Bioinformatics, 34(7), 2018, 1081–1085
doi: 10.1093/bioinformatics/btx729
Advance Access Publication Date: 8 November 2017
Discovery Note

OXFORD

Genome analysis

Complex analyses of inverted repeats in mitochondrial genomes revealed their importance and variability

Jana Čechová[1], Jiří Lýsek[2], Martin Bartas[1,3] and Václav Brázda[1,*]

[1]Department of Biophysical Chemistry and Molecular Oncology, Institute of Biophysics, The Czech Academy of Sciences, 612 65 Brno, Czech Republic, [2]Department of Informatics, Mendel University in Brno, 613 00 Brno, Czech Republic and [3]Department of Biology and Ecology/Institute of Environmental Technologies, Faculty of Science, University of Ostrava, 701 03 Ostrava, Czech Republic

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on August 4, 2017; revised on October 31, 2017; editorial decision on November 3, 2017; accepted on November 7, 2017

Abstract

Motivation: The NCBI database contains mitochondrial DNA (mtDNA) genomes from numerous species. We investigated the presence and locations of inverted repeat sequences (IRs) in these mtDNA sequences, which are known to be important for regulating nuclear genomes.
Results: IRs were identified in mtDNA in all species. IR lengths and frequencies correlate with evolutionary age and the greatest variability was detected in subgroups of plants and fungi and the lowest variability in mammals. IR presence is non-random and evolutionary favoured. The frequency of IRs generally decreased with IR length, but not for IRs 24 or 30 bp long, which are 1.5 times more abundant. IRs are enriched in sequences from the replication origin, followed by D-loop, stem-loop and miscellaneous sequences, pointing to the importance of IRs in regulatory regions of mitochondrial DNA.
Availability and implementation: Data were produced using Palindrome analyser, freely available on the web at http://bioinformatics.ibp.cz.
Contact: vaclav@ibp.cz
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Although most DNA of eukaryotic organisms is localized in chromosomes within the nucleus, mitochondrial DNA (mtDNA) is a very important part of the vast majority of eukaryotes. Mitochondria are double membrane-bound subcellular organelles which play a central role in metabolism (Brand, 1997), apoptosis (Kroemer *et al.*, 1998) and ageing (Kauppila *et al.*, 2017; Wei *et al.*, 2001). Moreover, defective mitochondrial dynamics play important roles in various human diseases including cancer (Srinivasan *et al.*, 2017). Cells usually contain hundreds to thousands of mitochondria in the cytoplasm. Mitochondria produce energy through oxidative phosphorylation production of adenosine triphosphate (ATP), the main source of energy in the cell. According to the endosymbiotic theory, mitochondria are derived from bacteria that were engulfed by the ancestors of today's eukaryotic cells (Archibald, 2015; Martin *et al.*, 2015). In higher eukaryotes, mtDNA codes for a small but crucial part of oxidative phosphorylation pathway proteins and independent translation machinery RNAs, compatible with bacterial translation and differing from translation of the nuclear genome. These data suggested that mitochondria evolved from bacteria that were endocytosed before animals and plants separated when oxygen entered the atmosphere about 1.5×10^9 years ago (López-García *et al.*, 2017). The majority of mitochondrial proteins are encoded currently in the cell nucleus. Even if the present organelle genomes are stable, extensive transfer of genes from organelle to nuclear DNA must have occurred during eukaryote evolution. For example, human mtDNA (and mtDNA of most animals) encodes 13 proteins

and 24 RNAs [transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs)] (Boore, 1999). However, there are many longer mitochondrial genomes that contain additional genes compared to animal and yeast mitochondrial genomes (Barr *et al.*, 2005; Gualberto *et al.*, 2014).

