Analysis of numts in sixteen different mice strains

Bálint Biró[1], Zoltán Gál[1], Michael Brookman[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] 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 (Wheeler et al., 2007; Cunningham et al., 2022). The mitochondrion of Mus musculus musculus was annotated using MITOS server hosted by the University of Leipzig (Bernt et al., 2013). For the numt comparisons 12 inbred strains (129S1/SvImJ, A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6NJ, CBA/J, DBA/2J, FVB/NJ, LP/J, NZO/HlLtJ, and NOD/ShiLtJ) and four ”wild derived” strains (CAST/EiJ, PWK/PhJ, WSB/EiJ, SPRET/EiJ) were selected based on a previously published genome comparison analysis (?, ?). Double mtDNA and gDNA were aligned using LASTAL (v1219) (Kiełbasa et al., 2011) 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 (Tsuji et al., 2012). 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) (Virtanen et al., 2020). Whether tRNA genes are expressed were derived from their structures using free energy values which were calculated with seqfold (Ouyang et al., 2013) based on a previous research paper (Liu & Zhao, 2007).
Pairwise divergences were calculated using the modified Kimura 2 parameter (Nishimaki & Sato, 2019) which tolerates gaps in the alignments. The modified Kimura 2 formula is described by Equation 2.1.

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

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 (Li et al., 2009).

The phylogenetic analysis was conducted in R with phangorn (Schliep, 2011). 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 (Rodríguez et al., 2007) with rat as outgroup.

All figures were created in matplotlib (v3.4.3) and seaborn (v0.11.2) (Hunter, 2007; Waskom, 2021).

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.1/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.1/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.

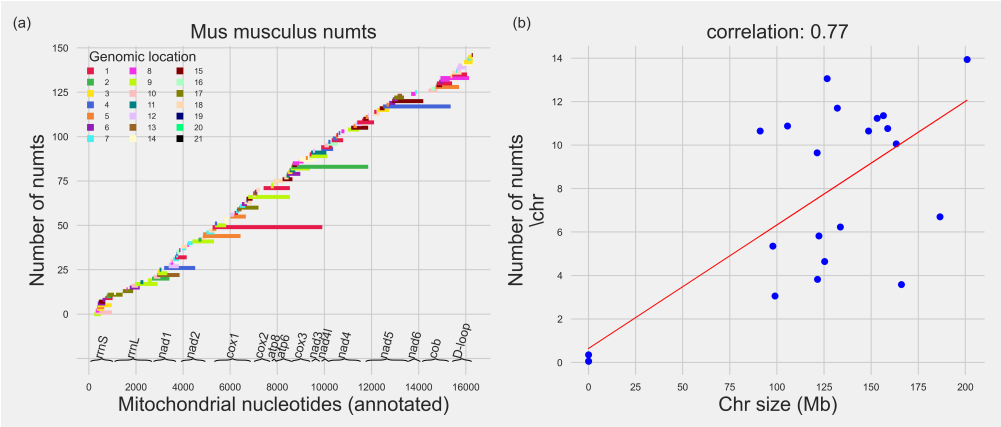

Fig. 1. Patterns of Mus musculus numts. Distribution of Mus musculus numts with the genomic locations (a) and the correlation between the number of numts and the size of the corresponding chromosome (b). Small tRNA genes are not part of the annotation.

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.2.).

