Re-investigating the Molecular Evidence of Muller's Ratchet with lager data-set and different statistic

Yu Mo, Aazin Asif Shaikh, Sarthak Mishra, Renu Jaiswal [1]*

Abstract
A search for molecular evidence of Mullers Ratchet has been going on in Molecular Biology and Evolutionary Biology community since very long time. Notably, Lynch1996 [1] has claimed to have found the molecular evidence of Mullers Ratchet1.1 by comparing the mitochondria tRNA with nuclear tRNA. But results of later study by Cooper et al. [2] implies quite contradictory results. We propose to re-examine the problem with a larger data-set and better statistics. In doing so we are planning to compare the summary statistics for synonymous regions of mitochondrial tRNA and with the tRNA from Nuclear genome.

Keywords
Mullers Ratchet, tRNA , Synonymous,Non-synonymous, Mitochondrial,Nuclear

[1] *Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA*

Contents

1. Problem statement

1.1 Re-examining the Molecular evidence of Mullers Ratchet

As-sexual population can undergo extinction due to irreversible mutational degradation in the absence of segregation and re-combination. This phenomenon is called Muller's Ratchet (Muller1964) [1]. We are re-examining the molecular evidence of Mullers ratchet provided in the literature.

1.2 Literature Review

1. Lynch 1996 [1] showed the molecular evidence of Mullers Ratchet in mitochondrial Genome (mtDNA). The paper compared the evolution of Transfer RNA (tRNA) in mitochondrial and nuclear genomes. It shows that the t-RNA has a higher substitution rate in Mitochondrial Genome compared to Nuclear Genome. The author has provided a variety of reasons for this difference including the physical arrangement of strands of tRNA and varying loop sizes in the structure of Mitochondrial tRNA and Nuclear tRNA.

2. Cooper et al. [2] used *Drosophila Melanogaster*'s mitochondrial genome and nuclear genome. They could not find the evidence of recombination in mitochondrial genome but found the neutrality index is weakly significantly different between the mitochondrial and nuclear loci. They have attributed this difference to a larger proportion of beneficial mutations in X-linked relative to autosomes. But they could not find any difference whatsoever in the neutrality index of Mitochondria and autosome.

3. Popadin et al. [3] found that mammalian mitochondrial genomes differ from the nuclear genomes by maternal inheritance in the absence of recombination, and

a higher mutation rate. They found Mitochondrial accumulate at least 5-folds more deleterious mutations compared to nuclear genome. This causes irreversibly degradation leading to decrease of organisms fitness in assexual lineages with low effective population size. They reach this concluding by comparing Kn/Ks in mitochondrial and nuclear regions. Kn/Ks are summary statistics similar compared to Dn/Ds but they need very strong signal to detect selection [4] also they cannot distinguish between positive and negative nonsynonymous substitution [5].

2. Data description

The data sets that we used are publised in several genomic databases First, we started with an aligned genome of 30 primates [6]. The genome is aligned with human (hg38). As a direct result of this, genes from non-human species are eliminated if they are absent from the human genome. Meanwhile, it may lead to in-dels in non-human species, located in the position where genes are present in the human genome only. In the context of this study, our primary research interest is the protein-coding and tRNA genes present in mitochondrial and autosome (non-sex chromosomes) genomes. To estimate the data size briefly, we use human as an example. The length of the mtDNA is more than 16,000 base pairs (bp), and it includes 13 protein-coding genes, 22 tRNAs, 2 ribosomal components, and little noncoding DNA. The length of the nuclear genome is more1 than 3×10^9 bp, including approximately 20,000 protein-coding genes, and over 400 tRNA genes. The total number of genes in mtDNA and autosome vary. Therefore, we anticipate less than 13 protein-coding genes and no greater than 22 tRNAs in mtDNA, and more than 500 protein-coding genes and approximately 400 tRNAs in the autosome [7].

3. Methods

3.1 Data filtering

Due to the aligned techniques employed to construct the dataset, it is not suitable to include all species in subsequent analyses. There are few overlapping genes across all species, resulting in incredibly short sequences that invalidate our test. We expect closely related species to have more overlapping genes. Another notable factor for selecting species is to avoid incomplete lineage sorting (ILS). ILS indicates the genealogical history of some genes within a group of species is not concordant with the species tree, leading to biased estimation [8].