Local DNA structures such as cruciforms, left-handed DNA (Z-DNA), triplexes and quadruplexes play critical roles in regulating many fundamental biological functions (Cer *et al.*, 2012; Chasovskikh *et al.*, 2005; Paleček, 1991). Cruciform formation requires inverted repeats (IRs) of six or more nucleotides in the nucleic acid sequence (Mikheikin *et al.*, 2006). IRs are distributed non-randomly in the genomes of all living organisms. Although cruciforms are unstable in linear naked DNA because of branch migration (Shlyakhtenko *et al.*, 2000), cruciform formation has been identified in both prokaryotes and eukaryotes in vivo (Panayotatos and Fontaine, 1987; Yamaguchi and Yamaguchi, 1984). A number of proteins with preferential affinity for cruciforms have been identified, including 14-3-3 proteins and tumor suppressor protein p53 (Brazda *et al.*, 2012, 2017; Brazda and Coufal, 2017) and nuclear DNA cruciforms can regulate DNA replication, gene expression and DNA recombination (Bikard *et al.*, 2010; Brázda *et al.*, 2011). The potential role of cruciforms in mtDNA has not been well studied. We analyzed IRs in all sequenced mitochondrial genomes to determine frequencies, localization and similarities. The data show IRs in mtDNA that have been conserved through evolution, pointing to the importance of IRs in mitochondrial as well as nuclear genomes.

2 Materials and methods

2.1 mtDNA sequences

Complete mtDNA sequences were downloaded from the genome database of the National Center for Biotechnology Information (NCBI).

2.2 Data analysis

We used computational core of our DNA analyser software written in Java (Brazda *et al.*, 2016). We did not use the web frontend of DNA analyser tool for this task. The program was modified to read NCBI identifiers of sequences. There was one text file for each group of species. After the file containing mtDNA sequence was downloaded from NCBI, an analysis process was launched to find IRs using recommended parameters for Palindrome analyser. IR size was set from 6 to 30 bp, spacer size 0 to 10 bp and maximally one mismatch was allowed. An example IR identified using such criteria is provided in Supplementary Figure S1. We produced a separate list of IRs found in each of the 7135 mtDNA sequences available in NCBI and overall reports for each of the 18 species groups. Raw results for each sequence contained IR signature and position, but we did not find these useful for further processing. Results for each species group contained a list of species with size of mtDNA sequence and number of IRs found in that sequence. We also counted IRs grouped by their individual size (6–30 bp individually and sum of IRs longer than 8, 10 and 12 bp).

2.3 Analysis of IRs around annotated NCBI features

We downloaded the so called feature tables containing annotations of known features in mtDNA sequences; see Supplementary Figure S2. We analyzed IR occurrence inside, before and after features grouped by name to obtain a file with numbers of IRs inside and around features for each group of species. Search for IRs took place in predefined feature neighbourhood (we used ±100 bp – this figure is important for calculating IR frequency in feature neighbourhood) and inside feature boundaries. We calculated the amount of all IRs and those longer than 8, 10 and 12 bp in regions before, inside and after features. Categorization of an IR according to its overlap with a feature or feature neighbourhood is shown in Supplementary Figure S3. Further processing was performed in Microsoft Excel.

2.4 Phylogenetic tree construction

Exact taxid IDs of all analyzed groups [obtained from Taxonomy Browser via NCBI Taxonomy Database (Federhen, 2011)] were downloaded to phyloT: a tree generator (http://phylot.biobyte.de) and a phylogenetic tree was constructed using function 'Visualize in iTOL' in Interactive Tree of Life environment (Letunic and Bork, 2016). The resulting tree is shown in Supplementary Figure S4.

2.5 Statistical analysis

Cluster dendrogram of IR incidence (Supplementary Table S1) was made in R v. 3. 4. 0 (R Core Team, 2014) using *pvclust* (Shimodaira, 2006) with the parameters: cluster method 'average', distance 'uncentered' and number of bootstrap replications '10 000'. Cluster method and distance choice was validated using function *seplot*. The resulting cluster dendrogram is shown in Supplementary Figure S5. Principle component analysis (PCA) interactive plots were made in R with ggplot2 (Wickham, 2016) and plotly (Sievert *et al.*, 2016). The R code is available in Supplementary Code S1. Incidence of IRs (categorized by length) in individual species groups were used as input data, so for each species group one PCA plot was constructed to display intragroup variability.

