Update

Genome Analysis

Repeats, longevity and the sources of mtDNA deletions: evidence from ‘deletional spectra’

Xinhong Guo[1,7*], Konstantin Yu. Popadin[2*], Natalya Markuzon[3], Yuriy L. Orlov[4], Yevgenya Kraytsberg[1], Kim J. Krishnan[5], Gábor Zsurka[6], Douglas M. Turnbull[5], Wolfram S. Kunz[6] and Konstantin Khrapko[1]

[1] Harvard Medical School/Beth Israel Deaconess Medical Center, Boston, MA USA
[2] Russian Academy of Sciences Institute for Information Transmission Problems and Moscow State University, Moscow, Russia
[3] Draper Laboratory, Cambridge MA, USA
[4] Genome Institute of Singapore, Singapore
[5] The Institute of Ageing and Health and Newcastle University, Newcastle upon Tyne, UK
[6] Department of Epileptology and Life&Brain Center, University Bonn Medical Center, Bonn, Germany
[7] Hunan University, Changsha, Hunan, People’s Republic of China

Perfect direct repeats and, in particular, the prominent 13 bp repeat, are thought to cause mitochondrial DNA (mtDNA) deletions, which have been associated with the aging process. Accordingly, individuals lacking the 13 bp repeat are highly prevalent among centenarians and overall number of perfect repeats in mammalian mitochondrial genomes negatively correlates with species’ longevity. However, detailed examination of the distribution of mtDNA deletions challenges a special role of the 13 bp repeat in generating mtDNA deletions. Instead, deletions appear to depend on long and stable, albeit imperfect, duplexes between distant mtDNA segments. Furthermore, significant dissimilarities in breakpoint distributions suggest that multiple mechanisms are involved in creating mtDNA deletions.

Direct repeats in the mitochondrial genome and longevity

The premise that accumulation of mtDNA mutations [1] and, in particular, of large-scale deletions in mtDNA is one of the possible causes of aging has received substantial support from biochemical and longevity studies [2–5]. mtDNA deletions are usually flanked by direct repeats, implying that these repeats are involved in the generation of deletions. Recombination [6,7], slip-replication [8], and double-strand break repair [9] have been suggested as potential alternative mechanisms involving direct repeats. In corroboration of the connection between mtDNA deletions and aging, the number of direct repeats in mtDNA of various mammal species is inversely correlated with longevity [10,11]. Of particular interest is the so-called ‘common deletion’ [6], the deletion most frequently detected in humans, which is flanked by a prominent 13 bp perfect direct repeat. Interestingly, carriers of the well-studied D4a mitochondrial haplogroup, who are significantly enriched among Japanese centenarians [12], lack the 13 bp direct repeat in their mtDNA, and thus presumably lack the common deletion, and this seems to support the premise that deletions are involved in the aging process [13]. It should be noted, however, that although the ‘common’ deletion is the most abundant mtDNA deletion, it typically constitutes no more than 10% of all deletions in aging tissues [5]. Therefore D4a individuals would have at most 10% fewer deletions, which perhaps is too moderate a change to affect longevity. There is another possibility, however [13]. According to an elegant hypothesis of Samuels, Schon and Chinnery, the 13 bp repeat might be responsible for the formation of nearly all mtDNA deletions, not just the common deletion [14]. Thus, absence of this repeat could result in a reduction of overall deletion burden and, if deletions indeed are involved in the aging process, this might constitute a realistic cause of exceptional D4a longevity. The importance of this question prompted us to test the Samuels, Schon, Chinnery hypothesis.

mtDNA deletions, in general, are not related to the 13 bp repeat

The Samuels, Schon, Chinnery hypothesis rests on the observation that the distribution of deletion breakpoints across the mitochondrial genome consists of two broad peaks centered around the 5’ and 3’ arms of the 13 bp repeat – nucleotide positions 8469–8482 and 13447–13459 [14]. The hypothesis would therefore predict that loss of the 13 bp repeat would result in the disappearance of this characteristic pattern. To test this prediction, we compared the distribution of breakpoints in individuals with and without the 13 bp repeat. Using single molecule PCR [15] we determined breakpoints of 242 unique mtDNA deletions from the frontal cortex tissue of four individuals. mtDNAs of two of these individuals (controls) contain the common repeat, whereas the other two belong to the mtDNA haplogroup N1b in which the 5’ arm of the repeat is altered by the 8472C>T polymorphism, which is in a neighboring position to the D4a polymorphism (8473T>C) (Supplement 1 in the supplementary material online; for methods see Supplement 2). In accord with Samuels *et al.* [14], the distribution of breakpoints of these deletions in control individuals is bimodal with maxima roughly coincident with the arms of the repeat (Figure 1a and

Corresponding author: Khrapko, K. (kkhrapko@gmail.com)

* X.G. and K.P. contributed equally to this work.