Therefore, we selected a subset of species to eliminate the factors mentioned above. Specifically, we used the phylogeny of all species as a reference, with the Jukes-Cantor substitution model. We extracted 1 mega-base chunk consisting of several continuous blocks in a single chromosome. Then, we calculated a summary statistic called site Concordant Factor (sCF) [9] for each branch. The sCF ranges from zero to a hundred and is used to assess the level of ILS. Intuitively, a smaller sCF implies fewer sites are concordant with the inferred branch, indicating greater ILS. We would thus eliminate branches with sCF less than 90. The remaining phylogeny is utilized to estimate parameters with ML.

3.2 Parameter estimation

With protein-coding genes, we have the ability to identify synonymous and nonsynonymous substitutions. As natural selection primarily functions at the level of proteins, both mutations are fixed at vastly different rates. Thus comparison of their rates provides a means to understanding the effect of natural selection on the protein level [10]. For protein-coding genes, codon substitution models are implemented. Codon is a triplet of nucleotides that encodes a specific amino acid or a stop signal in a protein. The codon substitution model describes the relation of codons transition instead of nucleotides. To be specific, we are more interested in synonymous substitution rates (d_S) in this project. To estimate d_S, we are using a codon substitution model [11] with protein-coding regions in mtDNA and autosome respectively. Besides, we focused on the divergence of tRNAs in mtDNA and autosome, denoted as d_N . We could only use the nucleotide substitution model with tRNAs. Therefore, we estimated d_N with the Kimura 2-parameter (K80) model.

3.3 Test of selection

Inspired by McDonald-Kreitman test [12] for neutrality, we construct a two-by-two contingency table of d_N and d_S, as shown in Table 2. Assuming mtDNA has a higher evolutionary rate compared to autosome [1], and the relative evolutionary rate within mtDNA and autosome would be the same, our null hypothesis is that the ratio of d_N to d_S is identical for mtDNA and autosome. We test for significant deviations from null hypothesis with the Fisher's exact tests.

	mtDNA	autosome
tRNA	d_N^{mt}	d_N^A
protein-coding regions	d_S^{mt}	d_S^A

Table 1. Estimate for the hypothesis test.

4. Data Exploration

Reference ID	Definition
sCF	Site concordance factor averaged over 1000 quartets
s_N	Number of informative sites averaged over 1000 quartets
Length	Branch length

Table 2. Definitions used in sCF

Due to ILS, it is impossible to utilize all species for further analysis. As a matter of that, we would select taxa that are consistent with phylogeny inferred with the whole genome. A chunk with approximately 1×10^6 sites is extracted from a single chromosome to reduce the computational time. To measure the concordance, We assume the evolution process is under the Jukes-Cantor model, then infer the concordance of all internal branches with a score called site concordance factor in IQ-TREE[9]. The sCF ranges from 0 to 100. Higher sCF indicates associated branch is more concordant with the input phylogeny. Species associated with higher sCF branches are selected for the next procedure.

The set of species is shown in Fig. 1 where the left panel displays the original set of species and the right panel shows the set of species we used for analysis. The branch length represents the expected substitution rate inferred from previous study and the number above an internal branch is the associated sCF. We observed short internal branch tends to have lower sCF, consistent with the observation of ILS. Therefore, we remove *ponAbe2, cebCap1, papAnu3, aotNan1, macFas5, panTro5, HLpilTep1, saiBol1, panPan2*. Then we checked the complete mitochondrion genome for each species in GenBank, removing *cebCap1, otoGar3* due to unavailability. The sCF was re-examined with the remaining species, ensuring the metrics is greater than 60 for each internal branch.

To obtain the alignment in different categories, we used multiple existing databases. For the reference genome, we obtained the coordinates for tRNA and coding region of mtDNA with NCBI[13], coordinates of tRNA in the nuclear genome with GtRNAdb[14], coordinates of the coding region in the nuclear genome from Gencode[15]. All of tRNA from mtDNA are included and 31 tRNA from autosomes are considered. For coding genes in mtDNA, all of them are used. For coding genes in the nuclear genome, we stratified sampled over all autosomes and utilized approximately 1800 genes at first, dropping to 309 genes after quality control

	mtDNA	autosome
tRNA	22	31
coding gene	13	309

Table 3. Size of genes used in the analysis.

