Risk of mitochondrial deletions is affected by the global secondary structure of the human 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], Jonathan B. Overdevest[6], Samuel W. Lukowski[7], Chia-Wei Liou[8], Tsu-Kung Lin[8], Wolfram S. Kunz[9,10], Alexandre Reymond[11], Ilya Mazunin[12,21], Georgii A. Bazykin[12,13], Konstantin Gunbin[1,14], Jacques Fellay[15], Masashi Tanaka[16-18], Konstantin Khrapko[19], Konstantin Popadin[1,11,15,20]*

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 Otolaryngology, Columbia University Irving Medical Center, 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.
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.
20. Swiss Institute of Bioinformatics, Switzerland.
21. Fomin Clinic, Moscow, Russia.

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

Abstract

Aging is often associated with clonal expansion of somatic mitochondrial deletions, while their origin is still poorly understood. Deletions are often flanked by direct nucleotide repeats, however, repeats alone do not provide an exhaustive explanation of deletion distribution. Here, we hypothesized that spatial proximity of repeats alters the rate of conversion into deletions. Analyzing distribution of human deletions, we observed a hot spot of deletions with one breakpoint located within the region of 6-9kb and another within 13-16kb of the mitochondrial genome. This deletion hot spot is not explained by the distribution of the direct repeats and might be driven by close contacts of these two regions during mtDNA replication. Using Hi-C data derived from several samples of two different human tissues we observed the increased density of contacts in the potential contact zone. Using several in silico approaches we reconstructed the secondary structure of mitochondrial DNA and confirmed that it can be organized as a large-scale hairpin-like loop with a center close to 11 kb and stem between 6-9 kb and 13-16 kb. We demonstrated that repeats within the contact zone are 3-times more mutagenic as compared to repeats outside the contact zone, which would also explain the well known increased mutagenicity of the common repeat (8470-8482 bp and 13447-13459 bp). An analysis of the largest human deletion dataset with 470,000 unique deletions showed that the contact zone describes better age-associated, not disease-related, deletions, emphasizing that our model explains the mechanism of deletion manifestation that commonly occurs with the process of healthy aging in the human population. The proposed topological model improves our understanding of the mechanisms of deletion formation in the human mitochondrial genome and opens a possibility to predict more precisely a deletion burden in different human haplogroups and mammalian species.

Introduction

Aging is associated with accumulation of DNA damage. The mitochondrial genome (mtDNA), existing within a cell in a large number of copies with continuous turnover [1] and high mutation rate [2] is strongly predisposed to accumulate an age-related damage. The coexistence of different mtDNA variants within the same cell (heteroplasmy) [3] assure an intracellular mtDNA competition, which is especially influential in slowly-dividing tissues, where the so-called "selfish" mtDNA mutants have a time for clonal expansion. One of the best studied examples of the selfish mtDNA mutations are deletions. In substantia nigra neurons, for example, the first mtDNA deletions appear at around 50 years of age [4,5]. Each year, this fraction of heteroplasmy increases by 1-2% until, after several decades, a phenotypically essential threshold of 50-80% is reached [6] leading to neurodegeneration. Skeletal muscle is another tissue that is predisposed to the accumulation of mtDNA somatic deletions: an expansion of somatic mtDNA deletions within myofibrils is associated with sarcopenia - loss of muscle weight and strength with age [7,8]. Other tissues with slow-dividing cells that are also affected by mtDNA deletions include extraocular muscles [9] and oocytes [10–12]. In the case of oocytes, the expansion of mtDNA deletions could potentially manifest itself across all tissues, including proliferative ones, leading to multisystem disorders [13,14].

Somatic mtDNA deletions are considered to be causative for the host cell degeneration and several corresponding age-related phenotypes. The proof-reading-deficient version of mtDNA polymerase causes somatic accumulation of point mutations and deletions in mice, leading to reduced lifespan and premature onset of aging-specific phenotypes ([15][16], but see also [17,18]). The occurance of mtDNA deletions in regions of myofiber breakage [19] and respiratory chain-deficient neurons (Kraytsberg et al 2006) supports the causative effects of mtDNA deletions on aging. A deficit of neurons carrying an extremely high (> 80%) deletion burden suggests that such cells are degraded and are no longer present in the analyzed tissue [6]. Altogether, 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 rate of expansion, is of primary importance [20,21].

It has been shown that the majority of somatic mtDNA deletions are flanked by direct nucleotide repeats [22] or by long imperfect duplexes consisting of short stretches of direct repeats [23]. 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 harmful in late life [24]. The negative correlation between the abundance of direct repeats in mtDNA and species-specific lifespan of mammals [25,26] has been interpreted as evidence of the deleterious effect of repeats in mtDNA of long-lived mammals. Similarly to a deficit of direct repeats in mtDNA of long-lived mammals, we previously hypothesized 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 aging and postponing the aging process [27].