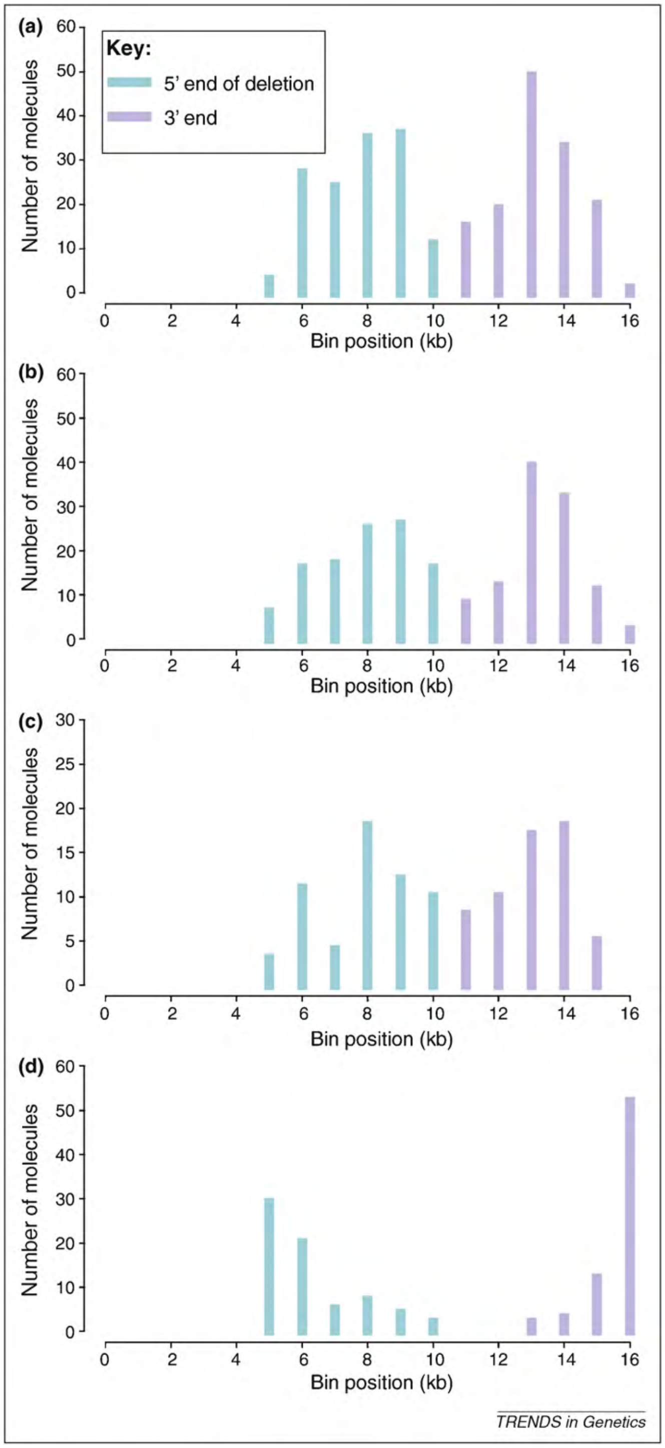


Figure 1. 'Deletional spectra' challenge the 'common' 13 bp repeat and perfect direct repeats in general as sources of mtDNA deletions. Deletional spectra were constructed by plotting the number of breakpoints discovered in each 1 kb interval of the mitochondrial genome ('bin') as a function of the bin position (see Supplement 3 for details). Breakpoint distributions in individuals with **(A)** or without **(B)** the 'common' 13 bp repeat are very similar, a finding which contradicts any special role of the 'common' repeat in generating most mtDNA deletions. **(C)** A subset of deletions without direct repeats at breakpoints (class II deletions, Ref. [7]) appear to be distributed similarly to class I deletions (with repeats), implying that the mechanism that shapes breakpoint distribution in cortex does not depend on the presence and/or absence of perfect repeats at breakpoints. **(D)** An example of strikingly dissimilar spectrum of deletions from epileptic hippocampi suggests that a distinct mechanism is causing mtDNA deletions in this tissue.

Supplement 3). Unexpectedly, the distribution of deletions in individuals without the 13 bp repeats (Figure 1b) is essentially the same as in controls. Thus, lack of the 13 bp repeat has no effect on the distribution of deletion breakpoints, in contradiction with the hypothesis prediction. Note that the 8472C>T polymorphism disrupts the common repeat, and significantly reduces occurrence of the common deletion. Indeed, we found only one 'pseudocommon' deletion (i.e. a deletion utilizing the remaining 10 bp repeat) in individuals without the 13 bp repeat compared to 12 common deletions found in control individuals. We conclude that there is no evidence that the common 13 bp repeat causes the formation of a significant proportion of mtDNA deletions.

Deletion breakpoints coincide with distant segments of mtDNA capable of forming stable imperfect duplexes with each other

What else (if not the common repeat) could have created the characteristic distribution of breakpoints with maxima around 8–9 kb and 13–14 kb (Figure 1a–c)? Basic local alignment search tool (BLAST) analysis of our breakpoint database revealed that many deletions are flanked by long (up to 50 nt) regions of partial or interrupted homology (data not shown). Regions of imprecise homology at breakpoints have been also noticed by others [7]. The formation of long imperfect duplexes between these regions of homology might have brought together distant segments of the mitochondrial genome, which is necessary for formation of deletions by any of the proposed mechanisms, i.e. slip replication, recombination, and double strand break repair [6–8]. We therefore explored *in silico* the distribution of possible secondary structures between 100 bp long segments of mtDNA from the region of 5' breakpoints (bp 5700–10737) and the region of the 3' breakpoints (bp 11400–16100) (Supplement 4). The resulting best-fit duplexes typically include several short double stranded regions separated by loops of various sizes (Figure 2a and Supplement 5). Stability of all such duplexes can be conveniently depicted by a 'matrix of free energies' (Figure 2b). Analysis of this matrix reveals a strong correlation between the stability of inter-segment duplexes and the distribution of mtDNA deletions (Figure 2b–e).

Perfect repeats or stable imperfect duplexes?

Because perfect repeats are widely considered to be the culprit of mtDNA deletions, we also performed regression analysis of the distribution of breakpoints versus the distribution of perfect direct repeats ≥5 bp (Supplement 7) and found a significant correlation between the probability of a deletion in a matrix element and the number of direct repeats in the same element. So what causes deletions: perfect repeats or long stable albeit imperfect duplexes? Precise sequence homology or duplex stability? Distinguishing between the two is difficult because long duplexes actually consist of appropriately arranged perfect repeats, although not every arrangement of short perfect repeats results in a long stable duplex. To sort out these possibilities, we used multiple logistic regression analysis (Supplement 7) of the probability to find at least one deletion within an element of the matrix of free energies

