MTALTCO1: a 259 amino-acid long mtDNA-encoded alternative protein that challenges conventional understandings of mitochondrial genomics

Francis Robitaille[1,a,*], Aziz Ben Hadj[1,a], Amy Campbell[1], Ludovic Nadeau-Lachance[1], Thierry Niaison[1], Thierry Choquette[1], Xavier Roucou[2], Annie Angers[1], Sophie Breton[1,*]

[1]Département de Sciences Biologiques, Université de Montréal, Montréal, Canada
[2]Département de Biochimie, Université de Sherbrooke, Sherbrooke, Canada

[a]Co-first authors

*Corresponding authors:

Sophie Breton, Département de Sciences Biologiques, Université de Montréal, Montréal, QC H3C 3J7, Canada
Email: s.breton@umontreal.ca

Francis Robitaille, Département de Sciences Biologiques, Université de Montréal, Montréal, QC H3C 3J7, Canada
Email: francis.robitaille.2@umontreal.ca

ABSTRACT

Mitochondrial-derived peptides significantly expand the coding potential of the human mitochondrial genome. Here, we report the discovery of MTALTCO1, a 259 amino-acid protein encoded by a mitochondrial alternative open reading frame found in the +3 reading frame of the cytochrome oxidase 1 gene (CO1). Using a custom antibody, we confirmed the mitochondrial expression of MTALTCO1 in human cell lines. Sequence analysis revealed high arginine content and an elevated isoelectric point that were not contingent on CO1's amino acid sequence, suggesting selective pressures might be acting on this protein. MTALTCO1 displays extensive fusion-fission dynamics at the interspecies level, yet produces a full-length protein throughout human mitochondrial lineages. Our findings highlight the importance of identifying novel mitochondrial-derived peptides and proteins in expanding our understanding of the mitochondrial proteome.

INTRODUCTION

Human mitochondria rely on a proteome of over 1,100 proteins distributed across tissues and conditions[1]. Although most of these proteins are encoded by nuclear genes, translated in the cytoplasm and subsequently imported into mitochondria[1,2], the mitogenome contributes to a small but essential portion of its resident proteome[3]. The human mitogenome is a circular DNA molecule (mtDNA) of 16,569 base pairs commonly thought to encode 37 genes, of which 13 produce key subunits of the electron transport chain and ATP synthase, two are ribosomal RNAs (12S and 16S rRNAs), and 22 are transfer RNAs (tRNAs) that decipher the mitochondrial genetic code[4]. Because annotated regions of the mtDNA are practically contiguous, it was believed that the mitogenome had no space for additional coding sequences and that its coding potential was elucidated[5]. The discovery of *humanin* in 2001, a small open reading frame (ORF) encoding a 24 amino acid (AA) long peptide found in the 16S rRNA gene challenged this view and revealed the first of a novel class of mitochondria-derived peptides termed MDPs, whose coding sequences are nested in or overlapping previously annotated regions of the mitogenome[6]. Around 2016, additional small functional ORFs were identified within ribosomal RNAs, producing the MDPs MOTS-c and small humanin-like peptides 1 through 6 (SHLP 1-6) ranging from 16 to 38 AA in length[7,8].

MDPs had until then been primarily investigated in non-coding mitochondrial genes like rRNAs, where relaxed selection pressures offer an environment more conducive to the evolution

of protein function than coding regions. The discovery of SHMOOSE in 2023[9], a 58 AA protein whose coding sequence overlaps tRNA Serine and ND5, which influences mitochondrial gene expression and regulates oxygen consumption, marked the onset of a series of larger mitochondrial-derived proteins found within coding regions. The same year, MTALTND4, a 99 AA protein—larger than some canonical mitochondrial proteins like ATP8 and ND4L—was identified in a mtaltORF nested in the ND4 gene and induces a quiescent state-like metabolic depression[11]. In 2024, CYTB-187AA, a 187 AA long protein whose coding sequence nests in the cytochrome b gene was confirmed, and promotes the pluripotent state in the primed-to-naive transition[12].

Here, we confirm the existence of MTALTCO1, the longest human mitochondrial alternative protein to date at 259 amino acids, encoded in the reading frame shifted two nucleotides downstream of the cytochrome oxidase 1 gene.

MATERIAL AND METHODS

Custom rabbit anti-MTALTCO1 antibody generation

Kienzle et al. (2023) interrogated the CPTAC_Breast_Cancer database, a large repository of more than 32 million mass spectrometry MS/MS scans. Using the PepQuery tool, they retrieved peptide-spectrum-matches (PSM) scores of peptides drawn from an in silico digestion of the theoretical mitochondrial alternative proteome generated with the mitochondrial genetic code using AGG and AGA (AGR) as arginine. A MTALTCO1-derived peptide (ELLPPWSLR) was identified and a custom rabbit anti-MTALTCO1 was ordered from MediMabs (Montreal, QC, Canada). The signal obtained was of expected weight, but did not disappear in Rho0 cells lysates obtained from Abcam (ab154479), casting doubt on its possible mitochondrial origin (Kienzle et al. 2023). In the present study, osteosarcoma 143B-ρ0 cells (Kerafast [ESA106]) were used, together with chloramphenicol and actinonin treatment, to show that MTALTCO1 clearly derives from the mtDNA (see below).

Cell cultures and treatments

HEK-293T and HeLa cells were cultured in Dulbecco’s modified eagle medium (DMEM) supplemented with 10% of fetal bovine serum (FBS), antibiotics (1% streptomycin and penicillin)

in 10 cm petri dishes maintained at 37°C with 5% CO_2. Actinonin treatment was realized on HeLa cells split into 2 groups, one serving as control and left untouched for 48h, and the other incubated with actinonin (150 µM) for 48h. Chloramphenicol treatment was realized on HeLa cells as in Kienzle et al. (2023), i.e. by adding 2 mL of chloramphenicol (50 µg/mL in 95% ethanol) in a 10-cm petri dish of 80% cell confluency. Cells were treated for 48h.

Western blotting

Cells were first washed twice with PBS, harvested and sonicated 3 X 3 seconds with a dismembrator sonicator (Fisher) in a homogenizer buffer (pH 7.4) containing 10mM HEPES, 150mM NaCl and a cocktail of protease inhibitors (benzamidine, PMSF, aprotinin and leupeptin). After sonication, Triton X-100 (10%) was added to a final concentration of 1% and samples were incubated for 20 minutes at 4°C with agitation. They were centrifuged at 10 000 rpm for 10 minutes at 4°C and the supernatant was collected and kept on ice. Total protein concentration was assessed using the Bradford test method and aliquots containing 100µg of proteins were prepared and stored at -80°C for later use. Various tissue lysates (kidney, heart, testis, brain, liver, skeletal muscle, spleen) were also obtained from Novus Biologicals (NB820-59231; NB820-59217; NB820-59266; NB820-59177; NB820-59232; NB820-59253; NB820-59259).

Samples were suspended in Laemmli Sample Buffer (LSB) 1X containing DTT and ß-mercaptoethanol and heated for 5 minutes at 95°C before being loaded at equal protein concentration (100 µg) on Tricine-SDS-PAGE 3-part gels (separating, spacer, and compaction gels), ideal for protein separation ranging from 1 to 100 kDa. Samples were migrated at low voltage (~43 V) overnight at room temperature and then transferred onto a previously hydrated PVDF membrane with a 0.2 µm pore size for an hour at 4°C and 1000 mA in electrophoresis transfer buffer (Towbin buffer). The membrane was then incubated for an hour at room temperature in a PBS blocking buffer with 5% non-fat dry milk and 0.05% tween-20. Membranes were then incubated with primary antibodies for 1 hour and a half at room temperature. The primary antibodies used were diluted in a PBS solution with 0.05% tween-20 as follows: “rabbit anti-MTALCO1” (1:1000) (MediMabs), “mouse anti-ATP5” (1:2000) (Abcam [ab14748]), “mouse anti-CO1” (1:1000) (LifeSpan BioSciences, Inc. [LS-C172110]), “rabbit anti-ND4” (1:1000) (Boster Biological Technology Co., Ltd. [A04180-1]), and “mouse anti-actin” (1:2000) (Abcam [ab1801]). Following primary incubation, membranes were washed 3 X 10 minutes in TBS-T.

They were then incubated with goat anti-rabbit IgG (1:2000) or goat anti-mouse IgG (1:2000) secondary antibodies coupled to horseradish peroxidase (HRP). These were diluted in PBS-T and membranes were incubated for 30 minutes at room temperature. They were washed 3 X 10 minutes in TBS-T and 1 X 5 minutes in TBS at room temperature. Equal parts of chemiluminescent substrate and enhancer solutions (Montreal Biotech Inc.; western blot detection kit) were added to the membranes and signals were captured using a FUSION FX chemiluminescence imaging system. To verify the specificity of the anti-MTALTCO1 antibody, the synthetic antigenic peptide was added to the primary antibody solution at a 10X concentration before the incubation, to competitively chelate every antigenic site of the primary antibody and to show the attenuation of the specific MTALTCO1 band.

