Supplementary Materials

In order to compare the evolutionary rates of mtDNA genomes containing major variant of disrupted common repeat with those having both arms of common repeat intacted we made a simple case-control comparative study of evolutionary rates on sister case-control clades. We based our study on the most recent and comprehensive human mtDNA genome sample containing all known mtDNA genome sequences which extracted from various sources. The evolutionary rates comparison can be biased by selection influences. Therefore, we should ensure that the analysed sample have no sequences under strong selection pressure - containing sites with hardly deleterious mutations or sites under strong positive selection, which are normally quickly selected out during recent mitochondrial evolution and/or having limited variation. In order to ensure that we do not analyse any deleterious mtDNA mutations we used mtDNA genome sequences classification from HmtDB (updated: October 2018) [1]. This database classify mtDNA sequences into two groups: containing deleterious mutations and without such mutations. We used multiple alignment from HmtDB containing 43437 mtDNA genome sequences that have no deleterious mutations which have 16908 alignment sites in total. In order to make sure that we analysed neutral or nearly neutral mtDNA sites for the evolutionary rates comparison we extracted variable sites from this multiple alignment. In this extraction procedure we exclude from consideration any ambiguous characters (not a A, T, G or C). We used three site variation thresholds: first, >0.5% of variation (relating to >220 sequences having nucleotide other than consensus; 791 alignment sites), second >0.1% of variation (>40 sequences have nucleotide other than consensus, 1941 alignment sites) and third, >0.05% of variation (>20 sequences have nucleotide other than consensus, 2778 alignment sites). Roughly speaking, these three thresholds reflect the simplified deviation spectra from nearly neutral site expectation - from mostly neutral category of sites describing by the first threshold to the sites having ability to variate due to neutral evolution or even moderate selection. We did not based our analysis on third codon positions because the HmtDB multiple alignment of mtDNA genomes of healthy cases have at least a one third of such sites in a strictly evolutionary conserved form, indicating the presence of hard selection load on at least one third of potentially variable positions. In parallel with variable site selection in normal (or nearly normal) mtDNAs, for each mtDNA we predict a haplogroup by the HaploGrep v. 2.1.20 software [2] and based on the whole multiple alignment of complete mtDNA genomes of healthy cases we reconstructed general phylogenetic tree by the IQ-TREE v. 1.6.1 software [3] using general time reversible (GTR) model of base substitutions [4], FreeRate site variation model [5] allowing for a proportion of invariable sites (option: -m GTR+F+I+R6). GTR model of base substitution characterized by unequal rates and unequal base freq; thus it requires big data for accurate estimation of all parameters. The simplified tree (obtained by clade collapsing) is shown on the Figure 3. The reconstructed general phylogenetic tree with marked haplogroups for each sequence allowed us to clearly discriminate five subsamples for the case-control comparative studies of evolutionary rates of mtDNA genomes containing disrupted and normal common repeat sequences (Figure1). These subsamples are composed of background sequences (control) having normal common repeat arms and foreground sequences (case) having disrupted common repeat by m.8473T>C mutation, background and foreground sequences have one recent common ancestor and the number of sequences in background and foreground groups are nearly equal. Additionally, based on the extracted tree topologies (using Archaeopteryx v. 0.9928 beta software [6]) of these five subsamples we discriminate root sequences for each subsample. In order to check the inequality of evolutionary rates between the foreground (case) and background (control) sequences

