RNA-DNA differences in human mitochondria restore ancestral form of 16S ribosomal RNA

Dan Bar-Yaacov, Gal Avital, Liron Levin, et al.

Genome Res. published online August 2, 2013
Access the most recent version at doi:10.1101/gr.161265.113

RNA-DNA Differences in Human Mitochondria Restore Ancestral Form of 16S Ribosomal RNA

Dan Bar-Yaacov[1]†, Gal Avital[1]†, Liron Levin[1], Allison Richards[2], Naomi Hachen[3], Boris Rebolledo Jaramillo[4], Anton Nekrutenko[4], Raz Zarivach[1] and Dan Mishmar[1*]

[1]Department of Life Sciences, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel

[2]Cell and Molecular Biology Graduate Program, University of Pennsylvania, Pennsylvania, USA

[3]Bioenginneering Program, University of Pennsylvania, Pennsylvania, USA

[4]Department of Biochemistry and Molecular Biology, Penn State University, Pennsylvania, USA

*Correspondence to:

Dan Mishmar

Department of Life Sciences

Ben-Gurion University of the Negev

Beer Sheva 84105, Israel

Tel: +972-8-6461355

Email: dmishmar@bgu.ac.il

†These authors contributed equally to this work.

Running Title: RDD in Human mtDNA Restore 16S rRNA Ancestral Form

Abstract

RNA transcripts are generally identical to the underlying DNA sequences. Nevertheless, RNA-DNA differences (RDDs) were found in the nuclear human genome and in plants and animals but not in human mitochondria. Here by deep sequencing of human mitochondrial DNA (mtDNA) and RNA, we identified three RDD sites at mtDNA positions 295 (C-to-U), 13710 (A-to-U, A-to-G) and 2617 (A-to-U, A-to-G). Position 2617, within the 16S rRNA, harbored the most prevalent RDDs (more than 30% A-to-U and ~15% A-to-G of the reads in all tested samples). The 2617 RDDs appeared already at the precursor polycistrone mitochondrial transcript. Using traditional Sanger sequencing we identified the A-to-U RDD in 6 different cell lines and representative primates (*Gorilla gorilla*, *Pongo pigmaeus and Macca mulatta*), suggesting conservation of the mechanism generating such RDD. Phylogenetic analysis of more than 1700 vertebrate mtDNA sequences supported a thymine as the primate ancestral allele at position 2617, suggesting that the 2617 RDD recapitulates the ancestral 16S rRNA. Modeling U or G (the RDDs) at position 2617 stabilized the large ribosomal subunit structure in contrast to destabilization by an A (the pre-RDDs). Hence, these mitochondrial RDDs are likely functional.

Introduction

Mitochondrial DNA (mtDNA) is one of the most variable coding sequences in humans, and many of the genetic sequence variants alter mitochondrial function and disease susceptibility (Wallace 2011). In addition, mtDNA exhibits intracellular variability (heteroplasmy), the extent of which differs across individuals (Goto et al. 2011; Avital et al. 2012). Moreover, in many organisms (vertebrates and invertebrates), RNA editing contributes a third layer of sequence diversity in the mitochondria (Reichert et al. 1998). Advances in sequencing technology have enabled comparison of DNA and RNA sequences which revealed extensive RNA-DNA sequence differences (RDDs) in the human nuclear genome (Ju et al. 2011; Li et al. 2011; Pachter 2012; Peng et al. 2012), although the identification of non-canonical RDDs (i.e. non A-to-G or C-to-U) was subjected to ongoing discussion (Kleinman and Majewski 2012; Lin et al. 2012; Peng et al. 2012; Pickrell et al. 2012; Piskol et al. 2013). Even though RNA editing is common in the mitochondria of many organisms (Knoop 2011), it has not been explored in humans. Here, by using stringent parameters to analyze mtDNA and matching RNA from different individuals and tissues, we uncovered sites in the human mitochondria where the RNA sequences do not match the underlying DNA sequences.

Results

High throughput sequence analysis detects RDDs in human mitochondria DNA: DNA and mRNA samples from cultured B-cells of five Caucasians were sequenced using Illumina technology (Bentley et al. 2008). We analyzed uniquely mapped mtDNA sequence reads that aligned to the revised Cambridge Reference Sequence (rCRS) but not to the nuclear genome sequences (hg19). We excluded sites that mapped to low complexity regions (filter A), that are

frequently misaligned among the reads due to sequencing errors. We required that all nucleotide position in our analyses have at least 1,000 sequence reads coverage (filter B) (Avital et al. 2012). Out of the 16,569 mtDNA bases, on average 16,435 and 14,752 nucleotide positions in our DNA and RNA samples, respectively, met these criteria (Supplementary Fig .1, Supplementary Table 1). We also excluded sites corresponding to heteroplasmic changes (i.e. either present in corresponding RNA and DNA reads or present only in DNA reads, but not in RNA) and excluded sample having in their secondary reads combination of mutations forming known haplotypes. By comparing RNA sequences to their corresponding DNA sequences, we found RDD sites. To be considered an RDD site, the RNA sequence that differs from the corresponding DNA sequence must be covered by reads in both directions. At least 1.6% of the reads (i.e. 0.8% from the reads of each of the strands, filter C) (He et al. 2010) and a minimum of five reads per strand at a given site (filter D) must have the RDD base. These stringent filters were imposed to minimize our false discovery rate; however, most likely they led us to exclude some true RDD sites (Avital et al. 2012). By these criteria, we uncovered 3 mitochondrial RDD sites; two of these sites were present in all five individuals (positions 2617: A-to-U and A-to-G; 13710: A-to-U and A-to-G) and the third was found in two of the five individuals (position 295: C-to-U). Manual inspection of the C295U site revealed that it is found in all individuals, but was filtered out in three of the five individuals (Table 1).

