Analysis of mitochondrial codon usage

Shelina Begum
University at Albany, State University of New York, sbshelina.begum@gmail.com

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Biology Commons

Recommended Citation
https://scholarsarchive.library.albany.edu/legacy-etd/2862

This Master's Thesis is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive. Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
Analysis of Mitochondrial Codon Usage

By
Shelina Begum

A Thesis
Submitted to the University at Albany, State University of New York
in Partial Fulfillment of
the Requirements for the Degree of
Master of Science

College of Arts and Sciences
Department of Biological Sciences
2022
# Table of Contents

1. Abstract ........................................................................................................................................... iii

2. Introduction ......................................................................................................................................... 1

3. Materials and Methods .................................................................................................................. 4

4. Results and Discussion, Figures ..................................................................................................... 7
   - Human Mitochondrial Codon Usage ............................................................................................... 7
   - Mouse Mitochondrial Codon Usage .............................................................................................. 8
   - Yeast Mitochondrial Codon Usage ................................................................................................. 9
   - Fruit Fly Mitochondrial Codon Usage .......................................................................................... 10
   - Comprehensive Mitochondrial Codon Usage ................................................................................ 11

5. Conclusions ....................................................................................................................................... 13

6. Next Steps ......................................................................................................................................... 15

7. Tables ............................................................................................................................................... 16

8. Appendix Materials ....................................................................................................................... 16

9. Reference .......................................................................................................................................... 17
1. Abstract

The human mitochondrial genome is circular and encodes for its own two ribosomal RNAs, 22 tRNAs and 13 proteins needed for oxidative phosphorylation. Although the mitochondria require nuclear proteins and machinery for transcription and translation, the translation of 13 mRNAs occurs within the mitochondria. The dysregulation of mitochondrial translation is associated with diseases like MELAS (Mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes) and MERRF (Myoclonic epilepsy with ragged-red fibers), which can be caused by tRNA defects and lead to decreased translation of mitochondrial complex I proteins (MELAS).

We have analyzed the codon usage of human mitochondrial genes and compared them to themselves and to nuclear counterparts using normalized Z-scores. We have determined that the codon usage in mitochondrial gene MT-ND6 (NADH dehydrogenase 6), a subunit of complex I, is opposite to the codon usage of all other mitochondrial genes. Codon analytics of mouse mitochondrial genes was also performed, and we observed a similar trend of opposing codon usage relative to other mitochondrial genes in mouse MT-ND6. When comparing to nuclear encoded genes, both mouse and human mitochondrial genes overused the codon LeuCTA and underused GlnCAG. When comparing to nuclear genome, all four organisms (human, mouse, yeast, and fruit fly) mitochondrial genes had certain codons that were overused or underused across all genes. Understanding both the baseline codon usage patterns and codon bias in the mitochondria can be used to model mitochondrial translation and understand how tRNA defects could cause disease, and potentially be used to develop targeted therapeutics for patients with mitochondrial diseases.
2. Introduction

The central dogma of biology is that DNA is transcribed into RNA and RNA is translated into proteins. At each stage of transcription and translation, DNA, RNA, proteins can modified to regulate the process. Modification of RNAs is known as the epitranscriptome and these modifications can range from being simple methylation to complex modifications carried out through multiple enzymatic steps. They can determine when and how much of translation occurs without changing the core ribonucleotide sequence, allowing for post transcriptional control of gene expression by effecting the stability, folding, and processing RNAs. Out of them all, tRNAs are the most abundantly modified RNAs and play a significant role in the codon-anticodon interactions during translation.¹

During translation, the mRNA is “read” by the ribosome in 3-letter codons and the transfer RNA (tRNA) carries and incorporates amino acids into the growing polypeptide chain. With 64 codons encoding for only 20 amino acids, there are multiple codons that encode for the same amino acid using isoacceptor tRNAs (tRNAs that have different anticodons but carry the same amino acid). In the anticodon loop of a tRNA there are 3 anti-codons at the 34, 35, and 36th position that are complimentary to the codons on mRNA. These codon and anticodon follow Watson-Crick base pairing rules; Cytosine pairs with Guanine and Uracil pairs with Adenine. Under different circumstances, however, the tRNAs can interact with codons that are outside of the predetermined genetic code.¹⁻³

