Analysis of numts in sixteen different mice strains

Bálint Biró[1], Zoltán Gál[1], Giuseppina Schiavo[2], Anisa Ribari[2], Valerio Joe Utzeri[2], Michael Brookman[3], Luca Fontanesi[2], Orsolya Ivett Hoffmann[1]

[1] Hungarian University of Agricultural and Life Sciences, Institute of Genetics and Biotechnology, Szent-Györgyi Albert Str. 4, H-2100, Gödöllö, Hungary
[2] University of Bologna, Department of Agricultural and Food Sciences, Division of Animal Sciences, Viale Fanin 46, 40127 Bologna, Italy
[3] Hanze University of Applied Sciences, Department for Biology and Medical Laboratory Research, Zernikeplein 7, 9747 AS Groningen, Netherlands

1 Introduction

2 Materials and Methods

Nuclear and mitochondrial DNAs were acquired from Ensembl and NCBI databases (?, ?, ?). The mitochondrion of Mus musculus musculus was annotated using MITOS server hosted by the University of Leipzig (?, ?). Double mtDNA and gDNA were aligned using LASTAL (v1219) (?, ?) with the scoring scheme of + 1 for matches, −1 for mismatches, 7 for gap-open penalty and 1 for gap-extension penalty. Using doubled mtDNA, we were able to identify numts that are located on the linearization point of the mtDNA. After that, alignments were filtered based on their e-values as proposed by Tsuji (?, ?). False positive alignments that were the results of using double mtDNA were also discarded.

To examine normality, Anderson-Darling test was performed at 0.05 p value. t-test or Wilcoxon signed rank test was performed based on the result of normality testing. All the statistical calculations were conducted in Scipy (v1.6.2) (?, ?). The free energy values of RNA structures were calculated with seqfold (?, ?).
Pairwise divergences were calculated using the modified Kimura 2 parameter (?, ?) which tolerates gaps in the alignments. The modified Kimura 2 formula is described by Equation **??**.

$$K = \frac{3}{4} w \log w - \frac{w}{2} \log(S - P)\sqrt{S + P - Q} \qquad \text{(Figure. 2.1)}$$

where w=the probability that the given position contains a nucleotide,
S=n_1/n,
n_1=the number of positions where the two aligned sequences contain the same nucleotide,
n=total number of nucleotides,
P=n_2/n,
n_2=number of transition type mutations,
n_3=number of transversion type mutations.

Pairwise divergence values were calculated using a sliding window approach with 1kb window size and 10bp step size.

For the free energy calculations, the corresponding tRNA sequences for each strains were acquired with SAMTOOLS's (v1.6) faidx function (?, ?).

The phylogenetic analysis was conducted in R with phangorn (?, ?). The Maximum-Parsimony tree was constructed using Jukes-Cantor distance with 100 bootstrap. For tree construction cytochrome b (*CYTB*) and corresponding numt sequences were used as described by a previous study (?, ?) with rat as outgroup.

All figures were created in matplotlib (v3.4.3) and seaborn (v0.11.2) (?, ?, ?).

3 Results

152 numts can be identified in the Mus musculus genome. There is a huge variability in terms of numt numbers on the chromosomes. For example, chr1 contains 14 numts, while no numt can be found on chr19. The numts cover the total mitochondrion. The longest numt (4654 bp) covers 6 protein coding genes (Fig.**??**/a). There is a strong correlation (Pearson correlation coefficients: 0.77) between the chromosome size and the number of numts on a given chromosome (Fig.**??**/b). The shortest chromosome contains the smallest number of numts while the longest chromosome contains the highest number of numts. However this relationship is not true for any of the investigated strains.