At position 2617, within the 16S rRNA of the large ribosomal subunit, we identified both A-to-U and A-to-G RDDs. In each of the tested samples, the A-to-G levels were about ~15%, which is lower than the A-to-U levels (>30%). First, Sanger sequencing of PCR products confirmed the A-to-U RDD in all five tested individuals (Fig. 1, Supplementary Fig. 2). Since we suspected that the A-to-G RDD was not detected due to the low sensitivity of direct Sanger

sequencing of the PCR fragment we cloned the PCR product encompassing position 2617 from a representative sample (GM14447) and Sanger sequenced 24 independent plasmid clones. This analysis revealed 9 clones with an adenine (no RDD), 13 with a thymine (A-to-U) and 2 with a guanine (A-to-G), thus validating both the A-to-U and A-to-G RDDs at this site.

The mtDNA RDDs do not stem from mapping errors: To exclude the possibility that the RDD site represent erroneous pseudogene sequence encoded in the nuclear genome (NUMTs), we examined RNA-sequences of isolated mitochondria (mitoplasts) from the Mattick lab (Mercer et al. 2011). The results confirmed the A-to-U and A-to-G RDD at position 2617 as well as the RDDs identified at positions 295 and 13710 (Table 1). Since the A-to-U RDD at position 2617 was found at high levels (>30%) in all subjects we focused on this site for further analyses. Examination of 9,868 publicly available whole human mtDNA sequences (www.phylotree.org) revealed only an adenine (A) at that position.

When we used mtDNA fragment positions 2567-2666 in a BLAT screen against the entire human genome (hg19), we identified 30 hits of which all had mutations in additional positions that were not identified in our entire dataset (both RNA and DNA reads), thus excluding contamination of our data by NUMTs (Supplementary Fig. 3). Moreover, 26 of the BLAT hits harbored an Adenine at the mtDNA homologous position 2617, suggesting that if some of the reads did represent NUMT leakage then these reads would not influence the presence of our identified RDDs. Finally, we obtained paired DNA and RNA samples corresponding to 6 different human cell types including normal brain cortices, colon, skeletal muscle and liver, pre-adipocyte (Chub-S7) as well as a neuroblastoma cell line (SHSY5Y). Sanger sequencing showed the A-to-U RDD in all the samples (Supplementary Fig. 2); thus the results confirmed

the findings from our deep sequencing data and showed that this RDD is found in mtDNA transcripts from different human cell types.

The 2617 RDD appeared at the precursor polycistrone mitochondrial transcript: We aimed to assess when during transcription do the RDDs form by analyzing polycistronic transcripts. MtDNA is transcribed first as precursor polycistronic transcripts and then cleaved into mature transcripts following the tRNA punctuation model (Ojala et al. 1981; Bestwick and Shadel 2013). We amplified four fragments of these precursor transcripts that encompassed the 3' and 5' junctions of the 16S rRNA gene (Supplementary Fig. 4) using cDNA from two double DNase1 treated purified RNA samples (GM14447 and GM14381). We also included a 'no-reverse transcriptase' control to exclude DNA contamination. Deep sequencing (>100,000 reads per base per sample, no strand bias) showed that the precursor polycistronic transcripts had significantly lower RDD level (GM14447 - 5.3% A-to-U, 0.4% A-to-G; GM14381 - 3.2% A-to-U, 0.2% A-to-G) compared to the deep sequencing data which encompass both mature and polycistronic mtDNA transcripts (χ^2, $p<10^{-12}$). This shows that the A-to-U and A-to-G RDDs at position 2617 likely start as early as the emergence of the polycistronic transcript. Although the sequence reads stem from our purely amplified polycistronic fragments, we cannot disregard the possibility that the RDDs reflect remains of the mature transcript. However notably, we did not detect any sequence reads corresponding to mature transcripts from any mtDNA regions outside of our polycistronic fragment (supplementary Table 2).

The 2617 RDD is present in other primates and restores ancestral form of the 16S rRNA: We next assessed the evolutionary conservation of position 2617. Alignment of DNA sequences from 1,755 vertebrates revealed that the homologous positions to human mtDNA position 2617 have either an adenine (A) or a thymine (T) in 1,752 species (supplementary Figs. 5 and 6).

Within primates, most simians have an adenine (A) except prosimians (slow loris, tarsier and various lemur species) who have a thymine (T) and one lemur sub-species with a cytosine (C) (Fig. 2, Fig. 3 and supplementary Fig. 6). Non- primate mammals that are phylogenetically closest to primates have a thymine at the corresponding position (Supplementary Fig. 6). Together, these data show that thymine is the primate ancestral allele at this position. We then asked if RDD is found at this site in other species besides human. Sanger sequencing of regions homologous to human mtDNA position 2617 in representative non-human primates revealed an A-to-U RDD in organisms where their mtDNA harbors an A (*Pongo pigmaeus*, *Gorilla gorilla* and *Macaca mulatta*) but no RDD in *Lemur katta* and *Nycticebus coucang* where their genomic sequence is a T at this site (Fig. 4, supplementary Fig. 7). Thus, RDD at position 2617 occurred in organisms where the genomic DNA sequence is A, but not in those with a T in their DNA; this suggests that the RDD event converts the A to recapitulate the ancestral T state.