Modifications on the RNA are regulated by epitranscriptomic readers, writers, and erasers. Writers are enzymes that add a mark on the RNA, erasers are enzymes that
take away the marks, and readers are enzymes that recognize the marks to facilitate a response. These modifications can occur on the D loop, T loop, as well as the anticodon loop of the tRNA. In tRNAs, the modifications in the anticodon loop as on well on the other distal sites of the tRNA affect the interactions between codons and anticodons. The method by which the modifications affect the interactions vary from modification to modification and well as the position in which the modification occurs. Similar to the nuclear genome, the mitochondrial genome also exhibits a wide range of modifications that affect translation.

Mitochondria have their own separate genome which encodes for its own genes. It encodes for two ribosomal RNAs, 22 tRNAs and 13 proteins for the oxidative phosphorylation system. Despite having their own genome, some protein components like translational factors and factors required for processing and modifying RNAs are encoded in the nuclear genome and then transported in. Unlike the linear nuclear DNA, mitochondrial DNA is circular with two strands: the heavy strand on the outside and a light strand on the inside. Certain organisms (like mouse and human) do not have introns in the mitochondrial genome. In human mitochondria, all mRNAs except for one (ND6) also go through 3’ polyadenylation. The mitochondrial genetic code is different from the nuclear genetic and also varies from organism to organism. For human mitochondria, there are more than one start codon and well as more stop codons than nuclear (Table 1).

Transcription and translation are both also different in the mitochondria, for one the DNA is transcribed as polycistronic transcripts. For human mitochondria the 3 transcribed mRNAs are: the H1 H2 and LSP. H1 contains the genes for the
phenylalanine and leucine tRNA, H2 contains majority of tRNAs, two ribosomal subunits, and 12 out of 13 protein coding genes, LSP (on the light strand) only contains the protein coding ND6. Structurally, the mouse mitochondrial genome is very similar to the human mitochondrial genome; it is circular and also has 13 protein coding genes, with only two polycistronic RNAs. The yeast mitochondrial genome is also circular, however it contains introns similar to the nuclear genome. Unlike mouse and human, yeast mitochondria encode for eight proteins as opposed the 13 and have 11 polycistronic RNAs. Similar to the rest, fruit flies also have circular mitochondrial DNA but only have five polycistronic RNAs (Table 2).

Despite having a separate genome, the dysregulation of mitochondria is associated with various diseases that can affect nearly every organ in the body, including age related metabolic diseases and degenerative diseases and cancer. The most notable mitochondrial diseases are Mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes (MELAS) and Myoclonic epilepsy with ragged red fibers (MERFS), and according to previous studies approximately 80% of MELAS cases have a mutation on the MT-TL1 gene which encodes for the leucine tRNA (decodes UUA/UUG). 5-taurino methyluridine modification is a tRNA modification at uridine 34 of leucine and tryptophan. The two known writers for this taurine modification are MTO1 and GTPBP3; MTO1 is a mitochondrial protein called tRNA translation optimization 1 and GTPBP3 is a GTP binding protein localized in the mitochondria. In the mitochondria, to decrease the number of tRNAs required, in general an unmodified uridine can read any on the 4 bases at position 34, which is known as 4 way wobbling
or super wobbling. With the taurine modification it restricts decoding the third position to purines only.\textsuperscript{21}

In MELAS patients the taurine modification is rare or nonexistent in the mitochondrias of the affected tissues.\textsuperscript{18} In one study, researchers used what they called an operated tRNA (i.e., tRNA without the modification, which is common in MELAS patients) to demonstrate that without the taurine modification, the codon-anticodon interaction for UUG was very weak in comparison to the UUA codon and therefore this modification was crucial for UUG decoding. As these modifications can affect the codons being translated.\textsuperscript{22}

I have analyzed codon usage in mitochondria of 4 different organisms (human, mouse, yeast, fruit fly) and found that there is distinct codon usage for the mitochondrial gene ND6 in both human and mouse, and identified codons that are overused or underused across all genes in all organisms. Knowing the baseline or standard frequency of codons can be used to figure out if there is codon biased translation in the mitochondria and can be compared to the codon usage in mitochondria of diseased patients.

