA deletion risks and associated variation in mtDNA-related phenotypes. Because of the high mutagenicity of the spatially close to each other mtDNA regions (equation 2) we expect that mtDNA secondary structure, yet poorly known molecular phenotype, may play an especially important role in explanation of haplogroup-specific risks of encephalomyopathies.

The possibility of predicting mtDNA deletion burden and spectrum based on mtDNA sequence represents an important step forward for mitochondrial medicine. Haplogroups with low expected deletion burden can provide a donor mtDNA in mitochondrial donation (Kang et al. 2016; Hyslop et al. 2016; Gorman et al. 2018) and mitochondrial transplantation (Gollihue, Patel, and Rabchevsky 2018; Ali Pour, Kenney, and Kheradvar 2020) approaches. Additionally, predicted haplogroup-specific spectrum of deletions can help to establish targeted systems eliminating expected deletions in

neurons and muscle cells of aged individuals (Bacman et al. 2013; Gammage et al. 2014; Bacman et al. 2018; Gammage et al. 2018; Loutre et al. 2018).

Evolutionary, comparative species data might be used to test this hypothesis. 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). 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 demonstrated also that abundance of both types of repeats, direct and inverted, correlate with each other (Shamanskiy et al. 2019). If the mutual effect of both direct and inverted repeats is associated with mammalian lifespan - is an interesting anticipated direction of the future research.

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). Additionally we folded pairs of 100 bp windows of the single-stranded heavy chain using ViennaRna Package 2 (Lorenz et al. 2011) with DNA settings and analyzed resulting Gibbs Energies as a metric of a potential interaction between two single-stranded DNA regions.

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 cluster 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 minimal level of similarity 80%) with length 10 bp and more has been used for all analyses (Shamanskiy et al. 2019).

In this dataset there are 2957 direct and 764 inverted repeats in the major arc. 207 direct repeats with both arms located close to each other were deleted from the analysis because they do not have any no nested inverted repeats. The rest 2750 direct repeats (DD), containing at least one pair of nested inverted repeats, were used in downstream analyses. For each such DD pair we called all pairs of inverted repeats, located between the direct ones and found the proximal gap as the minimal distance between the first arm of the direct repeat and the first arm of the inverted repeat. The median of the proximal gaps is 21. Similarly, for each such DD pair we found the distal gap as the minimal distance between the second arm of an inverted repeat and the second arm of the direct repeat. The median of the distal gaps is 25.

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]. 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 oval marks a potential contact zone within the major arc. **A.** potential hairpin structures formed by a single-stranded parental heavy chain during mtDNA replication ; **B.** *In silico* global foldings. bottom left: contact matrix derived from the *in silico* folding of the whole single-stranded heavy chain of the major arc; top-tight: contact matrix derived from the *in silico* foldings of 100 bp windows of the single-stranded heavy chain of the major arc. **C.** Clusters of deletions within the major arc. Majority of clusters are located within the potential contact zone; **D.** Centers of deletions. The median of the distribution is close to the expected center, 11 kb, of the loop of the major arc; **E.** bottom-left: heatmap of the microhomologies between 100 bp windows within the major arc. Microhomology alone poorly

explains the distribution of the deletions (empty circled) within the major arc; upper - right: Heatmap of the data-driven contact zone, based on AIC of the compared models. **F.** Realized (red) versus non-realized (gray) repeats tend to be enriched in the potential contact zone; **F1.** Mosaic-plot of repeats (realized versus non-realized) within and outside the potential contact zone;

Supplementary Figure 1: Hi-C contact matrix of mtDNA. Dotted circles mark the potential contact zone.

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.

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.

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.

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.

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.

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.

Lorenz, Ronny, Stephan H. Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. 2011. "ViennaRNA Package 2.0." *Algorithms for Molecular Biology: AMB* 6 (November): 26.

Loutre, Romuald, Anne-Marie Heckel, Damien Jeandard, Ivan Tarassov, and Nina Entelis. 2018. "Anti-Replicative Recombinant 5S rRNA Molecules Can Modulate the mtDNA Heteroplasmy in a Glucose-Dependent Manner." *PloS One* 13 (6): e0199258.

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.

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.

Poovathingal, Suresh Kumar, Jan Gruber, Lakshminarayanan Lakshmanan, Barry Halliwell, and Rudiyanto Gunawan. 2012. "Is Mitochondrial DNA Turnover Slower than Commonly Assumed?" *Biogerontology* 13 (5): 557–64.

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.