Even as the most well understood contributors to mtDNA deletion development, direct nucleotide repeats (or long imperfect duplexes) still explain a small fraction of observed deletion distributions. Why are some repeats very often realized into deletions, while others (with the same length and motif) - never? What are the additional factors behind the mtDNA deletion formation? Here, by drawing parallels between deletions in bacteria [28], mtDNA common repeat [29] and nuclear DNA [30], where in all cases topological proximity of DNA regions with repeats increases their chances to interact and to cause a deletion, we hypothesize that direct repeats may cause deletions with higher likelihood of occurrence based on their spatial proximity to each other. In order to uncover the main factors behind the formation of mtDNA deletions it is reasonable to start with analysis of the most mutagenic direct repeat of the human genome - the common repeat ([22], [23], [27]). An 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" [22]. 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. To test this hypothesis we analyzed the mtDNA deletion spectrum in the frontal cortex of N1b samples with disrupted common repeat (the proximal arm was acTtccctcacca versus acCtccctcacca as it is in the vast majority of the human population). If there is a special structural role of the common repeat we expected to see that the distribution of all deletions within N1b samples will differ from the controls (haplogroups with the perfect common repeat). Within our sample size (2 cases and 2 controls), we observed a near complete absence of the common deletion *per se* in N1b samples, however we didn't find any changes in distribution of other deletions [23]. Thus we rejected the hypothesis that the common repeat is the main factor behind the formation of the majority of deletions [22]. 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 novel explanation. Here, we hypothesized that the increased probability of deletions appearing 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. In order to test this hypothesis, using Hi-C approach, several in silico tools, as well as large scale data from recent ultrasensitive method to call mtDNA deletions, we reconstructed the potential secondary structure of the whole mitochondrial major arc - region with the vast majority of deletions - and analyzed how it may affect the deletion spectra. Altogether we propose that mtDNA is organized as a large-scale hairpin-like loop with a center close to 11 kb and stem between 6-9 kb and 13-16 kb. This infinity-symbol-like structure affects realization of direct repeats and thus shapes a deletion formation.

Results

1. The major arc of mtDNA may be folded into large-scale loops: Hi-C data

To reconstruct the spatial structure of mtDNA, we first used the publicly available densest contact matrix from Hi-C experiments on human lymphoblastoid cells with available resolution of at least 1 kb [31]. We observed high-density contacts between 0-1 kb and 15-16.5 kb which most likely reflect the circular nature of mtDNA (nucleotides with positions 1 and 16569 are neighbour nucleotides) and confirms that the quality of the spatial reconstruction of mtDNA from the Hi-C data is meaningful (Supplementary figure 1, ovals mark the contacts reflecting the circular nature of mtDNA). The next strongest contact was observed within the major arc (Supplementary figure 1, dotted squares mark the potential contact zone): 6-9 kb versus 13-16 kb. This mtDNA contact suggests that the major arc is shaped as a loop, and the whole mtDNA may resemble an "infinity symbol" with the potential contact zones 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.

To check the robustness of the publicly available mtDNA Hi-C matrix (31), we obtained six Hi-C contact matrixes of mtDNA derived from olfactory receptors of controls and Covid patients [32]. Analysis of these contact matrixes confirmed an existence of both high-density contacts between 0-1 kb and 15-16.5 kb, reflecting circular nature of mtDNA, and the contact zone between 6-9 kb and 13-16 kb, supporting our "infinity symbol" model (Supplementary Fig 2). No apparent difference between Covid patients and controls was observed (Supplementary Fig 2).

According to a recent report [29], mtDNA deletions originate during mtDNA replication when a long stretch of the parental heavy chain stays single-stranded. The aforementioned Hi-C contact matrixes reflect mainly non-replicating double-stranded mtDNA with a minor fraction of replicated molecules with single-sranded heavy chain. This expectedly low fraction of replicating mtDNA molecules can make results of the Hi-C analysis less relevant for the reconstruction of the structure of replicating mtDNA and in parallel can explain a weak structural signal obtained from the Hi-C data. Future Hi-C experiments with enriched fraction of replicating mtDNA will shed light on the structure of the molecule that is more relevant to deletion formation. In our downstream analyses we focused on the reconstruction of the secondary structure of the single-stranded major arc during replication.

2. The major arc of mtDNA may be folded into large-scale loops: in silico folding

In order to reconstruct the potential secondary structure of the single-stranded parental heavy chain of the major arc, we performed an *in silico* folding using Mfold (see Methods). Using the Mfold output obtained for the whole molecule, we derived a contact matrix as the number of hydrogen bonds between two DNA regions with 1kb resolution (Fig 2A, bottom-left triangle). Interestingly, we observed unexpected diagonal contacts from the lower left to the upper right part of the matrix,

which also overlapped with the potential contact zone from the Hi-C matrix (Supplementary Fig 1). This cross-like contact matrix graph resembles bacterial Hi-C data [33], 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 formed by the 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 found at 6.5 kb in front of 14.5 kb (see several alternative schemes on Fig 1).

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 contacting elements of the *in silico* folding matrix (Fig 2A, bottom-left triangle) and observed an expected positive trend (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 was used to generate the result in Fig 2A, bottom-left triangle, which may have computational limitations and artificially force the origin of the hairpin-like structures (Fig 1). To avoid this potential problem we split the major arc into short (100 bp) windows, folded all pairwise combinations (see Methods), estimated Gibbs Energies for each pair, and finally reconstructed the contact matrix (Fig 2A, upper-right triangle). The contact matrix graph 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 overlaps well with a potential contact zone. Altogether, the three different in silico folding approaches (Gibbs Energy of long regions of DNA, distribution of inverted repeats, Gibbs Energy of short regions of DNA) all support the existence of a potential contact zone between 6-9 kb and 13-16 kb of mtDNA.

3. single-stranded heavy chain of the major arc can be shaped like a hairpin: evidence from deletion spectra

Hi-C and *in silico* folding experiments described above suggested that the single-stranded heavy chain of the major arc might be shaped like a hairpin (Fig 1). If the formation of deletions depends on a 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 [34], 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 test this, we grouped individual deletions into several clusters (Fig 2B clusters are presented by colored dots, see Methods) and confirmed that the majority of clusters are indeed located within the potential contact zone (upper panel on Fig

2B: several biggest clusters of deletions are marked by colored dotted lines, connecting regions of the contact zone).

4. distribution of deletions is poorly explained by the direct repeats solely

It is widely accepted that the distribution of deletions is partially explained by direct repeats within the human mtDNA [22,23] and bacterial genomes [28]. 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. To do it we compared the distribution of the perfect direct repeats (see Methods) and deletions from MitoBreak database. We observed approximately uniform distribution of the direct repeats within the major arc versus the strongly biased distribution of the deletions (Supplementary Fig 3). This result is in line with the previous observation of Samuels et al [22] and confirms that direct repeats alone poorly explain the distribution of deletions.