3. Materials and Methods

Mitochondrial Genomic Data

Nucleotide reads for human (\textit{Homo sapiens}; locus: \url{NC_012920}), mouse (\textit{Mus musculus}; locus: \url{NC_005089}), fruit fly (\textit{Drosophila melanogaster}; locus: \url{NC_024511}), and yeast (\textit{Saccharomyces cerevisiae}; locus: \url{NC_001224}) were gathered from NCBI Nucleotide database. Nucleotide sequence files were downloaded for individual genes
and for incomplete sequences, stop codons for multiple genes were added according to the comments on NCBI page (i.e. “TAA stop codon is completed by the addition of 3' A residues to the mRNA”). This process was done on Microsoft Word and the complete sequences were copied into Microsoft Excel.

**Excel Macro**

All genomic sequence were compiled on excel similar to previous studies for nuclear genome. Excel macro codes from the past studies were used and modified according to each organism's mitochondrial genetic code to calculate codon count, codon frequencies, and z-score. 

**Codon analytics**

Individual codon counts were calculated using macro; the individual codon triplets were all assigned a value of “0” and as the program read through the sequence in 3’s it added the count to the Excel cell the codon was assigned to. Codon frequency was calculated by counting the number of times a specific codon of an amino acid was used (codon count, i.e Ala$^{GCA}$) and dividing by the number of times the amino acid was used in a gene (count sum of all the codons of an amino acid, i.e Ala$^{GCA}$ + Ala$^{GCC}$ + Ala$^{GCG}$ + Ala$^{GCT}$).

$$Codon\ frequency = \frac{target\ codon\ count}{total\ codon\ count\ for\ amino\ acid}$$

Normalized z-score-mitochondria was calculated by subtracting codon frequency minus the average codon usage of the target codon in the mitochondria and dividing by standard deviation of the target codon.
\[ z - \text{score} = \frac{\text{codon frequency} - \text{average usage of target codon}}{\text{standard deviation of target codon}} \]

For human and mouse, z-score-nuclear was also calculated in relation to the nuclear genome. The codon frequency from mitochondrial genome was used against the average and standard deviation calculated for nuclear genome in a different study.

\[ z - \text{score} = \frac{\text{mitochondrial codon frequency} - \text{nuclear average usage of target codon}}{\text{nuclear standard deviation of target codon}} \]

**Heatmap**

Morpheus, an online matrix visualization and analysis software from Broad Institute, was used to generate heatmaps to visualize z-score distribution. Excel files with z-scores were uploaded to the website and used hierarchical clustering analysis (One minus Pearson correlation) was used. The hierarchical clustering algorithm finds the similarity and dissimilarity between data and groups (or clusters) them together based on their similarity. Information about their methods and algorithms for hierarchical clustering and Pearson correlation can be found here and here, respectively.\textsuperscript{24,25}
4. Results and Discussion

Human

Codon frequencies and z-scores for the 13 protein coding sequences were calculated, the z-scores were plotted as a heatmap to visualize whether certain codons were overused or underused compared to their isoacceptor codons. In human mitochondria, codon usage of ND6 was opposite of all other genes: ND6 was found to overuse codons that were generally underused in the remaining 12 genes, and vice versa. (Fig. 1A). This was the case for when compared to the codon usage of both mitochondrial and nuclear genomes. When compared to the nuclear genome, GlnCAA, LeuCTA, and LysAAA are overused and GlnCAG, LysAAG, and ValGTG are underused across all genes (Fig. 1B).

**Figure 1.** Human Mitochondrial Codon Usage

---

**Fig. 1A.** Codon usage in the 13 human mitochondrial genes based on Z-scores-mitochondria. No codons are universally overused or underused across all genes. Gene ND6 has a distinct codon usage pattern compared to the remaining 12 genes.

