Comparative analysis of mitochondrial genomes in different species reveals the evolution of G-quadruplexes

Hiroshi Sugiyama (✉ hs@kuchem.kyoto-u.ac.jp)  
Kyoto University  https://orcid.org/0000-0001-8923-5946

Vinodh Sahayasheela  
Kyoto University

Zutao Yu  
Kyoto University

Ganesh Pandian  
Kyoto University

Biological Sciences - Article

Keywords:

Posted Date: December 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1126357/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Comparative analysis of mitochondrial genomes in different species reveals the evolution of G-quadruplexes

Vinodh J Sahayasheela\textsuperscript{a}, Zutao Yu\textsuperscript{a}, Ganesh N. Pandian\textsuperscript{b,*}, Hiroshi Sugiyama\textsuperscript{a,b,*}

\textsuperscript{a} Department of Chemistry, Graduate School of Science, Kyoto University, Sakyo, Kyoto 606-8502, Japan

\textsuperscript{b} Institute for Integrated Cell-Material Science (WPI-iCeMS), Kyoto University, Sakyo, Kyoto 606-8501, Japan

Abstract

G-quadruplexes (G4s) are noncanonical structures that can form in the genomes of a range of organisms and are known to play various roles in cellular function. G4s can also form in mitochondrial DNA (mtDNA) because of their high guanine content, and these G4s may play roles in regulating gene expression, DNA replication, and genome stability. However, little is known regarding the evolution and dissemination of G4s in mitochondria. Here we analyzed the potential G4-forming sequences in mtDNA of 16 species from various families and demonstrated that the heavy strand of mtDNA of higher-order organisms contained higher levels of G4 regions than that of lower-order organisms. Analysis of the codons in the light strand revealed enrichment of guanine/cytosine-rich regions in higher eukaryotes and of adenine/thymidine-rich regions in lower-order organisms. Our study showed the diversity of G4s in species ranging from lower to higher orders. In particular, mammals such as humans, chimpanzees, and monkeys display a greater number of G4s than lower-order organisms. These potentially play a role in a range of cellular functions and assist in the evolution of higher organisms.

Introduction

Mitochondria are known as the powerhouse of the cell because of their ability to generate adenosine triphosphate, which is essential for normal cellular function. They also play a vital part in various cellular functions including calcium homeostasis, apoptosis, stem cell generation, and heme...
Most eukaryotic mitochondrial DNAs (mtDNAs) are double-stranded, circular molecules that typically occur at several hundreds to thousands of copies per cell. In the majority of eukaryotes, nearly 90% of mtDNA comprises coding regions, unlike nuclear DNA, and its genetic code differs slightly from that of nuclear DNA.

Most mammalian mtDNAs encode 37 genes, including 13 genes that form the essential subunits of the mitochondrial respiratory chain complexes, while the organism’s remaining genes are encoded by the nuclear genome.

MtDNA consists of two strands that can be distinguished by their nucleotide composition and are termed heavy (H) and light (L) strands. Typically, mtDNA strands are separated on the basis of density using the classical biochemical technique ultracentrifugation, producing the H strand, which has a high guanine (G) + thymidine (T) content, and the L strand, which contains low G + T content. The number of coding sequences in the H and L strands is often confused, with the H strand being assumed to contain a large number of coding sequences and L strand fewer. However, a detailed analysis of most vertebrate mitochondrial sequences demonstrated that the L strand is the main coding strand.

Mitochondria are proposed to have evolved from endosymbiotic bacteria, and phylogenetic analysis confirms that the lineage of the mtDNA is closely related to that of alphaproteobacteria. This has raised a number of intriguing questions for the fields of mitochondrial and evolutionary research, such as how mitochondria integrate and adapt within the host, and whether mitochondria played an important role in the transition from prokaryote to eukaryote. The advent of the genomics era has allowed researchers to try to solve these mysteries.