3 Results

MtDNAs in NCBI database are stored in five groups (Animals, Fungi, Other, Plants and Protists) and 18 taxonomy subgroups. We downloaded all 7135 mtDNA sequences available (listed in Supplementary list of sequences), which vary from 1136 to 1 999 602 bp (Basu *et al.*, 2016). We firstly compared mtDNA lengths in the 18 subgroups. Length variability is lower in animals than in fungi, plants and protists (Fig. 1). Contrast between larger groups

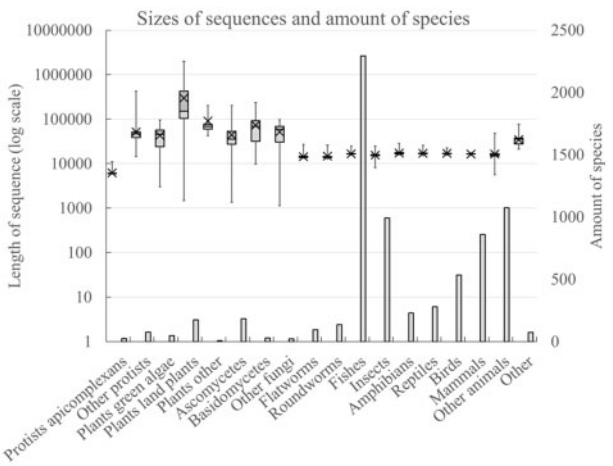

Fig. 1. Variability of length and amount of mtDNAs. Box plots show sequence length interquartile ranges for different species groups. The whiskers represent the minimum and maximum values. Numbers of species in each group is visualized with bars (scale is on the secondary vertical axis)

with low variability (e.g. fishes or insects) and smaller groups with large variability in length of sequence is clearly observable. Length variability generally correlated with evolutionary age. The largest variability is observed in the group Plantae and Fungi while mtDNA lengths are relatively constant in Animalia. The longest mtDNAs are typical for *Land plants*, the shortest for *Protists apicomplexan*.

3.1 Analyses of IRs

The parameters of analysis by Palindrome analyser were; IR length of 6–30 bp, spacer size 0–10 bp and maximally one mismatch. Totally we analyzed 179 624 234 bases and found 7 540 694 IRs; the overall IR frequency is therefore 41.9 IR/Kbp. The differences between organisms are significant; 50% of mtDNAs have a frequency of 27–47 IR/Kbp, but IR frequencies range from 9.47 IR/Kbp in a unicellular red alga found in hot sulphur springs—*Galdieria sulphuraria* 074W, while *Candida castellii* CBS 4332 (Ascomycetes fungi, class Sacharomycetes) has a frequency of 248.50 IR/Kbp. Values for all groups are shown in Figure 2.

The highest IR frequencies are in the groups *Insects* (89.33) and *Ascomycetes* (85.04) and the lowest in the group *Birds* (22.36). Statistics for all groups are provided in Supplementary Table S2. Statistical evaluations for each mtDNA are summarized in Supplementary Table S3. Comparing IRs in individual organisms and subgroups shows a general decrease in frequency with increasing IR length, except for IRs 24 and 30 bp long, which are 1.5 times more abundant than expected by approximation from neighbouring values (Table 1). We performed an additional analysis to distinguish