Deletions might be induced by special motifs within the direct repeats - for example, it has been shown that repeats inducing deletions are more C-rich [29]. Thus, there is a possibility that 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 [35] 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 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 a 2000 bp shorter distance between arms of non-realized repeats versus realized (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 2E) can explain the shifted distribution of deletions. Because 80% of realized repeats start in the interval 6465-10954 bp and end at the interval 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 Hi-C contact matrixes and contact matrixes of our in silico folding experiments (see Fig 2A) as frequent contacts between 6-9 kb and 13-16 kb. Indeed, there is a strong excess of realized repeats within the potential contact zone (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 2B,C,E) is that the regions with the high density of deletions, i.e. realized repeats, are located in spatial proximity to each other during mtDNA replication (Fig 1, Fig 2A).

5. probability of deletions is 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 alone (Fig 2E) while the potential secondary structure of the single-stranded mtDNA can be an extra factor affecting deletion formation (Fig 1, Fig 2A, 2B, 2C). Here, we aim to build a multiple model including both repeats and secondary structure as major factors affecting deletion development. Instead of direct repeats, we derived a more biologically meaningful "microhomology similarity" metric, which is (i) an integral metric of similarity between two regions of DNA and (ii) it is fixed to 100 bp windows to facilitate downstream analyses (see Methods). We estimated all pairwise microhomology similarities between all 100 bp windows inside the major arc (Fig 2D bottom - left triangle) and, as expected, obtained a positive correlation 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). As a next step, to analyze an association between deletions and the microhomology similarity, we performed logistic regression where for each cell (100x100bp) of our matrix we derived a value for presence (coded as 1, N = 484) or absence (coded as 0, N = 4466) of deletions from the MitoBreak database [34]. We estimated the logit function of deletions as a function of the Microhomology Similarity (MS):

log(p/(1-p)) = -2.25264 + 0.27442 * MS, all p values (intercept, coefficient) less or equal 5.13e-09, N = 4005. (equation 1)

This result confirms our previous findings [23] and demonstrates that a high microhomology similairty is associated with a higher probability of deletion. Next, we added a second independent variable to our model, the Contact Zone (CZ) coded as one within the zone (6-9kb and 13-16kb) and zero - outside:

log(p/(1-p)) = -2.38296 + 0.32592 * MS + 0.90579 * CZ, all p values less or equal 3.8e-10, N = 4005. (equation 2)

We can see that the variable of potential contact zone significantly and positively affects the probability of deletions. Because we used standardised variables, we can compare coefficients and conclude that the potential contact zone variable affects the odds of probabilities 3-fold more than microhomology similarity.

Until now the potential contact zone area (6-9 kb and 13-16kb) was derived arbitrarily from our (i) Hi-C data, (ii) in-silico folding experiments (Fig 2A), (iii) non-uniform distribution of deletions and (iv) weak association of deletions with repeats (Fig 2B, 2C, 2E, 2D lower-left triangle). In order to derive a more exact location of the potential contact zone we performed additional data-driven analyses. 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 ran 4005 logistic regressions and tested 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 of 7550 bp (5') and 15150 bp (3'). Plotting the heatmap with AIC for each contact point, we demonstrated that the data-driven contact zone was similar to our visually-derived 6-9kb vs 13-16kb potential contact zone (fig 2D upper right triangle). If instead of AIC as a metric for model selection

we use residual deviance, P-value or the coefficient of the second independent variable - the results are qualitatively similar (data not shown). Altogether, using several orthogonal sources of information: Hi-C data (Supplementary Figures 1 and 2), *in silico* foldings (Fig 2A), empirical distribution of deletions (Fig 2C, 2E) and microhomology similarities (Fig 2D 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 1) and the local similarity between DNA regions, i.e. microhomologies.

6. the contact zone describes the age-related deletions - the most common deletions in the human population

Recently, an ultrasensitive method for quantifying deletions in the human mtDNA helped to uncover about 470,000 unique deletions [36]. Unsupervised bioinformatic analyses of this big dataset revealed three principal components, describing the main properties of deletions: the first component discriminated disease- versus healthy- associated deletions; the second component split deletions according to the location into the minor and the major arcs; the third component was aging related since the scores of deletions increased with age at biopsy. Deletions with high positive scores on the third principal component were (i) not only associated with advanced age, but also (ii) were mainly located within the major arc, (iii) had high microhomology similarity between the breakpoints and (iv) were located within the major arc in a manner that breakpoints near origins of replication and in the middle of the major arc were mostly penalized (fig 4C in the paper [36]). This location pattern of the age-associated deletions in the major arc of healthy samples strongly resembles our potential contact zone. Using the metadata available to our team, we confirmed that the scores of the third principal component for each bin (207 x 207 bp) of the major arc were higher within contact zone (p-value $< 4.48^{-13}$, Mann-Whitney U test, Supplementary Fig 4). This strong association of the contact zone with age-related deletions corroborates our results and emphasizes that the secondary structure of mtDNA mainly affects the healthy age-related deletions, which are the most common in the human population.

Interestingly, this class of deletions, with high scores on the third PC, is also in line with the mechanism of deletion formation as a primer slippage during replication (class 2 deletion in terms of the paper Lujan et al. 2020). This mechanism in turn, implies the asynchronous strand displacement DNA replication model [36] - a classical model of mtDNA replication, recently unambiqously confirmed by several papers [37,38].