Nucleic acids are known to form structures other than the traditional double helix. A G-quadruplex (G4) is a stable secondary structure of nucleic acid that can arise from single-stranded G-rich DNA and RNA sequences. G4s are formed by Hoogsteen hydrogen bonding between four Gs to form a planar G-tetrad. They exhibit extremely high stability under physiological conditions in the presence of monovalent metal cations (such as Na⁺, K⁺, and Li⁺) and are resistant to degradation by nucleases. G4 regions in the nuclear genome form in G-rich sequences and have been visualized in human cells. The formation of such G4s plays a role in various biological functions such as transcription, DNA replication, DNA damage repair, and telomere maintenance. Most importantly, many studies have shown that G4 structures are prevalent in the promoter regions of genes and play an important role in the regulation of gene expression. Many in vitro techniques and
computational analysis methods have been developed to explore the background characteristics of G4 sequences and their distribution in the genome of various organisms. One study mapped the formation of G4s in the DNA of multiple species of bacteria, plants, and eukaryotes including human, mouse, and Drosophila, validating the vast majority of G4 structures predicted by computational analysis and demonstrating the existence of G4s in the genome of a range of species.  

The role of DNA secondary structures and their contribution to the evolution of species is poorly explored. One study comparing the genomes of multiple species showed that DNA secondary structures appeared in genomes from lower organisms and gradually evolved from lower to higher organisms. It is hypothesized that the increase in G4 structures might facilitate the development of new gene regulatory mechanisms to achieve increasingly complex cellular, physiological, and behavioral activities. Another study using a comparative bioinformatic analysis of seven species of Saccharomyces revealed that the G4 structures are relatively conserved throughout its evolution. MtDNA is known to form more G4 secondary structures than genomic DNA because of the high G content of the H strand, but the evolution of G4s in mitochondria from lower- to higher-order organisms has not been investigated.

In this study, we performed a sequence-wide predictive analysis of G4 motifs in the mtDNA of different species ranging from unicellular to multicellular species and analyzed their evolutionary pattern. We found that the number of G4 DNA motifs increased from lower- to higher-order organisms, suggesting that mitochondria have evolved over time. The increase in the number of G4s with increasing complexity of the organism suggests that they have a potential functional role in facilitating this increasing complexity.

Results and Discussion

Identification of G4s in mitochondria of multiple species

Using the QGRS mapping tool, we identified G4 motifs in the mtDNA of 16 different species belonging to different families ranging from unicellular to multicellular. Most mitochondrial genomes are between 14 and 17 kb, although the size of the plasmodium mtDNA is relatively small at 6 kb. We analyzed both the H and L strands and found that most of the G4-forming sequences lie
on the H strand because of its greater G content. The mitochondrial reference sequences for different species were downloaded from the University of California Santa Cruz genome browser, and a phylogenetic tree was generated by performing multiple-sequence alignment using ClustalW followed by maximum likelihood analysis. The comparative analysis of the mitochondrial sequences of organisms from Saccharomyces to *Homo sapiens* showed an increasing number of G4s in the H strand of higher-order organisms (Figure 1). The number of G4 motifs ranged from a minimum of two in Drosophila to a maximum of 172 in *H. sapiens*.

Figure 1. A phylogenetic tree based on the mitochondrial sequences of the 16 species together with their mitochondrial genome size, GC content, and the number of potential G4-forming sequences (PQS) in the heavy (H) and light (L) strands.

G4s are widely present in the telomere and gene promoter regions of the nuclear genome, and play an important role in gene transcription and regulation. Earlier studies showed no clear correlation between the number of G4-forming regions and the complexity of the organism. However, an analysis of the location of the G4-forming regions in various organisms identified an
enrichment from lower- to higher-order organisms of G4 motifs in transcription start sites.\textsuperscript{15} It was suggested that this characteristic plays a key role in various epigenetic mechanisms and helps the organism to evolve to meet physiological and biological challenges. The analysis of G4-forming sites in the mtDNA of various species showed an increase in G4 motifs with greater complexity of the organism (Figure 2b). The GC content was also found to increase with the complexity of the organism, suggesting that it correlates with increased G4 formation and plays a role in evolution (Figure 2a). We then evaluated how the individual nucleobases and GC content influence the increase in G4 by considering the average measurement in all species. We found out that PQS was positively influenced by the G richness (R=0.9425) and GC content (R=0.8557) using pearson and spearman correlation coefficient (Figure S1) while other factors such as adenine(R=-0.8814), cytosine(R=-0.4666) has negatively influenced the formation of G4 and cytosine has very small effect(R=-0.1568).