Modeling the 2617 RDDs reveal stabilization of the ribosome structure in contrast to destabilization by the DNA original base: Finally, we tested whether the RDD at position 2617 affect the structure of 16S rRNA. In the absence of the human mitochondrial ribosomal high resolution structure, we analyzed the closely related bacterial and yeast 23S rRNA (Mears et al. 2006). We found striking structural conservation of a stem and loop structure (H71) of the large ribosomal subunit lying within the interaction interface with the small ribosomal subunit and at the tRNA entrance channel (Fig. 5). In *Escherichia coli*, position 1954 which is the homologous position to human mtDNA position 2617, harbors a guanine and in the determined structures of the nuclear *Homo sapiens* (Anger et al. 2013) and *Saccharomyces cerevisiae* (Ben-Shem et al. 2011) rRNA, it is a uracil. We modeled a cytosine, a guanine or an uracil in this position and found that all could be accommodated without changing the local rRNA fold. This

accommodation could be explained either by direct formation of a hydrogen bond between the guanine and the rRNA backbone or by indirect hydrogen bonds that are mediated by water or ion molecules between a pyrimidine and the rRNA backbone at this position (Fig. 5). However, the model suggests that an adenine at this position will abolish the potential hydrogen bond to H64 backbone. Thus, the A-to-U and A-to-G RDDs at this position recapitulate the secondary structure of the bacterial rRNA loop and therefore likely stabilize the ribosome structure. Interestingly, although the human nuclear DNA encoded rRNA harbors a structurally conserved stem and loop to H71 (Fig. 5), the homologous position to mtDNA 2617 harbors a uracil, as in the mtDNA RDD. Moreover, the RNA reads of this position in our five human analyzed individuals were identical to the DNA template, i.e. harboring a Thymine (100,000 sequence reads coverage, 99.95% T, 0.05% of the reads could be regarded as sequencing errors). These results further support the need for adenine replacement in the 16S rRNA at position 2617.

Discussion

This is the first report of RDDs in human mitochondria. We showed that the RDDs in position 2617 were present already in the polycistronic RNA molecule, though in lower levels (an order of magnitude) as compared to the total mtDNA transcript analysis. Hence, we suggest that the RDDs start either co-transcriptionally or immediately after the synthesis of the RNA molecule. We interpret the increase in RDD levels in the total RNA sample as the result of either increased stability of the RDD-containing transcript or continuation of RDD generation during the maturation process of the 16S rRNA molecule. Sequence analysis of over 1,700 organisms revealed that while the human mtDNA sequence at position 2617 is an adenine, the primate ancestral base is a thymine. Thus, mtDNA RDD formation in humans recapitulates the primate ancestral 16S rRNA. In other primates where the DNA base at this position is an adenine, an RDD changed the RNA bases to uracil but no RDD was found in organisms where the DNA base was thymine. Functionally, position 2617 is embedded within a very important region in the ribosome, harboring the position where the small and large subunits of ribosomal RNA interface with the tRNA. Thus our observed importance of the RDD for the 16S rRNA secondary structure is likely functional. Recently we learned that a non-canonical A-to-U RDD is essential for intron processing of the tRNA-Tyr gene in the nuclear genome of *Trypanosoma brucei*, for the translation process and for cell life (Rubio et al. personal communication). Thus, similar to our results, the A-to-U RDD is functionally important. Moreover, this finding corroborates our identification of a non-canonical A-to-U RDD in a phylogenetically distant species, thus supporting the existence of previously unidentified RNA processing machinery.

Heated discussion about the false discovery rate of RDDs mainly stem from mapping errors (i.e. false interpretation of sequence alterations in pseudogenes and gene paralogues as RDDs), false

RDD identification at the end of sequence reads, misalignment of exon-intron boundaries and misinterpretation of rare polymorphisms as RDDs (Kleinman and Majewski 2012; Lin et al. 2012; Peng et al. 2012; Pickrell et al. 2012; Piskol et al. 2013). Firstly, unlike the nuclear genome human mtDNA-encoded genes have no active paralogues in other loci. Second, we used only uniquely mapped reads to exclude mapping errors. Furthermore, although the region harboring our identified 2617 RDDs had sequence similarity to 30 nuclear DNA loci (which likely are NUMTs), all of these loci had mutations in additional positions that were not identified in our entire dataset (both RNA and DNA reads), thus excluding NUMT contamination and mapping errors. Thirdly, strand bias was addressed by the usage of our filters C and D (also see supplementary Fig. 8 and supplementary Table 3). Fourth, since the human mtDNA genes do not undergo splicing, exon intron boundaries cannot explain false discovery in our study. Finally, we show that neither our identified mtDNA RDDs occur at the end of the reads (supplementary Fig. 8). We thus conclude that our approach identified true human mtDNA RDDs.