**Fig. 1B.** Codon usage in the 13 human mitochondrial genes based on Z-scores in reference to the human nuclear genome (z-score-nuclear). Codons GlnCAA, LeuCTA, and LysAAA are most overused and GlnCAG, LysAAG, and ValGTG are most underused. Gene ND6 has a distinct codon usage pattern compared to the remaining 12 genes.
Mouse

Codon frequencies and z-scores for the 13 protein coding sequences were calculated, the z-scores were plotted as a heatmap. Similar to the codon usage in the ND6 gene in human, the codon usage of ND6 in mouse mitochondria also showcased a similar pattern; the codons that were overused in ND6 were generally underused in the remaining 12 genes, and vice versa (Fig. 2A). Mouse mitochondria also had certain codons that were either overused or underused across all genes, save for ND6. GlnCAG LysAAG and GluGAG being the most underused while MetATA LeuCTA and GlnCAA were the most overused. A similar pattern was observed when codon usage was compared to the nuclear genome (Fig. 2B).

Figure 2. Mouse Mitochondrial Codon Usage

| Fig. 2A. Codon usage in the 13 mouse mitochondrial genes based on Z-score-mitochondrial. GlnCAG LysAAG and GluGAG are the most underused while MetATA LeuCTA and GlnCAA are the most underused. Gene ND6 has a distinct codon usage pattern compared to the remaining 12 genes. |
| Fig. 2B. Codon usage in 13 mouse mitochondrial gene based on Z-scores in reference to the mouse nuclear genome (z-score-nuclear). Codons GlnCAG, LysAAG, GluGAG are most underused while MetATA and LeuCTA are most overused. Gene ND6 has a distinct codon usage pattern compared to the remaining 12 genes. |
Yeast

Unlike mouse and human, yeast mitochondria encode for eight proteins only. Two arginine codons (ArgCGA, ArgCGC), two End codons (EndTAG and EndTAA), three serine codons [SerTCG, SerTGG, SerAGC], and three threonine codons (ThrCTG, ThrCTC, ThrACG) are not utilized at all (Fig. 3A). VAR1, a mitochondrial ribosomal protein, was found to have a small set of codons that are overused in comparison to the remaining 7 genes, and a small set of codons that were underused in comparison (Fig. 3A). When compared to the nuclear genome (LeuTTA) are most overused and (LysAAG, LeuTTG, ArgAAG) are most underused across all genes (Fig. 3B).

**Figure 3. Yeast Mitochondrial Codon Usage**

![Fig. 3A. Codon usage in 8 Yeast mitochondrial genes based on Z-score mitochondrial. No codons are universally overused or underused across all genes. Two arginine codons, two End codons, three serine codons and three threonine codons are not utilized at all. Ribosomal protein VAR1 has a distinct codon usage in comparison to the 7 other genes.](image)

![Fig. 3B. Codon usage in 8 Yeast mitochondrial gene based on Z-scores in reference to the human nuclear genome (Z-score-nuclear). LeuTTA and SerTCA are most overused and LysAAG, LeuTTG, ArgAAG are most underused across all genes.](image)
Fruit fly

Similar to yeast, *Drosophila* had multiple codons that are not used at all: one arginine (ArgCGC), two serine (SerAGG, SerAGC) and one leucine (LeuCTG) (**Fig. 4A**). When compared to nuclear genome (LeuTTA) is the most overused and (GlnCAG, LysAAG, GluGAG) are the most underused across all genes (**Fig. 4B**).

**Figure 4.** Fruit fly Mitochondrial Codon Usage

**z-score**

**Fig. 4A.** Codon usage in the 13 Fruit fly mitochondrial genes based on Z-score-mitochondrial. One arginine (ArgCGC), two serine (SerAGG, SerAGC) and one leucine (LeuCTG) are not used by any genes. No codons are universally overused or underused across all genes.