Figure 2 a) The GC content of the mitochondrial genomes of the 16 species. b) Potential G4-forming motifs in the mitochondrial heavy strand of the 16 species. The x-axis represents the ranking of organisms based on their complexity by phylogeny analysis, going from lower- to higher-order organisms.
Potential role of mtG4s in different species

The G4 distribution in the mtDNA of the 16 species from yeast to mammals was plotted with respect to their genome size (Figure 3). The circos plot shows the varying GC content of the organism also the GC skew was very high in the higher organism evidently from the plot. The density of the G4 motifs in these species indicated an even distribution in the mitochondrial genome, but they were clustered densely in the higher mammalian species with a stepwise reduction to lower organisms. Interestingly, most of the G4s are located on the H strand, which acts as a template for very few genes, rather than the L strand, which has a high C content and acts as the template for most genes. Mitochondria are well known to accumulate reactive oxygen species (ROS), which are known to oxidize DNA. Of the four DNA bases, G is most frequently oxidized (to 8-oxo-7,8-dihydroguanine [8-OG]) by ROS because of its low redox potential. Because G4s are composed of several runs of Gs, they become an easy target for ROS that can affect DNA stability and in turn can modify cellular processes. The increase of G4 in the H strand of mtDNA may act as a cellular stress to control mitochondrial replication. The formation of 8-OG in G4s has several implications: it is known to act as an on/off switch for transcription and may play an important role in the processes of mitochondrial replication and repair.
Human mtDNA contains 37 genes coding for two ribosomal RNAs, 22 transfer (t)RNAs, and 13 polypeptides. All 37 genes have homologs in most other species, although the arrangement of the genes varies between species. Mitochondrial translation differs from the cellular process, the key features being a reduced number of tRNAs and a reassignment of codons. To evaluate the consequences of increasing the level of G in the H strand compared with the major coding (L) strand, we compared the usage of 60 codons other than the four stop codons and ATG in the highly translated genes of the L strand from 11 species and normalized the data. As shown in Figure S2 we
found that the lower eukaryotes have high levels of A- or T-rich codons, whereas the higher
eukaryotes are rich in G or C. The levels of codons without C and G, such as TTT, ATT, and TTA, were
unusually high in the lower-order organisms Drosophila, Caenorhabditis elegans, and
Saccharomyces. We next compared the frequency of amino acids by normalizing using eight major
coding genes from the L strand of the 11 species (Figure S3). We found high levels of leucine in
higher-order organisms where it is encoded by a CG-rich codon, whereas the lower-order organisms
had high levels of phenylalanine, which is encoded by AT-rich codons. In addition, the lower-order
organisms showed a strong preference for codons low in G or C content for all amino acids,
indicating a strong codon bias, while the higher-order organisms used the codons equally with a
slight preference for C-rich codons.

We then compared the codon usage of human mitochondrial genes with that of species
such as chimpanzee, mouse, and C. elegans. We calculated the increase of C in the mtDNA across
eight different protein-coding genes by aligning with the respective amino acid encoded, and
devised a score for comparison. Based on the total C increase in humans, the respective amino acid
codons in other organisms were given scores of 1, 2, 3 and −1, −2, −3 for the relative frequency of C
(Table 1). This ideally should show the overall change in C content in the protein-coding genes
compared with that in higher-order species. Most interestingly, we identified a greater increase in
the C content of human mtDNA genes when compared with that of C. elegans. The overall increase
in C content of the human mtDNA genes was less relative to mouse than compared with C. elegans.
The chimpanzee, which is closely related to human, had only a slight reduction in the C content
compared with human, which indicated an increase in C compared with other species. The increase
in C content from the lower-order to higher-order organisms may contribute to a functional role for
the mitochondria in the development of the organism to adapt to its physiological needs and evolve.
| Gene  | Mutation | Codon change | Mutation | Codon change | Mutation | Codon change |
|-------|----------|--------------|----------|--------------|----------|--------------|
| ND1   | 1        | 11           | 35       | 56           | 153      | 55           |
| COX1  | 1        | 8            | 6        | 92           | 115      | 163          |
| COX2  | 2        | 6            | 17       | 22           | 95       | 45           |
| ATP6  | -3       | 7            | 17       | 38           | 75       | 20           |
| COX3  | 2        | 2            | 30       | 50           | 113      | 48           |
| ND4   | 0        | 6            | 54       | 63           | 186      | 80           |
| ND5   | 9        | 11           | 90       | 48           | 152      | 73           |
| CYTB  | -1       | 2            | 32       | 50           | 161      | 101          |
| Total | 11       | 53           | 281      | 419          | 1050     | 585          |

Table 1: The increase in cytosines in the coding strand with respective codon usage compared between human and chimpanzee, mouse, and C. elegans.