How were the 2617 RDDs formed? Since this position harbors both A-to-U and A-to-G RDDs three major possibilities come to mind: (A) First, it is possible that two separate enzymes modify the adenosine at this position, one replacing it for an uracil and the other (possibly an ADAR-like enzyme, reviewed in: (Knoop 2011)) for an inosine (or a guanine). (B) Replacing the adenine for an uracil in two steps: an ADAR-like enzyme replaces the adenine for an inosine, which in turn is replaced for an uracil by another enzyme. (C) A single enzyme that replaces the 2617 adenine at the 16S rRNA for an unknown modified nucleotide which is read by the resultant sequence mainly as an uracil but also as a guanine. With this in mind, although our sequence analysis detected an A-to-U and A-to-G RDDs it is possible that these RDDs constitute unknown base

alterations that are read ultimately as T and G, respectively. These alternatives also apply to position 13710 which harbored the same RDDs.

Although three mtDNA positions harbored RDDs (positions 295, 13710 and 2617), the functional potential of the RDDs within positions 295 and 13710 is not easy to interpret. That is, since the first position (295) lies within the non-coding D-loop and the latter (13710) alter a 3rd codon position of a relatively prevalent amino acid (alanine) in the ND5 gene. In contrast, as mentioned above, the high sequence and structural conservation of position 2617 within the 16S rRNA, its high RDD level in all tested samples and its occurrence in all tested species underlined the functional potential of RDDs at this position. We speculate that the functional importance of the 2617 RDDs could be further investigated in conditions when the RDD occurrence is perturbed, possibly in patients exhibiting mitochondrial translation defects (Rotig 2011).

If a U but not an A at position 2617 is important for mitochondrial ribosome function, why did an A become fixed at the mtDNA of so many vertebrate nodes? This question raises the possibility that the fixation of 2617A was due to a ribosome independent negative selection, i.e. selective pressure acting on the DNA sequence at this position, independent of the selection which acts on that position at the RNA level. This further suggests another dimension for the functional importance of this position.

Materials and Methods:

Cell culture and tissue samples: Lymphoblastoid cell lines derived from 5 female Caucasian individuals from an apparently healthy collection (GM14432/452/468/447/381) were grown in suspension in RPMI 1640. Chub S7 preadipocytes (Darimont et al. 2003) were grown in DMEM/F12 culture media 1:1 (v/v); Human neuroblastoma cell line T-Rex SHSY5Y (Lee et al.

2007) were grown in DMEM (High glucose). All growth media were supplemented with 10% fetal calf serum (FCS), 2mM L-glutamine, 100 U/ml penicillin, and 100µg/ml strepotomycin and were grown in 5% CO_2 at 37^{o}C. Seven normal colon RNA and DNA sample pairs were purchased from Asterand (samples catalog numbers: 107807B1, 110476B1, 1112467F, 1118987F, 112964A1, 113003A1, 126828A1). Human brain cortex, skeletal muscle and liver tissue samples were obtained from National Disease Research Interchange from six individuals (64998, 65080, 65288, 65699, 65777 and 65914). Tissues were collected 7-12 hours post-mortem during routine autopsies of donors that suffered from respiratory or cardiac failure. All individuals were Caucasians between ages 62 and 79 years, including both males and females. Samples were snap-frozen and kept at -80^{o}C until DNA/RNA extraction.

DNA and RNA extraction from cell lines and tissues: DNA was extracted using the Genomics DNA Extraction Mini Kit (RBC BIOSCIENCE) and RNA was extracted using the PerfectPure RNA Cell & Tissue Kit (5 Prime), following Manufacturer Protocol.

DNA from brain cortex, skeletal muscle and liver tissue was extracted using Gentra Puregene Tissue Kit (Qiagen). RNA was extracted from brain cortex using MaXtract High Density Kit (Qiagen), from skeletal muscle using RNeasy Maxi Kit (Qiagen), and from liver using RNeasy Lipid Tissue Mini kit (Qiagen). All extractions followed the manufacturer protocol.

cDNA synthesis: 1µg of total RNA was subjected to cDNA synthesis using iScript cDNA Synthesis Kit (BIO-RAD), following manufacturer protocol.

Total RNA from brain cortex, skeletal muscle and liver tissues were converted into cDNA using Taqman reverse transcription reagents with random hexamer priming following manufacturer protocol (Applied Biosystems).

Massive parallel deep sequencing: DNA was extracted from the 5 lymphoblastoid cell lines (see 'cell culture') and libraries were prepared using TruSeq Paired End Kit (Illumina). RNA was collected from the same cell lines and libraries were prepared according to the manufacturer protocol of TruSeq RNA Kit (Illumina). Both DNA and RNA were sequenced using HiSeq 2000 instrument (Illumina). DNA libraries were sequenced using 100 nucleotide, paired end reads. RNA libraries were sequenced with 100 nucleotide, single end reads.

Analysis of Illumina data: Illumina sequencing reads were aligned against the hg19 (Genbank, GCA_000001405.1). In order to identify mtDNA sequences we utilized the rCRS (revised Cambridge reference sequence, Genbank NC_012920). BWA sequence alignment tool was used (Li and Durbin 2009) following default protocol of the 1000 genome sequence analysis (ftp.1000genomes.ebi.ac.uk/vol1/ftp/README.alignment_data). Only reads that were uniquely aligned to the rCRS were used for further analyses. SAMtools (Li et al. 2009) was used to convert the SAM to BAM sequence format. MitoBam Annotator (Zhidkov et al. 2011) was used to identify secondary read changes either in DNA or both in corresponding RNA and DNA samples; these changes were considered heteroplasmic and were excluded from further analyses; RNA specific secondary reads were considered RDDs. Secondary read changes were considered high quality only if they occurred outside of low complexity regions (filter A), if they were identified by at least a 1000 high quality sequence reads (filter B), if their minimal read fraction

was at least 1.6% (i.e. 0.8% from the reads of each of the strands, filter C) (He et al. 2010) and if their minimal sequence read count per strand was more than 5 (filter D).