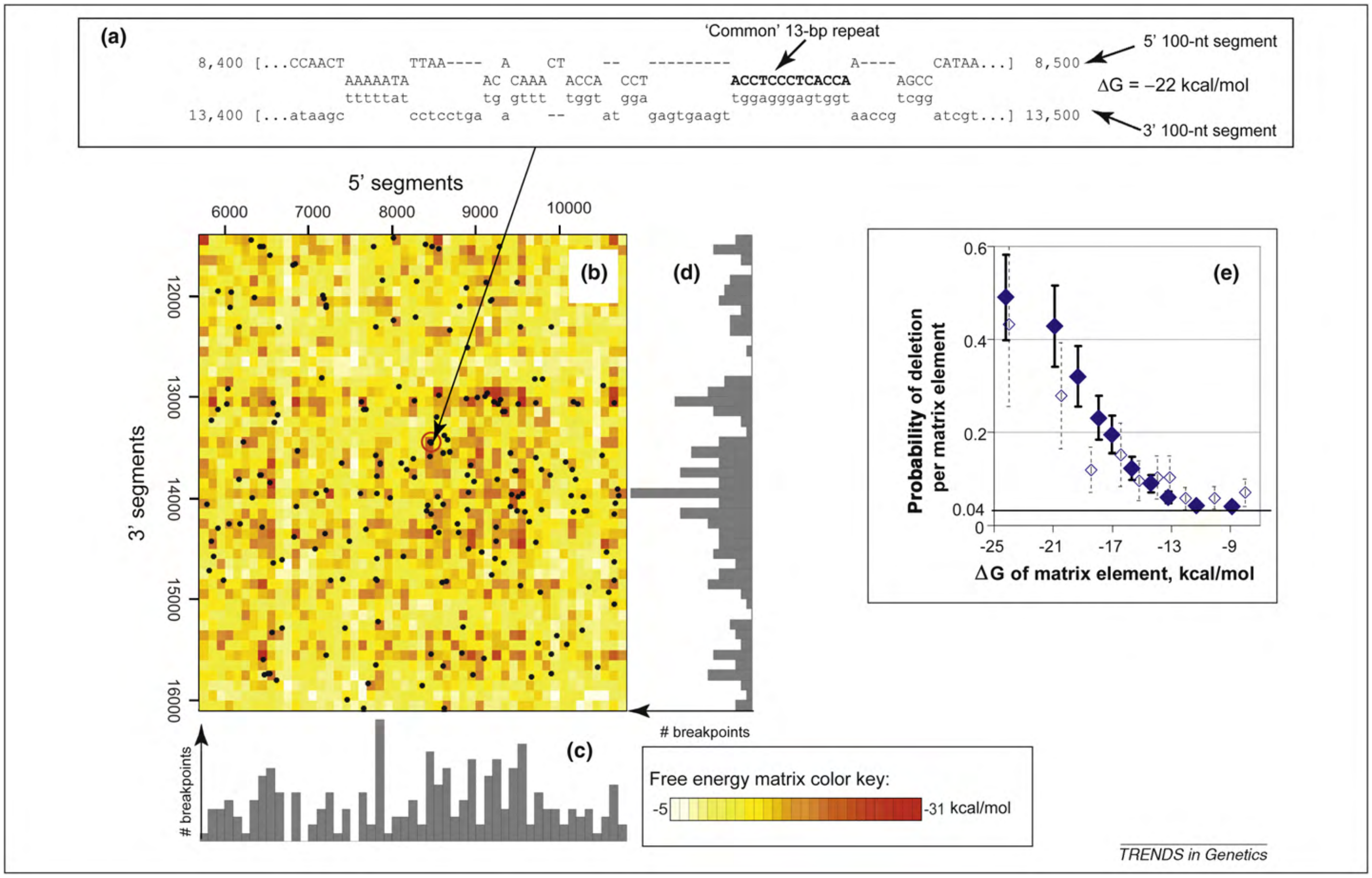


Figure 2. Breakpoints of mtDNA deletions coincide with positions of potential long and stable imperfect duplexes between distant segments of the mitochondrial genome. **(A)** An example of a long imperfect duplex between two distant 100 bp long segments of mtDNA discovered by an *in silic*o search (see also Supplement 4). For demonstration purposes we selected the duplex between the segments containing the arms of the common 13 bp repeat. The 5′ and 3′ 100 nt segments of mtDNA that participate in the formation of this duplex and the corresponding free energy ΔG are indicated. **(B)** The distribution and stabilities of all possible intersegment duplexes is depicted by the matrix of free energies (see also Supplement 4). The color of each matrix element represents the free energy (ΔG) of the best possible DNA duplex between the corresponding 100 bp 5′-segment and the 100 bp 3′-segment (the lower the ΔG, the higher the stability, the more intense the color; see color key). Black dots depict the breakpoint positions of the real cortical deletions which apparently correlate with dark-colored matrix elements: deletions 'prefer' to form between segments that have the potential to form a stable duplex with each other. Similarly, maxima or minima of the distributions of 5′ **(C)** and 3′ **(D)** deletion breakpoints appear to correlate with dark or pale 'stripes' that run vertically or horizontally across the matrix, correspondingly. In other words, the probability that a deletion falls within a certain matrix element increases rapidly and significantly ($P = {\sim}2 \times 10^{-16}$) with increasing duplex stability (decreasing ΔG), as quantitatively shown **(E)** (filled diamonds; see also Supplement 6). This finding strongly implicates long imperfect duplexes in mtDNA deletion formation. Interestingly, deletions without repeats at breakpoints are also much more likely to fall within stable matrix elements (empty diamonds), implying that short perfect repeats *per se* do not significantly affect the generation of deletions.

(Figure 2b) as a function of two variables: the ΔG (the free energy of the best duplex) and the number of perfect repeats within the same element. It appears that the contribution of the ΔG of the best imperfect duplex remains as highly significant as in the single-variable analysis (*P* values $\sim 10^{-16}$), whereas the contribution of the number of perfect repeats completely loses its significance ($P = 0.2$) in the presence of ΔG as a variable. We therefore conclude that mtDNA deletion formation depends primarily on the availability of stable imperfect duplexes rather than on the presence of perfect repeats.

