bioRxiv preprint doi: https://doi.org/10.1101/2022.12.12.520138; this version posted December 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Reanalysis of mtDNA mutations of human primordial germ cells (PGCs) reveals significant contamination with NUMTs, and challenges predominantly purifying selection in late PGCs.

Zoë Fleischmann[1]*, Auden Cote-L'Heureux[1]*, Melissa Franco[1]*, Sergey Oreshkov[4], Sofia Annis[1], Mark Khrapko[1], Dylan Alden[1], Konstantin Popadin[2,3,4], Dori C. Woods[1], Jonathan L. Tilly[1], and Konstantin Khrapko** [1].
* equal contribution ** correspondence: k.khrapko@northeastern.com
[1]Department of Biology, Northeastern University, Boston, Massachusetts, USA.
[2]School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland;
[3]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland;
[4]Center for Mitochondrial Functional Genomics, Institute of Living Systems, Immanuel Kant Baltic Federal University, 236040 Kaliningrad, Russia.

Abstract:

The resilience of the mitochondrial genome (mtDNA) to a high mutational pressure depends, in part, on negative purifying selection against detrimental mutations in the germline. Floros et al. reported a sharp increase in the synonymity of mtDNA mutations between early and late-stage primordial germ cells (PGCs), concomitant with a shift from glycolytic to oxidative metabolism, as evidenced by comparing data from pooled samples of early vs late PGCs. They thus asserted that this metabolic shift exposed deleterious mtDNA mutations to negative selection. We re-analyzed Floros' data to resolve a perceived inconsistency of the pattern of synonymity change and discovered a significant contamination of pooled PGC mutations with nuclear sequence derived from mtDNA (NUMT). We determined that contamination was caused by co-amplification of the NUMT sequence by cross-specific PCR primers. Importantly, when we removed *NUMT-derived sequence variants from pooled PGC data, the evidence of purifying selection in late PGCs was abolished.* We then turned to mutations of *single* PGCs, also from the same group, for which no synonymity analysis was reported. We found no evidence of NUMT contamination of single PGCs mutations. This is consistent with the use of a different set of PCR primers which are unable to amplify NUMT since they were positioned outside the NUMT sequence. Importantly, we further demonstrated that single PGC mutations show a significant *decrease* of synonymity with increased mutant fraction. This observation is incompatible with predominantly purifying selection of mtDNA mutations in PGCs at these mutant fractions and suggests that selection may be predominantly positive.

Abbreviations:

MAF minor allele frequency, same as mutant fraction, as long as mutant fraction is below 50%
CS – Carnegie Stage, (conventional staging of human embryonic development) followed by the stage number.
PGC – Primordial Germ Cell
NUMT – a pseudogene of a mitochondrial DNA sequence residing in the nuclear genome
SNV – single nucleotide variant. Used instead of 'mutation', especially where it is not clear whether the mutation is real or artificial.

Introduction.

Mitochondrial DNA (mtDNA) is known for having a high mutational rate and a high density of crucial genes. Therefore, mtDNA mutations not only underlie a spectrum of mitochondrial diseases[1,2] but also, if left unpurged, could result in long-term detrimental effects on species evolution, a.k.a. Muller's ratchet [3]. This question has been raised by most papers that study germline selection of mtDNA mutations. It is thus vital to understand how mtDNA mutations are kept at a sustainable level across generations. A seminal study from 2008 demonstrated that significant levels of purifying selection against detrimental mutations occur in the germline: mutations are depleted of nonsynonymous variants as early as the second generation after mutations are generated *de novo* in an mtDNA 'mutator' mouse line[4]. Similarly, purifying selection is reported to shape human mtDNA diversity[12]. However, the mechanism(s) and precise developmental timing of germline purifying selection remain controversial. Recently, Floros and colleagues[5] reported changes in the transcriptome and mtDNA mutations during the late-stage development of human primordial germ cells (PGCs). The importance of this research cannot be overestimated, and these data are crucial for understanding the dynamics and mechanisms of germline selection. Floros et al. reported a sharp increase in synonymity of mtDNA mutations in late-stage (CS20/21: Carnegie Stage 20/21) PGCs, compared to early-stage (CS12), concomitant with transcription changes that imply a shift from glycolytic to oxidative metabolism[5]. They thus asserted that purifying selection was caused by exposure of mutations, expanded by intracellular segregation, to the metabolic shift. Surprisingly, the reported increase in synonymity in pooled late-stage PGCs appeared to result not from the depletion of nonsynonymous mtDNA mutations as would be expected under negative purifying selection, but instead from an increase in the frequency of synonymous mutations. In search of an explanation, we turned to the primary mutational data and re-analyzed pooled PGC mutations (Floros et al., Table S5). Additionally, we analyzed synonymity of single PGC mutations, which were provided to us by Dr. Chinnery and which were not analyzed by Floros et ai. Our analyses of primary data of the Floros study data convinced us that these data do not support Floros's conclusions of purifying selection in late PGC; moreover, they suggest that nonsynonymous mutations may be subject to positive selection in PGCs.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.12.520138; this version posted December 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results and discussion.