PCR amplification and direct sequencing of mtDNA and corresponding RNA fragments encompassing positions 2617: PCR reaction contained primers 1 and 2 for human samples *or* 3 and 4 for primate samples (supplementary Table 4), 0.5 unit Phusion Taq polymerase and 1 x reaction buffer (FINNZYMES), 2.5mM dNTPs mix, and either 30 ng DNA or 1µl of cDNA used as templates. Reaction conditions: 98°C for 5 minutes, 30 cycles including denaturation (98°C, 15 sec), annealing (70°C - humans or 67°C – primates, 30 sec) and elongation (72°C, 30 sec – humans, 10 sec primates), and a final extension step (72°C, 7 min). The reaction was stored at -20˚C until usage. PCR products were visualized on an EtBr-stained 1% Agarose gel, purified using Wizard SV Gel and PCR Clean-up system (Promega), following manufacture protocol and were sequenced (ABI 3100) using the amplification primers (BGU sequencing facility). The brain cortex, skeletal muscle and liver samples were PCR amplified using primers 5 and 6 (supplementary Table 4), and Phusion HotStart DNA Polymerase (Thermo Scientific). Reaction conditions: 98^{o}C for 30 seconds, followed by 35 cycles including denaturation (98^{o}C, 10 sec), annealing (58^{o}C, 30 sec.), elongation (72^{o}C, 90 sec), and a final extension step (72^{o}C, 10 min.). PCR products were sequenced using 3730 DNA analyzer (Applied Biosystems) with the amplification primers. All sequences were aligned against the rCRS (Sequencher 4.10.1, GeneCodes inc.)

Cloning and sequencing: The purified PCR product from sample GM14447 was ligated into pGEMT-Easy vector (Promega). The ligation reaction was performed using T4 ligase (Promega) with vector/insert ratio of 1:3 at room temperature for one hour according to

the instruction of the DNA Ligation Kit (Promega). Then, 5 µl of the ligation reaction were mixed with 50 µl of competent *E.coli* cells (DH5α) and were subjected to electric shock using GenePulserXcell (BIO-RAD). Following electroporation, 500 µl of LB were added, and cells were shaken gently at 37°C for 1 hour. The bacteria were plated onto LB Petri dishes containing 50 µg/ml ampicillin, 40 µl of 0.1 M IPTG and 40 µl of 2%-X-gal, and grown overnight at 37°C. Following 'blue/white' colony selection, white insert-containing colonies were isolated, and grown in 5 ml liquid LB with 100 mg/ml ampicillin at 37°C for 12 hours, shaking. Plasmid DNA was purified using Wizard plus SV minipreps DNA purification system (Promega), according to the manufacturer protocol. Each plasmid was sequenced in the ABI 3100 sequencing machine using SP6 standard primer (BGU sequencing facility). Sequences generated from each plasmid were aligned against the rCRS using Sequencher 4.10.1 (GeneCodes inc.)

Phylogenetic analysis: Whole mtDNA sequences from 1755 vertebrates were downloaded from NCBI organelle resources (www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi) and the 16S rRNA gene sequences were extracted and aligned using MAFFT (mafft.cbrc.jp/alignment/server/). The same approach was used to align whole mtDNA sequences from 334 *Eutheria* (placental mammals). Maximum likelihood phylogenetic tree was constructed, with 1000 bootstrap replicates; this enabled us to predict the ancestral nucleotide of all major phylogenetic clades.

Second DNase treatment and RNA purification: In order to exclude DNA contamination from our RNA samples, we subjected RNA to a second round of DNase 1 treatment according to manufacturer instructions (5 prime #FP-2500120), in addition to the treatment performed during

the RNA purification protocol. RNA was then isolated by isopropanol precipitation: 3M sodium acetate was added up to 10% of the RNA solution (5 µg RNA) volume. Then room-temperature isopropanol (0.7 reaction volume) were added to the solution, following by 30 minutes centrifugation in 4ºC in 14,000g. The supernatant was carefully removed, and 180 µl of 70% ethanol was added, followed by another round of centrifugation for 15 minutes in 4ºC in 14,000g. The supernatant was removed carefully, and 25 µl of RNA elution buffer (5 prime #FP-2500120) was added.