4.1 Github

`https://github.com/smishra677/Mol_evol`

5. Algorithm and Methodology

We have two phase to our project. In the first part we extracted the genes for mitochondrial tRNA, Nuclear tRNA, mitochondrial coding region and Nuclear coding region . In the second phase we compute the summary statistic using Phylogenetic Analysis by Maximum Likelihood (PAML). PAML is a package of programs for phylogenetic analyses of DNA or protein sequences using maximum likelihood[16]. We are using CODEML and BASEML for computing the statistic.

1. **PAML CODEML**: We are using CODEML to compute the synonymous divergence rate, *dS*, for our studies. The input is the a codon alignment file and the corresponding tree and CODEML outputs the Maximum likelihood estimate for the *dS* on each branch. It will then compute the total divergence for all the given species. This value is our *dS*. The MLE is estimated by using a substitution matrix. You can find a brief explanation of the substitution matrices in our study in the Experiments and Results Section.

2. **PAML BASEML**: Similar to CODEML , BASEML is used to compute to compute the non-synonymous divergence rate, *dN*, for our studies. The input is the a Nucleotide alignment file and the corresponding tree and BASEML outputs the Maximum likelihood estimate for the *dN* for each branch. The branches are then summed to get the height of the Tree. This height is the estimator of the divergence of the non-synonymous region. Similar to CODEML, BASEML also uses a substitution matrix. You can find a brief explanation of the substitution matrices in our study in the Experiments and Results Section.

6. Experiments and Results

We have five phases to the experiment:

1. Firstly we extracted the sequence and the phylogeny for Mammalian clade from the NCBI. Then we used IQ-Tree [11] to filter out the clade with lower site concordance factor (sCF). By doing so we are trying to increase our confidence on the results.

2. Now that we have filtered out the clades with low sCF we now find out the region of genes that are present in human genome. We are doing this to make sure we have a region of alignment which is comparable on the all the species. With the filtering and the selection of common region , we are left with 16 species. For the mitochondria coding region we were able to get 13 genes where as for nuclear coding region we are able to get 309 genes.

3. Similarly for the non-synonymous regions(tRNA) we concatenated all the genes to make a single sequence. Then we aligned it against the human genome for the same region.

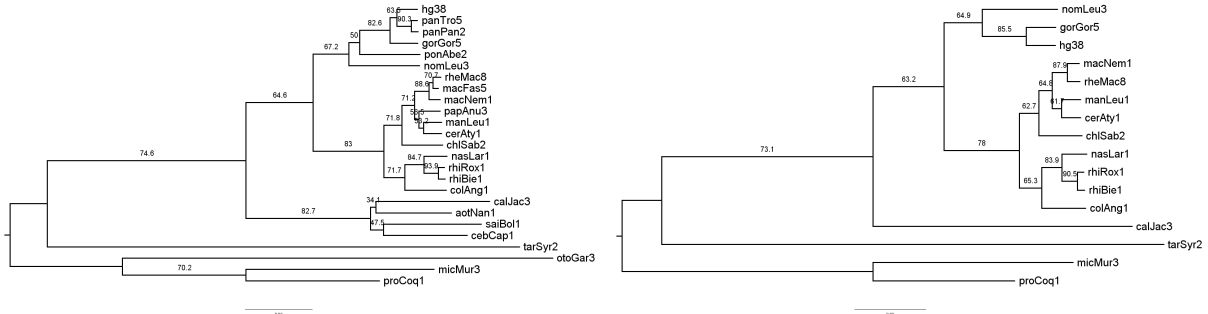

Figure 1. sCF comparison. Left panel: 27 primates are displayed in the phylogeny. Right panel: species we used in our analysis including 16 primates. Species from the right panel are a subset of ones in the left panel, removing species leading to high discordance or without a complete mitochondrion genome. The number above the internal branch is sCF.