Immunofluorescence

Coverslips were incubated in 95% ethanol for 1h and then placed in a 24-well PCA detachable microscope slide and covered with 50µg/mL Poly-D-Lysine (PDL) solution for 2h. HeLa cells were cultured on these slides in DMEM supplemented with 10% calf bovine serum, fungizone, penicillin and streptomycin and were kept at 37˚C with 5% CO2. Each well was washed with PBS and fixed with paraformaldehyde (4%, 20 minutes duration) before another three consecutive 5 minutes washes with PBS. HeLa cell membranes were then permeabilized by adding Triton (0.2%) for 4 minutes, followed by three consecutive 5 minutes washes in PBS. Ensued a 30 minutes (room temperature) blocking using 5% NGS in PBS followed by an incubation period of 2 hours in a combination of 1:50 anti-MTALTCO1 rabbit and 1:250 anti-ATP5 mouse antibodies in PBS (1% NGS). After 3 X 5 minutes PBS washes, cells were incubated in PBS (1% NGS) for 1 hour at room temperature along with the secondary IgG antibodies (respectively Alexa Fluor TM 488 anti-rabbit [1:500] and Alexa Fluor TM 488 anti-mouse [1:500]). Finally, cells were washed 3 X 5 minutes in PBS before adding a mounting drop containing DAPI and visualized with a confocal microscope Zeiss-LSM800.

MTALTCO1 sequence analysis and *in silico* characterization

Using GenBank's web-server, nucleotide basic local alignment search tool (BLASTn) was employed to determine if the MTALTCO1 sequence could be found in the nuclear genome. This search was done to look for potential nuclear mtDNA sequences (NUMTs), which are fragments

of mtDNA that have been integrated in the nuclear genome[21,22]. The resulting matches were aligned and translated in MEGA11[14] using the universal code to see if they could produce a sequence that our antibody would recognize. The antigenic peptide sequence of MTALTCO1 was then blasted (BLASTp) against the entire human reference proteome.

To explore whether characteristics of the MTALTCO1 sequence were the coincidental outcome of the amino acid sequence of the CO1 sequence in humans, 50 000 sequences that adhere to the CO1 sequence were randomly generated by leveraging the degeneracy of the mitochondrial genetic code in a Jupyter Notebook. Next, each of these sequences in the +3 reading frame (corresponding to MTALTCO1's reading frame) were translated with the mitochondrial genetic code (with AGR as arginine) and the isoelectric point as well as the hydropathy score were collected using the Bio.SeqUtils.ProtParam python module. These were pooled and compared to the actual value of MTALTCO1 for each of these metrics using a p-value two-sided hypothesis testing. The number of generated sequences was much higher than needed to obtain a stable distribution.

Phylo-P conservation scores were sourced from UCSC for 30 primates and 100 vertebrates and smoothed using a fast Fourier transform with a sliding window of 25 nucleotides.

To monitor the fission-fusion dynamics of MTALTCO1 throughout hominids and vertebrates we imported sequences from NCBI using the Entrez module. Every hominid sequence was sourced along with those of extinct *H. neaderthalensis* and *H. heidelbergensis.* 100 vertebrate sequences were queried in GenBank and we ensured equal representation of major taxa. These were then read and translated in the +3 reading frame using the mitochondrial genetic code with AGR encoding arginine. The resulting proteins were mapped for each species across the CO1 sequence, with only a selection of these being shown in the supplementary files. To monitor whether the comparisons of MTALTCO1 to sequences generated and read in the +3 reading frames held throughout hominids, this analysis was redone with each hominid's CO1 sequence and the resulting Z-scores were graphed for each metric.

Consensus sequence graphs were made in WebLogo[13] following alignment in MEGA11[14] of hominids' MTALTCO1 sequences. Structure predictions were made using RoseTTAFold[15] in their web-server and sub-cellular localization predictions were made using Deeploc 2.0's web-servers[16].

RESULTS AND DISCUSSION

MTALTCO1 is a 259AA protein found in an alternative reading frame nested in CO1

The mitogenome was investigated for unannotated ORFs with the PepQuery tool to interrogate spectra from the published CPTAC_Breast_Cancer dataset (Kienzle et al. 2023). A mtaltORF-derived peptide (ELLPPWSLR) was found in MS data and a custom antibody was produced to confirm the existence of its corresponding protein in HEK-293T and Hela cells via western blotting (Fig. 1a,b). This protein, dubbed MTALTCO1 (Mitochondrial Alternative CO1) in keeping with the nomenclature introduced by Kienzle et al. (2023) was confirmed as a 259 AA protein whose 777 base pair long coding sequence nests in the sense +3 reading frame, shifted two nucleotides downstream of the mitochondrial cytochrome oxidase 1 gene. It is initiated by the mitochondria-specific start codon ATT and terminated by the mitochondrial and cytoplasmic stop codon TAG. Like other mtaltORFs, MTALTCO1 exhibits a tissue-specific expression pattern[17,11,9], showing the highest expression level in the kidneys, testis, and spleen (Fig. 1c). In contrast, heart and skeletal muscle show only faint expression, while the liver and brain display little to no detectable expression, an expression pattern that does not conform to CO1's[19]. Confocal microscopy data suggests that MTALTCO1 might localize in the cytoplasm, with partial co-localization with mitochondria (Fig. 1e). The high arginine content of MTALTCO1, known for its membrane-penetrating properties[20], could account for the diffuse localization signal observed.

MTALTCO1 is translated in the mitochondria

We investigated the mitochondrial origin of MTALTCO1 using two approaches (Fig. 2). First, we examined whether MTALTCO1 could originate from a nuclear mitochondrial DNA transfer (NUMT)[21,22]. Despite numerous CO1-derived NUMTs, none code for a complete MTALTCO1 nor provide an antigenic sequence that is not flanked by stop codons (Fig. 2a). Because other nucleotide sequences could in theory produce a similar antigenic peptide sequence, we also blasted the antigenic peptide sequence against the human reference proteome and found that the only matches were those described in CO1-derived NUMTs. Next, we assessed the loss of western blot signal following three treatments that specifically impact mitochondrial expression. In Rho0 cells, which lack endogenous mtDNA, the signal loss was evident (Fig. 2b). Cells treated with chloramphenicol, which inhibits protein synthesis in mitochondria, similarly showed a clear loss of signal (Fig. 2c). Others treated with actinonin, which inhibits mitochondrial transcription,

showed a reduction in signal (Fig. 2d), though less pronounced than with chloramphenicol, possibly due to the continued translation of persistent mRNAs after treatment. We thus conclude that MTALTCO1 is translated only in mitochondria, and that 31.2 kDa western blot signals correspond to MTALTCO1.

CO1 sequence does not predict important features of MTALTCO1

We explored whether characteristics of the MTALTCO1 sequence were the coincidental outcome of the amino acid sequence of the reference frame (Fig. 3). Given the severe impact of mutations in mitochondrial reference proteins[23,24], we assumed that sequences in alternative reading frames must maintain CO1 sequence integrity. We will refer to the pool of nucleotide sequences that conform to this constraint as the Degenerate Sequence Space (DSS). Demonstrating that confirmed mtaltORFs adopt unlikely traits within proteins translated in alternative frames from DSS makes a convincing case that selection is acting specifically on these mtaltORFs to promote certain protein properties.

To test this, we generated 50 000 sequences within the DSS and compared them across six metrics with the observed values in MTALTCO1: the number of start codons, stop codons, arginine content, AGR content, isoelectric point, and hydropathy. Using two-tailed p-value hypothesis testing, we examined the null hypothesis, which proposed that the observed characteristics of MTALTCO1 would fall within the expected range from a distribution of these traits in alternative proteins of the +3 reading frame generated randomly from DSS sequences. Rejecting the null hypothesis thus posits that the observed characteristics were unlikely to occur at random.

Our results show that the number of start codons in MTALTCO1 is within the expected range, suggesting that these codons might not be under direct selection for quantity. Whether some of these codons initiate translation remains uncertain but we did not detect other proteins in our western blot analysis. A start codon one site upstream of the antigenic peptide sequence would have produced, if recognized by the translation apparatus, a band of 17.6 kDa. CO1 also contains numerous in-sequence start codons, the recognition of which would create spurious CO1 fragments that are not seen. Regulatory factors, likely involving mRNA folding patterns that mask certain initiating codons from the translation machinery[25,4], may cause these codons to be overlooked, suggesting they might not be under strong selection pressure for removal. The number of in

sequence stop codons—zero by definition— is significantly lower ($p<0.05$) than expected, meaning that MTALTCO1 is unexpectedly large with respect to DSS. The isoelectric point is higher than expected ($p<0.001$) with hydrophily (negative hydropathy score) being also significantly higher than expected ($p<0.001$). Both relations were not solely driven by AGR content as they retain significance (respectively $p<0.01$ and $p<0.05$ - not shown) upon removal of AGRs from the mitochondrial code. Arginine and AGR content are both significantly higher than expected with p-values of respectively $p<0.01$ and $p<0.001$. High AGR content was solely responsible for the high arginine content as removing them from the mitochondrial code brought observed arginine content within the expected range ($p>0.05$ - not shown).