Inspection of position 2617 in the polycistronic RNA molecule: We PCR amplified four fragments encompassing the 16S rRNA gene and flanking coding sequences, representing pre-cleavage polycistronic mtDNA molecule (Supplementary Fig. 4). Template for the fragments amplification was cDNA generated from GM14447 and GM14381 RNA samples (double DNase1 treatment to avoid residual DNA contamination – see above) in which we already identified the 2617 A-to-U and A-to-G RDDs. A 'no-reverse transcriptase' control of the above mentioned RNA samples was included. PCR reactions were as described above using the following primer couples (supplementary Table 4): Fragment 1 - primers 7 and 8; Fragment 2: primers 7 and 9, Fragment 3 – primers 10 and 11; Fragment 4 - primers 11 and 12. Reaction conditions: 98°C for 5 minutes followed by 30 cycles, each including denaturation (98°C, 15 sec), 20 sec annealing (Fragment 1,2 - 68°C; Fragment 3,4 - 63°C), elongation (72°C, 30 sec) and a final extension step (72°C, 7 minutes). The reactions were concluded at 10˚C and stored in -20˚C until usage. As a negative control for these amplification reactions we directly used RNA (not cDNA) extracted from sample GM14447. PCR products were visualized on an Agarose gel and purified. Phosphate was added to the 5' ends of all fragments using T4 Polynucleotide

Kinase (Fermentas - #EK0031) according to manufacturers' instructions to allow adaptors ligation for sequencing in Illumina MiSEQ platform without shearing (Technion Genome Center, Israel). A total of 404,509 and 242,395 reads were generated for individual GM14447 and GM14381, respectively. The reads were mapped to the rCRS and analyzed using MITOBAM annotator as described above.

Ribosome structure visualization: Six available unique large ribosomal subunits structures (3R8S, 1JJ2, 2ZJR, 2J01, 3J3F and 3U5D) were overlapped by coot (Emsley et al. 2010) using Least Squares Quadratic method (LSQ) with the sequence range of (1900-2000ec) onto *Dianoccocus radiodurance* 50S ribosomal subunit structure 1NKW (Harms et al. 2001). Structural visualization and figure preparation were performed using PyMole (DeLano 2002).

Data Access: The deep sequencing raw data from this study have been submitted to the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena/) under study accession nos. ERP001523 (DNA) and ERP002075 (RNA).

Acknowledgments: This work was funded by the Israeli Science Foundation grant to D.M. (ISF 387/08). The authors deeply thank Prof. Vivian G Cheung for providing samples and data, critical reading, editing and discussion of the manuscript. We are grateful to Dr. Andrey Krasilnikov and Dr. Craig Cameron for their advice on rRNA chemistry and mitochondrial transcription. We would also like to thank Prof. Ada Yonath (Weizmman institute) and Prof. Batsheva Kerem (Hebrew University, Jerusalem) for critical reading of the manuscript.

Author Contributions: DM and DBY – conceived the study; DBY and GA – analyzed the samples, data, and participated in study design; LL – analyzed the data; NH, AR, BRJ and AN – analyzed the samples; RZ – analyzed the structural data; DM wrote the manuscript.

Disclosure Declaration: The authors declare no conflict of interest with any party regarding the data and conclusions presented in this manuscript.

Figure Legends:

Figure 1: Validation of the RDD at position 2617 by Sanger sequencing. The numbers from each side of the sequences correspond to mtDNA positions. Red arrow points at position 2617. Presented is a representative sample - cultured B-cells from GM14468.

Figure 2: Phylogenetic analysis of mitochondrial DNA sequences of 62 Primates and 1 Dermoptera (*Galeopterus variegatus*). The primate portion of the phylogenetic analysis of mtDNA *Ehutherian* sequences from 334 organisms; the full vertebrate tree is shown in Supplementary Fig. 6. Numbers on the branches are scores from 1000 bootstrap replicates. Ancestral state of position 2617 is indicated for each branch.

Figure 3: Multiple sequence alignment of mitochondrial DNA sequences of 62 Primates and 1 Dermoptera (Galeopterus variegatus). Shown are primate orthologs of human mtDNA positions 2608-2624 (framed is the nucleotide at position 2617). Stem, Loop, Stem columns – mtDNA sequence corresponding to the stem-and-loop structure of 16S rRNA around position 2617. The full vertebrate sequence alignment is in Supplementary Fig. 5.

Figure 4: Sanger sequence analysis of position 2617 from DNA and RNA samples of mtDNA from various primates: *Pongo pigmaeus*: (a) DNA and (b) RNA; *Nycticebus coucang* (c) DNA and (d) RNA. Red arrow indicates nucleotide position corresponding to position 2617 in human mtDNA. Numbers on top of each base indicates positions according to the rCRS.

Figure 5: Structure of the ribosome section corresponding to region orthologous to that of position 2617. Left - the large ribosomal subunit from a bacterium, *Deinococcus radiodurans,* represented as ribbon. The A-, P- and E- binding sites of tRNA on the ribosomal large subunit are shown. Right - sticks and ribbon representation of H71 and H64 interaction in the bacterial (blue) or human (yellow) orthologous ribosomes. The hydrogen bond that is disrupted by an adenine in position 2617 is represented as a dashed line. Numbers represent positions of *E.coli* ribosomal RNA.

Tables

Individual	GM14468		GM14452		GM14381		GM14447		GM14432		Mitoplast (Mercer et al. 2011)	Locus
Position in rCRS	RNA	DNA	RNA	DNA	RNA	DNA	RNA	DNA	RNA	DNA	RNA	
295	C 97.15% T 2.85%	C 100%	C# 99.40% T# 0.60%	C 100%	C* 92.63% T* 7.37%	C 100%	C 90.37% T 9.63%	C 100%	C# 98.35% T# 1.65%	C 100%	C* 96.66% T* 3.34%	Dloop
2617	A 35.57% T 47.39% G 15.89% C 1.15%	A 100%	A 41.00% T 43.23% G 14.60% C 1.17%	A 100%	A 43.81% T 41.17% G 14.10% C 0.92%	A 100%	A 39.38% T 46.55% G 13.06% C 1.01%	A 100%	A 54.07% T 32.81% G 12.26% C 0.86%	A 100%	A 11.70% T 56.85% G 28.90% C 2.56%	16S rRNA
13710	A 93.49% T 2.82% G 3.55% C 0.14%	A 100%	A 92.36% T 3.60% G 3.93% C 0.1%	A 100%	A 95.50% T 1.93% G 2.45% C 0.12%	A 100%	A 94.90% T 2.27% G 2.74% C 0.09%	A 100%	A 96.20% T 1.69% G 2.03% C 0.08%	A 100%	A* 93.17% T* 5.12% G* 1.71% C* 0%	ND5

Table 1: RDDs and their levels in B-cells from 5 individuals. * RDDs removed by filter B; # RDDs removed by filter C.