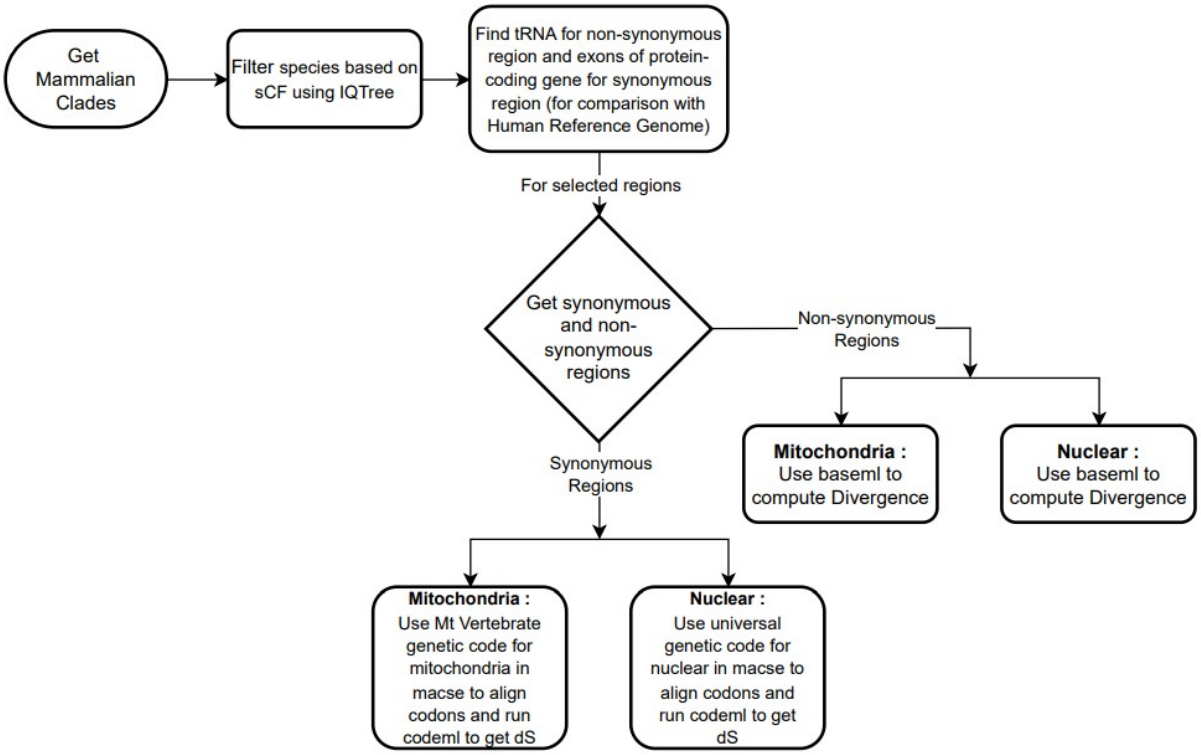

Figure 2. Project pipeline

	Nuclear	Mitochondria
dN	0.046	1.23
dS	0.68	15.8
$\frac{dN}{dS}$	0.067	0.077

Table 4. Summary statistics for 13 mitochondrial regions and 10 nuclear regions

4. We used *macse* -a program to do multiple sequence alignment [17] ,to align the synonymous region. We used universal genetic substitution[18] code to align the synonymous nuclear region, whereas the Vertebrate Mitochondrial Code [18] to align the codons of mitochondrial regions. After this we used *macse* [17] pipeline to remove the stop codons. Removal of stop codon is essential for the working of the PAML.

5. Now we run BASEML with K80 substitution model to compute the total tree height of the non-coding regions for both mitochondrial and nuclear genes. And we use CODEML on the genes processed by *macse*. This gives us an estimate of dS.

6.1 Results

We have presented our results in Table 4. Contrary to Lynch [1] our initial result suggests there is not much deviation in evolution of mitochondria region compared to the nuclear region but we need to perform the test on bigger data set.