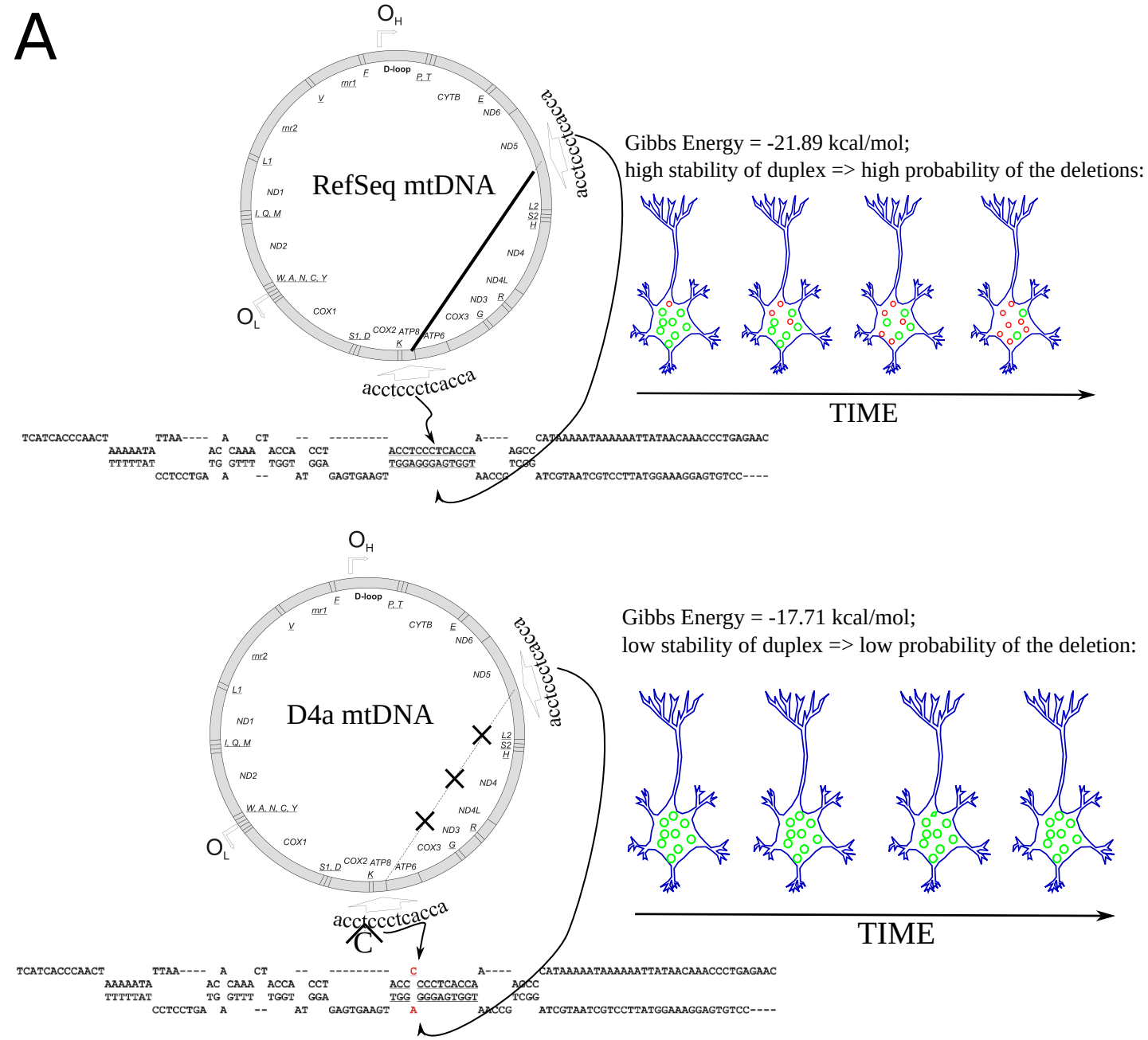
Popadin, Konstantin, Adeel Safdar, Yevgenya Kraytsberg, and Konstantin Khrapko. 2014. "When Man Got His mtDNA Deletions?" *Aging Cell*.

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.

Rebolledo-Jaramillo, Boris, Marcia Shu-Wei Su, Nicholas Stoler, Jennifer A. McElhoe, Benjamin Dickins, Daniel Blankenberg, Thorfinn 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 of the United States of*

America 111 (43): 15474–79.
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.
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.
Wilton, Peter R., Arslan Zaidi, Kateryna Makova, and Rasmus Nielsen. 2018. "A Population Phylogenetic View of Mitochondrial Heteroplasmy." *Genetics* 208 (3): 1261–74.
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