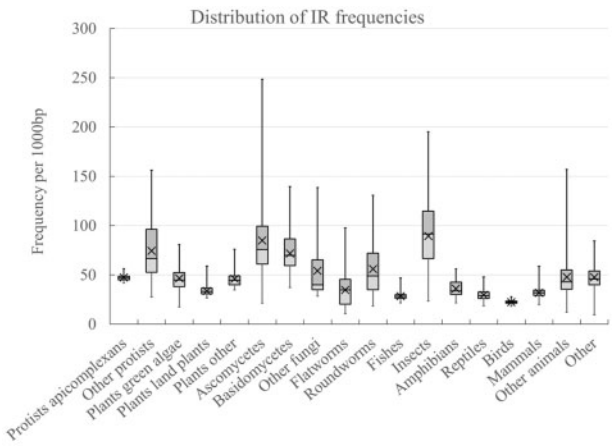

Fig. 2. Frequency of IRs in mtDNAs for subgroups and numbers of mtDNAs. The box plot shows the interquartile ranges of IR frequencies per 1000 bp in different species groups. Whiskers represent the minimum and maximum values

IRs of 30 and 31 bp or longer. IRs longer than 30 bp were found in only 180 of 7135 mtDNA sequences.

The detailed results for all groups are summarized in Table 2. The most common longest IR varied from 11 (in mammals) to 18 (in plants). IRs longer than 30 bp are rare, but their presence is interesting and we made additional analyses of these IRs (see Supplementary comments).

The NCBI genome database contains mtDNA annotations. The best described are 'gene' (163 443), 'tRNA' (152 631), 'rRNA' (14 570) and 'regulatory regions' such as D-loop, replication origins and stem loops. Numbers of annotations at the time of analysis are given in Supplementary Table S4. The annotations used are those defined in the sequence metadata and may not be entirely accurate, however most are validated by several methods and we obtained very similar results with smaller subsets of well-characterized mitochondrial genomes. To compare IR frequencies at different locations we used the most commonly described location 'gene' as a standard for comparison with other locations. There are significant differences in IR frequency in diverse segments of mtDNAs. The largest relative increase of IR frequency is for replication origin sequences followed by D-loop, stem-loop and misc sequences (Fig. 3).

The frequency of IRs located in replication origins is double that of IRs located in genes. Frequency changes are more distinct for longer IRs; 4-fold higher for IRs 8 bp and longer, 8-times for 10 bp or longer and 15-times for IRs 12 bp and longer (Fig. 3, orange). There are also changes in frequency in the neighbourhoods of annotated sequences (Fig. 3). The highest enrichment is not only within replication origins and stem loops, but also 100 bp before and after these sequences. Overall statistics of IRs in near neighbourhood and overlapping with annotations are shown in Supplementary Table S5. The ratios of IR frequencies of different annotation classes to gene class are given in Supplementary Table S6.

4 Discussion

In this paper, we analyzed all available mitochondrial genomes for the presence and localization of IRs. The typical maximal IR length was 12–14 bp, although many mtDNAs contain longer IRs. For statistical purposes, we compared IRs of 6 to 30 bp, which can be bound by DNA binding proteins and can form cruciform structures (Brázda *et al.*, 2011). Surprisingly, substantial numbers of longer IRs are detected in some mtDNAs. See supplementary comments for details of these extended IR sequences.

Homo sapiens has one of the lowest mtDNA IR frequencies (21.67 IR/Kbp), with only 359 IRs identified. Furthermore, only 24 are perfect (the other 335 IRs have one mismatch). The two longest IRs are 10 bp long, one with the sequence CCCCTTCGAC (one mismatch and CTT spacer) located in the middle of the *ND1* gene

Table 1. Numbers and frequencies of IRs according to size

IR size	Number in dataset	IR/1000bp	IR size	Number in dataset	IR/1000bp	IR size	Number in dataset	IR/1000bp
6	4460126	24.8303	15	4359	0.0243	24	254	0.0014
7	1841110	10.2498	16	2849	0.0159	25	113	0.0006
8	717399	3.9939	17	1807	0.0101	26	108	0.0006
9	289601	1.6123	18	1177	0.0066	27	91	0.0005
10	117709	0.6553	19	889	0.0049	26	80	0.0004
11	52939	0.2947	20	621	0.0035	29	58	0.0003
12	26048	0.1450	21	490	0.0027	30	65	0.0004
13	14252	0.0793	22	297	0.0017	>30	477	0.0027
14	7556	0.0421	23	228	0.0013			