AGR codons might code for arginine in mtaltORFs translated in the mitochondria

Although AGR codons have been considered stop codons for lack of cognate tRNAs in the mitochondrial code[4], the exact mechanism underlying AGR-related translation termination are still debated[22,11] and mitochondria are known to import cytosolic tRNAs[27,28]. MTALTND4 is another mtaltORF that is translated in the mitochondria yet contains AGR codons[11]. These are not read through and effectively code for arginine as evidenced by an MS-derived 18 AA tryptic peptide found directly after an AGR encoded arginine in MTALTND4's sequence (see supplementary files in Kienzle et al., 2023). The MTALTCO1-derived peptide (ELLPPWSLR) identified in MS data using PepQuery represents the 73d to 82nd AA of MTALTCO1, with the 72nd AA being an AGG encoded arginine and the 82nd AA a AGA encoded arginine. Because the only NUMT (MTCO1P12) that preserved the peptide sequence could not initiate translation upstream of the peptide (see supplementary files, Fig. I), this adds to the evidence that AGRs might code for arginine in the mitochondria. The other overlapping mtaltORFs challenge conventional understandings of mitochondrial expression, albeit in different ways. The SHMOOSE ORF overlaps tRNA Serine and ND5, implying for the first time that translation can occur before excision of tRNAs from the full-length mRNA transcript, an early step in the maturation of mitochondrial mRNAs[9]. CYTB-187AA also presents a novel case of mitochondrial mRNA export followed by cytoplasmic translation and subsequent mitochondrial re-entry via a mitochondrial targeting sequence[29,12]. Recent mitochondrial ribosome profiling studies have also suggested that non-canonical start codons might initiate translation in alternative reading frames[30].

As even recent models of mitochondrial expression[31,4] outright ignore mtaltORFs, they cannot be expected to fully capture their unique expression dynamics. We propose that the confident identification of additional mtaltORFs, particularly those that challenge established rules, will offer insight into these conflicts and serve as catalysts for refining and expanding current models of mitochondrial expression.

MTALTCO1's unique protein properties are functionally important

Isoelectric point

MTALTCO1 has a very high isoelectric point (pI) at 11.68, a value which is of functional relevance[32]. A notable bias in pI distribution was observed in specific sub-cellular compartments among certain functional groups of proteins[33], with high pI values being often associated with mitochondrial and nuclear localization[34]. These findings support the idea that MTALTCO1's sequence has undergone selection for specific physico-chemical properties indicative of function.

Arginine content

MTALTCO1 has a total arginine content nearing 15%, a value that is associated with proteins considered arginine enriched[35]. Arginine-rich proteins play important roles in shaping protein structure, DNA binding, and function by facilitating electrostatic interactions and internalization mechanisms[36,37]. These roles can be attributed to its planar side group, a guanidinium, that can form up to five hydrogen bonds, helping in protein and protein-complex stability and which is a key feature of cell penetrating peptides[16]. Deeploc 2.0[38], a sub-cellular localization prediction tool predicted extracellular localization mostly on account of an arginine rich motif of 11 amino acids (HLLRPRRRRRP - Fig. 4). The biological importance of this motif should be further studied. The sub-cellular localization of MTALTCO1 remains uncertain, with theoretical, *in silico* and fluorescence insights not reaching a clear consensus and should receive further study.

Dynamic fission and fusion of MTALTCO1 at the cross-species level

MTALTCO1 is preserved in humans

We interrogated mitochondrial variants in the MitoMap[39] database to assess whether MTALTCO1 was preserved throughout mitochondrial lineages. It was found that only three in

>60k sequences are affected at MTALTCO1's start codon. Although its prevalence was quite rare across lineages (0.58%) we did find one mutant that in 5.19% of the lineage L pseudogenized MTALTCO1 with a nonsense mutation Q154*. If other mutations pseudogenize MTALTCO1, they were less prevalent than 0.5% over all mitochondrial lineages (supplementary files, Fig. II)

The base pair conservation signature of MTALTCO1 is not different than other regions in CO1

We next assessed whether Phylo-P signals in primates and vertebrates yielded a conservation signal similar to the nuclear alt-FUS[40], for which nucleotides are less conserved if they host an altORF compared to regions that do not, and found no such evidence (supplementary files, Fig. III,IV). This is probably due to the strong conservation of CO1 sequence and the lack of negative controls, i.e. regions confirmed to host no alternative proteins.

Fusion-fission dynamics of MTALTCO1

MTALTCO1 is unique to *Homo sapiens and H. heidelbergensis* and is pseudogenized by a nonsense mutation in neanderthals and other Great Apes (Fig. 5). The widespread presence of start codons in the +3 reading frame in vertebrates indicates that although pseudogenized, MTALTCO1-derived proteins and peptides might still get translated and be under selection (supplementary files, Fig. V). Since MTALTCO1 is a fast-evolving protein, we focused on the Great Apes and asked whether the unlikeliness of characteristics of MTALTCO1 with respect to proteins generated from DSS in *Homo sapiens* showed some measure of conservation. For every species of Great Apes we generated DSS from the CO1 sequence that were overlapped by MTALTCO1 or its fission products. We compared the latter's sequence to proteins translated from the +3 frame in DSS again using a two-tailed p-value test, with Z-value results from each species being then pooled together and compared for each metric (Fig. 6). In these graphics, a high significance does not suggest that a particular value is high or low in its biological context but only that it is unlikely to arise at random.

The number of start codons conforms throughout these species to the distribution of start codons in random DSS-generated protein sequences in the +3 reading frame, and the distribution of stop codons is skewed towards the left, indicating fewer stop codons than expected, with significance levels hovering around values similar to those in MTALTCO1. This means that throughout *Hominidae*, MTALTCO1 and related fission products were often unexpectedly large

with respect to DSS. AGR content was very skewed to the right, mostly beyond the $p<0.05$ threshold and sometimes reaching significance levels beyond $p<0.0001$, with global arginine content dispersed on both sides of the $p=0.05$ mark of positive Z-scores. The isoelectric point was likewise skewed to the right with about half of the points hovering around significance levels of $p=0.05$. The hydropathy was very skewed to the left indicating that proteins are persistently more hydrophilic in regard to DSS. These findings suggest that although MTALTCO1 is pseudogenized by nonsense mutations in *Hominidae*, certain important protein properties are consistently over or under-represented with respect to DSS in MTALTCO1 or its fission products. Despite rapid mutation rates in mtDNA[41], the persistence of these functionally relevant traits suggests that fragments of MTALTCO1 might retain functionality in other species.

As mentioned, recent mitochondrial ribosome profiling studies have suggested that mitochondrial translation of mtaltORFs can be initiated by non-canonical start codons[30]. Until there is enough evidence to accurately predict the full set of initiating codons, we cannot rule out the possibility that gaps in the expression profile of MTALTCO1-fission products in hominids are artifactual and that a higher percentage of the +3 reading frame is expressed in other species.

A cancer-associated mutation affects MTALTCO1 and not CO1

A study[42] examined the genetic factors contributing to prostate cancer (PCa) in African American men, who are disproportionately affected by the disease, the most diagnosed cancer and second leading cause of death among men in the United States[42]. Notably, researchers identified two mutations in the reference gene CO1 significantly associated with PCa: one silent mutation (T6221C, $p = 0.02$) in the reference frame and another missense mutation (T7389C, $p = 0.03$). These mutations were in moderate linkage disequilibrium with an r^2 value of 0.6, indicating the potential for T6221C to have an independent effect. T6221C produces a shift from Serine to Proline at the 45th amino acid of MTALTCO1, a mutation that MitoMap's[39] variants database finds prevalent in the African L lineage[42] at 13.32% compared to around 1% in other lineages. The impact of this mutation along with that of the nonsense mutation Q154*, also specific to the L lineage, should receive further attention.

Because of its modest size, mtDNA is often unjustifiably downplayed in its contribution to complex diseases[44]. Mitochondria have a very high mutation rate that causes extensive age-related mtDNA mosaicism[45,46]. As missense mutations in the reference proteome are highly detrimental

and thus less likely to accumulate[47], the alternative proteome may drift significantly more than its counterpart, potentially contributing to mtDNA mutation-related diseases. Mutations that were previously overlooked and framed as inconsequent because they did not affect the protein sequence in the reference frame can impact the protein sequence of ORFs in alternative reading frames. In the nuclei, synonymous mutations in the reference ORF are the second most frequent in cancer samples after missense mutations, and up to 30% of cancer related mutations alter unannotated ORFs[48,49]. Furthermore, although 80% of rare diseases are thought to be of genetic origin, only 39% have a described gene association[50]. Current disease annotation models in the mitochondrial genome often exclude synonymous mutations in the reference frame, thereby overlooking critical data for the screening of disease-associated mtaltORF mutants. We strongly argue for the systematic inclusion of mutations that are synonymous in mitochondrial reference proteins in these models to improve the detection of clinically relevant mtaltORF-specific variants. Beyond clinical implications, these variants can serve as evidence of function, guiding efforts to identify additional mtaltORFs as argued by Miller et al. (2023) and Breton (2021).