Discussion

We proposed that the secondary structure of the human single-stranded heavy chain of mtDNA has a hairpin structure with a contact zone between 6-9 and 13-16 kb and demonstrated that the mtDNA deletion spectrum is affected not only by the DNA similarity between the breakpoint regions, but also by the contact zone (Fig 1, Fig 2). 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 [29]; [28] where the nested pattern of direct and inverted repeats (hereafter DIID: Direct Inverted Inverted Direct) leads to the formation of deletions.

Recently reported experimental validations on mtDNA further support our hypothesise. An interesting control experiment has been recently performed on mtDNA of Nematomorpha, characterized by high enrichment for perfect inverted repeats of considerable length [39]. The authors validated the capability of inverted repeats to form hairpins and affect DNA replication in the PCR amplification and restriction experiment. It has been shown that DIID - inverted repeats nested within the direct repeats disappeared during PCR (see suppl Fig 2 in the paper) meaning that the shorter products are most likely a result of PCR jumping facilitated by the presence of direct repeats flanking the main hairpin. This demonstrates that DIID pattern indeed is highly mutagenic and leads to deletion formation. Until now, there is no experimental reconstruction of the spatial structure of a single stranded parental heavy chain of the major arc during human mtDNA replication. However, another recent experiment on human mtDNA showing that replicative polymerases cause deletion by copy-choice recombination between direct repeats and that this effect is stimulated by secondary structures [29] is aligned with our results. Altogether, we used six complimentary analytic approaches that provide convergent support for our proposed model, which has externally validated biological relevance in published studies on the formation of bacterial deletions [28], DIID experiments with Nematomorpha [39] and DIID experiments with mitochondrial deletions [29]. We would like to emphasize also that our model ranks mutagenicity of direct repeats according to their location (3-times higher mutability of direct repeats in the contact zone) and thus it is testable in future experiments. Additionally, our model suggests the development of a contact zone puts in close spatial proximity two origins of mtDNA replication (heavy and light) that will facilitate their cross-talk and co-regulation, where this theory requires further experimental validation.

Deeper understanding of the deletion formation process opens a possibility of predicting 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 aids in the uncovering of haplogroup-specific mtDNA disease risks. Our topological model adds one more argument to the increased mutagenicity of the common repeat, because it is located within the contact zone, and predicts that the disruption of any repat, including the common one, in the contact zone will have stronger molecular phenotypes as compared to repeats outside the contact zone. For example, in the haplogroups with the disrupted common repeat (D4a, D5a, N1b), we expect to observe the decreased somatic mtDNA deletion burden [23] and consequently: postponed aging [40,41], decreased rate of neurodegeneration [6], frequency of Parkinson diseases [5], skeletal muscle myopathies [7,8], extraocular muscle weakness [9], rate of mitochondrial aging in HIV patients [42] and rate of early-onset “mtDNA deletion syndromes” classically consisting of Kearns-Sayre syndrome (KSS), Pearson syndrome and progressive

external ophthalmoplegia (PEO) [13,14]. 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 spatially proximate mtDNA regions (see equation 2) we expect that mtDNA secondary structure may play an important role in the explanation of haplogroup-specific risks of encephalomyopathies and other human phenotypes [43].

The possibility of predicting mtDNA deletion burden and spectrum based on mtDNA sequence would offer an important step forward for mitochondrial medicine. Haplogroups with low expected deletion burden can provide a preferable donor mtDNA in mitochondrial donation [44–46] and mitochondrial transplantation [47]; [48] approaches. Additionally, a predicted haplogroup-specific spectrum of deletions can potentially help to establish a way of using of targeted systems for elimination of expected deletions in neurons and muscle cells of aged individuals ([49,50]; [51,52]; [53]).

In the future it would be advantageous to use comparative species data in order to extend our hypothesis into evolutionary scale and prove that DIID patterns increase occurrence of deletions in mtDNA of all species. Initially, it was reported that the mammalian lifespan negatively correlates with an abundance of direct repeats in mtDNA [25,26], assuming that direct repeats lead to the formation of mtDNA deletions, limiting the lifespan. Later it was demonstrated that inverted repeats show an even stronger negative correlation with mammalian lifespan [54]. Recently, it has also been estimated that the abundance of direct and inverted repeats shows strong correlation between those two types [35]. If the combined effect of both direct and inverted repeats, i.e. abundance of DIID patterns, is the best predictor of the mammalian lifespan - it would open up an exciting potential direction for future research in the field.

Methods

Hi-C mtDNA contact matrix: The publicly available mtDNA matrix was visualized using Juicebox.

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

The corresponding paper describing the methogolody of obtaining Hi-C data derived from the human lymphoblastoid cell line is Rao et al [31].

Additionally we obtained six Hi-C mtDNA contact matrixes from olfactory receptors of covid patients and controls. Details of the in situ Hi-C protocol as well as bioinformatics analyses are described in the original paper [32]. Matrices were visualized using Juicebox [55].

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 [29]. Using Mfold [56] 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 [57] 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 [35].

Clusterization of deletions: For clusterization, we used all MitoBreak [34] 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 [58] 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 [23].

Realized and non-realized direct degraded repeats: We used our database of degraded mtDNA repeats [35] 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 [59] 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 [35].

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/, with the key scripts, deposited within the branch HumanMtDnaDeletions, 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]. I.M. was supported by the Russian Science Foundation [No. 21-75-10081]. 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. We acknowledge Scott Lujan and Bill Copeland for providing the metadata of the principal component analysis from their paper, Natalia Ri, Maxim Ri and Irina Gostimskaya for editing and improving the manuscript. Contributions from JBO supported by grant K23DC019678 from the National Institute on Deafness and Other Communication Disorders and the National Institutes of Health as well as through grant UL1TR001873 from the National Center for Advancing Translational Sciences, National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Author contributions

KP designed the study, KG, 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 contributed towards writing of the manuscript.