An initial review of primary data of *pooled* PGCs revealed that a majority (~80%) of 'endorsed' sequence variants (i.e., those above the 1% threshold) were found in multiple unrelated samples (**Suppl. Table,** orange shading). Such profuse recurrence is normally not observed with real somatic mutations, most of which are individually rare events and therefore are negligibly likely to occur in multiple samples. One widely recognized cause of such inter-sample repetition of low fraction sequence variants is contamination with nuclear DNA of mitochondrial origin (NUMTs).

This is a well-recognized problem of mtDNA mutational analysis, which we and others first encountered decades ago [6,7]. NUMTs are nuclear pseudogenes derived from fragments of mtDNA inserted into the nuclear genome. Many were inserted millions of years ago and since then nuclear and mitochondrial sequences have been diverging from each other. Typical NUMT differs from mtDNA by multiple changes, which appear as blocks of SNVs, which constitute a 'haplotype' specific to each NUMT.

We tested the NUMT contamination hypothesis and were able to confirm it with three lines of evidence. First, we aligned the Floros et al. variants to several NUMT sequences known to reside in the human genome. This analysis revealed that at least ***30%*** of protein-coding single-nucleotide variants (SNVs) reported by Floros et al. as real mutations, mapped to a NUMT located on chromosome 5 (CNVs marked 'ch5' in **Suppl. Table**). Coincidentally, we have previously used this NUMT as a marker of human evolution (**Popadin 2022**)[8]. We determined that this NUMT contamination was introduced by co-amplification of the NUMT sequences by one of the PCR primer pairs used in their study. Indeed, all NUMT-mapping SNVs were located in the mtDNA sequence between the primers of a primer pair that was used by Floros et al. to amplify one of their PCR fragments (**Figure 1, Suppl. Note 1**).

Second, NUMT origin of NUMT-mapping SNVs was confirmed by the analysis of mutations derived by Floros et al. from single PGCs**.** Floros et al. amplified single PGC samples using different sets of primers. We determined that none of the two primer pairs used for single PGCs amplification were able to amplify any portion of this NUMT (**Suppl. Note 2**). In full agreement with NUMT co-amplification being the source of NUMT-mapping SNVs in pooled PGC data, no NUMT - mapping SNVs were detected in single PGC samples.

Third, the strongest evidence of NUMT origin of NUMT-mapping SNVs was obtained using raw next-generation sequencing (NGS) data (provided by Floros et al in their Supplement). We discovered that NGS reads of the pooled PGCs fell into two classes – the 'mtDNA-derived reads' and the 'NUMT-derived reads'. The latter were clearly distinguished by the NUMT 'haplotype', i.e., a full set of NUMT-specific nucleotide changes in a single NGS read.