The above conclusion, however, seems to contradict an observation that almost 75% of deletions in the dataset, many more than expected by chance, contain perfect repeats exactly at breakpoints. To reconcile these observations, we note that there is no shortage of perfect repeats in mtDNA: an average 100 x 100 matrix element contains about 70 direct repeats ≥ 5 bp (Supplement 8). Therefore it is tempting to speculate that although deletions might form preferentially at perfect repeats, the vast abundance of such repeats renders this requirement non-limiting. Instead, deletions probably depend on the formation of long and stable, albeit imperfect, duplexes, which are probably needed to hold distant mtDNA segments together. Once such a duplex has formed, a deletion most likely (but not necessarily) will be created precisely at one of the many perfect repeats available nearby. This explains both the presence of perfect repeats at breakpoints and the weakness of the correlation between the distributions of breakpoints and perfect repeats. In support of this view, deletions without perfect repeats at breakpoints are distributed similarly to deletions with repeats (Figure 1a,b versus c; also Ref. [14]) and, just like deletions with repeats at breakpoints, deletions without repeats are much more likely to form in elements with stable intersegment duplexes than elsewhere in the genome (Figure 2e).

Other factors affecting the formation of mtDNA deletions: multiple mechanisms?

It is worth noting, however, that the proposed relationship between the stability of intersegment duplexes and the distribution of deletion breakpoints in mtDNA from normal cortex is by no means all-encompassing. Other tissues or individuals might present with highly dissimilar deletional spectra. A remarkable example is given by the distribution of mtDNA deletion breakpoints in the hippocampi of patients with temporal lobe epilepsy and Ammon's horn sclerosis (Figure 1d). These hippocampi contain over 10-fold more mtDNA deletions than normal hippocampus, and this could be related to increased oxidative stress [16]. The significant differences in the distribution of deletion breakpoints to the normal frontal cortex pattern are particularly evident at the 3'-end because there is a large excess of breakpoints at the long-known breakpoint hot spot at position 16 070 [17]. This deletion hot spot is suggested to be generated by double-strand breaks [18] that can be formed by attacks of reactive oxygen species [19]. It is tempting to speculate that differences in the distribution of breakpoints indicate that different mechanisms can direct the generation of mtDNA deletions.

Concluding remarks

The formation of deletions in the mitochondrial genome appears to be strongly dependent on stable secondary structures that might potentially form between distant segments of the genome, rather than by relatively short perfect direct repeats as is widely assumed. This observation will certainly affect thinking related to mechanisms of deletion formation. Furthermore, previously discovered associations between the presence or absence of the common repeat or other perfect direct repeats in mtDNA and longevity [10,11,13] might need to be revisited. Methodologically, this work demonstrates the power of analysis of deletion breakpoint distributions – 'deletional spectra'. Spectral dissimilarities between different tissues imply the existence of multiple factors that can cause mtDNA deletions and/or shape their distribution. This opens up promising research avenues where one could explore mechanisms of deletion formation by comparing deletional spectra created by various potential sources such as oxidative stress, double-strand breaks and other types of DNA damage, or by defective mtDNA maintenance proteins, to delineate the various factors responsible for the accumulation of mtDNA deletions in aging and disease.

Acknowledgements

K.K. has been supported in part by National Institutes of Health grants AG019787 and NS058988, X.G. by a visiting scholar fellowship from the China Scholarship Council Program, W.S.K. by Deutsche Forschungsgemeinschaft grant SFB TR3 A11, D.M.T by the Newcastle University Centre for Brain Ageing and Vitality supported by the Biotechnology and Biological Sciences Research Council, the Engineering and Physical Sciences Research Council, the Economic and Social Research Council, and the Medical Research Council as part of the cross-council Lifelong Health and Wellbeing Initiative, and K.P. by the Russian Academy of Sciences Molecular and Cellular Biology Program, Russian Foundation for Basic Research Grants 08-04-01651- and 10-04-00276-, Russian Ministry of Science and Education (P2442), Russian Science Support Foundation, and by a Grant of the President of the Russian Federation (MK-5423.2008.4).

Appendix A. Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2010.05.006.

References

1 Linnane, A.W. *et al.* (1989) Mitochondrial DNA mutations as an important contributor to ageing and degenerative diseases. *Lancet* 1, 642–645

2 Wanagat, J. *et al.* (2001) Mitochondrial DNA deletion mutations colocalize with segmental electron transport system abnormalities, muscle fiber atrophy, fiber splitting, and oxidative damage in sarcopenia. *FASEB J.* 15, 322–332

3 Herbst, A. *et al.* (2007) Accumulation of mitochondrial DNA deletion mutations in aged muscle fibers: evidence for a causal role in muscle fiber loss. *J. Gerontol. A Biol. Sci. Med. Sci.* 62, 235–245

4 Bender, A. *et al.* (2006) High levels of mitochondrial DNA deletions in substantia nigra neurons in aging and Parkinson disease. *Nat. Genet.* 38, 515–517

5 Kraytsberg, Y. *et al.* (2006) Mitochondrial DNA deletions are abundant and cause functional impairment in aged human substantia nigra neurons. *Nat. Genet.* 38, 518–520

6 Schon, E.A. *et al.* (1989) A direct repeat is a hotspot for large-scale deletion of human mitochondrial DNA. *Science* 244, 346–349

7 Mita, S. *et al.* (1990) Recombination via flanking direct repeats is a major cause of large-scale deletions of human mitochondrial DNA. *Nucl. Acids Res.* 18, 561–567

8 Shoffner, J.M. *et al.* (1989) Spontaneous Kearns–Sayre/chronic external ophthalmoplegia plus syndrome associated with a mitochondrial DNA deletion: a slip-replication model and metabolic therapy. *Proc. Natl. Acad. Sci. U. S. A.* 86, 7952–7956

9 Krishnan, K.J. *et al.* (2008) What causes mitochondrial DNA deletions in human cells? *Nat. Genet.* 40, 275–279

10 Samuels, D.C. (2004) Mitochondrial DNA repeats constrain the life span of mammals. *Trends Genet.* 20, 226–229

11 Khaidakov, M. *et al.* (2006) Direct repeats in mitochondrial DNA and mammalian lifespan. *Mech. Ageing Dev.* 127, 808–812