Figures

Figure_1. Potential secondary structures formed by a single-stranded parental heavy chain during mtDNA replication. The lower panel shows that direct repeats, marked by black arrows, have different chances to be realized into deletions as a function of a spatial structure. The close spatial proximity of repets (bold dotted lines) increases the probability of deletion occurance, while for repeats that are spatially separated by greater distance this probably is decreased (thin dotted line).

Figure_2. Secondary structure of mtDNA. The dotted oval marks a potential contact zone within the major arc.

A. *In silico* approach for global folding prediction. Bottom left: contact matrix derived from the *in silico* folding estimation of the major arc's whole single-stranded heavy chain; top right: contact matrix derived from the *in silico* folding estimation of 100 bp windows of the single-stranded heavy chain of the major arc.

B. Clusters of deletions within the major arc. Majority of clusters are located within the potential contact zone;

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. 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; top right: heatmap of the data-driven contact zone, based on AIC of the compared models.

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;

Supplementary Figures:

Supplementary Figure 1: Hi-C contact matrix of mtDNA obtained from the human lymphoblastoid cells. Dotted squares mark the potential contact zone. Ovals mark the contacts, emphasising the circularity of mtDNA.

Supplementary Figure 2: Hi-C contact matrix of mtDNA obtained from the human olfactory epithelium autopsies. A and B - are contact matrixes from covid patients, C, D, E, F - from controls. Dotted squares mark the potential contact zone. Ovals mark the contacts, emphasising the circularity of mtDNA.

Supplementary Figure_3: Distribution of the perfect direct repeats (red dots) and deletions from MitoBreak (grey dots) in the major arc.

Supplementary Figure 4: The third principal component scores, associated with aging-relalted deletions of healthy samples from paper [36]. Contact zone, marked by the pink square, is characterized by the increased scores (p-value $< 4.48^{-13}$, Mann-Whitney U test).

Literature:

1. Poovathingal SK, Gruber J, Lakshmanan L, Halliwell B, Gunawan R. Is mitochondrial DNA turnover slower than commonly assumed? Biogerontology. 2012;13:557–64.

2. Rebolledo-Jaramillo B, Su MS-W, Stoler N, McElhoe JA, Dickins B, Blankenberg D, et al. Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. Proc Natl Acad Sci U S A. 2014;111:15474–9.

3. Wilton PR, Zaidi A, Makova K, Nielsen R. A Population Phylogenetic View of Mitochondrial Heteroplasmy. Genetics. 2018;208:1261–74.

4. Kraytsberg Y, Kudryavtseva E, McKee AC, Geula C, Kowall NW, Khrapko K. Mitochondrial DNA deletions are abundant and cause functional impairment in aged human substantia nigra neurons. Nat Genet. 2006;38:518–20.

5. Bender A, Krishnan KJ, Morris CM, Taylor GA, Reeve AK, Perry RH, et al. High levels of mitochondrial DNA deletions in substantia nigra neurons in aging and Parkinson disease. Nat Genet. 2006;38:515–7.

6. Kraytsberg Y, Kudryavtseva E, McKee AC, Geula C, Kowall NW, Khrapko K. Mitochondrial DNA deletions are abundant and cause functional impairment in aged human substantia nigra neurons. Nat Genet. 2006;38:518–20.

7. Herbst A, Pak JW, McKenzie D, Bua E, Bassiouni M, Aiken JM. Accumulation of Mitochondrial DNA Deletion Mutations in Aged Muscle Fibers: Evidence for a Causal Role in Muscle Fiber Loss [Internet]. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences. 2007. p. 235–45. Available from: http://dx.doi.org/10.1093/gerona/62.3.235

8. Herbst A, Wanagat J, Cheema N, Widjaja K, McKenzie D, Aiken JM. Latent mitochondrial DNA deletion mutations drive muscle fiber loss at old age [Internet]. Aging Cell. 2016. p. 1132–9. Available from: http://dx.doi.org/10.1111/acel.12520

9. Yu-Wai-Man P, Lai-Cheong J, Borthwick GM, He L, Taylor GA, Greaves LC, et al. Somatic mitochondrial DNA deletions accumulate to high levels in aging human extraocular muscles. Invest Ophthalmol Vis Sci. 2010;51:3347–53.

10. Blakely EL, He L, Taylor RW, Chinnery PF, Lightowlers RN, Schaefer AM, et al. Mitochondrial DNA deletion in “identical” twin brothers. J Med Genet. 2004;41:e19.

11. Barritt JA, Brenner CA, Cohen J, Matt DW. Mitochondrial DNA rearrangements in human oocytes and embryos. Mol Hum Reprod. 1999;5:927–33.

12. Chan CCW, Liu VWS, Lau EYL, Yeung WSB, Ng EHY, Ho PC. Mitochondrial DNA content and 4977 bp deletion in unfertilized oocytes. Mol Hum Reprod. 2005;11:843–6.

13. Schon E, Rizzuto R, Moraes C, Nakase H, Zeviani M, DiMauro S. A direct repeat is a hotspot for large-scale deletion of human mitochondrial DNA [Internet]. Science. 1989. p. 346–9. Available from: http://dx.doi.org/10.1126/science.2711184

14. Goldstein A, Falk MJ. Mitochondrial DNA Deletion Syndromes. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Stephens K, et al., editors. GeneReviews. Seattle (WA): University of Washington, Seattle; 2003.

15. Vermulst M, Wanagat J, Kujoth GC, Bielas JH, Rabinovitch PS, Prolla TA, et al. DNA deletions and clonal mutations drive premature aging in mitochondrial mutator mice. Nat Genet. 2008;40:392–4.

16. Trifunovic A, Wredenberg A, Falkenberg M, Spelbrink JN, Rovio AT, Bruder CE, et al. Premature ageing in mice expressing defective mitochondrial DNA polymerase. Nature. 2004;429:417–23.