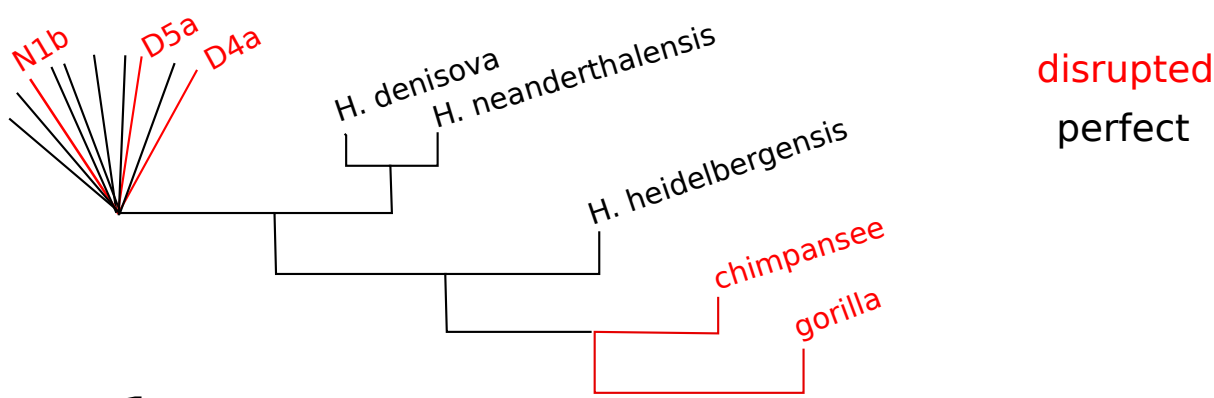
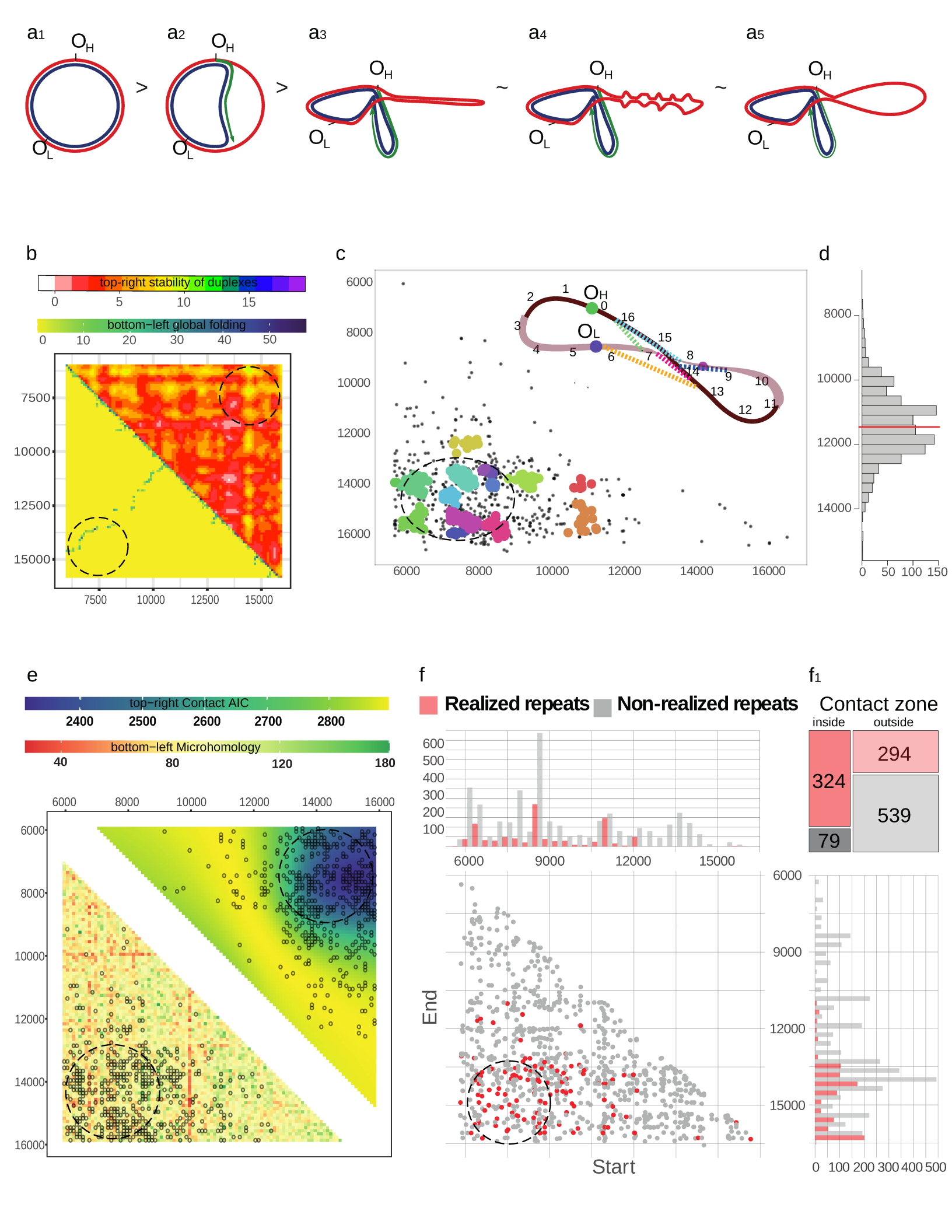

N1b
D5a
D4a
H. denisova
H. neanderthalensis
H. heidelbergensis
chimpansee
gorilla
disrupted
perfect

figure 1


a1
a2
a3
a4
a5
OH
OL
b
top-right stability of duplexes
bottom−left global folding
c
d
e
top−right Contact AIC
bottom−left Microhomology
f
Realized repeats
Non-realized repeats
End
Start
f1
Contact zone
inside
outside
324
294
539
79