12 Bilal, E. *et al.* (2008) Mitochondrial DNA haplogroup D4a is a marker for extreme longevity in Japan. *PLoS One* 3, e2421

13 Popadin, K. and Bazykin, G. (2009) Nucleotide repeats in mitochondrial genome determine human lifespan. *Nat. Prec.*, http://precedings.nature.com/documents/2399/version/2391

14 Samuels, D.C. *et al.* (2004) Two direct repeats cause most human mtDNA deletions. *Trends Genet.* 20, 393–398

15 Kraytsberg, Y. and Khrapko, K. (2005) Single-molecule PCR: an artifact-free PCR approach for the analysis of somatic mutations. *Expert Rev. Mol. Diagn.* 5, 809–815

16 Kudin, A.P. *et al.* (2009) Mitochondrial involvement in temporal lobe epilepsy. *Exp. Neurol.* 218, 326–332

17 Zeviani, M. *et al.* (1989) An autosomal dominant disorder with multiple deletions of mitochondrial DNA starting at the D-loop region. *Nature* 339, 309–311

18 Srivastava, S. and Moraes, C.T. (2005) Double-strand breaks of mouse muscle mtDNA promote large deletions similar to multiple mtDNA deletions in humans. *Hum. Mol. Genet.* 14, 893–902

19 Imlay, J.A. and Linn, S. (1988) DNA damage and oxygen radical toxicity. *Science* 240, 1302–1309

doi:10.1016/j.tig.2010.05.006 Trends in Genetics xx (2010) 1–4

ON-LINE SUPPLEMENTARY MATERIAL

Repeats, longevity and the sources of mtDNA deletions: evidence from "deletional spectra".

Xinhong Guo*[1,7], Konstantin Yu. Popadin*[2], Natalya. Markuzon[3], Yuriy L. Orlov[4] Yevgenya Kraytsberg[1], Kim. J. Krishnan[5], Gabor. Zsurka[6], Douglas. M. Turnbull[5], Wolfram S. Kunz[6], and Konstantin Khrapko[1].

1 Harvard Medical School/Beth Israel Deaconess Medical Center (BIDMC), Boston, MA USA
2 Russian Academy of Sciences Institute for Information Transmission Problems and Moscow State University, Moscow, Russia
3 Draper Laboratory, Cambridge MA, USA
4Genome Institute of Singapore, Singapore
5 The Institute of Ageing and Health and Newcastle University, Newcastle upon Tyne, UK
6 Department of Epileptology and Life&Brain Center, University Bonn Medical Center, Bonn, Germany
7 Hunan University, Changsha, Hunan, People's Republic of China

*XG and KP equally contributed to the study
Corresponding author: Khrapko, K (kkhrapko@gmail.com)

Supplement 1.

Tissues.

Frontal cortex tissue. We have screened a large number of archived fresh frozen autopsy tissue samples for polymorphisms in the 13-bp repeat (8469-8482 and 13447-13459) of the mitochondrial genome. We were able to locate two individuals with the 8472C>T polymorphism (frontal cortex, 86 years old female and 91 year old male). We also selected two age- and tissue-matched control samples (frontal cortex, 86 years old male and 82 year old male). All samples were originally provided by the Massachusetts Alzheimer's Tissue and Resource Center at Massachusetts General Hospital, Charlestown, MA. All studies were carried in accordance with IRB of each respective institution.

Hippocampus tissue from epilepsy subjects. Fresh hippocampal tissues from two male patients (ages 37 and 34 years old) with temporal lobe epilepsy (TLE) and Ammon's horn sclerosis were obtained after epilepsy surgery according to the procedure described in (Vielhaber et al. 2008). Different regions of the hippocampus were dissected from 400 µm vibratome slices, snap frozen and stored in liquid nitrogen for further processing. The research protocol was approved by the Ethical Committee of the University Bonn. Area CA3 was used for analysis.

Supplement 2.

Identification of mtDNA deletions their breakpoints.

General description of the approach: single molecule PCR (smPCR) First, wild-type mtDNA molecules were rendered non-amplifiable by restriction digestion with SnaBI, an enzyme with a single recognition site in the middle of the major arch (bp 10737) then DNA was diluted and subjected to long range single molecule PCR (Kraytsberg 05). Extent of dilution was empirically adjusted so that a typical PCR reaction contained one amplifiable molecule (hence the name "single molecule PCR"), which in most cases was a deletion, because wild type molecules were removed by restriction digestion. Each amplified molecule was then mapped by an appropriate set of restriction enzymes which determined the approximate position of the deletion breakpoint (within ~700bp) and an appropriate sequencing primer was chosen. Then PCR fragment was sequenced across the breakpoint. The resulting sequences were BLASTed against the mtDNA sequence, which identified the position of the deletion breakpoint and also any nucleotide homologies present in the vicinity of the breakpoint

Details of the procedure:

Total genomic DNA was restricted by SnaBI and Nae I restriction enzymes with single cuts in the human mtDNA at nucleotide positions 10,737 and 933, correspondingly. Our previous research showed that a great majority of mtDNA deletions remove the SnaBI site, which means that our assay was capable of detecting most of mtDNA deletions. Note that we do not claim to have detected all possible deletions, and that detecting all deletions is not essential for the conclusions that we draw from our data. Nae I cut was made to avoid potential artifacts (see below). PCR conditions were: primers 2999F and 16450R, 45 cycles (20 sec 94C, 13 min 68C). LaTaq thermostable DNA polymerase from TaKaRa was used according to manufacturer's recommendations. PCR was performed in 96-well PCR plates, PCR products were visualized using agarose gel electrophoresis; products shorter than 13 kb (which is the length of the wild type PCR product) were considered deletions and were mapped by restriction enzymes (XhoI, BamHI, DraI and XcmI) which produced a restriction fragment pattern allowing the position of the 3' breakpoint to be determined with sufficient precision (the first 3'-most restriction fragment lacking from the restriction ladder is the one including the breakpoint). A set of appropriate sequencing primers was used, covering the region from approximately bp 11400 to bp 16400. In other words, positions of the 3' breakpoints were determined if and only if they fell anywhere between bp11400 and bp16400. With respect to 5' breakpoints, these were detected as they were attached to the 3'-breakpoints. This means that there were no restrictions on the detectable positions of 5'-breakpoints other than the position of the forward PCR primer 2,999 and the SnaBI restriction site 10,737 from the 5' and 3' sides, correspondingly. Standard BLAST analysis (blastn at http://blast.ncbi.nlm.nih.gov/Blast.cgi) has been used to align sequences to the human mtDNA reference sequence NC_012920, search parameters were as listed below ("somewhat similar" setting):

Program: blastn Word size 16, Expect value 10, Hitlist size 100. Match/Mismatch scores 2,-3, Gapcosts 5,2.