17. Khrapko K, Vijg J. Mitochondrial DNA mutations and aging: a case closed? [Internet]. Nature Genetics. 2007. p. 445–6. Available from: http://dx.doi.org/10.1038/ng0407-445

18. Vermulst M, Bielas JH, Kujoth GC, Ladiges WC, Rabinovitch PS, Prolla TA, et al. Mitochondrial point mutations do not limit the natural lifespan of mice. Nat Genet. 2007;39:540–3.

19. Hiona A, Sanz A, Kujoth GC, Pamplona R, Seo AY, Hofer T, et al. Mitochondrial DNA mutations induce mitochondrial dysfunction, apoptosis and sarcopenia in skeletal muscle of mitochondrial DNA mutator mice. PLoS One. 2010;5:e11468.

20. Khrapko K. The timing of mitochondrial DNA mutations in aging. Nat. Genet. 2011. p. 726–7.

21. Popadin K, Safdar A, Kraytsberg Y, Khrapko K. When man got his mtDNA deletions? Aging Cell. 2014. p. 579–82.

22. Samuels DC, Schon EA, Chinnery PF. Two direct repeats cause most human mtDNA deletions. Trends Genet. 2004;20:393–8.

23. Guo X, Popadin KY, Markuzon N, Orlov YL, Kraytsberg Y, Krishnan KJ, et al. Repeats, longevity and the sources of mtDNA deletions: evidence from "deletional spectra." Trends Genet. 2010;26:340–3.

24. Cortopassi GA. A neutral theory predicts multigenic aging and increased concentrations of deleterious mutations on the mitochondrial and Y chromosomes. Free Radic Biol Med. 2002;33:605–10.

25. Samuels DC. Mitochondrial DNA repeats constrain the life span of mammals. Trends Genet. 2004;20:226–9.

26. Khaidakov M, Siegel ER, Shmookler Reis RJ. Direct repeats in mitochondrial DNA and mammalian lifespan. Mech Ageing Dev. 2006;127:808–12.

27. Popadin K, Bazykin G. Nucleotide repeats in mitochondrial genome determine human lifespan [Internet]. Nature Precedings. 2008. Available from: http://dx.doi.org/10.1038/npre.2008.2399.1

28. Albertini AM, Hofer M, Calos MP, Miller JH. On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. Cell. 1982;29:319–28.

29. Persson Ö, Muthukumar Y, Basu S, Jenninger L, Uhler JP, Berglund A-K, et al. Copy-choice recombination during mitochondrial L-strand synthesis causes DNA deletions. Nat Commun. Nature Publishing Group; 2019;10:1–10.

30. Canela A, Maman Y, Jung S, Wong N, Callen E, Day A, et al. Genome Organization Drives Chromosome Fragility. Cell. 2017;170:507–21.e18.

31. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.

32. Zazhytska M, Kodra A, Hoagland DA, Frere J, Fullard JF, Shayya H, et al. Non-cell-autonomous disruption of nuclear architecture as a potential cause of COVID-19-induced anosmia. Cell. 2022;185:1052–64.e12.

33. Le TBK, Imakaev MV, Mirny LA, Laub MT. High-resolution mapping of the spatial organization of a bacterial chromosome. Science. 2013;342:731–4.

34. Damas J, Carneiro J, Amorim A, Pereira F. MitoBreak: the mitochondrial DNA breakpoints database. Nucleic Acids Res. 2014;42:D1261–8.

35. Shamanskiy VA, Timonina VN, Popadin KY, Gunbin KV. ImtRDB: a database and software for mitochondrial imperfect interspersed repeats annotation. BMC Genomics. 2019;20:295.

36. Lujan SA, Longley MJ, Humble MH, Lavender CA, Burkholder A, Blakely EL, et al. Ultrasensitive deletion detection links mitochondrial DNA replication, disease, and aging. Genome Biol. 2020;21:248.

37. Sanchez-Contreras M, Sweetwyne MT, Kohrn BF, Tsantilas KA, Hipp MJ, Schmidt EK, et al. A replication-linked mutational gradient drives somatic mutation accumulation and influences germline polymorphisms and genome composition in mitochondrial DNA. Nucleic Acids Res. Oxford University Press (OUP); 2021;49:11103–18.

38. Mikhaylova AG, Mikhailova AA, Ushakova K, Tretiakov EO, Iliushchenko D, Shamansky V, et al. A mitochondria-specific mutational signature of aging: increased rate of A>G substitutions on a heavy chain [Internet]. bioRxiv. 2022 [cited 2022 Jun 4]. p. 2021.12.03.460832. Available from: https://www.biorxiv.org/content/10.1101/2021.12.03.460832v2

39. Mikhailov KV, Efeykin BD, Panchin AY, Knorre DA, Logacheva MD, Penin AA, et al. Coding palindromes in mitochondrial genes of Nematomorpha. Nucleic Acids Res. 2019;47:6858–70.

40. Bilal E, Rabadan R, Alexe G, Fuku N, Ueno H, Nishigaki Y, et al. Mitochondrial DNA haplogroup D4a is a marker for extreme longevity in Japan. PLoS One. 2008;3:e2421.

41. Alexe G, Fuku N, Bilal E, Ueno H, Nishigaki Y, Fujita Y, et al. Enrichment of longevity phenotype in mtDNA haplogroups D4b2b, D4a, and D5 in the Japanese population. Hum Genet. 2007;121:347–56.

42. Payne BAI, Wilson IJ, Hateley CA, Horvath R, Santibanez-Koref M, Samuels DC, et al. Mitochondrial aging is accelerated by anti-retroviral therapy through the clonal expansion of mtDNA mutations. Nat Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011;43:806.