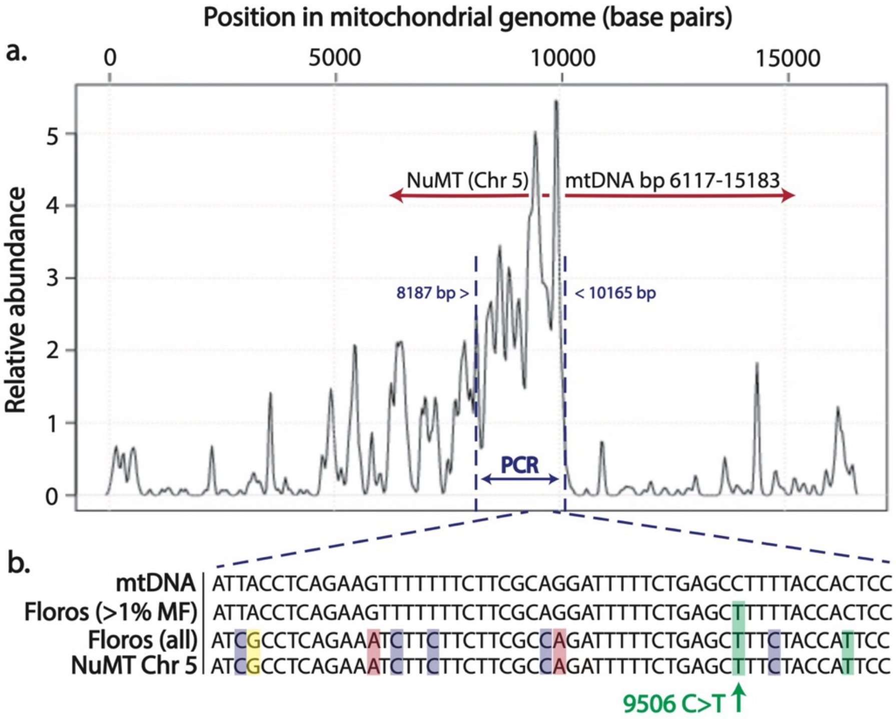

Figure 1. NUMT co-amplification contaminates mtDNA mutations from pooled PGCs: evidence from spatial distribution and sequence alignment.

A*. Spatial density distribution along the mitochondrial genome of SNVs from the dataset provided by Floros et al. The unexpected dense patch of variants (between ~8kb and 10Kb, dashed vertical blue lines) corresponds to one of the PCR fragments used in Floros et al. (bp:8187-10165, horizontal blue double arrow). The PCR primers of this fragment co-amplify a portion of the nuclear pseudogene of mtDNA (NUMT) on Chromosome 5 (hg38(100045938-100055045)) that is homologous to this region of mtDNA (red double arrow; see also Suppl. Note 2).*

B*. A representative section of alignment of the reference mtDNA sequence ('mtDNA') with the NUMT sequence on Chromosome 5 ('NuMT Chr 5') and two mtDNA sequences with SNVs from Table S5 of Floros et al., including those that are under the 1% cutoff ("Floros (all)"), and just the 'endorsed' SNVs that are above 1% threshold ("Floros (>1% MF)"). This alignment shows that a great majority of Floros et al. SNVs in this genome region are identical to the NUMT-derived SNVs, which is consistent with their putative NUMT origin. Of 10 variants located in the section of the genome shown here, one (9506C>T, green arrow) exceeded the threshold and thus was erroneously considered a 'endorsed' mutation. This illustrates the mechanism of 'leakage' of NUMT-driven variants into the pool of "endorsed" SNVs with >1% MF. This contamination is very extensive: in total, NUMT-derived SNVs with >1% MF account for 10 of 28 protein coding SNVs in late stage PGCs (36%) and 1 of 7 in early PGCs (14%).*

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.12.520138; this version posted December 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