Discontinuity in the alignment indicated the breakpoint, and overlap of the 5' and 3' portions of the alignment defined the homology at the breakpoints. If no overlap was discovered by the BLAST procedure, the deletion was considered not carrying a repeat at breakpoints .

Control for artifacts.

Any PCR-based procedure is potentially subject to PCR artifacts. Artificial deletions may be conceivably formed via mispriming by the 3' end of one arm of a broken mtDNA molecule (or the corresponding prematurely terminated PCR product) at a distant site on the other arm of the same molecule or on a different molecule/PCR fragment. As discussed elsewhere (Kraytsberg & Khrapko, 2005), smPCR is a procedure designed to minimize these potential artifacts. Extreme dilution of smPCR reactions makes intermolecular interactions very unlikely if at all possible, and intramolecular process (priming of one arm of a native DNA molercule on the other arm of the same molecule) is prevented by the additional NaeI cut, which separates any broken wild type mtDNA molecule into two, thus making it impossible for the two arms to meet. Mispriming events may become quite frequent at later PCR cycles, when higher concentration of PCR fragments make intermolecular interactions much more likely. These later mispriming events, though they certainly do result in artificially deleted PCR products, do not interfere with our analysis because these later generated molecules are present at low fractions, and are disregarded because in smPCR we pay attention only to the main product (this issue is extensively discussed in (Kraytsberg & Khrapko, 2005).

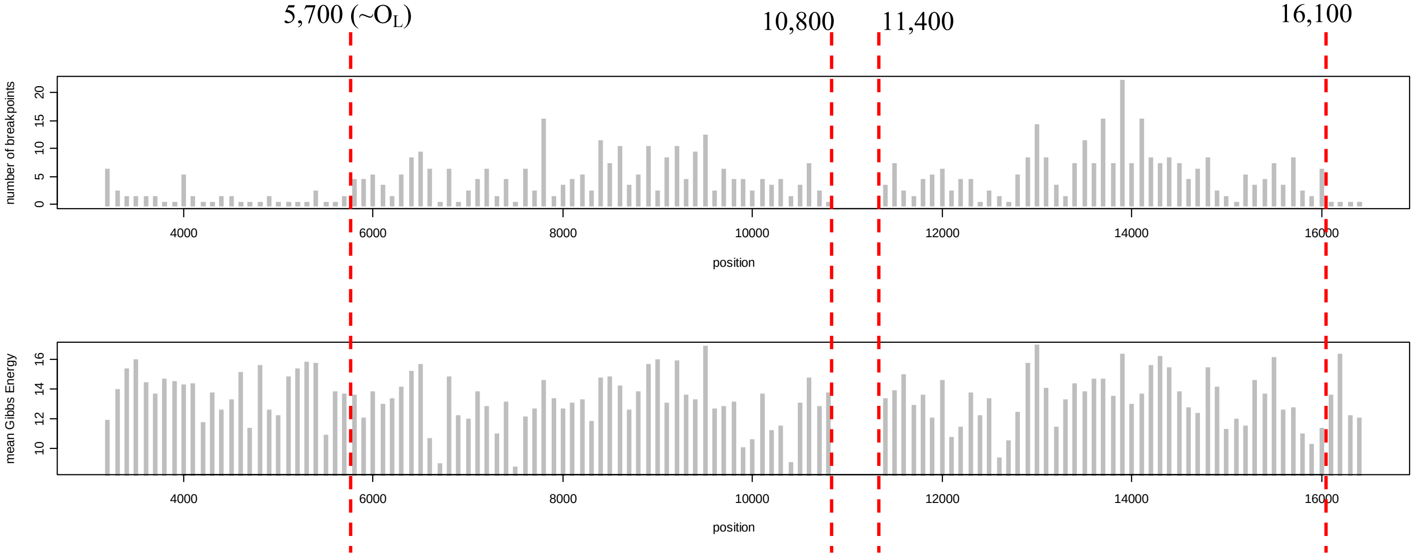
The theoretical reasons presented above are confirmed by the hard data. Note that our detection method contains an internal control: our deletions are distributed extremely non-uniformly: there is a sharp decrease in the number of deletions at light strand origin of replication and at around bp 16,100 (see Supplementary Figure 1 upper panel, below). The frequency decrease is even more dramatic for the epileptic hippocampus (data not shown), where very high frequency of breakpoints at site 16070 and around 5,800 abruptly turn into zero outside these limits. At the same time there is no decrease in the availability of potential mispriming sites, as distribution of stable duplexes shows no drop-off at these positions (Supplementary Figure 1, lower panel). These limits are thus of purely biological nature (see Supplement 3 for further discussion of this issue) and the fact that we see them with such clarity implies that artificial deletions do not constitute any significant fraction of deletions in our database.

Supplement 3.

Construction of "deletional spectra".