43. Yonova-Doing E, Calabrese C, Gomez-Duran A, Schon K, Wei W, Karthikeyan S, et al. An atlas of mitochondrial DNA genotype–phenotype associations in the UK Biobank [Internet]. Nature Genetics. 2021. p. 982–93. Available from: http://dx.doi.org/10.1038/s41588-021-00868-1

44. Kang E, Wu J, Gutierrez NM, Koski A, Tippner-Hedges R, Agaronyan K, et al. Mitochondrial replacement in human oocytes carrying pathogenic mitochondrial DNA mutations. Nature. 2016;540:270–5.

45. Hyslop LA, Blakeley P, Craven L, Richardson J, Fogarty NME, Fragouli E, et al. Towards clinical application of pronuclear transfer to prevent mitochondrial DNA disease. Nature. 2016;534:383–6.

46. Gorman GS, McFarland R, Stewart J, Feeney C, Turnbull DM. Mitochondrial donation: from test tube to clinic. Lancet. 2018;392:1191–2.

47. Gollihue JL, Patel SP, Rabchevsky AG. Mitochondrial transplantation strategies as potential therapeutics for central nervous system trauma. Neural Regeneration Res. 2018;13:194–7.

48. Ali Pour P, Kenney MC, Kheradvar A. Bioenergetics Consequences of Mitochondrial Transplantation in Cardiomyocytes. J Am Heart Assoc. 2020;9:e014501.

49. Bacman SR, Williams SL, Pinto M, Peralta S, Moraes CT. Specific elimination of mutant mitochondrial genomes in patient-derived cells by mitoTALENs. Nat Med. 2013;19:1111–3.

50. Gammage PA, Rorbach J, Vincent AI, Rebar EJ, Minczuk M. Mitochondrially targeted ZFNs for selective degradation of pathogenic mitochondrial genomes bearing large-scale deletions or point mutations. EMBO Mol Med. 2014;6:458–66.

51. Bacman SR, Kauppila JHK, Pereira CV, Nissanka N, Miranda M, Pinto M, et al. MitoTALEN reduces mutant mtDNA load and restores tRNA levels in a mouse model of heteroplasmic mtDNA mutation. Nat Med. 2018;24:1696–700.

52. Gammage PA, Viscomi C, Simard M-L, Costa ASH, Gaude E, Powell CA, et al. Genome editing in mitochondria corrects a pathogenic mtDNA mutation in vivo. Nat Med. 2018;24:1691–5.

53. Loutre R, Heckel A-M, Jeandard D, Tarassov I, Entelis N. Anti-replicative recombinant 5S rRNA molecules can modulate the mtDNA heteroplasmy in a glucose-dependent manner. PLoS One. 2018;13:e0199258.

54. Yang J-N, Seluanov A, Gorbunova V. Mitochondrial inverted repeats strongly correlate with lifespan: mtDNA inversions and aging. PLoS One. 2013;8:e73318.

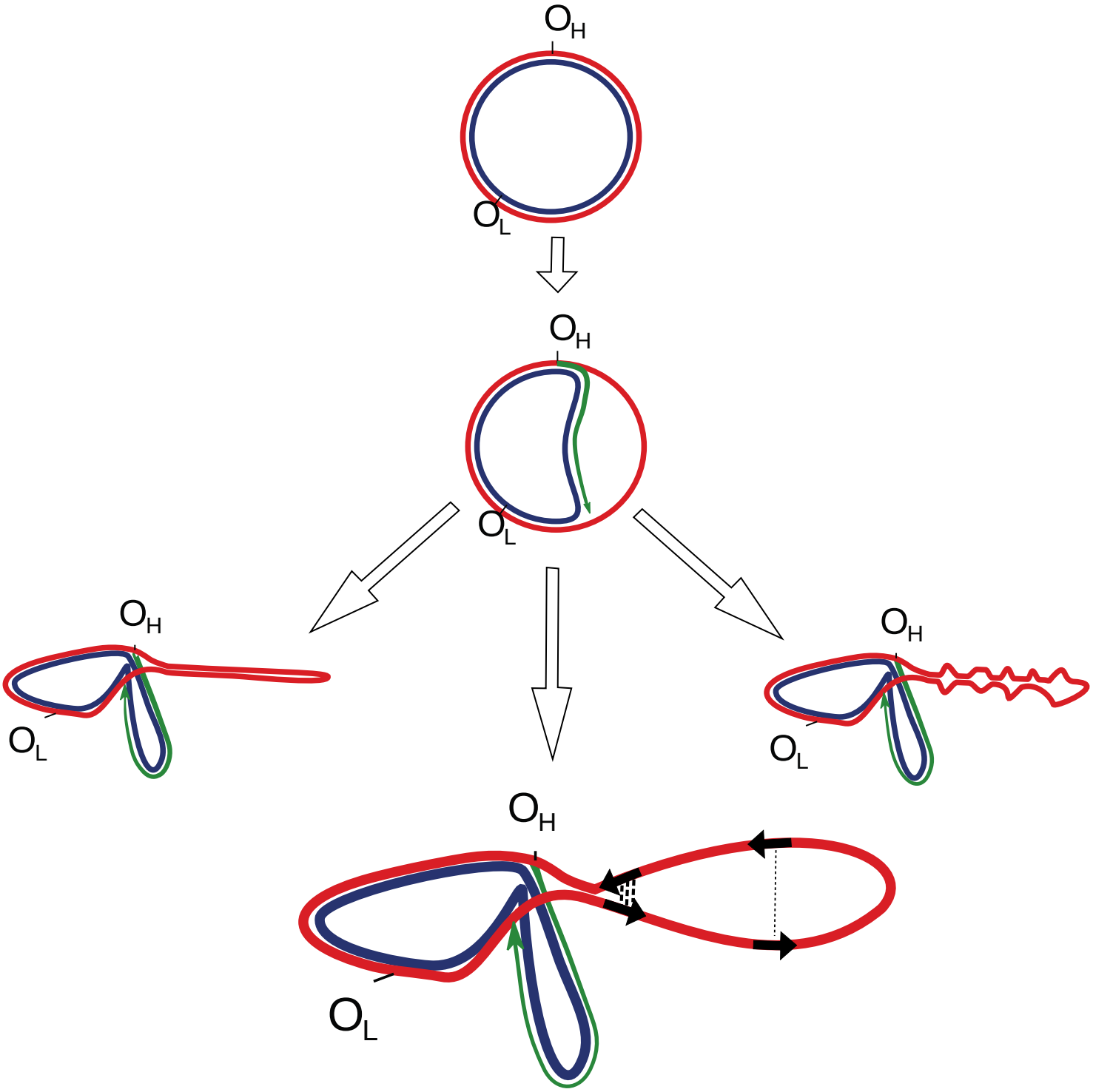
55. Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. Cell Syst. 2018;6:256–8.e1.