Table 2. MtDNA sizes and IR frequencies and lengths

Group name	Number of seq.	Median size [bp]	Shortest sequence	Longest sequence	IR/Kbp – mean range	Longest IR for 50% of seq. [bp]
Protists-apicomplexans	24	5 977	Plasmodium vivax (5 882 bp)	Babasia microti (11 109 bp)	47 42–56	14
Other protists	76	46 840	Physarum polycephalum (14 503 bp)	Chromera velia (430 597 bp)	74 28–156	17
Plants green algae	48	45 175	Polytomella parva (3 018 bp)	Pseudendoclonium akinetum (95 880 bp)	47 17–81	18
Plants land plants	174	151 983	Vicia faba (1 478 bp)	Corchorus capsularis (1 999 602 bp)	34 27–59	18
Other plants	8	69 465	Mesostigma viride (42 424 bp)	Chlorokybus atmophyticus (201 763 bp)	46 35–76	17
Ascomycetes	183	35 655	Cryphonectria parasitica (1 364 bp)	Sclerotinia borealis (203 051 bp)	85 21–249	17
Basidiomycetes	29	69 195	Moniliophthora roreri (9 745 bp)	Rhizoctonia solani (235 849 bp)	72 37–140	15
Other fungi	23	58 788	Spizellomyces punctatus (1 136 bp)	Gigaspora rosea DAOM 194757 (97 350 bp)	54 28–138	15
Flatworms	96	13 968	Taenia pisiformis (13 383 bp)	Schmidtea mediterranea (27 133 bp)	35 11–98	12
Roundworms	137	13 960	Xiphinema americanum (12 626 bp)	Romanomermis culicivorax (26 194 bp)	56 19–131	15
Fishes	2 294	16 595	Gadus ogac (15 564 bp)	Rhinochimaera pacifica (24 889 bp)	28 22–47	12
Insects	992	15 534	Anaticola crassicornis (8 118 bp)	Hydropsyche pellucidula (25 004 bp)	89 23–195	15
Amphibians	231	17 175	Gegeneophis ramaswamii (15 897 bp)	Breviceps adspersus (28 757 bp)	36 22–56	13
Reptiles	279	17 107	Sphenodon punctatus (15 181 bp)	Heteronotia binoei (25 972 bp)	30 19–48	12
Birds	534	16 826	Malurus melanocephalus (15 568 bp)	Penelopides panini (22 737 bp)	22 18–28	12
Mammals	860	16 543	Macrotis lagotis (15 289 bp)	Lepus timidus (17 755 bp)	32 20–59	11
Other animals	1 074	15 754	Clathrina clathrus (5 596 bp)	Anadara sativa (48 161 bp)	48 12–157	12
Other	73	35 594	Galdieria sulphuraria (21 428 bp)	Phaeodactylum tricornutum (77 356 bp)	48 9–84	13

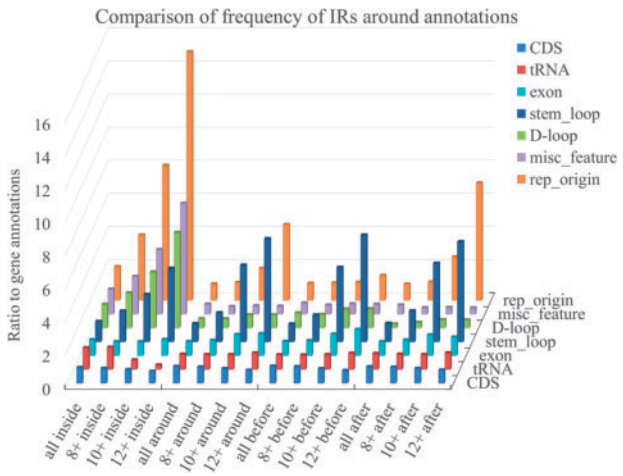

Fig. 3. Differences in IR frequency by DNA locus. The chart shows IR frequencies comparison per 1000 bp between 'gene' annotation and other annotated locations from the NCBI database. We analyzed frequencies of all IRs (all) and of IRs with lengths 8 bp and longer (8+), 10 bp and longer (10+) and 12 bp and longer (12+) within annotated locations (inside) and before and after annotated locations (Color version of this figure is available at *Bioinformatics* online.)