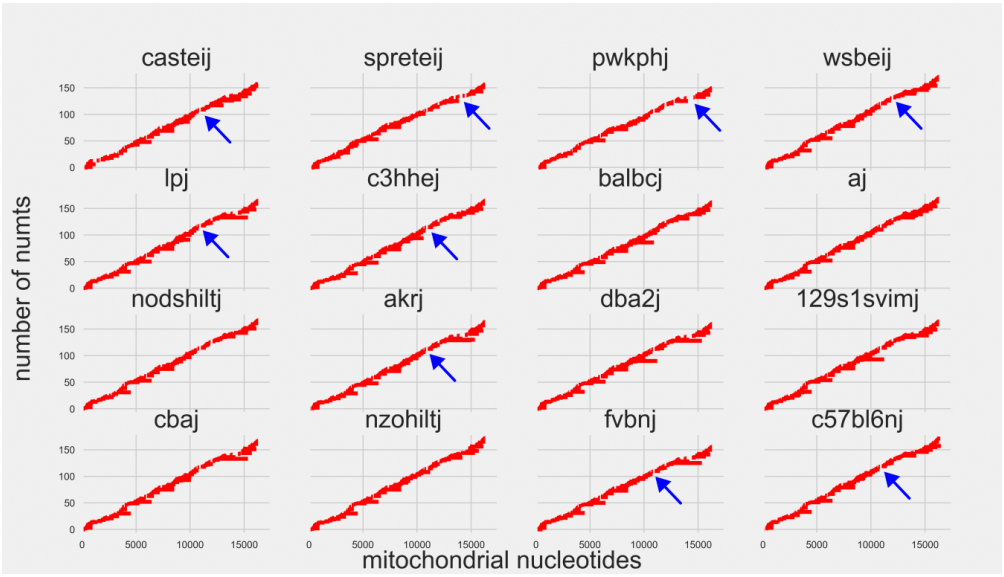

Fig. 2. Mitochondrial origin of numts in different mice strains. The upper row contains the wild derived strains. Blue arrows indicate mitochondrial regions where numts do not originate from.

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.3.).

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.4.).

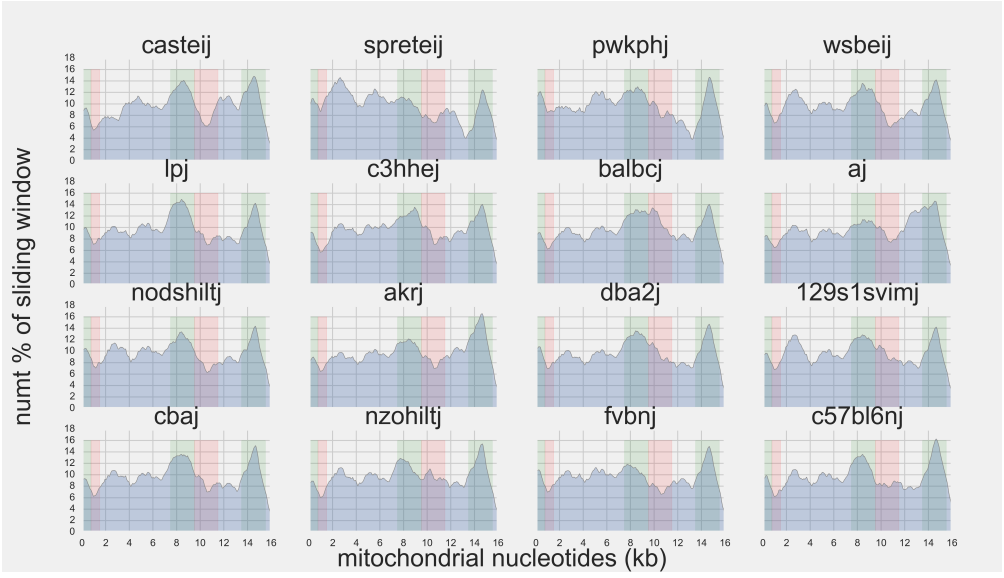

Fig. 3. Numt contents along mitochondria. Colored areas show common patterns. Green coloring is corresponding for dense numt region while red coloring is corresponding for numt sparse region.

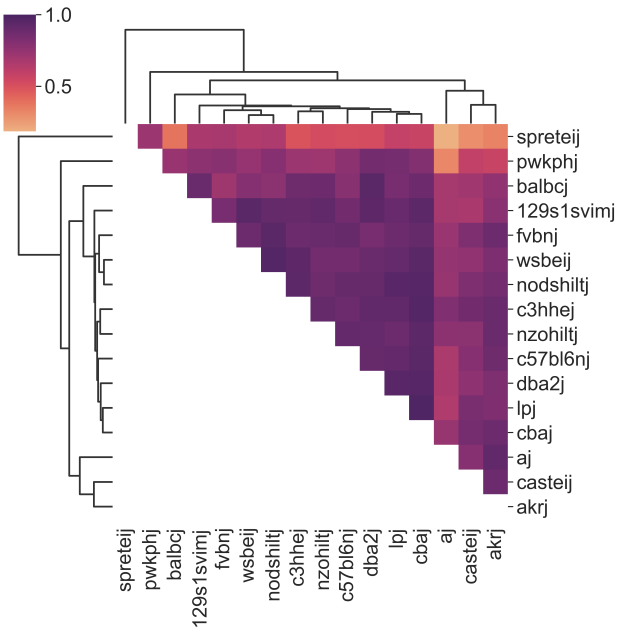

Fig. 4. Pearson correlation matrix of nuclotides involved in numtogenesis along mitochondria.

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.5.).