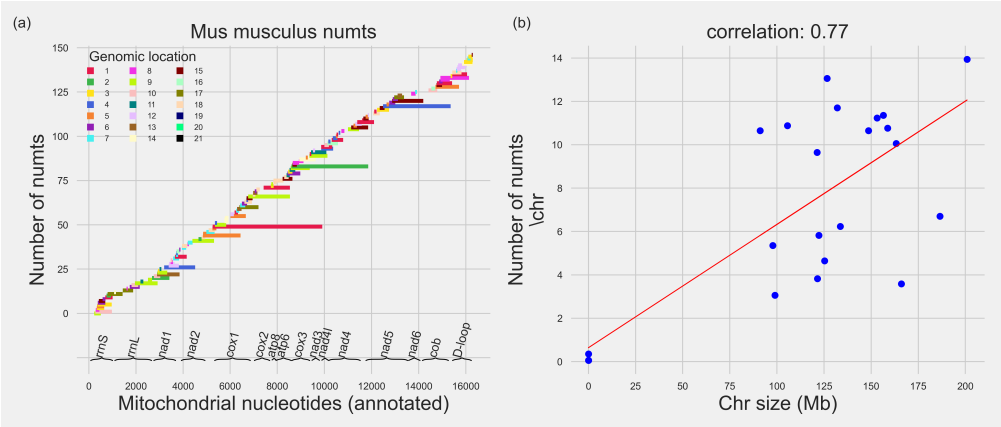

Fig. 1. Patterns of Mus musculus numts. Distribution of Mus musculus numts with the genomic locations (a) and the correlation between the number of numts and the size of the corresponding chromosome (b). Small tRNA genes are not part of the annotation.

When investigating the distribution of numts along the mitochondria in different strains, it turns out that the majority of the numts are the same as in the case of Mus musculus musculus. However exceptions do exist. For example as we have already described above, numts originate from the whole mitochondrion in case of Mus musculus musculus but not in the case of several inbred and wild derived strains. Differences are also present in the lengths of the numts. For instance, the longest numt in the Mus musculus musculus genome which locates on chr1 is missing from all the strains investigated (Fig.**??**.).

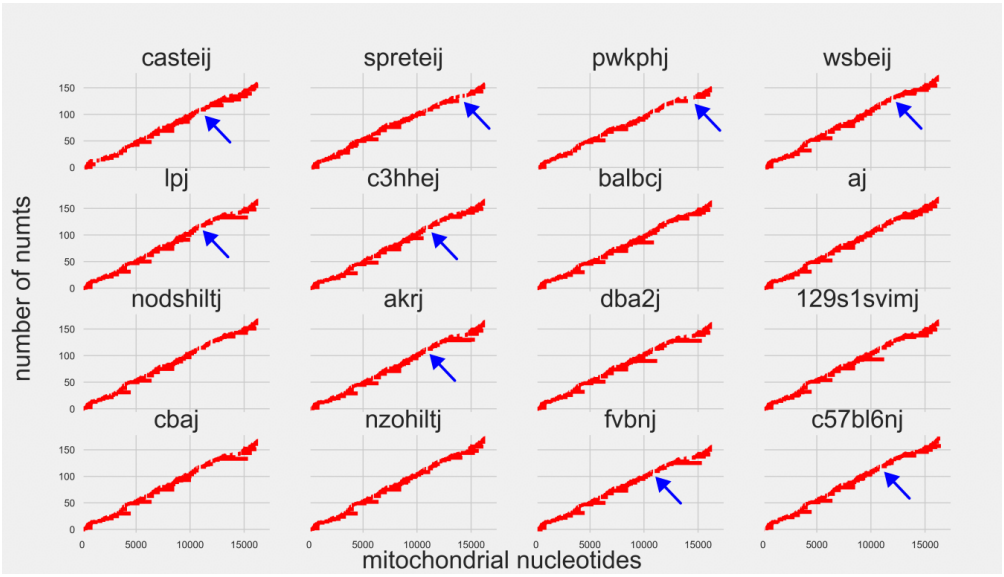

Fig. 2. Distributions of numts in different mice strains. The upper row contains the wild derived strains. Blue arrows indicate mitochondrial regions where numts do not originate from.