**z-score**

**Fig. 4B.** Codon usage in 13 Fruit fly mitochondrial gene based on Z-scores in reference to the human nuclear genome (z-score-nuclear). LeuTTA is the most overused and GlnCAG, LysAAG, and GluGAG are the most underused across all genes.
Comprehensive

A comprehensive heatmap of all the genes from the four organisms was created to visualize the usage of codon triplets. As the organisms have different genetic codes, only the three-letter codons were used and not which amino acid they encode for. The heatmaps were created using the z-scores of the codon triplet in every gene of all four organisms (i.e, z-score-mitochondria of CGA across all genes of all four organisms). A cluster of overused and underused codons from the mouse genome stand out as well as the human and mouse ND6 gene. (Fig. 5A). The comprehensive heatmap of mitochondrial codons in reference to nuclear genome shows distinct trends. There are codon triplets that are underused across all four organisms (CAG, CTG, AAG, AGG, GAG, GTG, TTG, TGG), underused only in fruit flies and yeast, and codons that are either over used only in mouse and human, or only in yeast and fruit flies (Fig. 5B).

Figure 5. Mitochondrial Codon Usage for All Four Organisms

![Figure 5A](image-url) Comprehensive heatmap of mitochondrial codon usage of all 4 organisms based on z-score-mitochondrial. A cluster of overused and underused codons in mouse genes can be seen as well as the distinct codon usage of ND6 in human and mouse.
Fig. 5B Comprehensive heatmap of mitochondrial codon usage of all 4 organisms when compared to nuclear genome (z-score-nuclear). There are codons that are (1) underused across all 4 organisms (2) underused only in fruit flies and yeast and (3) codons that are either over used only in mouse and human or (4) only in yeast and fruit flies.
5. Conclusions

In human and mouse, ND6 (NADH dehydrogenase 6, a subunit of complex I) exhibits distinct codon usage. The codons that are overused in the remaining 12 genes are underused in ND6, and vice versa (Fig 1. And Fig. 2). This was seen when compared to nuclear genome as well as mitochondrial genome, although it is not as pronounced in comparison to the nuclear genome. This could be due to the fact that for both human and mouse the ND6 gene is the only non-tRNA gene on the light strand. As mentioned previously, both human and mouse mitochondria genes are transcribed in 3 and 2 polycistronic strands, respectively. For both organisms, the ND6 gene is on the light strand of the genome and is transcribed by the light strand promoter. It is also the only gene that does not get polyadenylated and therefore gets processed differently.\textsuperscript{26,27} Being from a different strand and being the only mRNA to not get polyadenylated could be a possible reason why ND6 showcases distinct codon usage. Unlike the human mitochondria, the codon usage (in reference to mitochondrial genome) in mouse mitochondria either overuses or underuses certain codons across all genes, save for ND6 (Fig. 2A). Interestingly, in mouse, the same set of codons were overused and underused even when they were being compared to the nuclear average.

The codon usage in mammals (human and mouse) was also distinct when compared to the codon usage in the other organisms (yeast and fruit fly). One of the genes in the mitochondrial oxidative phosphorylation system (ND6) in both human and mouse showcased opposite codon usage from the remaining genes. However, the ND6 gene in fruit flies did not have any distinct codon usage (Fig. 4). Yeasts, on the other hand, have a smaller mitochondrial protein coding gene count and do not have ND6 or
a homolog of ND6. Out of the four organism analyzed in this study, yeast is also the only organism that have a ribosomal protein coding gene (along with the oxidative phosphorylation system). When compared to its own mitochondrial genome, this gene (VAR1) had a distinct codon usage where its codon usage was opposite the remaining 7 genes. The yeast and fruit fly mitochondrial genomes also have codons that are not utilized in any genes (Fig. 3 and Fig. 4). Yeast have upwards of 8 codons that encode for the same amino acid, and it is possible such usage pattern can contribute to a clear bias towards some codons over others. This can also be seen in Fig. 5B, where codon triplets’ usage (in reference to human nuclear genome) was compared across all four organisms; there are codon triplets that underused only in fruit flies and yeast, triplets that are overused only in mouse and human, or overused only in yeast and fruit flies.