References

Anger AM, Armache JP, Berninghausen O, Habeck M, Subklewe M, Wilson DN, Beckmann R. 2013. Structures of the human and Drosophila 80S ribosome. *Nature* **497**(7447): 80-85.

Avital G, Buchshtav M, Zhidkov I, Tuval Feder J, Dadon S, Rubin E, Glass D, Spector TD, Mishmar D. 2012. Mitochondrial DNA heteroplasmy in diabetes and normal adults: Role of acquired and inherited mutational patterns in twins. *Hum Mol Genet* **21**(19): 4214-4224.

Ben-Shem A, Garreau de Loubresse N, Melnikov S, Jenner L, Yusupova G, Yusupov M. 2011. The structure of the eukaryotic ribosome at 3.0 A resolution. *Science* **334**(6062): 1524-1529.

Bentley DR Balasubramanian S Swerdlow HP Smith GP Milton J Brown CG Hall KP Evers DJ Barnes CL Bignell HR et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218): 53-59.

Bestwick ML, Shadel GS. 2013. Accessorizing the human mitochondrial transcription machinery. *Trends Biochem Sci* **38**(6): 283-291.

Darimont C, Zbinden I, Avanti O, Leone-Vautravers P, Giusti V, Burckhardt P, Pfeifer AM, Mace K. 2003. Reconstitution of telomerase activity combined with HPV-E7 expression allow human preadipocytes to preserve their differentiation capacity after immortalization. *Cell Death Differ* **10**(9): 1025-1031.

DeLano WL. 2002. *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos.

Emsley P, Lohkamp B, Scott WG, Cowtan K. 2010. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**(Pt 4): 486-501.

Goto H, Dickins B, Afgan E, Paul IM, Taylor J, Makova KD, Nekrutenko A. 2011. Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. *Genome Biol* **12**(6): R59.

Harms J, Schluenzen F, Zarivach R, Bashan A, Gat S, Agmon I, Bartels H, Franceschi F, Yonath A. 2001. High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell* **107**(5): 679-688.

He Y, Wu J, Dressman DC, Iacobuzio-Donahue C, Markowitz SD, Velculescu VE, Diaz LA, Jr., Kinzler KW, Vogelstein B, Papadopoulos N. 2010. Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature* **464**(7288): 610-614.

Ju YS, Kim JI, Kim S, Hong D, Park H, Shin JY, Lee S, Lee WC, Kim S, Yu SB et al. 2011. Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nat Genet* **43**(8): 745-752.

Kleinman CL, Majewski J. 2012. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* **335**(6074): 1302; author reply 1302.

Knoop V. 2011. When you can't trust the DNA: RNA editing changes transcript sequences. *Cell Mol Life Sci* **68**(4): 567-586.

Lee YJ, Miyake S, Wakita H, McMullen DC, Azuma Y, Auh S, Hallenbeck JM. 2007. Protein SUMOylation is massively increased in hibernation torpor and is critical for the cytoprotection provided by ischemic preconditioning and hypothermia in SHSY5Y cells. *J Cereb Blood Flow Metab* **27**(5): 950-962.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. 2011. Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333**(6038): 53-58.

Lin W, Piskol R, Tan MH, Li JB. 2012. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* **335**(6074): 1302; author reply 1302.

Mears JA, Sharma MR, Gutell RR, McCook AS, Richardson PE, Caulfield TR, Agrawal RK, Harvey SC. 2006. A structural model for the large subunit of the mammalian mitochondrial ribosome. *J Mol Biol* **358**(1): 193-212.

Mercer TR, Neph S, Dinger ME, Crawford J, Smith MA, Shearwood AM, Haugen E, Bracken CP, Rackham O, Stamatoyannopoulos JA et al. 2011. The human mitochondrial transcriptome. *Cell* **146**(4): 645-658.

Ojala D, Montoya J, Attardi G. 1981. tRNA punctuation model of RNA processing in human mitochondria. *Nature* **290**: 470-474.

Pachter L. 2012. A closer look at RNA editing. *Nat Biotechnol* **30**(3): 246-247.

Peng Z, Cheng Y, Tan BC, Kang L, Tian Z, Zhu Y, Zhang W, Liang Y, Hu X, Tan X et al. 2012. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol* **30**(3): 253-260.

Pickrell JK, Gilad Y, Pritchard JK. 2012. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* **335**(6074): 1302; author reply 1302.

Piskol R, Peng Z, Wang J, Li JB. 2013. Lack of evidence for existence of noncanonical RNA editing. *Nat Biotechnol* **31**(1): 19-20.