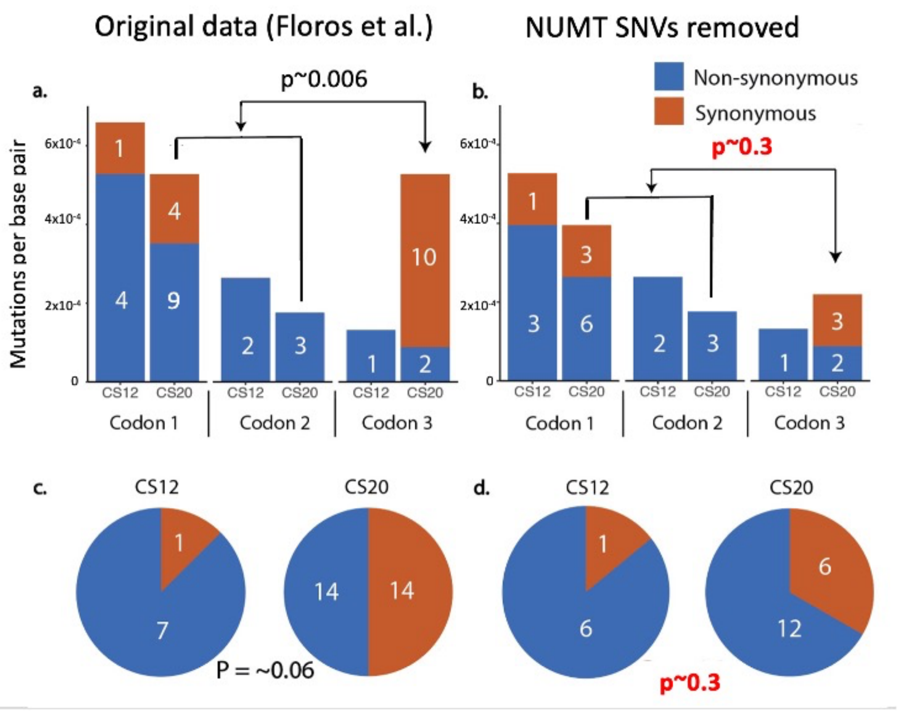

Figure 2. Removal of NUMT contamination fully abolishes statistical evidence of purifying selection in late stage PGCs. a: *a recreation of Floros et al., 2018 Figure 2b from the original data (Floros et al. Table S5).* ***b:*** *The same graph after removal of NUMT-mapping SNVs.*
c: *A recreation of Figure 2c from Floros et al., 2018 using the original data and* ***d:*** *pie charts after removal of NUMT-derived SNVs. We note that removal of NUMT contamination illustrated in this figure is very conservative. P-values were calculated using Fisher Exact Test. White numbers in bars and pie charts represent the absolute number of mutations in each category; note that these numbers are not proportional to the size of the bars because the bars are scaled by the number of embryos in each category.*

The most conclusive finding pertaining to determining the source of contamination, was that 'endorsed' NUMT-mapping SNVs reported by Floros et al. were present exclusively on the NUMT-derived NGS reads, which carries the full NUMT haplotype, definitively proving their NUMT origin (see Fig S1 and its caption for details of the analysis).

Having established that NUMT-mapping SNVs are indeed NUMT-derived, we asked whether contamination with NuMT-derived SNVs affected the key conclusions of the study and in which way. The effect of the removal of NuMT contamination on the calculated synonymity change between early and late PGCs reported by Floros et al. is shown in **Figure 2**. In figure 2, we recreated original data plots from Figures 2b and 2c of the Floros study (these are left-side panels 2a and 2c of our **Figure 2**, correspondingly). Then we removed NuMT-derived SNVs and re-plotted the data (right side panels Figure 2: b and d, correspondingly).

As seen in Fig. 2, the removal of NUMT-derived SNVs results in a *complete loss* of statistical significance of the synonymity change between early and late-stage PGCs both in overall synonymity (p-value increases from 0.06 to 0.34, Fig.2 c vs. d) and in codon bias (p-value increases from 0.006 to 0.3, Fig.2 a vs. b). We further explored what caused such an erroneous assessment of early-to-late synonymity change in the original data. We noted that the distribution of NUMT-derived SNVs in the original data is highly skewed: 10 of them are in late-stage PGC samples, and only one is in early-stage PGCs (see Suppl. Table). Moreover, this skew is principally a result of heavy NUMT contamination of a single late-stage PGC sample, CS20-11593, which contributed 8 NUMT-derived SMVs out of 10. This 'distribution skew' turns into 'synonymity skew' because NUMT-derived SNVs are highly synonymous (**Suppl. Note 3,** see also (Popadin et al., 2022)[8]). An excess of synonymous NuMT-derived SNVs in the contaminated late sample naturally led to a significant overestimation of the synonymity of late PGC mutations, and to an unsubstantiated claim of purifying selection in late PGCs. Additionally, this prevalence of synonymous NuMT-derived SNPs fully explains the puzzling surge of synonymous variants in late-stage PGCs (**Fig.2a**) which initiated this inquiry.