There are common patterns in terms of numt coverage. Interestingly the regions of mitochondria that contain the linearization point (the start and the end of the linearized mitochondrion) are highly covered by numts. This region contains the *tRNA-S* coding gene and the D-loop. Another numt dense region which is present in every strains, covers *atp6*, *cox3*, *nad3* and *nad4l* genes. There are two general numt sparse regions. The first one partially covers *tRNA-S* and *tRNA-L* genes while the second one covers *nad4* (Fig.**??**.).

From the wild derived strains casteij, spreteij and pwkphj while from inbred strains aj and akrj are clustered together when it comes to nuclotides along mitochondria involved in numtogenesis. At the same time wsbeij and the rest of the inbred strains are also clustered together (Fig.**??**.).

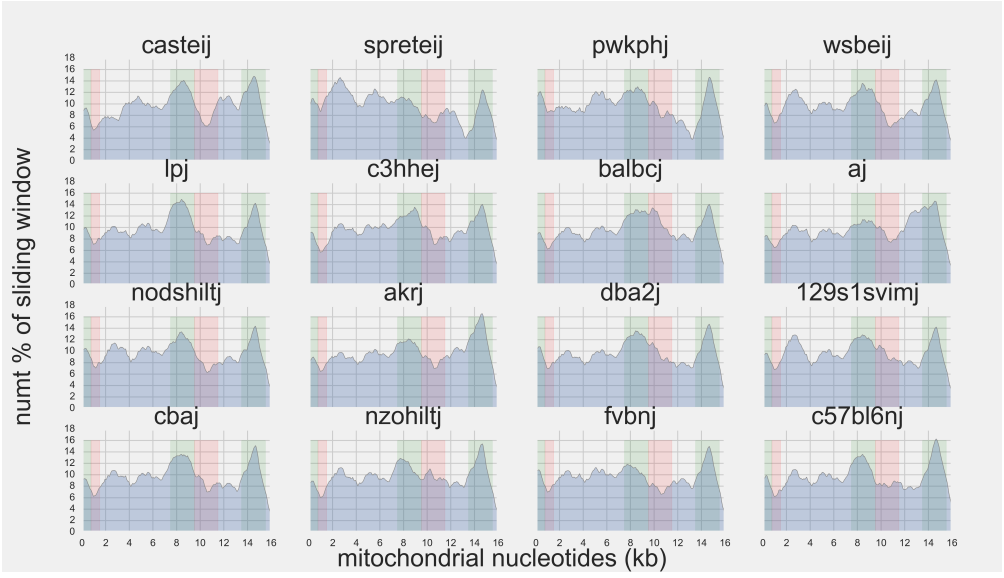

Fig. 3. Numt contents along mitochondria. Colored areas show common patterns. Green color is corresponding for dense numt region while red color is corresponding for numt sparse region.

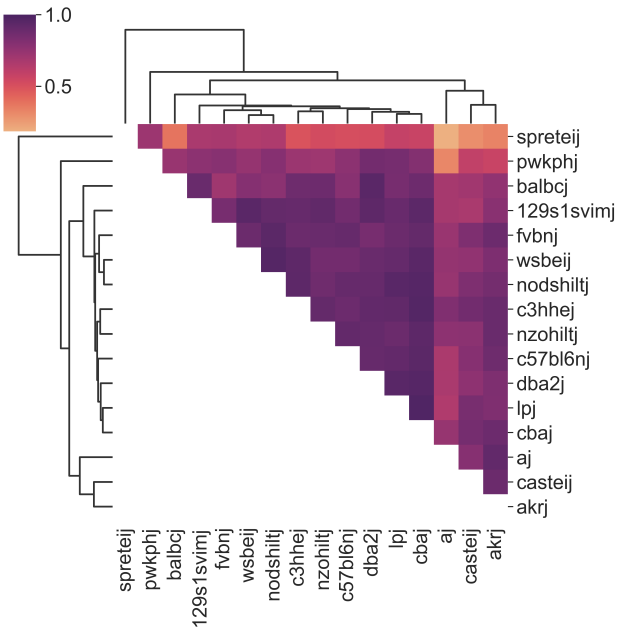

Fig. 4. Pearson correlation matrix of nuclotides involved in numtogenesis along mitochondria.