[NADH dehydrogenase, subunit 1 (complex I)] and the other with the sequence GTCCAAAGAG (no mismatch and GAACAG spacer) located within the *RNR2* gene (mitochondrially encoded 16S RNA). Interestingly, *Gorilla gorilla* mtDNA contains 410 IRs (25.06 IR/Kbp) and *Pan troglodytes* mtDNA contains 384 IRs (23.20 IR/Kbp). This IR reduction (*Gorilla* > *Pan* > *Homo*) is in congruence with phylogenetic relationships in hominidae (Pozzi *et al.*, 2014). In the lower primate group, Lemuriform primate *Lemur catta* has 554 IRs (32.52 IR/Kbp); tarsiformis primate *Tarsius bancanus* has 593 IRs with an average frequency 35.03 IR/Kbp.

PCA interactive plots intuitively represent similarities in pattern of IR length between all subgroups of organisms (Supplementary Plot P1) and between particular organisms within each subgroup (Supplementary plots P2–P19). The most distinct group is Protists Apicomplexans and all vertebrate subgroups are close together. Land Plants and Green Algae are also closely related by their IR incidence. Therefore, IRs in mitochondrial genomes are copying evolutionary trends and are relatively well conserved between organisms within each phylogenetic clade. From this point of view, IR pattern/incidence could be used as a new additional phylogenetic marker in the future.

Our analyses of all accessible mitochondrial genomes show that IR sequences are abundant and non-randomly distributed in the mitochondrial genomes of all living organisms. However, the frequencies of IRs differ between phylogenetic groups. The lowest average IR/Kbp was found in a unicellular polyextremophylic red alga *Galdieria sulphuraria* strain 074W, an acido-thermophile that can grow both autotrophically and heterotrophically in the dark. Other than living in extreme conditions of temperature and acidity, it also tolerates high metal ion concentrations. This mt genome of 21 428 bp has only 9.47 IR/Kbp and no IR is longer than 9 bp. Plastid and mitochondrial genomes of this organism show many extreme features, for example the mitochondrial genome is much smaller than other algae (Jain *et al.*, 2015). We have not found any mitochondrial genome without IRs. Most mitochondrial genomes have numerous IRs especially in regulatory regions such as replication origin and D-loop region. These results point to the importance of IRs in basic biological processes.

Acknowledgement

We thank Philip J. Coates for proofreading and editing the manuscript.

Funding

This work was supported by The Czech Science Foundation (15-21855S).

Conflict of Interest: none declared.

References

Archibald,J.M. (2015) Endosymbiosis and eukaryotic cell evolution. *Curr. Biol.*, **25**, R911–R921.

Barr,C.M. *et al.* (2005) Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. *New Phytol.*, **168**, 39–50.

Basu,T. *et al.* (2016) Organelle genetic diversity in a global collection of Jute (*Corchorus capsularis* and *C. olitorius*, Malvaceae). *South African J. Bot.*, **103**, 54–60.

Bikard,D. *et al.* (2010) Folded DNA in action: hairpin formation and biological functions in prokaryotes. *Microbiol. Mol. Biol. Rev.*, **74**, 570–588.

Boore,J.L. (1999) Animal mitochondrial genomes. *Nucleic Acids Res.*, **27**, 1767–1780.

Brand,M.D. (1997) Regulation analysis of energy metabolism. *J. Exp. Biol.*, **200**, 193–202.

Brazda,V. *et al.* (2016) Palindrome analyser – a new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. *Biochem. Biophys. Res. Commun.*, **478**, 1739–1745.