Long noncoding RNAs (lncRNAs), usually >200 nucleotides in length, are a class of nonprotein coding transcripts that have multiple functions in the cell including as activators, repressors, miRNA sponges, and epigenetic modulators. Human mitochondrial lncRNAs have been identified using deep sequencing and have been shown to be complementary to the coding RNA, which means they are rich in G and have the potential to form G4s. Although lncRNAs are noncoding, they are transcribed at very high levels, and the mechanism of their degradation is unknown. Recently, a dedicated mechanism for the degradation of G4-containing RNAs based on the mitochondrial degradosome and quasi-RNA recognition motif (qRRM) protein GRSF1 has been demonstrated. Interestingly, the GRSF1 protein is found only in vertebrates, which are known to have a high level of G4-forming sequences. During evolution, only the mtDNA has undergone a change from GC-poor to GC-rich involving some type of dedicated mechanism for control of G4-forming lncRNA. However, the significance of high levels of G4-forming lncRNA in the mitochondria has not been identified. The lncRNAs in the nuclear telomere repeat-containing RNA (TERRA), which is G rich and has the potential to form G4s, are known to play an important role in the regulation of telomerase activity and in the protection of the DNA. TERRA is also known to play a role in protecting the cell from...
oxidative stress by scavenging ROS.\textsuperscript{24} The lncRNAs may also play a role in scavenging ROS because the mitochondrion is a hot spot for various ROS that are deleterious to mtDNA.

**Summary**

In this study, the mtDNA of a range of species from lower orders to higher orders was selected and analyzed for G4 motifs using QGRS. Although this algorithm is widely used to predict G4 formation in nucleic acid sequences, it may have limitations in the identification of actual G4-forming motifs inside cells. However, it can be used as a preliminary tool to detect intracellular G4-forming sites, and with the advent of next-generation sequencing techniques, it is now possible to map these G4-forming sites in cells. Sequencing has confirmed that most of the computationally predicted sequences do form G4 in cells. The comparative analysis of the mtDNA of various species revealed an increased number of G4 sites in higher organisms. These may have played various roles in the evolution of multicellular organisms by regulating transcription and replication and scavenging ROS inside the cell. We also compared the C content of eight protein-coding mtDNA genes of humans with that of three other species (chimpanzee, mouse, and \textit{C. elegans}) that are related more or less closely based on a phylogenetic tree. This analysis clearly identified an increase of C in the coding genes of humans in comparison with \textit{C. elegans} and to a lesser extent with chimpanzees, the species most closely related to human. The increase in C in higher organisms may play a role in the development of the organism to meet changing physiological demands and other factors. The noncoding transcripts in mtDNA have unusually high levels of G compared with genomic transcripts. The increase in G in these noncoding transcripts means that they have a high likelihood of forming G4 structures, because it is well-known that RNA can also form secondary structures such as G4 that render it very stable. Because mitochondria are well known to produce ROS, the G4-forming RNA may act as a scavenger or sensor for ROS inside the cell.

In summary, over long-term evolution, the level of G4 has gradually increased from lower-to higher-order organisms. This increase in G4 may have facilitated many aspects of mitochondrial replication and transcription to allow the development of multicellular organisms to meet cellular needs. The increase in G content may also have played a key role in acting as a signal and scavenger for ROS in cells. Our analysis may assist in an initial understanding of mtDNA evolution with respect
to secondary structure, but further detailed analysis and confirmation of G4s in the mtDNA of various species is required to identify the potential role of G4.

References

1. Osellame, L. D., Blacker, T. S. & Duchen, M. R. Cellular and molecular mechanisms of mitochondrial function. *Best Pract. Res. Clin. Endocrinol. Metab.* **26**, 711–723 (2012).