CONCLUSION

The discovery of MTALTCO1, a 259 amino-acid protein encoded by an alternative reading frame in the CO1 gene, further challenges the long-held assumption that the human mitogenome codes for only 13 proteins. Our study not only confirms the mitochondrial origin of this protein but also highlights key structural features, such as its high arginine content and elevated isoelectric point which are of functional importance. Furthermore, the persistence of MTALTCO1 across human mitochondrial lineages and of important protein properties of MTALTCO1 or its fission products across hominids suggests that MTALTCO1 might be under functional selection. These findings open new avenues for the study of mitochondrial-derived proteins overlapping coding regions and underscore the importance of re-examining mitogenomic sequences for unannotated ORFs.

REFERENCES

1- Morgenstern, M., Peikert, C. D., Lübbert, P., Suppanz, I., Klemm, C., Alka, O., ... & Warscheid, B. (2021). Quantitative high-confidence human mitochondrial proteome and its dynamics in cellular context. *Cell metabolism*, *33*(12), 2464-2483.

2- Palmfeldt, J., & Bross, P. (2017). Proteomics of human mitochondria. *Mitochondrion*, *33*,

3- Butenko, A., Lukeš, J., Speijer, D., & Wideman, J. G. (2024). Mitochondrial genomes revisited: why do different lineages retain different genes?. *BMC biology*, *22*(1), 15.

4- Rackham, O., & Filipovska, A. (2022). Organization and expression of the mammalian mitochondrial genome. *Nature Reviews Genetics*, *23*(10), 606-623.

5- Grivell, L. A. (1983). Mitochondrial dna. *Scientific American*, *248*(3), 78-89.

6- Lee, C., Yen, K., & Cohen, P. (2013). Humanin: a harbinger of mitochondrial-derived peptides?. *Trends in Endocrinology & Metabolism*, *24*(5), 222-228.

7- Lee, C., Kim, K. H., & Cohen, P. (2016). MOTS-c: a novel mitochondrial-derived peptide regulating muscle and fat metabolism. *Free Radical Biology and Medicine*, *100*, 182-187.

8- Cobb, L. J., Lee, C., Xiao, J., Yen, K., Wong, R. G., Nakamura, H. K., ... & Cohen, P. (2016). Naturally occurring mitochondrial-derived peptides are age-dependent regulators of apoptosis, insulin sensitivity, and inflammatory markers. *Aging (Albany NY)*, *8*(4), 796.

9-Miller, B., Kim, S. J., Mehta, H. H., Cao, K., Kumagai, H., Thumaty, N., ... & Alzheimer's Disease Neuroimaging Initiative*. (2023). Mitochondrial DNA variation in Alzheimer's disease reveals a unique microprotein called SHMOOSE. *Molecular psychiatry*, *28*(4), 1813-1826.

10- Vanderperre, B., Lucier, J. F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., ... & Roucou, X. (2013). Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PloS one*, *8*(8), e70698.

11- Kienzle, L., Bettinazzi, S., Choquette, T., Brunet, M., Khorami, H. H., Jacques, J. F., ... & Breton, S. (2023). A small protein coded within the mitochondrial canonical gene nd4 regulates mitochondrial bioenergetics. *Bmc Biology*, *21*(1), 111.

12- Hu, Z., Yang, L., Zhang, M., Tang, H., Huang, Y., Su, Y., ... & Liu, X. (2024). A novel protein CYTB-187AA encoded by the mitochondrial gene CYTB modulates mammalian early development. *Cell Metabolism*.

13- Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome research*, *14*(6), 1188-1190.

14- Glen Stecher, Koichiro Tamura, and Sudhir Kumar (2020) Molecular Evolutionary Genetics Analysis (MEGA) for macOS. Molecular Biology and Evolution 37:1237-1239.

15- Thumuluri, V., Almagro Armenteros, J. J., Johansen, A. R., Nielsen, H., & Winther, O. (2022). DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic acids research*, *50*(W1), W228-W234.

16- Baek, M., Anishchenko, I., Humphreys, I. R., Cong, Q., Baker, D., & DiMaio, F. (2023). Efficient and accurate prediction of protein structure using RoseTTAFold2. *BioRxiv*, 2023-05.

17- Brunet, M. A., Brunelle, M., Lucier, J. F., Delcourt, V., Levesque, M., Grenier, F., ... & Roucou, X. (2019). OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic acids research*, *47*(D1), D403-D410.

18-Merry, T. L., Chan, A., Woodhead, J. S., Reynolds, J. C., Kumagai, H., Kim, S. J., & Lee, C. (2020). Mitochondrial-derived peptides in energy metabolism. American Journal of Physiology-Endocrinology and Metabolism, 319(4), E659-E666.

19- Herbers, E., Kekäläinen, N. J., Hangas, A., Pohjoismäki, J. L., & Goffart, S. (2019). Tissue specific differences in mitochondrial DNA maintenance and expression. *Mitochondrion*, *44*, 85-92.

20-Vazdar, M., Heyda, J., Mason, P. E., Tesei, G., Allolio, C., Lund, M., & Jungwirth, P. (2018). Arginine "magic": Guanidinium like-charge ion pairing from aqueous salts to cell penetrating peptides. *Accounts of chemical research*, *51*(6), 1455-1464.

21. Tourmen, Y., Baris, O., Dessen, P., Jacques, C., Malthièry, Y., & Reynier, P. (2002). Structure and chromosomal distribution of human mitochondrial pseudogenes. *Genomics*, *80*(1), 71-77.

22. Ricchetti, M., Tekaia, F., & Dujon, B. (2004). Continued colonization of the human genome by mitochondrial DNA. *PLoS biology*, *2*(9), e273.

23.Schaefer, A. M., McFarland, R., Blakely, E. L., He, L., Whittaker, R. G., Taylor, R. W., ... & Turnbull, D. M. (2008). Prevalence of mitochondrial DNA disease in adults. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, *63*(1), 35-39.

24. Johns, D. R. (1995). Mitochondrial DNA and disease. *New England journal of medicine*, *333*(10), 638-644.

25. LRPPRC- Siira, S. J., Spåhr, H., Shearwood, A. M. J., Ruzzenente, B., Larsson, N. G., Rackham, O., & Filipovska, A. (2017). LRPPRC-mediated folding of the mitochondrial transcriptome. *Nature communications*, *8*(1), 1532.

26. Krüger, A., Remes, C., Shiriaev, D. I., Liu, Y., Spåhr, H., Wibom, R., ... & Rorbach, J. (2023). Human mitochondria require mtRF1 for translation termination at non-canonical stop codons. *Nature communications*, *14*(1), 30.

27.Schneider, A. (2011). Mitochondrial tRNA import and its consequences for mitochondrial translation. *Annual review of biochemistry*, *80*(1), 1033-1053.

28. Alfonzo, J. D., & Söll, D. (2009). Mitochondrial tRNA import–the challenge to understand has just begun.

29. Janssen, K. A., & Song, A. (2024). From whence it came: Mitochondrial mRNA leaves, a protein returns. *Cell Metabolism*, *36*(7), 1433-1435.

30. Wakigawa, T., Mito, M., Yamashiro, H., Tomuro, K., Tani, H., Tomizawa, K., ... & Iwasaki, S. (2023). The complexity and dynamics of in organello translation assessed by high-resolution mitochondrial ribosome profiling. *bioRxiv*, 2023-07.

31. Tan, B. G., Gustafsson, C. M., & Falkenberg, M. (2024). Mechanisms and regulation of human mitochondrial transcription. *Nature Reviews Molecular Cell Biology*, *25*(2), 119-132.

32. Chasapis, C. T., & Konstantinoudis, G. (2020). Protein isoelectric point distribution in the interactomes across the domains of life. *Biophysical Chemistry*, *256*, 106269.

33. Wang, T., and Tang, H. (2017). The Physical Characteristics of Human Proteins in Different Biological Functions. *PloS one* 12 (5), e0176234. doi:10.1371/journal.pone.0176234

34.Tokmakov, A. A., Kurotani, A., & Sato, K. I. (2021). Protein pI and intracellular localization. *Frontiers in Molecular Biosciences*, *8*, 775736.

35. Chandana, T., & P Venkatesh, Y. (2016). Occurrence, functions and biological significance of arginine-rich proteins. *Current Protein and Peptide Science*, *17*(5), 507-516.