Maxmimum-Parsimony phylogenetic tree supports two monophyletic clades, namely the mitochondrial CYTB and the corresponding numt sequences. The result of the phylogenetic analysis resembles the whole mitochondria divergence values even though the sequences of only one gene were used. In the CYTB clade the wild derived strains casteij, spreteij and pwkphj plus the inbred nzohiltj strain differ from the other strains. However nzohiltj is very close to the other inbred strains. In the numt clade the wild derived strains casteij, spreteij and pwkphj are different than the rest of the strains. In this clade the inbred strain nzohiltj is clustered together with the inbred strains unlike in the case of CYTB clade, pairwise divergence and numt coverage analysis. Intriguingly wsbeij is clustered together with the inbred strains in both clades even though it is a wild dervied strain (Fig.6.).

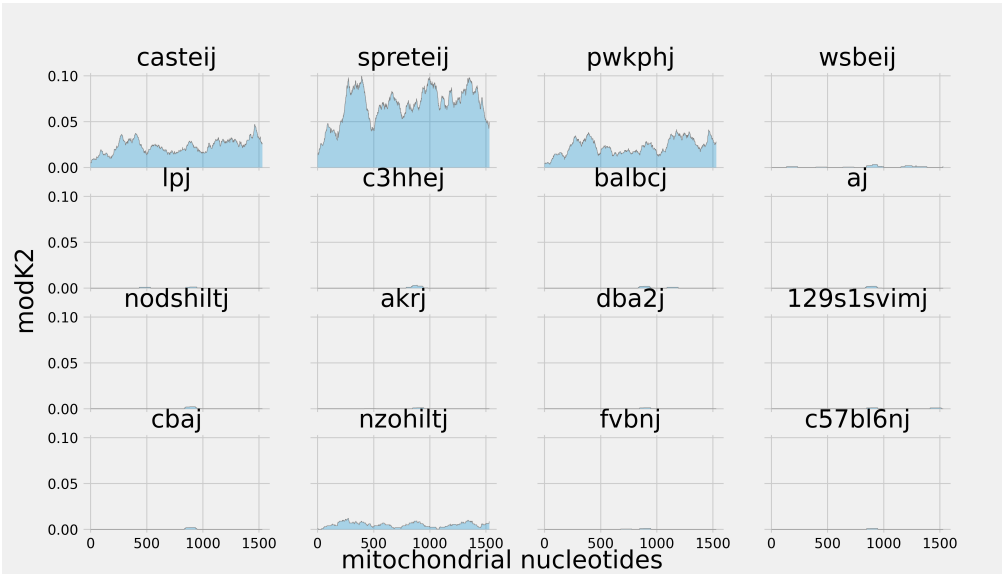

Fig. 5. Pairwise divergence values. The upper row contains the wild derived strains.

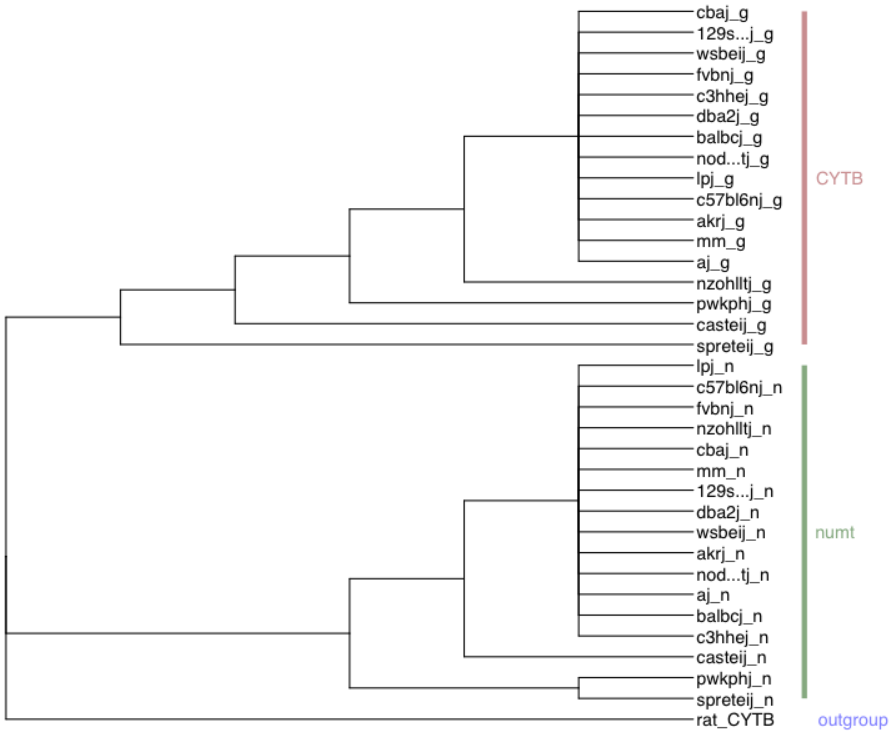

Fig. 6. Maximum-Parsimony tree based on Jukes-Cantor distance with 100 bootstrap and rat CYTB as outgroup.