In addition to protein-coding mutations, Floros et al. reported a decrease of RNA-coding mutations in late PGC with a p-value of 0.03 (Figure 2d of Floros et al). We were not able to reproduce this result using Floros et al. data. There is only one non-coding RNA mutation in CS12 and only two in CS20/21. Finally, Floros et al. reported convincing evidence of selection among mutations of the control region. We note, however, that D-loop mutations, although many of them are known to be prone to selection, are typically not detrimental, making the role of D-loop mutations in purifying selection unclear.

Because NUMT-contaminated pooled PGC data proved inconclusive concerning testing the late PGC selection hypothesis, we asked whether single PGC data, which are free from contamination with NUMT ch5, could provide a better insight. Single PGS dataset is comprised of late-stage PGCs only, so early vs. late comparison as it was done in pooled PGCs is not possible. We note, however, that, as Floros et al. has demonstrated, single PGC data contain a large number of clonally expanded mutations, which permits an alternative way to test purifying selection in late PGCs. According to Floros et al. (see Floros et al., Fig.5), purifying selection can proceed in two ways. Within cells, selection may preferentially remove mitochondria with detrimental mutations or prevent them from proliferation, either way, blocking their clonal expansion. On the cellular level, selection may preferentially remove cells with higher mutant

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.12.520138; this version posted December 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

fractions (i.e., larger clonal expansions) of detrimental mutations. Either scenario predicts an increased proportion of neutral mutations, i.e., increased synonymity, among mutations with higher intracellular heteroplasmy levels. Similarly, it predicts that mutant fractions of individual non-synonymous mutations, on average, should be of lower than mutant fractions of individual synonymous ones.

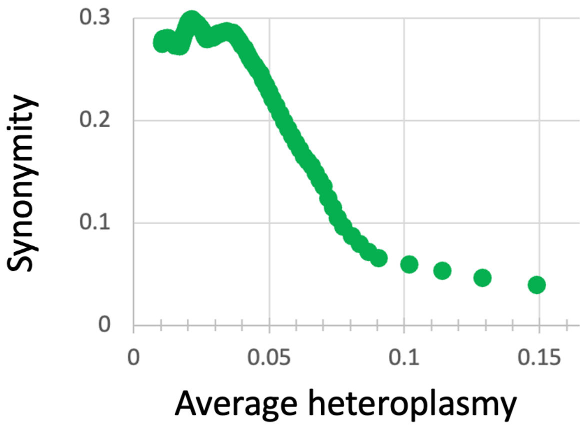

***Figure 3**. Average synonymity of mtDNA mutations in single PGCs within a sliding window 20 cells wide, calculated over cells ranked by the heteroplasmy level. This graph illustrates a consistent and dramatic decrease of synonymity in PGCs with higher heteroplasmy levels. This is the opposite of what is expected according to the purifying selection hypothesis. The statistical significance of this decrease is formally treated in Supplemental note 4. Decrease of synonymity at higher heteroplasmy level implies positive selection of non-synonymous mutations in PGCs.*

We therefore tested the late PGC purifying selection hypothesis by testing its predictions outlined above. First, we noted that, in direct disagreement to the predictions, average of fractions of individual synonymous mutations was significantly lower than that of nonsynonymous ones (0.028 vs. 0.042, *p<0.02*, two sample t-test). This effect is echoed by dramatic decrease of synonymity at higher heteroplasmy levels shown in **Figure 3**. Moreover, detailed analysis revealed an even stronger disagreement with the purifying selection hypothesis: we found that the difference in average mutant fraction between synonymous and nonsynonymous mutations was due to a group of statistical outliers: 12 nonsynonymous mutations with highest mutant fractions in the entire dataset (see Supplement 4 for details).