36. Futaki, S., Suzuki, T., Ohashi, W., Yagami, T., Tanaka, S., Ueda, K., & Sugiura, Y. (2001). Arginine-rich peptides: an abundant source of membrane-permeable peptides having potential as carriers for intracellular protein delivery. *Journal of Biological Chemistry*, *276*(8), 5836-5840.

37. Borders Jr, C. L., Broadwater, J. A., Bekeny, P. A., Salmon, J. E., Lee, A. S., Eldridge, A. M., & Pett, V. B. (1994). A structural role for arginine in proteins: multiple hydrogen bonds to backbone carbonyl oxygens. *Protein Science*, *3*(4), 541-548.

38. Thumuluri, V., Almagro Armenteros, J. J., Johansen, A. R., Nielsen, H., & Winther, O. (2022). DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic acids research*, *50*(W1), W228-W234.

39. Kogelnik, A. M., Lott, M. T., Brown, M. D., Navathe, S. B., & Wallace, D. C. (1996). MITOMAP: a human mitochondrial genome database. *Nucleic acids research*, *24*(1), 177-179.

40. Brunet, M. A., Jacques, J. F., Nassari, S., Tyzack, G. E., McGoldrick, P., Zinman, L., ... & Roucou, X. (2021). The FUS gene is dual-coding with both proteins contributing to FUS-mediated toxicity. *EMBO reports*, *22*(1), e50640.

41. Brown, W. M., George Jr, M., & Wilson, A. C. (1979). Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences*, *76*(4), 1967-1971.

42. Ray, A. M., Zuhlke, K. A., Levin, A. M., Douglas, J. A., Cooney, K. A., & Petros, J. A. (2009). Sequence variation in the mitochondrial gene cytochrome c oxidase subunit I and prostate cancer in African American men. *The Prostate*, *69*(9), 956-960.

43. Gonçalves, V. F. (2019). Mitochondrial genetics. *Mitochondria in Health and in Sickness*, 247-255.

44. Welch, D. R., Foster, C., & Rigoutsos, I. (2022). Roles of mitochondrial genetics in cancer metastasis. Trends in cancer, 8(12), 1002-1018

45. An, J., Nam, C. H., Kim, R., Lee, Y., Won, H., Park, S., ... & Ju, Y. S. (2024). Mitochondrial DNA mosaicism in normal human somatic cells. Nature Genetics, 1-13.

46. Wallace, D. C., & Chalkia, D. (2013). Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. Cold Spring Harbor perspectives in biology, 5(11), a021220.

47. Wong, L. J. C. (2007). Pathogenic mitochondrial DNA mutations in protein-coding genes. Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine, 36(3), 279-293.

48. Brunet, M. A., Levesque, S. A., Hunting, D. J., Cohen, A. A., & Roucou, X. (2018). Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. Genome research, 28(5), 609-624.

49. Sharma, Y., Miladi, M., Dukare, S., Boulay, K., Caudron-Herger, M., Groß, M., ... & Diederichs, S. (2019). A pan-cancer analysis of synonymous mutations. Nature communications, 10(1), 2569.

50. Ferreira, C. R. (2019). The burden of rare diseases. American journal of medical genetics Part A, 179(6), 885-892.

51. Breton, S. (2021). Mitochondrial Russian doll genes may explain some discrepancies in links between mtDNA mutations and mitochondrial diseases. *BioEssays*, *43*(6).

FIGURES

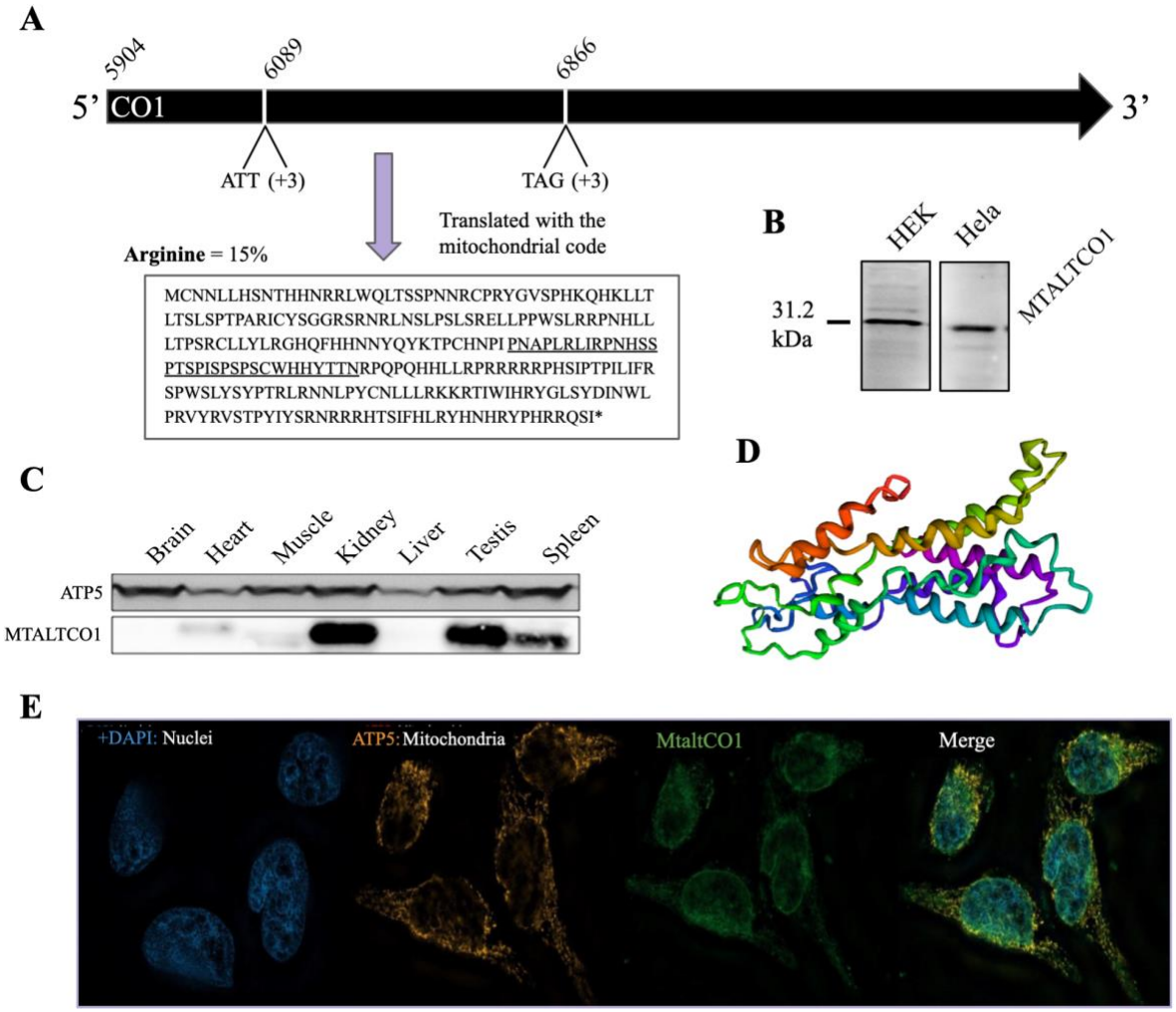

Figure 1: MTALTCO1 is a 259 amino acid protein encoded in the +3 reading frame of the COX1 gene.

(A) MTALTCO1 is initiated by ATT and terminated by TAG, translated from the mitochondrial genetic code to produce a 259 AA long protein that has a total arginine content of 15%. (B) MTALTCO1 detected by Western Blot in HEK293T cells as well as in HeLa cells. (C) MTALTCO1 and nuclear-encoded mitochondrial ATP5 detection by Western Blot in various tissue lysates. (D) RoseTTAFold prediction of structure with a confidence score of 0.43 (E) MTALTCO1 and nuclear-encoded mitochondrial ATP5 detection by immunofluorescence in HeLa cells stained with dapi.