Brazda,V. *et al.* (2012) Superhelical DNA as a preferential binding target of 14-3-3gamma protein. *J. Biomol. Struct. Dyn.*, **30**, 371–378.

Brazda,V. *et al.* (2017) The structure formed by inverted repeats in p53 response elements determines the transactivation activity of p53 protein. *Biochem. Biophys. Res. Commun.*, **483**, 516–521.

Brázda,V. *et al.* (2011) Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol.*, **12**, 33.

Brazda,V. and Coufal,J. (2017) Recognition of Local DNA Structures by p53 Protein. *Int. J. Mol. Sci.*, **18**,

Cer,R.Z. *et al.* (2012) Searching for non-B DNA-forming motifs using nBMST (non-B DNA motif search tool). *Curr. Protoc. Hum. Genet.*, **18**, 1–22.

Chasovskikh,S. *et al.* (2005) DNA transitions induced by binding of PARP-1 to cruciform structures in supercoiled plasmids. *Cytometry A*, **68**, 21–27.

Federhen,S. (2011) The NCBI taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.

Gualberto,J.M. *et al.* (2014) The plant mitochondrial genome: dynamics and maintenance. *Biochimie*, **100**, 107–120.

Jain,K. *et al.* (2015) Extreme features of the *Galdieria sulphuraria* organellar genomes: a consequence of polyextremophily? *Genome Biol. Evol.*, **7**, 367–380.

Kauppila,T.E.S. *et al.* (2017) Mammalian mitochondria and aging: an update. *Cell. Metab.*, **25**, 57–71.

Kroemer,G. *et al.* (1998) The mitochondrial death/life regulator in apoptosis and necrosis. *Annu. Rev. Physiol.*, **60**, 619–642.

Letunic,I. and Bork,P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.*, **44**, W242–W245.

López-García,P. *et al.* (2017) Symbiosis in eukaryotic evolution. *J. Theor. Biol*, **434**, 20–33.

Martin,W.F. *et al.* (2015) Endosymbiotic theories for eukaryote origin. *Philos. Trans. R. Soc. B Biol. Sci.*, **370**, 20140330.

Mikheikin,A.L. *et al.* (2006) Effect of DNA supercoiling on the geometry of holliday junctions. *Biochemistry*, **45**, 12998–13006.

Paleček,E. (1991) Local supercoil-stabilized DNA structures. *Mol. Biol.*, **26**, 151–226.

Panayotatos,N. and Fontaine,A. (1987) A native cruciform DNA structure probed in bacteria by recombinant T7 endonuclease. *J. Biol. Chem.*, **262**, 11364–11368.

Pozzi,L. *et al.* (2014) Primate phylogenetic relationships and divergence dates inferred from complete mitochondrial genomes. *Mol. Phylogenet. Evol.*, **75**, 165–183.

R Core Team. (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/; https://stat.ethz.ch/pipermail/r-help/2014-October/422975.html.

Shlyakhtenko,L.S. *et al.* (2000) A cruciform structural transition provides a molecular switch for chromosome structure and dynamics. *J. Mol. Biol.*, **296**, 1169–1173.

Sievert,C. *et al.* (2016) Plotly: Create Interactive Web Graphics via'plotly. js' R package version 3.6.0. https://cran.r-project.org/web/packages/plotly/.

Srinivasan,S. *et al.* (2017) Mitochondrial dysfunction and mitochondrial dynamics-The cancer connection. *Biochim. Biophys. Acta - Bioenerg.*, **1858**, 602–614.

Wei,Y.H. *et al.* (2001) Mitochondrial theory of aging matures – roles of mtDNA mutation and oxidative stress in human aging. *Zhonghua Yi Xue Za Zhi (Taipei)*, **64**, 259–270.

Wickham,H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer, Berlin, Germany.

Yamaguchi,K. and Yamaguchi,M. (1984) The replication origin of pSC101: the nucleotide sequence and replication functions of the ori region. *Gene*, **29**, 211–219.