Following the example of Samuels et al. 2004, we constructed "deletional spectra" (i.e. frequency distributions of breakpoints along the mitochondrial genome) by binning the breakpoints into 1-kb intervals along the mtDNA sequence. The numbers of breakpoints falling within each interval was counted and plotted (Fig 1). In Figure 2, panels B and C, breakpoints were binned into 100-bp intervals to obtain higher-resolution "spectra". Note that each deletion of a particular sequence was counted once. For example, there were 12 common deletions in the control cortex dataset, but all these deletions contributed one 5' and one 3' breakpoint to the distribution. On the other hand, if two different deletions shared the same breakpoint at one end (while the other 5' or 3' breakpoints were different), then the shared breakpoint was counted twice. This was done to make our analysis comparable with earlier study (Samuels et al.), who used this same approach. Furthermore, counting deletions only once helped to diminish the possible contribution of clonal expansion of deletions, which was not a focus of this study. This measure also decreased the influence of sampling error resulting from random clonal expansions in our tissue samples (i.e. a deletion might have achieved prominence not because it is generated more frequently but just because clonal expansion of this deletion happened by chance). Note that removal of replicate deletions decreased the our data set only by ~20% (299 to 242).

Note that deletional spectra in Figures 1 and 2 are spatially truncated. Our approach was designed to detect deletions with breakpoints anywhere between bp 2,999 and 10,737 (5' breakpoints) and between bp 11,400 and 16,400 (3' breakpoints) (see Supplement 2). However, we chose to restrict deletion analysis to two smaller regions, i.e. bp 5,700-10,800 for 5'- and bp 11,400-16,100 for 3'-breakpoints presented in Figure 1 and Figure 2,` panels B and C. This is because it has been known for some time that frequency of mtDNA deletional breakpoints drop sharply at the boundaries of the major arch of the mitochondrial genome, i.e. at light strand origin of replication at about bp 5,700 and at the other side at about position bp 16,100. Sharp decreases of breakpoints at these boundaries apparently are related to certain biological restrictions on mtDNA deletions. For example, lack of the light strand origin of replication may impede the ability of the deleted molecule to propagate itself. Our data clearly shows these declines in frequency distribution at the borders (**Supplementary Figure** below). As long as our goal was to evaluate certain models of deletion formation based merely on the properties of DNA sequence, we reasoned that it is important to exclude other possible factors (such as influence of an origin of replication), which could obscure the correlations we were studying. As we make clear in the paper, we fully adhere to the idea that there are several independent factors affecting the distribution of deletions. To be able to study each of these factors it is important to segregate them from each other. Had we not removed these parts of sequence from our analysis, we would not be able to observe the correlations reported in this study with sufficient clarity. We also point out that the majority of mtDNA deletions are located within the regions of mitochondrial genome assessed in this study, while regions excluded from our study contain almost no deletions, so the above restrictions almost do not limit the explanatory power of the models we discuss as far as a majority of mtDNA deletions is concerned.



Supplementary Figure. Reduction in the frequency of deletion breakpoints beyond the major arch boundaries. Upper panel: the distribution of deletions, lower panel: stability of long duplexes. The whole distribution spans the area where deletions could be detected with our experimental settings, red lines delineate the area which were used for all analyses. Note that although there is no decrease in the availability of stable long duplexes at the borders (nucleotide positions 5,700 and 16,100), the frequency of breakpoints decreases sharply.

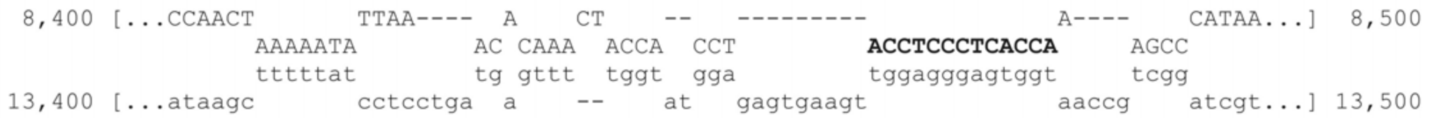
Supplement 4.

Search for best segment to segment duplexes and the construction of the free energy matrix.

100 bp segments (i.e., 5,700-5,800, 5,800-5,900, … etc.) were extracted from truncated major arc `bp 5,700-10,800 and bp 11,400-16,100` of Revised Cambridge Reference Sequence (Andrews et al 1999) using in house PERL script. All combinations of pairs of segment to segment sequences (one from the 5' region and the other from the 3' region) from major arc were obtained. To predict secondary structure and minimum free energy of these hybridized 100 bp DNA sequences we used hybrid-min operator in UNAFold 3.7 (Markham & Zuker 2008) package under the default parameters (t = 37°C etc). Each segment-to-segment hybridization have two potential structures – (i) hybridization of 5'-3' L strand with 3'-5' H strand and (ii) 3'-5' L strand with 5'-3' H strand, and the two resulting duplexes may have significantly different stabilities. The structure with minimal free energy among the two possible was selected.

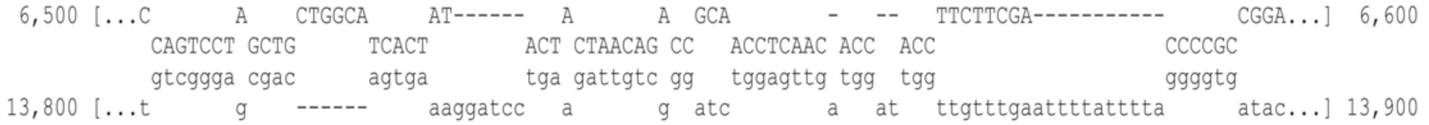
Supplement 5.