for each of the subsamples we extracted three subalignments of variable positions (for the three above described variation thresholds) of these sequences. Based on these subalignments unrooted phylogenetic trees were reconstructed. Phylogenetic tree reconstruction for each subsample was made by the IQ-TREE v. 1.6.1 software [3] and the JC [7], F81 [8], K80 [9] and HKY [10] models of base substitutions. After that with help of Newick Utilities v. 1.6 [11] we rooted these trees and compute branch lengths from the tree root. We select four simple models of base substitutions due to the simplicity of base changes count: JC model assumes equal base substitution rates and equal base frequencies [7], F81 model implies equal rates but unequal base frequencies [8], K80 model involves unequal transition/transversion rates and equal base frequencies [9], HKY - unequal transition/transversion rates and unequal base frequencies [10]. This simplicity allowed us to simplify comparison of the evolutionary rates to comparison of total branch lengths from recent common ancestor to the branch tips. Making the branch length comparison unbiased, before the comparison we subtract X/L from the branch lengths representing foreground sequence, where X is unity (in the case of JC and F81 models) or transition/transversion rate (in the case of K80 and HKY models) and L is the number of sites in alignment. To test the statistical significance of branch lengths inequality between background branches (control) and foreground branches (case) and its robustness to data variation we used 100 iteration of random-half-jackknife (by homemade Perl script) and Wilcoxon signed-rank test (by R version 3.4.1 [12]). Simply speaking, we selected randomly 100 times a half of branches from background and a half of branches from foreground and compare its lengths using Wilcoxon signed-rank test. The results of this analysis shown in Table XX. Additionally using the vioplot R package v. 0.3.0 [13] we visualize the compared background and foreground branch lengths.

Table 1. The case-control comparison of the evolutionary rates of mtDNA genomes containing disrupted common repeat (case) with those having both arms of common repeat intacted (control).

Substitution model	Sites variation, %	Stat. property of evolutionary rates in case, Wilcoxon signed-rank test	Clade composition				
			U6a	U2e	H1c	R/P	**D4**
F81	0.05	shift	greater	greater	greater	less	**less**
		average p	2.17E-05	9.73E-10	0.00176	0.062878	**7.75E-07**
		std.dev. of p	5.16E-05	1.48E-09	0.004146	0.091994	**3.22E-06**
	0.1	shift	greater	greater	greater	greater	**less**
		average p	0.00135	3.41E-12	0.04709	0.478095	**1.23E-06**
		std.dev. of p	0.001961	7.19E-12	0.06996	0.207913	**4.38E-06**
	0.5	shift	less	greater	less	greater	**less**
		average p	0.198667	0.058184	0.034564	0.281155	**2.88E-06**
		std.dev. of p	0.171724	0.060021	0.033213	0.142158	**6.9E-06**
HKY	0.05	shift	greater	greater	greater	less	**less**

		average p	4.27E-05	8.09E-12	0.04157	0.06907	**2.07E-06**
		std.dev. of p	9.7E-05	1.19E-11	0.057705	0.091293	**5.74E-06**
	0.1	shift	greater	greater	greater	greater	**less**
		average p	0.005236	5.22E-12	0.049723	0.293431	**7.91E-07**
		std.dev. of p	0.013786	6.58E-12	0.05822	0.172256	**2.48E-06**
	0.5	shift	less	less	less	greater	**less**
		average p	0.273149	0.284463	0.001448	0.418316	**5.53E-07**
		std.dev. of p	0.209097	0.141528	0.003522	0.236519	**1.02E-06**
JC	0.05	shift	greater	greater	greater	less	**less**
		average p	4E-05	6.89E-12	0.023705	0.0664	**4.62E-07**
		std.dev. of p	8.81E-05	9.34E-12	0.05136	0.071869	**1.22E-06**
	0.1	shift	greater	greater	greater	less	**less**
		average p	0.000618	2.35E-11	0.044148	0.5305	**2.71E-07**
		std.dev. of p	0.001676	4.55E-11	0.062387	0.214982	**6.51E-07**
	0.5	shift	less	greater	less	less	**less**
		average p	0.039706	0.237216	0.009861	0.461538	**8.35E-06**
		std.dev. of p	0.060181	0.186655	0.014167	0.266469	**1.69E-05**
K80	0.05	shift	greater	greater	greater	less	**less**
		average p	9.73E-06	1.66E-11	0.031127	0.087671	**1.14E-07**
		std.dev. of p	7.1E-06	5.27E-11	0.025872	0.146845	**2.04E-07**
	0.1	shift	greater	greater	greater	greater	**less**
		average p	0.000228	6.21E-10	0.060003	0.369783	**3.65E-07**
		std.dev. of p	0.000542	1.61E-09	0.106046	0.204312	**1.19E-06**
	0.5	shift	less	greater	less	greater	**less**
		average p	0.017182	0.189886	0.02839	0.554828	**3.98E-06**
		std.dev. of p	0.019592	0.11964	0.043765	0.255987	**9.78E-06**