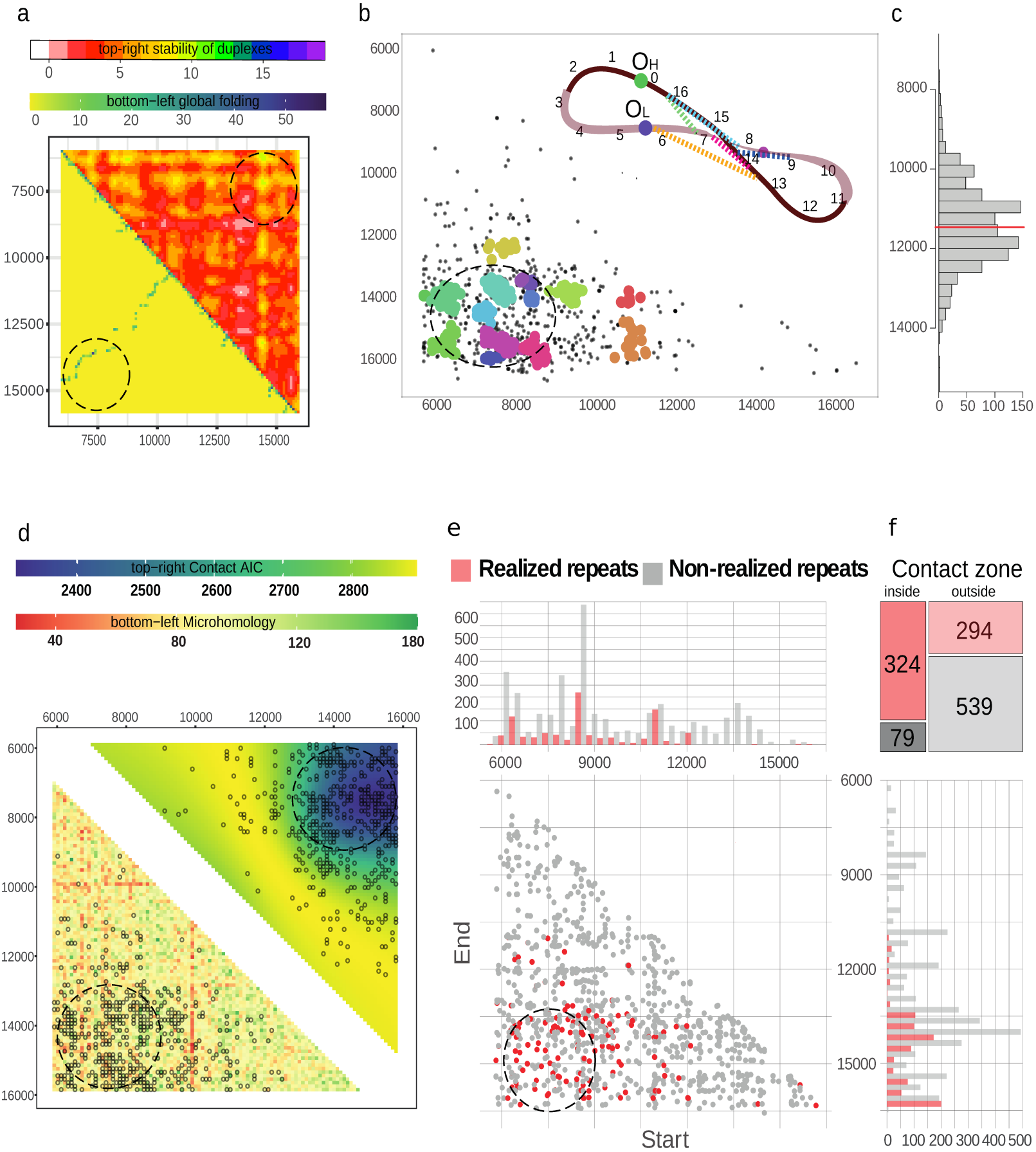
56. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31:3406–15.

57. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26.

58. Frey BJ, Dueck D. Clustering by passing messages between data points. Science. 2007;315:972–6.

59. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48:443–53.


O_H
O_L
O_H
O_L
O_H
O_L
O_H
O_L
O_H
O_L


a
top-right stability of duplexes
0 5 10 15
bottom−left global folding
0 10 20 30 40 50
7500 10000 12500 15000
b
6000 8000 10000 12000 14000 16000
O_H
O_L
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
c
8000 10000 12000 14000
0 50 100 150
d
top−right Contact AIC
2400 2500 2600 2700 2800
bottom−left Microhomology
40 80 120 180
6000 8000 10000 12000 14000 16000
e
Realized repeats
Non-realized repeats
600 500 400 300 200 100
6000 9000 12000 15000
End
Start
f
Contact zone
inside
outside
324
294
539
79
0 100 200 300 400 500