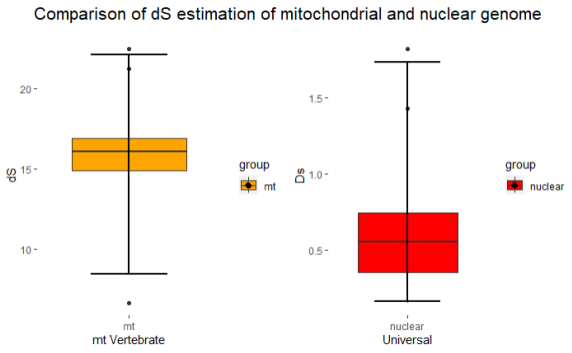

Figure 3. Comparison of dS for 13 Mitochondrial and 10 Nuclear Genome

7. Summary and Conclusions

To summarize we have extracted genes from a many regions. Then we have examined the evolutionary rate of the mitochondrial region- as-sexual lineage and compared it against the evolutionary rate of the Nuclear region- sexual region. The evolutionary rate is determined by the rate at which the non-synonymous region evolves in comparison to synonymous region. The data used in our study came from a filtered mammalian clade with 16 species. We have computed the dN/dS summary statistic on these regions. As to this point we have only computed the dS on 10 nuclear genomes. We are running for remaining genes. As of now our results shows not significant difference in the evolution of Mitochondrial region compared to Nuclear. To draw a conclusion from it will be naive as the number of nuclear genes used till this points is small. But our results are showing slight deviation from the results shown by [1] as an evidence for Muller's Ratchet.

Acknowledgments

We thank Matthew Hahn and Yadira Peña-Garcia for helpful discussion.

References

[1] M Lynch. Mutation accumulation in transfer RNAs: molecular evidence for muller's ratchet in mitochondrial genomes. *Mol Biol Evol*, 13(1):209–220, January 1996.

[2] Brandon S Cooper, Chad R Burrus, Chao Ji, Matthew W Hahn, and Kristi L Montooth. Similar Efficacies of Selection Shape Mitochondrial and Nuclear Genes in Both Drosophila melanogaster and Homo sapiens. *G3 Genes—Genomes—Genetics*, 5(10):2165–2176, 10 2015.

[3] Konstantin Yu Popadin, Sergey I Nikolaev, Thomas Junier, Maria Baranova, and Stylianos E Antonarakis. Purifying selection in mammalian mitochondrial protein-coding genes is highly effective and congruent with evolution of nuclear genes. *Mol Biol Evol*, 30(2):347–355, September 2012.

[4] Z Yang and J P Bielawski. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol*, 15(12):496–503, December 2000.

[5] Eduardo P.C. Rocha, John Maynard Smith, Laurence D. Hurst, Matthew T.G. Holden, Jessica E. Cooper, Noel H. Smith, and Edward J. Feil. Comparisons of dn/ds are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology*, 239(2):226–235, 2006. Special Issue in Memory of John Maynard Smith.

[6] 30 primates alignment. `http://hgdownload.cse.ucsc.edu/goldenpath/hg38/multiz30way/`. Accessed: 2010-09-30.

[7] Jamie A Abbott, Christopher S Francklyn, and Susan M Robey-Bond. Transfer rna and human disease. *Frontiers in genetics*, 5:158, 2014.

[8] Fábio K Mendes, Jesualdo A Fuentes-González, Joshua G Schraiber, and Matthew W Hahn. A multispecies coalescent model for quantitative traits. *eLife*, 7:e36482, jul 2018.

[9] Bui Quang Minh, Matthew W Hahn, and Robert Lanfear. New methods to calculate concordance factors for phylogenomic datasets. *Molecular biology and evolution*, 37(9):2727–2733, 2020.

[10] Ziheng Yang. *Computational molecular evolution*. OUP Oxford, 2006.

[11] Ziheng Yang and Rasmus Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular biology and evolution*, 17(1):32–43, 2000.

[12] John H McDonald and Martin Kreitman. Adaptive protein evolution at the adh locus in drosophila. *Nature*, 351(6328):652–654, 1991.

[13] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 35(suppl_1):D5–D12, 2007.

[14] Patricia P Chan and Todd M Lowe. Gtrnadb 2.0: an expanded database of transfer rna genes identified in complete and draft genomes. *Nucleic acids research*, 44(D1):D184–D189, 2016.

[15] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774, 2012.

[16] Ziheng Yang. Phylogenetic analysis by maximum likelihood (paml), 2000.

[17] Vincent Ranwez, Sébastien Harispe, Frédéric Delsuc, and Emmanuel J. P. Douzery. Macse: Multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLOS ONE*, 6(9):1–10, 09 2011.

[18] Andrzej (Anjay) Elzanowski and Jim Ostell. The genetic codes.