During the analysis of free energy values of the folding of tRNA sequences and their corresponding numts no case was shown where numt folding's ΔG value was the same as tRNA folding's. In most of the cases tRNA folding's ΔG was smaller than numt folding's ΔG (Fig.7.).

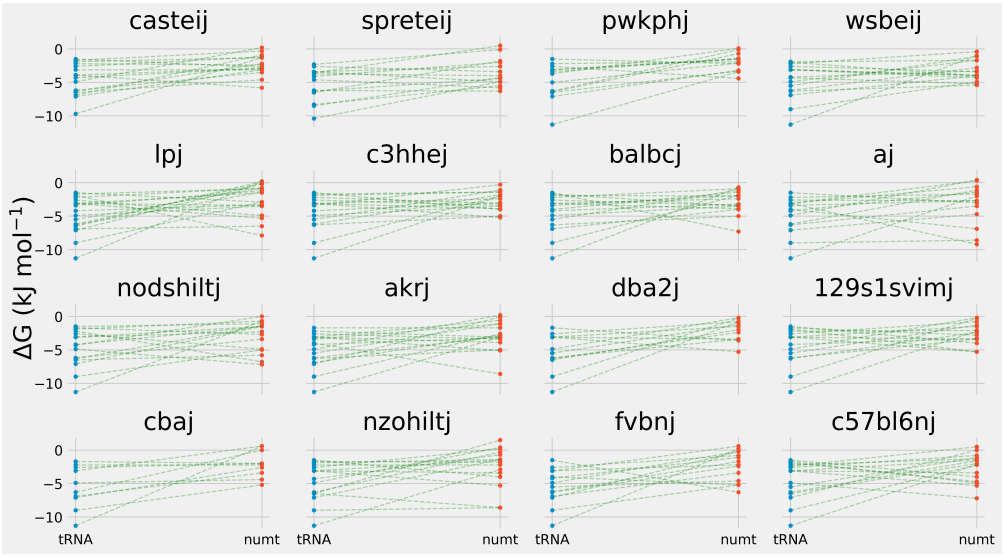

Fig. 7. The free energy values of the predicted structures of tRNAs and their corresponding numts. The upper row contains the wild derived strains.

4 Conclusions

Like in many other eukaryotic organisms (Hazkani-Covo et al., 2010; Calabrese et al., 2012; Tsuji et al., 2012; Wang et al., 2020), numts are also present in the genome of Mus musculus musculus. However the different patterns of numts in mice strains have not been described previously. Hence in this study numts of divergent mice strains were investigated.

We described 152 numts in the Mus musculus musculus genome which is comparable with previous studies (Calabrese et al., 2012; Tsuji et al., 2012).

There is a strong correlation in the Mus musculus musculus genome between chromosomal length and the number of the numts on the given chromosome. This correlation also exists in the human genome (Lascaro et al., 2008; Simone et al., 2011). In general, there is a higher gene density in case of shorter chromosomes. Since natural selection tries to avoid insertional mutagenesis (eg. intragenic numts), there is a higher possibility for less gene dense, longer chromosomes to tolerate a numt integration without serious consequences (Lascaro et al., 2008). However this is not the case in any of the strains investigated. No correlation was found between the above mentioned attributes also in the case of honey bee (Behura, 2007).

The majority of Mus musculus musculus numts were the same in the strains investigated. However the longest numt (4654 bp on chr1) could not be detected in any of the strains. This situation was found in a couple of organisms and it is originated from a fragmentation event after the given numt has been integrated into the nuclear genome (Behura, 2007; Rodríguez-Salinas et al., 2012).

The investigation of numt coverage along the mitochondria reveals that mitochondrial nucleotides are present in several copies. Numt coverage results also proved that the representation of the nucleotides along the mitochondria differ. Hence that over- as well as under-represented regions do exist. This kind of imbalanced representation of mitochondrial nucleotides were reported in several mammalian species (Simone et al., 2011; Tsuji et al., 2012).