Examples of long imperfect duplex structures.



```
 8,400 [...CCAACT       TTAA----  A    CT    --   ---------             A----    CATAA...]   8,500
                 AAAAATA        AC CAAA  ACCA  CCT         ACCTCCCTCACCA     AGCC
                 tttttat        tg gttt  tggt  gga         tggagggagtggt     tcgg
13,400 [...ataagc       cctcctga  a    --    at   gagtgaagt             aaccg    atcgt...]  13,500
```

Long duplex of the "common" 13 bp repeat matrix element (8400, 13400). ΔG= -22 kcal/mol, which is top 2nd percentile of all matrix elements by free energy. The common repeat itself is included in the duplex (highlighted bold in the light (upper) strand) and is actually responsible for significant fraction of long duplex stability (ΔG of the common repeat duplex alone is ~-16 kcal/mol).



```
 6,500 [...C       A    CTGGCA     AT------   A       A  GCA        -   --   TTCTTCGA----------       CGGA...]  6,600
            CAGTCCT GCTG      TCACT        ACT CTAACAG CC   ACCTCAAC ACC  ACC                   CCCCGC
            gtcggga cgac      agtga        tga gattgtc gg   tggagttg tgg  tgg                   ggggtg
13,800 [...t       g    ------     aaggatcc   a       g  atc        a   at   ttgtttgaattttatttta      atac...] 13,900
```

Element (6500,13800) ΔG= -28 kcal/mol, of the most stable of all matrix elements by free energy. This matrix element actually does contain a deletion, so this long duplex is "functional". This structure illustrates the fact that a functional and very stable long duplex may be composed of many approximately equally stable short stem-loop structures, without resorting to longer perfect duplexes (as in the case of the matrix element containing common deletion).

Supplement 6.

The probability of a deletion to be present in a matrix element as a function of free energy of the element (Figure 2D).

Perhaps the most convincing feature of the distribution of deletions (black dots) over the matrix of free energies is the visual correlation of deletions with the more stable (higher color intensity) matrix elements. Quantitatively, these correlations can be accounted for by the probability of a deletion to form between two particular segments of mtDNA (i.e. to fall within a certain matrix element) as a function of the stability of the best duplex between these segments (i.e. ΔG of this element). To quantify this apparent trait we ranged deletions by the ΔG value of the matrix elements to which they belonged and then binned ranged deletions (242) into 10 groups 24 or 25 elements in each group. Binning into groups with almost equal number of deletions ensured that each data point was estimated with equal confidence. For each group, the span of ΔG was determined, and the probability of deletion in an element within this span was estimated by dividing number of elements in the group (24) by total number of matrix elements falling within the same energy interval. This probability was plotted as a function of the average element ΔG within each group. Error bars represent the standard deviation of the count of 24 assuming Poisson distribution of the probability values within each group (i.e. approximately square root of 24), solely for illustrative purposes. Standard deviation is appropriately adjusted to fit the probability per element rather than the actual count. We then repeated exactly the same calculations exclusively for deletions without repeats at breakpoints (i.e. with 2 or less repeated nucleotides). There were much fewer such “repeatless” deletions (total of 66), so we binned them into 10 groups of 7 or 6 each. The results were plotted in the same way as for all deletions (unfilled diamonds in Figure 2D). Note that in this graph, probabilities of “repeatless” deletions were appropriately adjusted (i.e. multiplied by 242/66) to permit direct comparison.

Statistical significance of these correlations (P=2×10^{-16} and 0.005, for all deletions and for deletions without repeats, respectively) was estimated using logistic regression as described below.

For logistic regression analysis, we considered all 2397 elements of matrix with known free energy of duplexes as well as with known distribution of deletions (195 elements with at least one deletion and 2202 - without). Logistic regression revealed significant association between presence of deletion and high free energy of duplexes:

Z = -5.39872 -0.20146*Free energy of the element, P value $<2\times10^{-16}$.

For “repeatless” deletions analysis, all cells containing at least one deletion flanked by 3 or more repeated nucleotides, were eliminated from analysis. Totally 66 cells that contained “repeatless” deletions, as well as 2268 cells without deletions were analyzed. Logistic regression revealed significantly higher probability to observe ”repeatles” deletions in matrix element with high free energy:

Z = -4.74542 -0.08936*Free energy of the element, P value = 0.005.

(note that less prominent P value primarily reflects much lower number of data points rather than inherent inconsistency of data) So, even deletions not flanked by any substantial repeats depend on free energy of long imperfect duplexes.

Supplement 7.

Regression analysis.

Regression statistical analyses were done in R language (R Development Core Team (2009)). Logistic regression was performed using glm function (Generalized Linear Models) with settings “family = binomial()”.

Supplement 8.

Estimation of the number of direct repeats.

To estimate the number of repeats in mitochondrial genome we have developed an in-house computer program to map all perfect repeats greater than a fixed length starting at each position of genome sequence. The program used suffix tree presentation of all short words in sequence under analysis (i.e. mt genome). For each word suffix tree presentation keep positions of the word in genome sequence. Such initial presentation of short words allow fast search of longer repeated elements starting from these words. For fixed minimal repeat length program run time increases linearly with genome size. Repeats could be perfect (no mismatches) or with limited number of mismatches as defined by user. Maximal sequence size to be tested is up to 10Mb. The program outputs non-redundant list of direct and inverted (these were excluded from our present analysis) repeats (5' and 3' fragments) together with positional information in the genome. Maximal length of repeats is not limited. To avoid redundancy for each position we counted only repeats of maximal length. The program is available upon request from the authors (written in C++, has been tested for Windows and Unix).

References:

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet. 1999;23:147

Kraytsberg, Y., and Khrapko, K. (2005) Single-molecule PCR: an artifact-free PCR approach for the analysis of somatic mutations. Expert Rev Mol Diagn 5, 809-815

Markham, N. R. and Zuker, M. (2008) UNAFold: software for nucleic acid folding and hybriziation. In Keith, J. M., editor, *Bioinformatics, Volume II. Structure, Function and Applications*, number 453 in *Methods in Molecular Biology*, chapter 1, pages 3–31. Humana Press, Totowa, NJ. ISBN 978-1-60327-428-9.

Samuels, D. C., Schon, E. A., and Chinnery, P. F. (2004). Two direct repeats cause most human mtDNA deletions. Trends Genet *20*, 393-398.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Vielhaber S, Niessen HG, Debska-Vielhaber G, Kudin AP, Wellmer J, Kaufmann J, Schönfeld MA, Fendrich R, Willker W, Leibfritz D, Schramm J, Elger CE, Heinze HJ, Kunz WS. Subfield-specific loss of hippocampal N-acetyl aspartate in temporal lobe epilepsy. Epilepsia. 2008 Jan;49(1):40-50.