The existence of high-MAF nonsynonymous mutations exceeding MAFs of synonymous mutations strongly contradicts purifying negative selection on nonsynonymous mutations either on the mitochondrion or cellular level. However, the formal proof of statistical significance of the observation of these high-MAF nonsynonymous mutations requires non-parametrical analysis. The binomial probability of observing this set of nonsynonymous outliers is 0.017, i.e., this is a significant observation at least with p~0.017. In addition, we performed bootstrap simulation which demonstrated that the non-parametric p-value of the enrichment of nonsynonymous mutations at high mutant fraction is highly significant (p~0.01; see Supplement 4 for details). These findings imply that the hypothesis of predominant purifying selection in late PGCs at these mutant fractions is rejected. A statistically significant increase of non-synonymity among mutations expanded in individual PGCs implies positive selection in late PGCs, again at these mutant fractions. This conclusion is quite extraordinary, however, given the relatively small size of the study it should be considered preliminary and needs independent confirmation.

As counterintuitive as it may seem, positive germline selection of detrimental mtDNA mutations has been reported previously. We have recently reported that two fairly detrimental mutations, mouse 3875delC and human 3243A>G, are subject to positive selection (Fleischmann et al. 2021))[10] and (Franco et al. 2020))[9]. More recently, Chinnery's laboratory also proposed positive selection in mouse 5024C>T and human 3243A>G, 8344A>G, and 8993 T>G, though not in PGCs but rather in growing oocytes (Zhang et al. 2021)[11]. Thus, positive selection of detrimental mutations in the germline may be a fairly general phenomenon.

How can the putative positive selection of detrimental mtDNA mutations can be compatible with overall purifying selection in the germline, e.g., as reported by (Stewart et al. 2008)[4] or Wei et al. 2019)[12]? We note that the signature of positive selection inferred in our study pertains to mutations with moderate mutant fractions and at specific segment of germline development. The significance ceases at lower mutant fractions (see Supplementary Note 4). This is consistent with our previous findings of an 'arching' selection profile in human m.3243A>G mutation[10], where positive selection also ceases at lower fractions. More generally, mutations at other mutant fractions, mutations of other types (e.g., highly toxic mutations), and at different stages of germline development may be subject to purifying selection so that overall selection in germline is negative. Our research thus demonstrates the potential complexity of the inner works of purifying germline selection. This complexity is further corroborated by a recent report by the Chinnery group that certain detrimental mtDNA mutation in the mouse (*m.5024C>T)* is subject to positive selection during oocyte maturation[11.] More research is needed to reconcile this finding with ours. It is possible that the differences are attributable to the special properties of the m.5024C>T mutation. It is known for its unusual inheritance pattern: it was not possible to obtain heteroplasmy below 15% in the offspring (*Jim Stewart, personal communication*).

In conclusion, the pattern of germline selection appears to be potentially complex and mutation-specific. More research is needed to explore this critical phenomenon and in doing so great caution should be exercised to avoid dangerous artifacts inherent to mutation analysis of mtDNA, including NUMT contamination.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.12.520138; this version posted December 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods.

Synonymity and repeatability analysis of sequence variants of mtDNA.The data pertaining to the Floros et al. study have been imported from their supplemental table 5 at https://doi.org/10.1038/s41556-017-0017-8 (Floros et al., 2018). Synonymity was analyzed using the 'lookup' function on in-house Excel-based spreadsheets containing synonymity tables for all possible mutations in the mitochondrial genome. Identical mutations in different samples were identified using the 'Countif' function.