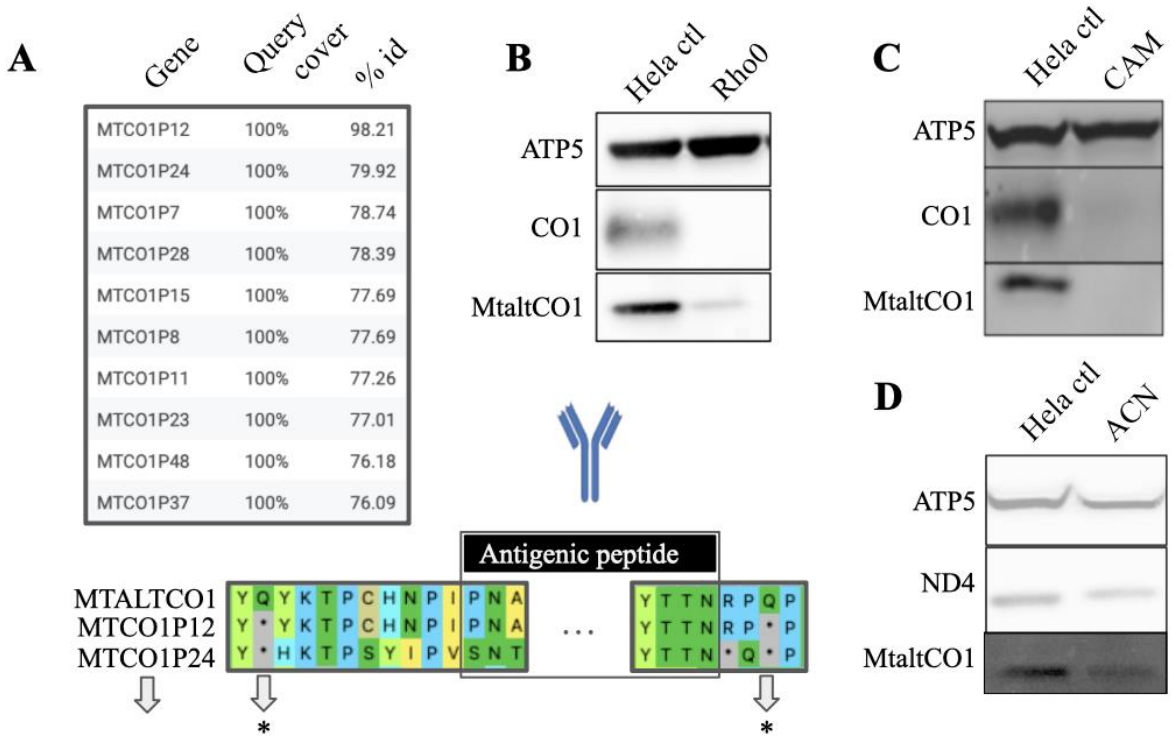

Figure 2: proof of the mitochondrial origin and translation of MTALTCO1.

(A) MTALTCO1 was blasted for nucleotides against the nuclear genome and the 10 best hits are shown (top), along with their translation in the +3 reading frame (bottom), showing that all of CO1 pseudogene's antigenic peptide sequences are flanked by stop codons. (B) MTALTCO1, mitochondrially encoded CO1 and nuclear-encoded mitochondrial ATP5 detection by Western Blot in HeLa cells (control) and Rho0 cells. (C) MTALTCO1, mitochondrially encoded CO1 and nuclear-encoded mitochondrial ATP5 detection by Western Blot in untreated HeLa cells (control) and in HeLa cells treated with chloramphenicol (CAM). (D) MTALTCO1, mitochondrially encoded NADH dehydrogenase subunit 4 (ND4) and nuclear-encoded mitochondrial ATP5 detection by Western Blot in untreated HeLa cells (control) and in HeLa cells treated with actinonin (ACN).

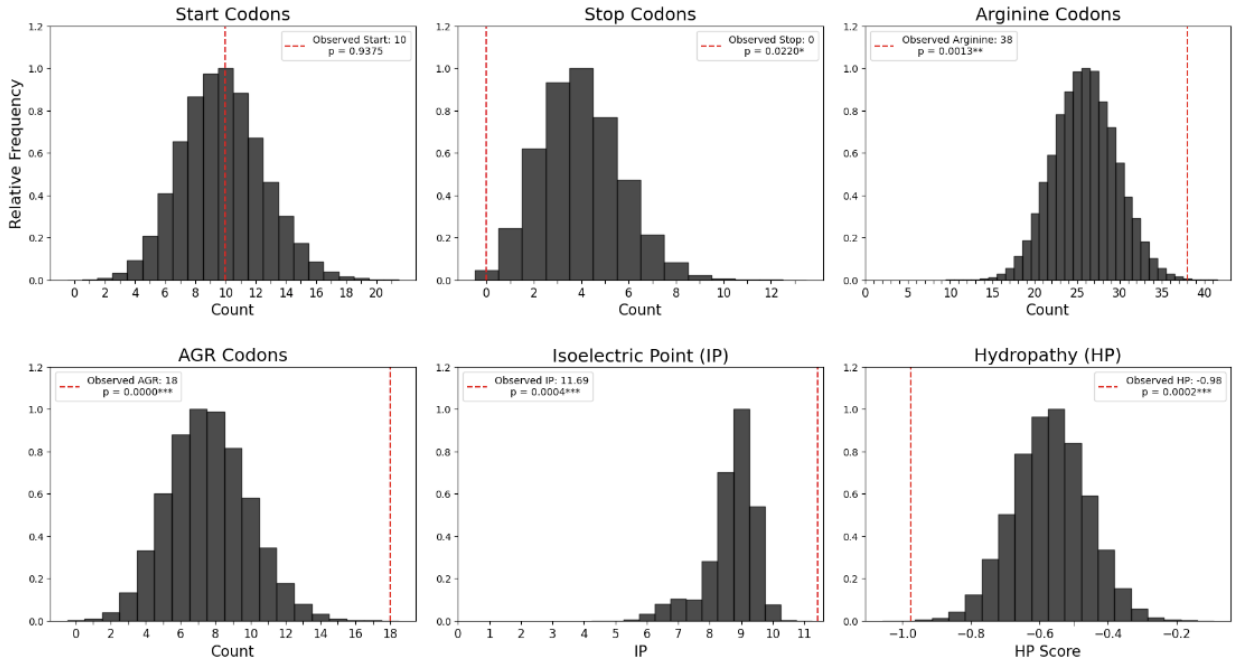

Figure 3: The CO1 sequence does not predict important characteristics of MTALTCO1. Six metrics of MTALTCO1 were compared to those of proteins translated from the +3 reading frames of generated nucleotide sequences that preserve CO1's sequence. The histogram bars represent the values from the generated sequences, while the observed value for MTALTCO1 is marked in red. The legend includes both the observed value and the p-value from the comparison.

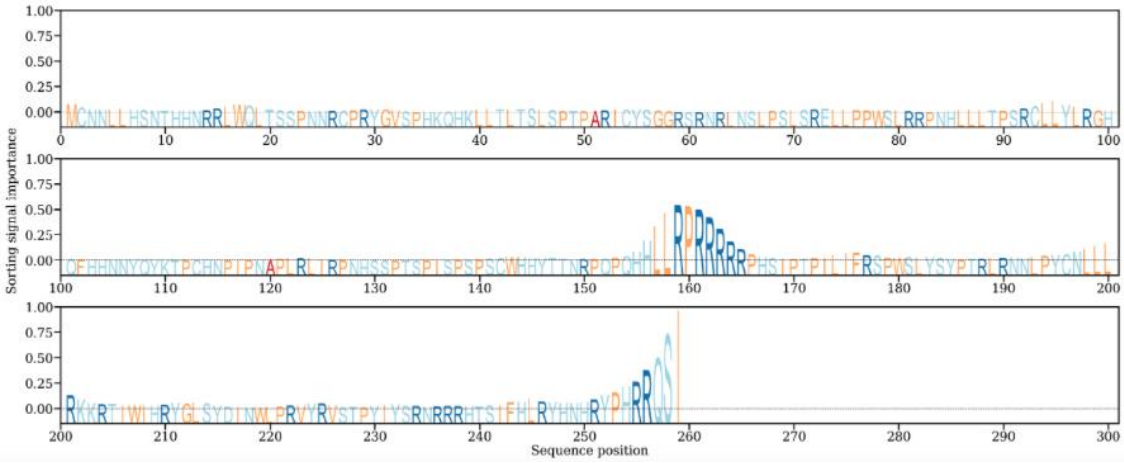

Figure 4: Deeploc predicts extracellular localization based on two motifs.

The Deeploc algorithm predicts sub-cellular localization patterns and provides the amino acid sequence of the queried protein, highlighting the importance (y-axis) of specific motifs in determining the localization prediction.

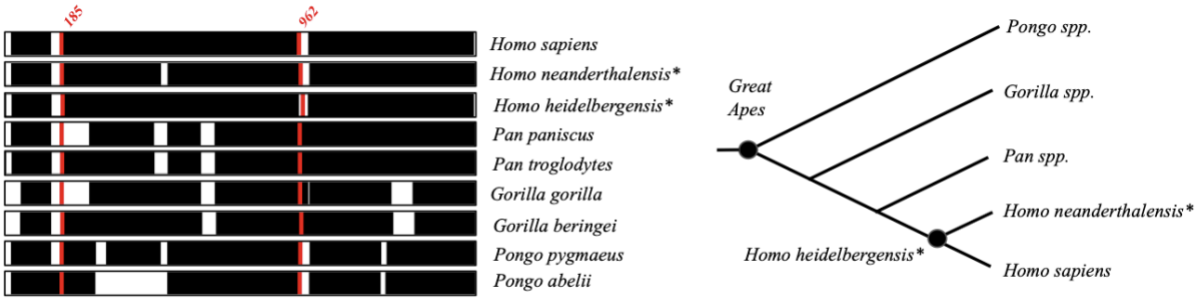

Figure 5: MTALTCO1 displays dynamic fission and fusion patterns in Hominids.