All of our experiments that somehow integrate data about the similarity of numts or mitochondria (numt content along mitochondria, pairwise divergence, phylogeny) resemble the relationship between Mus musculus musculus and spreteij as to our current knowledge. Namely that from the strains investigated, spreteij seems to be the most distantly related compared to Mus musculus musculus. The phenomenon that spreteij is the most distant strain from the strains examined also applies to when total nuclear genomes are being compared (?, ?).

Since the mitochondrial genetic code and the nuclear genetic code are different (Gonzalez et

al., 2012), the same DNA sequence would be resulted in different RNA molecules depending on the used genetic code. In addition, the nucleotide composition of numts are adjusted to the mitochondrial machinery. Hence numts are considered to be pseudogenes that are unable to code the genetic information that they used to code. However there is still a chance for numts to code functional tRNAs since tRNAS coded by the nucleus and tRNAs coded by mitochondrion would be identical (Pozzi & Dowling, 2019). Although we found no evidence for expressing numts that are corresponding to tRNA genes. All the tRNA numts were altered in sequence and the free energy of the folding of these numts were always higher than the free energy of the folding of corresponding tRNA genes. However the free energy values of the folding of numts were still negative, which means that the folding itself is a spontaneous process. But tRNA genes had the minimum free energy which means the most stable structure from a thermodynamic point of view (Su et al., 2019). Our results regarding tRNA expression are in agreement with the results in the case of cattle (Liu & Zhao, 2007).

5 Acknowledgements

B.B. received NTP-NFTÖ-21-B-0127 scholarship. O.I.H. was funded by NKFIH OTKA FK 124708, János Bolyai Research Scholarship of the Hungarian Academy of Sciences and New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund (ÚNKP-21-5).

6 References

Behura, S. K. (2007). Analysis of nuclear copies of mitochondrial sequences in honeybee (apis mellifera) genome. *Molecular biology and evolution*, *24*(7), 1492–1505.

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.

Calabrese, F. M., Simone, D., & Attimonelli, M. (2012). Primates and mouse numts in the ucsc genome browser. *BMC bioinformatics*, *13*(4), 1–9.

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.

Gonzalez, D. L., Giannerini, S., & Rosa, R. (2012). On the origin of the mitochondrial genetic code: towards a unified mathematical framework for the management of genetic information. *Nature precedings*, *7*, 1–1.

Hazkani-Covo, E., Zeller, R. M., & Martin, W. (2010). Molecular poltergeists: mitochondrial dna copies (numts) in sequenced nuclear genomes. *PLoS genetics*, *6*(2), e1000834.

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.

Lascaro, D., Castellana, S., Gasparre, G., Romeo, G., Saccone, C., & Attimonelli, M. (2008). The rhnumts compilation: features and bioinformatics approaches to locate and quantify human numts. *BMC genomics*, *9*(1), 1–13.

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.

Liu, Y., & Zhao, X. (2007). Distribution of nuclear mitochondrial dna in cattle nuclear genome. *Journal of Animal Breeding and Genetics*, *124*(5), 264–268.

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.

Pozzi, A., & Dowling, D. K. (2019). The genomic origins of small mitochondrial rnas: are they transcribed by the mitochondrial dna or by mitochondrial pseudogenes within the nucleus (numts)? *Genome biology and evolution*, *11*(7), 1883–1896.

Rodríguez, F., Albornoz, J., & Domínguez, A. (2007). Cytochrome b pseudogene originated from a highly divergent mitochondrial lineage in genus rupicapra. *Journal of Heredity*, *98*(3), 243–249.

Rodríguez-Salinas, E., Riveros-Rosas, H., Li, Z., Fučíková, K., Brand, J. J., Lewis, L. A., & González-Halphen, D. (2012). Lineage-specific fragmentation and nuclear relocation of the mitochondrial cox2 gene in chlorophycean green algae (chlorophyta). *Molecular phylogenetics and evolution*, *64*(1), 166–176.

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

Simone, D., Calabrese, F. M., Lang, M., Gasparre, G., & Attimonelli, M. (2011). The reference human nuclear mitochondrial sequences compilation validated and implemented on the ucsc genome browser. *BMC genomics*, *12*(1), 1–11.

Su, C., Weir, J. D., Zhang, F., Yan, H., & Wu, T. (2019). Entrna: a framework to predict rna foldability. *BMC bioinformatics*, *20*(1), 1–11.

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.

Wang, J.-X., Liu, J., Miao, Y.-H., Huang, D.-W., & Xiao, J.-H. (2020). Tracking the distribution and burst of nuclear mitochondrial dna sequences (numts) in fig wasp genomes. *Insects*, *11*(10), 680.

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.