Table 2 results of the multiple linear model: fraction of genome, covered by direct repeats as a function of generation length and nucleotide content. All values are PIC normalized (except model 0). All values (fraction of genome, covered by direct repeats; generation length; nucleotide composition) are log2 transformed. Number of analysed mammalian species is 705.

	Intercept (coefficient, p value) zero means that regression was forced to go through the origin	Generation length coefficient (p-value)	nucleotide: coefficient (p-value)
model 0 (without PIC norm.)	-1.18028 (< 2e-16)	**-0.03930** (0.000784)	NA
model 1.A	-0.20643 (0.0317)	**-0.01349** (0.2970)	NA
model 1.B	zero	**-0.01299** (0.317)	NA
model 2.A	-3.487e-02 (0.07895)	**-0.005679** (0.6641)	A: 0.947124 (0.0013)
model 2.B	zero	**-0.005294** (0.68642)	A: 0.930128 (0.00162)
model 3.A	-0.189388 (0.0490)	**0.001839** (0.9011)	T: 0.272742 (0.0347)
model 3.B	zero	**0.003499** (0.8132)	T: 0.294066 (0.0226)
model 4.A	-0.20438 (0.0337)	**-0.01170** (0.3868)	G: -0.08649 (0.6428)
model 4.B	zero	**-0.01082** (0.424)	G: -0.10474 (0.575)
model 5.A	-0.18158 (0.0556)	**0.02153** (0.1469)	C: -0.54527 (4.39e-06)
model 5.B	zero	**0.02280** (0.125)	C: -0.55808 (2.64e-06)

Literature

[1] Clima R, Preste R, Calabrese C, Diroma MA, Santorsola M, Scioscia G, Simone D, Shen L, Gasparre G, Attimonelli M. HmtDB 2016: data update, a better performing query system and human mitochondrial DNA haplogroup predictor. Nucleic Acids Res. 2017 Jan 4;45(D1):D698-D706. doi: 10.1093/nar/gkw1066.

[2] Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt HJ, Kronenberg F, Salas A, Schönherr S. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016 Jul 8;44(W1):W58-63. doi: 10.1093/nar/gkw233.

[3] Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015 Jan;32(1):268-74. doi: 10.1093/molbev/msu300.

[4] Tavaré S. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. Lectures on Mathematics in the Life Sciences. 1986 17:57–86.

[5] Soubrier J, Steel M, Lee MS, Der Sarkissian C, Guindon S, Ho SY, Cooper A. The influence of rate heterogeneity among sites on the time dependence of molecular rates. Mol Biol Evol. 2012 Nov;29(11):3345-58. doi: 10.1093/molbev/mss140.

[6] Han MV, Zmasek CM. phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinformatics. 2009 Oct 27;10:356. doi: 10.1186/1471-2105-10-356.

[7] Jukes TH, Cantor CR Evolution of protein molecules. In Munro HN, editor, Mammalian Protein Metabolism, (1969) pp. 21-132, Academic Press, New York.

[8] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17(6):368-76.

[9] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980 Dec;16(2):111-20.

[10] Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985;22(2):160-74.

[11] Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics. 2010 Jul 1;26(13):1669-70. doi: 10.1093/bioinformatics/btq243.

[12] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[13] Daniel Adler and S. Thomas Kelly (2018). vioplot: violin plot. R package version 0.3.0 https://github.com/TomKellyGenetics/vioplot