The horizontal bars represent the entire coding sequence of CO1, spanning from nucleotide 0 to 1541 in humans. White sections of the bars indicate regions not translated in the +3 reading frame, while black sections represent regions between a start and stop codon in the +3 reading frame. MTALTCO1's start and stop codons are marked in red to show the corresponding regions in other hominids. Species are arranged top to bottom in an approximately proximal-to-distal order, starting with humans, with the actual phylogeny displayed on the right. Extinct species are marked with an asterisk.

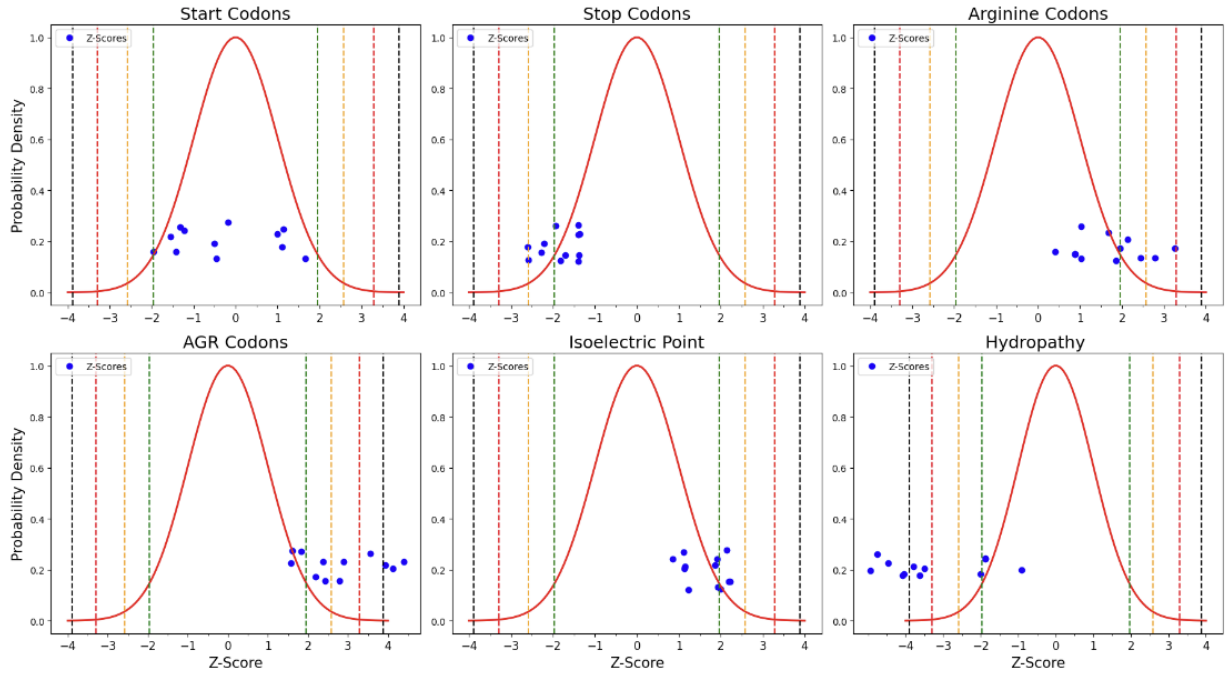

Figure 6: Some deviations of MTALTCO1 from DSS translated proteins in the +3 reading frames are conserved in Hominids.

For each metric, we displayed a generic bell curve to map Z-scores of deviations of MTALTCO1 or MTALTCO1 fission products depending on the species from DSS translated proteins in the +3 reading frames. Significance thresholds were laid out on both sides of the bell curve with green being p=0.05, yellow p = 0.01, red p = 0.001 and black p = 0.0001. Z-score points were displayed in a cloud on the vertical axis to prevent overlaps from obscuring their true density, their spread in y holds no significance beyond visibility.

>NC_000001.11:631074-632616 MTCO1P12 [organism=Homo sapiens] [GeneID=107075141] [chromosome=1]
AT + 3 GTT CGC CGA CCG TTG ACT ATT CTC TAC AAA CCA CAA AGA CAT TGG AAC ACT ATA CCT ATT ATT CGG CGC ATG AGC TGG AGT CCT AGG CAC AGC TCT AAG CCT CCT TAT TCG AGC CGA ACT GGG CCA GCC AGG CAA CCT TCT AGG TAA CGA CCA CAT CTA CAA CGT TAT CGT CAC AGC CCA TGC ATT TGT AAT AAT CTT CTT CAT AGT AAT ACC CAT CAT AAT CGG AGG CTT TGG CAA CTG ACT AGT TCC CCT AAT AAT CGG TGC CCC CGA TAT GGC GTT TCC CCG CAT AAA CAA CAT AAG CTT CTG ACT CTT ACC CCC CTC TCT CCT ACT CCT GCT TGC ATC TGC TAT AGT GGA GGC CGG CGC AGG AAC AGG TTG AAC AGT CTA CCC TCC CTT GGC AGG / **GAA CTA CTC CCA CCC TGG AGC CTC CGT AGA** /…

SUPPLEMENTARY FIGURES

Figure I: the MTCO1 pseudogene 12 does not produce a protein containing an MTALTCO1-derived peptide identified in MS data.

Start codons identified in green and stop codons in red. The ATT codon is a mitochondrial start codon but not a cytoplasmic one. The spacing between codons represents the +3 RF, the sequences in bold code for the MTALTCO1-derived peptide ELLPPWSLR.

Index	Locus	Position	rCRS nt	Variant nt	MTALTCO1 effect	verall Count in Sequence S	Overall frq	Frequency in Lineage L	Frequency in Lineage M	Frequency in Lineage N	Allele
4542	CO1	6524	T	C		272	0.44%	3.39%	0.03%	0.10%	6524C
4302	CO1	6216	T	C		321	0.53%	0.09%	0.19%	0.69%	6216C
4475	CO1	6446	G	A		345	0.56%	0.28%	2.07%	0.16%	6446A
4453	CO1	6413	T	C		351	0.57%	0.66%	0.19%	0.67%	6413C
4566	CO1	6548	C	T	Q154* pseudogenizing	356	0.58%	5.19%	0.01%	0.02%	6548T
4330	CO1	6257	G	A		407	0.67%	5.61%	0.20%	0.02%	6257A
4682	CO1	6752	A	G		417	0.68%	1.02%	2.34%	0.14%	6752G
4597	CO1	6587	C	T		428	0.70%	6.34%	0.00%	0.01%	6587T
4668	CO1	6734	G	A		426	0.70%	0.03%	0.10%	0.98%	6734A
4716	CO1	6827	T	C		426	0.70%	5.25%	0.23%	0.11%	6827C
4683	CO1	6755	G	A		435	0.71%	0.01%	0.39%	0.92%	6755A
4636	CO1	6671	T	C		451	0.74%	0.01%	1.73%	0.56%	6671C
4270	CO1	6152	T	C		471	0.77%	0.54%	0.03%	1.03%	6152C
4333	CO1	6261	G	A		496	0.81%	1.03%	0.30%	0.93%	6261A
4339	CO1	6266	A	G		545	0.89%	7.88%	0.05%	0.03%	6266G
4331	CO1	6260	G	A		576	0.94%	2.05%	0.58%	0.87%	6260A
4327	CO1	6253	T	C		592	0.97%	3.27%	1.80%	0.36%	6253C
4611	CO1	6620	T	C		659	1.08%	0.24%	3.61%	0.46%	6620C
4415	CO1	6371	C	T		682	1.12%	0.03%	0.01%	1.62%	6371T
4710	CO1	6815	T	C		893	1.46%	12.02%	0.33%	0.12%	6815C
4496	CO1	6473	C	T		1132	1.85%	0.01%	0.17%	2.64%	6473T
4641	CO1	6680	T	C		1219	1.99%	3.91%	7.46%	0.07%	6680C
4693	CO1	6776	T	C		1400	2.29%	0.15%	0.10%	3.28%	6776C
4483	CO1	6455	C	T		1679	2.75%	0.00%	12.55%	0.27%	6455T
4286	CO1	6185	T	C		1746	2.86%	25.16%	0.37%	0.05%	6185C
4305	CO1	6221	T	C	S45P, Cancer related	1774	2.90%	13.32%	0.71%	1.90%	6221C
4431	CO1	6392	T	C		2106	3.45%	0.12%	0.15%	4.95%	6392C
4660	CO1	6719	T	C		2164	3.54%	0.42%	0.12%	5.05%	6719C

Figure VI: MitoMap variants with frequencies greater than 0.50% over all lineages in MTALTCO1. The two mutants discussed in the article are shown in red.