When comparing codon usage in nuclear genomes across all the genes from all four organisms, there were certain codon triplets that were There are certain codons that are overused or underused across all genes and organisms. Codon triplets CAG, CTG, AAG, AGG, GAG, GTG, TTG, TGG are underused while CAA, AAA, GGA, GTA, GAA, are overused (Fig. 5B). When comparing these codon triplets (as well as other triplets are mostly overused or underused, with the exceptions of a few genes), it is clear that most underused codon triplets are G-ending while most overused codons triplets are A-ending. In a previous study done by Xu Wenjing et.al. looking at plant mitochondria, there was a preference of A and T ending genes and suggested it could be due to the AT richness in plant mitochondrial genome. This could very well be the reason for the overuse of A-ending codons in mitochondrial genome. There have been other studies that concluded that G/C ending codons are more preferential due to the
GC-rich content in genomes. Nuclear genome is GC-rich (G/C ending codons are preferential) and therefore it stands to reason that the G ending codons in mitochondrial will be underused when compared to the nuclear genome.

6. Next Steps

This project focused on the codon usage of the standard mitochondrial genome sequence of each organism. For future analysis, raw mitochondrial sequences from MELAS (or other mitochondrial disease) patients can be gathered and frequency and z-scores can be calculated in a similar fashion to this study. Since there have been studies done to show that modifications in tRNAs that are an integral part of diseases like MELAS, and that such modifications alter codon usage, this could be used to compare normal codon usage against that of diseased patients.

As this was a computational project to only analyze and compare the codon usage in different organisms, it does not delve why certain codons are overused or underused. As all the data for this project were collected from an NCBI database, a possible next step could be to look at published research where mitochondrial data is collected from samples of different organisms and look at codon abundance, codon frequencies, and other similar matrix to see if the patterns of those correlate with the patterns seen in this study (i.e. are all four organisms from this study AT-rich? Could that be a reason for the overuse of T-ending codons?)

One of the main highlight of this project has been the codon usage patterns of ND6. Although it is possible that being on the light strand and not being polyadenylated could be the reason for this distinct codon usage, looking at codon usage from more
organisms could further verify this hypothesis. This study could be extended to look
codon usage from a larger number of mammals and non-mammals to verify if ND6 (or
its homologs) have this distinct codon usage in mammals only.

7. Tables

<table>
<thead>
<tr>
<th>Codon</th>
<th>Standard</th>
<th>Mito</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATA</td>
<td>Ile</td>
<td>Start</td>
</tr>
<tr>
<td>AGT</td>
<td>Ser</td>
<td>Start</td>
</tr>
<tr>
<td>ATT</td>
<td>Ile</td>
<td>Start</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Codon</th>
<th>Standard</th>
<th>Mito</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGA</td>
<td>Arg</td>
<td>Stop</td>
</tr>
<tr>
<td>AGG</td>
<td>Arg</td>
<td>Stop</td>
</tr>
<tr>
<td>TAA</td>
<td>Stop</td>
<td>Stop</td>
</tr>
<tr>
<td>TAG</td>
<td>Stop</td>
<td>Stop</td>
</tr>
</tbody>
</table>

Table 1. Genetic code difference between human nuclear genome (standard) and human mitochondrial genome (Mito)

<table>
<thead>
<tr>
<th>DNA shape</th>
<th>Human</th>
<th>Mouse</th>
<th>Yeast</th>
<th>Fruit fly</th>
</tr>
</thead>
<tbody>
<tr>
<td>circular</td>
<td>circular</td>
<td>circular</td>
<td>circular</td>
<td>circular</td>
</tr>
</tbody>
</table>

| Total number of genes | 37 | 37 | 35 | 37 |
| Proteins Coding Sequence | 13 | 13 | 8 | 13 |
| Polycistronic transcripts | 3 | 2 | 11 | 5 |
| Introns | No | No | Yes | No |

Table 2. Information about human, mouse, yeast, and fruit fly mitochondrial genome.

8. Appendix Material: Mitochondrial gene sequences, codon count and codon frequency, Mitochondrial Z-score relative to mitochondria, Mitochondrial Z-score relative to nuclear genome, and heatmaps for all 4 organisms can be found can be found in the link below:
https://www.dropbox.com/sh/ghz7nwatour6uif/AAC5oxi_kPRWIWsYFj4eCVCla?dl=0
References


24. Morpheus by Broad Institute, https://software.broadinstitute.org/morpheus