2. Malinee, M., Pandian, G. N. & Sugiyama, H. Targeted epigenetic induction of mitochondrial biogenesis enhances antitumor immunity in mouse model. *Cell Chem Biol* (2021) doi:10.1016/j.chembiol.2021.08.001.

3. Chinnery, P. F. & Hudson, G. Mitochondrial genetics. *Br. Med. Bull.* **106**, 135–159 (2013).

4. Abhyankar, A., Park, H.-B., Tonolo, G. & Luthman, H. Comparative sequence analysis of the non-protein-coding mitochondrial DNA of inbred rat strains. *PLoS One* **4**, e8148 (2009).

5. Taanman, J. W. The mitochondrial genome: structure, transcription, translation and replication. *Biochim. Biophys. Acta* **1410**, 103–123 (1999).

6. Barroso Lima, N. C. & Prosdocimi, F. The heavy strand dilemma of vertebrate mitochondria on genome sequencing age: number of encoded genes or G + T content? *Mitochondrial DNA Part A* **29**, 300–302 (2018).

7. Alexeyev, M. Mitochondrial DNA: the common confusions. *Mitochondrial DNA A DNA Mapp Seq Anal* **31**, 45–47 (2020).

8. Gray, M. W. Mitochondrial evolution. *Cold Spring Harb. Perspect. Biol.* **4**, a011403 (2012).

9. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The Origin and Diversification of Mitochondria. *Curr. Biol.* **27**, R1177–R1192 (2017).
10. Bochman, M. L., Paeschke, K. & Zakian, V. A. DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.* **13**, 770–780 (2012).

11. Tateishi-Karimata, H. & Sugimoto, N. Structure, stability and behaviour of nucleic acids in ionic liquids. *Nucleic Acids Res.* **42**, 8831–8844 (2014).

12. Di Antonio, M. *et al.* Single-molecule visualization of DNA G-quadruplex formation in live cells. *Nat. Chem.* **12**, 832–837 (2020).

13. Asamitsu, S., Bando, T. & Sugiyama, H. Ligand Design to Acquire Specificity to Intended G-Quadruplex Structures. *Chemistry - A European Journal* vol. 25 417–430 (2019).

14. Marsico, G. *et al.* Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res.* **47**, 3862–3874 (2019).

15. Wu, F. *et al.* Genome-wide analysis of DNA G-quadruplex motifs across 37 species provides insights into G4 evolution. *Commun Biol* **4**, 98 (2021).

16. Falabella, M., Fernandez, R. J., Johnson, F. B. & Kaufman, B. A. Potential Roles for G-Quadruplexes in Mitochondria. *Curr. Med. Chem.* **26**, 2918–2932 (2019).

17. Capra, J. A., Paeschke, K., Singh, M. & Zakian, V. A. G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae. *PLoS Comput. Biol.* **6**, e1000861 (2010).

18. Kikin, O., D’Antonio, L. & Bagga, P. S. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.* **34**, W676–82 (2006).

19. Malinee, M. & Sugiyama, H. Impact of Reactive Oxygen Species and G-Quadruplexes in Telomeres and Mitochondria. in *Creative Complex Systems* (eds. Nishimura, K., Murase, M. & Yoshimura, K.) 249–274 (Springer Singapore, 2021).

20. Fleming, A. M., Ding, Y. & Burrows, C. J. Oxidative DNA damage is epigenetic by regulating
gene transcription via base excision repair. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 2604–2609 (2017).

21. Ghosh, A. *et al.* Alternative splicing modulation mediated by G-quadruplex structures in MALAT1 IncRNA. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab1066.

22. Rackham, O. *et al.* Long noncoding RNAs are generated from the mitochondrial genome and regulated by nuclear-encoded proteins. *RNA* **17**, 2085–2093 (2011).

23. Pietras, Z. *et al.* Dedicated surveillance mechanism controls G-quadruplex forming non-coding RNAs in human mitochondria. *Nat. Commun.* **9**, 2558 (2018).

24. Galigniana, N. M., Charó, N. L., Uranga, R., Cabanillas, A. M. & Piwien-Pilipuk, G. Oxidative stress induces transcription of telomeric repeat-containing RNA (TERRA) by engaging PKA signaling and cytoskeleton dynamics. *Biochim. Biophys. Acta Mol. Cell Res.* **1867**, 118643 (2020).
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MitoSI.pdf