The wild derived strains casteij, spreteij and pwkphj show higher pairwise divergence values when total mitochondria are compared with the mitochondrion of Mus musculus musculus. Surprisingly, the fourth wild derived strain wsbeij does not differ significantly from Mus musculus musculus while the inbred nzohiltj strain shows elevated pairvise distance values to some extent (Fig.**??**.).

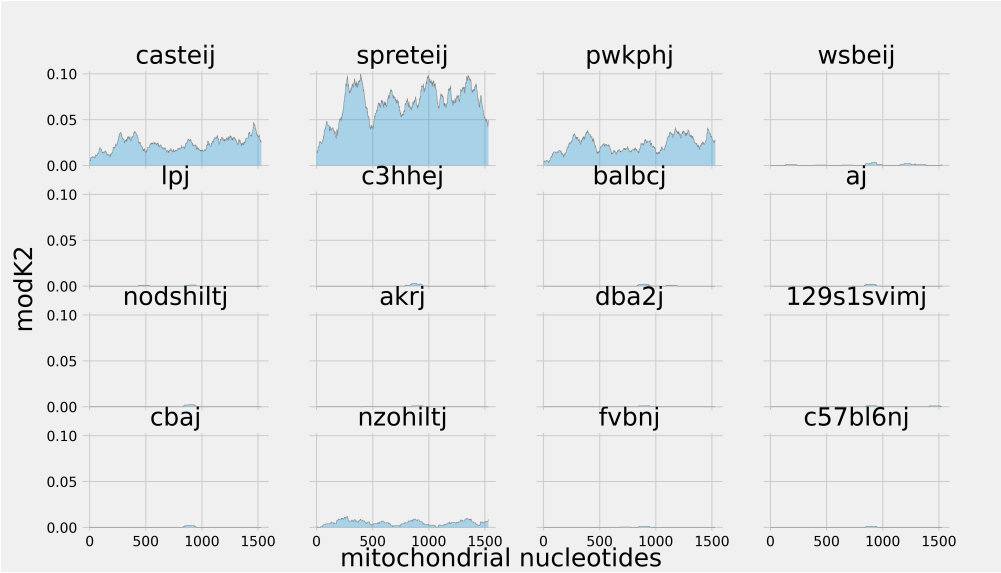

Fig. 5. Pairwise divergence values. The upper row contains the wild derived strains.