Extraction and multiple alignment of raw reads from the NGS dataset. We extracted raw read NGS data from NCBI Sequence Read Archive record SRR6288291, which corresponds to the late PGC sample CS20-11593. We aligned reads against an mtDNA reference sequence (bp 9379-9436) that corresponds to the Chr5 NuMT with the BWA-MEM tool (http://bio-bwa.sourceforge.net/) and then filtered the reads to only capture those with seven or fewer mismatches using samtools ("[NM]<=7") and had a mapping quality of ≥ 20; duplicate sequences were removed using a custom script. We again used samtools to convert .bam files to .fasta files.

Multiple alignment of the reads extracted thereby was performed using the NCBI BLAST server with default parameters, with the sequence of the human NuMT at chromosome 5 (hg38(ch5100045938-100055045)) as a query. We chose this method merely as a visually convenient approach to present the results.

Acknowledgements:

This study was supported by a grant from the U.S. National Institutes of Health (R01-HD091439 to J.L.T., D.C.W. and K.K.), KP was supported by the Ministry of Science and Higher Education of the Russian Federation (agreement no. 075-02-2022-872).

References:

1. Lander, E. S. & Lodish, H. Mitochondrial diseases: gene mapping and gene therapy. *Cell* **61,** 925–926 (1990).

2. Elliott, H. R., Samuels, D. C., Eden, J. A., Relton, C. L. & Chinnery, P. F. Pathogenic mitochondrial DNA mutations are common in the general population. *Am. J. Hum. Genet.* **83,** 254–260 (2008).

3. Popadin, K., Polishchuk, L. V., Mamirova, L., Knorre, D. & Gunbin, K. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proceedings of the National Academy of Sciences* **104,** 13390–13395 (2007).

4. Stewart, J. B., Freyer, C., Elson, J. L. & Larsson, N.-G. Purifying selection of mtDNA and its implications for understanding evolution and mitochondrial disease. *Nat Rev Genet* **9,** 657–662 (2008).

5. Floros, V. I. *et al.* Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. *Nature Cell Biology 2018 20:2* **20,** 144–151 (2018).

6. Khrapko, K., Andre, P. C., Cha, R., Hu, G. & Thilly, W. G. Mutational spectrometry: means and ends. *Prog Nucleic Acid Res Mol Biol* **49,** 285–312 (1994).

7. Hirano, M. *et al.* Apparent mtDNA heteroplasmy in Alzheimer's disease patients and in normals due to PCR amplification of nucleus-embedded mtDNA pseudogenes. *Proceedings of the National Academy of Sciences* **94,** 14894–14899 (1997).

8. Popadin, K., et al. Mitochondrial Pseudogenes Suggest Repeated Inter-Species Hybridization among Direct Human Ancestors. *Genes 13*, 810 (2022).

9. Franco, M., Pickett, S., Fleischmann, Z., Khrapko, M., Annis, S., Woods, D., Markuzon, N., Turnbull, D., and Khrapko, K. (2020). Can detrimental mtDNA mutations be under positive selection in the germline? The FASEB Journal *34*, 1–1. 10.1096/fasebj.2020.34.s1.09461.

10. Fleischmann, Z., Pickett, S.J., Franco, M., Aidlen, D., Khrapko, M., Stein, D., Markuzon, N., Popadin, K., Braverman, M., Woods, D.C., et al. (2021). Bi-phasic dynamics of the mitochondrial DNA mutation m.3243A&gt;G in blood: An unbiased, mutation level-dependent model implies positive selection in the germline. bioRxiv, 10.1101/2021.02.26.433045.

11. Zhang, H., Esposito, M., Pezet, M.G., Aryaman, J., Wei, W., Klimm, F., Calabrese, C., Burr, S.P., Macabelli, C.H., Viscomi, C., et al. (2021). Mitochondrial DNA heteroplasmy is modulated during oocyte development propagating mutation transmission. Sci. Adv. *7*, eabi5657. 10.1126/sciadv.abi5657.

12. Wei, W., Tuna, S., Keogh, M.J., Smith, K.R., Aitman, T.J., Beales, P.L., Bennett, D.L., Gale, D.P., Bitner-Glindzicz, M.A.K., Black, G.C., et al. (2019). Germline selection shapes human mitochondrial DNA diversity. Science 364, eaau6520.