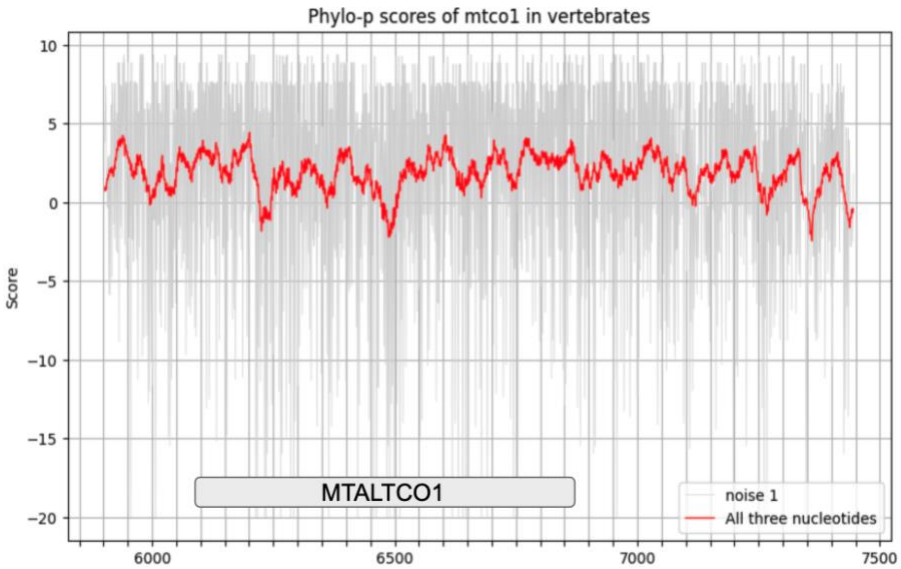

Figure III: Phylo-P scores of CO1 in 100 vertebrates smoothed with a Fast Fourier Transform over a 25-nucleotide window.

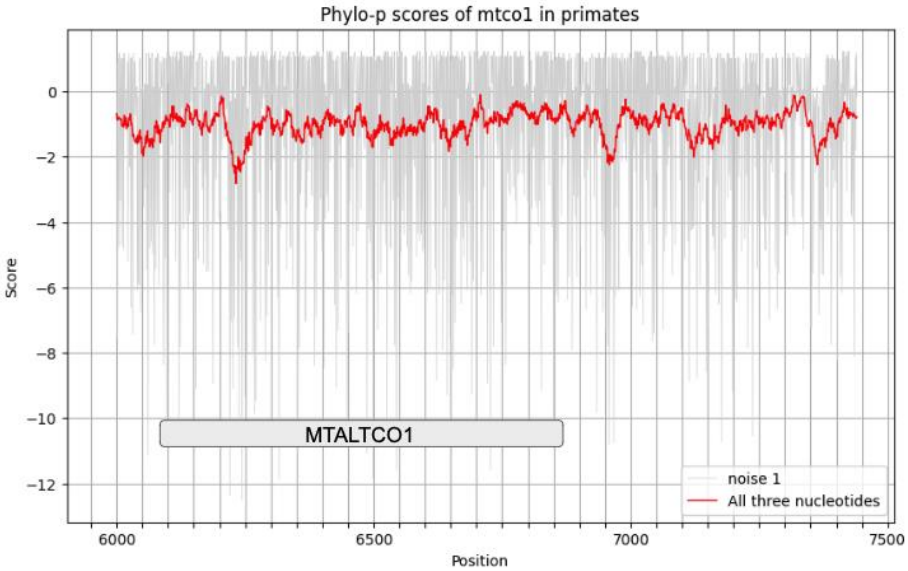

Figure IV: Phylo-P scores of CO1 in 30 primates smoothed with a Fast Fourier Transform over a 25-nucleotide window.

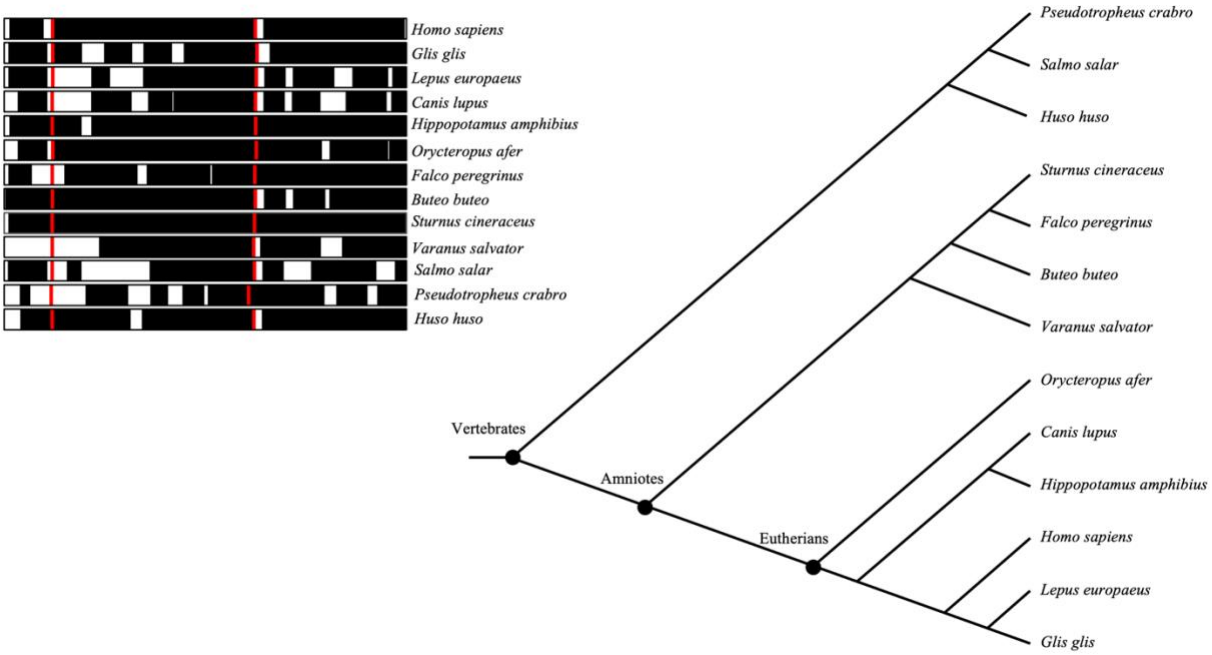

Figure V: MTALTCO1 displays dynamic fission and fusion patterns in Vertebrates.

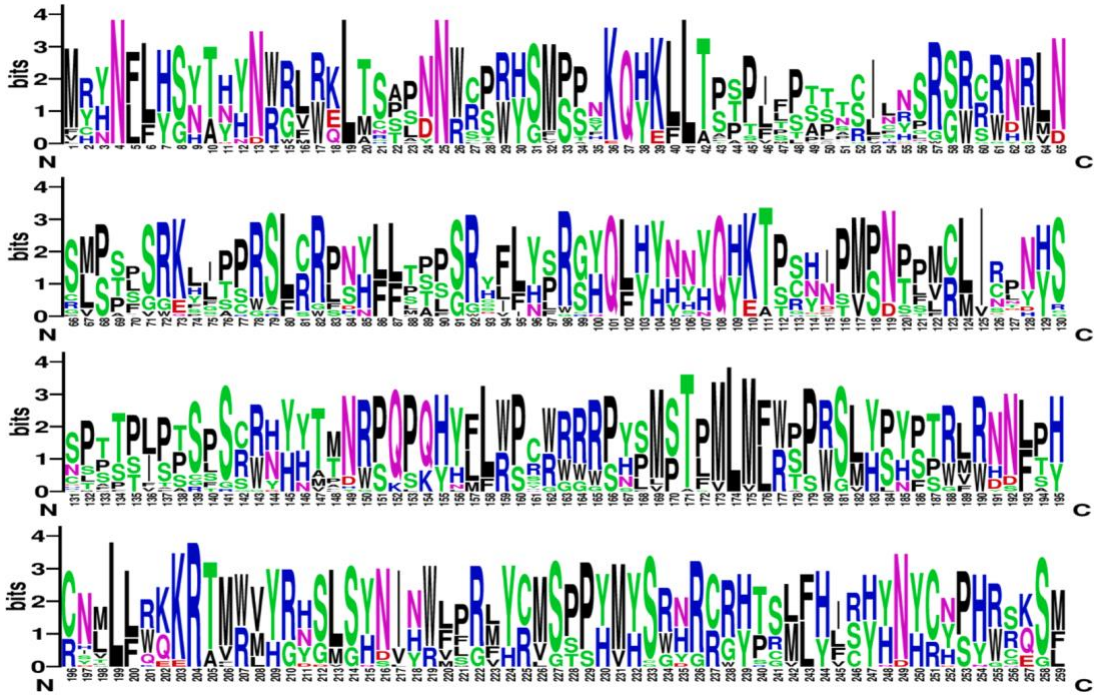

Figure VI: WebLogo of hominids sequence for MTALTCO1