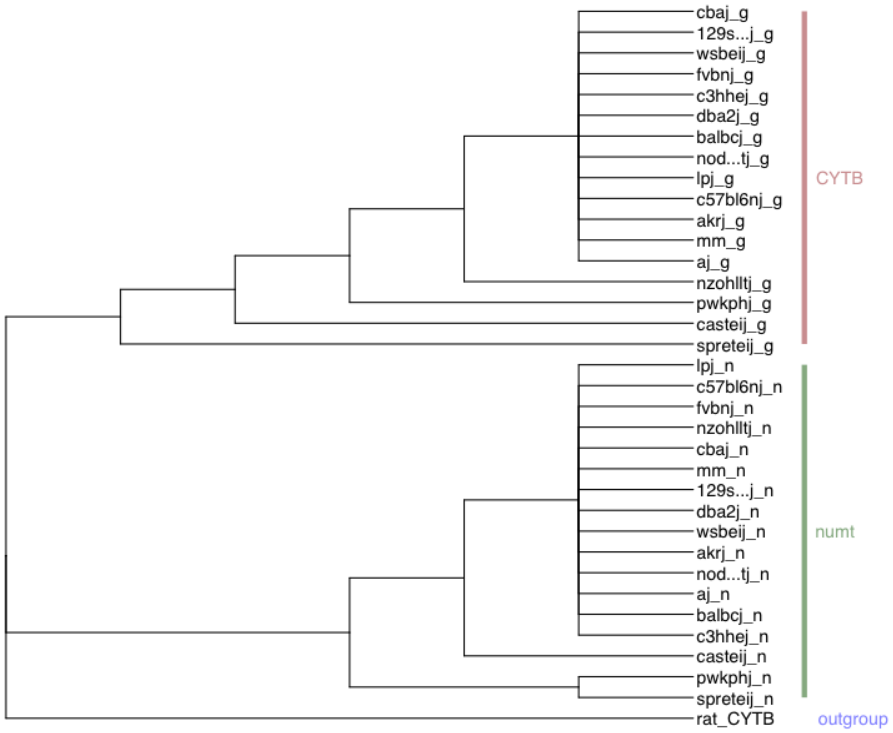

Fig. 6. Maximum-Parsimony tree based on Jukes-Cantor distance with 100 bootstrap and rat CYTB as outgroup.

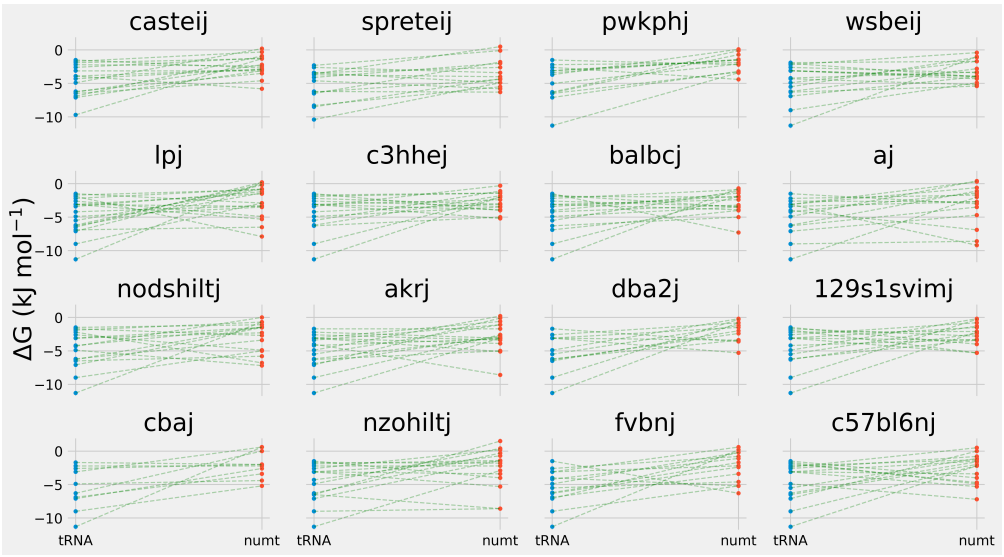

Fig. 7. The free energy values of the predicted structures of tRNA numts. The upper row contains the wild derived strains.

4 Conclusions

5 Acknowledgements

6 References

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., … Stadler, P. F.

(2013). Mitos: improved de novo metazoan mitochondrial genome annotation. *Molecular phylogenetics and evolution*, *69*(2), 313–319.

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., … others (2022). Ensembl 2022. *Nucleic acids research*, *50*(D1), D988–D995.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, *9*(03), 90–95.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome research*, *21*(3), 487–493.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, *25*(16), 2078–2079.

Nishimaki, T., & Sato, K. (2019). An extension of the kimura two-parameter model to the natural evolutionary process. *Journal of molecular evolution*, *87*(1), 60–67.

Ouyang, Z., Snyder, M. P., & Chang, H. Y. (2013). Seqfold: genome-scale reconstruction of rna secondary structure integrating high-throughput sequencing data. *Genome research*, *23*(2), 377–387.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in r. *Bioinformatics*, *27*(4), 592–593.

Tsuji, J., Frith, M. C., Tomii, K., & Horton, P. (2012). Mammalian numt insertion is non-random. *Nucleic acids research*, *40*(18), 9073–9088.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … others (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, *17*(3), 261–272.

Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., … others (2007). Database resources of the national center for biotechnology information. *Nucleic acids research*, *36*(suppl_1), D13–D21.