Reichert A, Rothbauer U, Morl M. 1998. Processing and editing of overlapping tRNAs in human mitochondria. *Journal of Biological Chemistry* **273**(48): 31977-31984.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**(1): 24-26.

Rotig A. 2011. Human diseases with impaired mitochondrial protein synthesis. *Biochim Biophys Acta* **1807**(9): 1198-1205.

Ruiz-Pesini E, Wallace DC. 2006. Evidence for adaptive selection acting on the tRNA and rRNA genes of human mitochondrial DNA. *Hum Mutat* **27**(11): 1072-1081.

Wallace DC. 2011. Bioenergetic origins of complexity and disease. *Cold Spring Harb Symp Quant Biol* **76**: 1-16.

Zhidkov I, Nagar T, Mishmar D, Rubin E. 2011. MitoBamAnnotator: A web-based tool for detecting and annotating heteroplasmy in human mitochondrial DNA sequences. *Mitochondrion* **11**(6): 924-928.

Supplementary Material:

Supplementary Table

DNA					
ID	**# of reads**	**# of mapped reads to rCRS**	**% mapped reads to rCRS**	**# of positions covered**	**Mean coverage**
GM14381	677294985	1121921	0.17	16340	5013
GM14452	709074221	1906023	0.27	16529	9587
GM14468	746560279	1660250	0.22	16494	8164
GM14432	1173348066	1302342	0.11	16440	5835
GM14447	587181365	1028243	0.18	16376	4189
RNA					
ID	**# of reads**	**# of mapped reads to rCRS**	**% mapped reads to rCRS**	**# of positions covered**	**Mean coverage**
GM14381	88,561,952	3,436,982	3.9	14,377	15,916
GM14452	108,180,232	7,591,065	7.0	15,196	34,961
GM14468	131,293,549	5,524,946	4.2	14,825	25,112
GM14432	100,107,660	4,894,867	4,9	14,792	23,022
GM14447	89,433,748	4,600,282	5.1	14,570	21,636

Supplementary Table 1: Summary statistics of deep sequencing data. The percentage of mapped reads to rCRS is the percentage of reads that uniquely mapped to the revised Cambridge Reference Sequence (rCRS, NC_012920); Number of positions covered: the number of positions within the mtDNA that had a minimal coverage of 1,000 sequence reads. In addition the mean coverage per position is indicated.

Supplementary Table 2: Mapping of MiSEQ reads derived from purely amplified polycistronic fragments to the rCRS. The reads are derived from the purely amplified polycistronic fragments, as evident from the exclusive sequence coverage between positions 1580 – 3414. Notice the extreme high coverage around position 2617 and adjacent to the beginning of each of the PCR amplified fragments – sequencing was performed using already existing PCR fragment edges without shearing (see Materials and Methods). Notably, sequence reads were virtually absent outside of our amplified fragments (representing either polycistronic or mature transcripts from other mtDNA regions), thus reducing the possibility that the RDDs in

position 2617 were derived from leakage of mature 16S rRNA into our sequencing data. Position 2617 is highlighted in yellow.

Supplementary Table 3: RNA annotation table of all 5 tested individuals.

pos – mtDNA position according to the rCRS; hq – high quality coverage; indels – number of indels identified at this position; best – the best base identified at a given position; best count – high quality reads harboring the best base; second_best – second identified best base; second_best_count - high quality reads harboring the second best base; second_best_for - high quality reads harboring the second best base forward reads; second_best_rev - high quality reads harboring the second best base reverse reads; fraction_percent – percent of the second best base out of the entire high quality reads; A or C or T or G – number of high quality reads harboring an adenine, a cytosine, a thymine or a guanine respectively in a specific position; A.for (or C.for or T.for or G.for) - number of forward high quality reads harboring an adenine, a cytosine, a thymine or a guanine respectively in a specific position; A.rev (or C.rev or T. rev or G. rev) number of reverse high quality reads harbor Adenin, a cytosine, a thymine or a guanine respectively in a specific position; best_count.for - high quality forward reads harboring the best base; best_count.rev - high quality reverse reads harboring the best base; LCR – low complexity repeat regions; gene – mtDNA gene annotation at a given nucleotide position; coding – detailed mtDNA gene annotation; Filter A – removed by filter A (1=yes, 0=no); Filter B – removed by filter B (1=yes, 0=no); Filter C – removed by filter C (1=yes, 0=no); Filter D – removed by filter D (1=yes, 0=no). Short explanation of color highlighted positions is included at the bottom of the table.

Primer Number	Position (corresponding to the rCRS)	Orientation	Sequence
1	2070 - 2090	forward	AATTTGCCCACAGAACCCTC
2	2920 - 2941	reverse	GACTCTAGAATAGGATTGCGC
3	2499 - 2520	Forward	TACCAAAAACATCACCTCTAGC
4	2937 - 2916	reverse	TATCCCTAGGGTAACTTGTTCC
5	2207 - 2227	forward	TCAAGCTCAACACCCACTACC
6	2670 - 2651	reverse	GGCAGGTCAATTTCACTGGT
7	1580 - 1601	Forward	TGGAAAGTGCACTTGGACGAAC
8	2640 - 2621	reverse	GGAGCCATTCATACAGGTCC
9	2655 - 2636	reverse	CAGCTGAACCCTCGTGGAGC
10	2578 - 2597	forward	CCTAACCGTGCAAAGGTAGC
11	3414 - 3395	reverse	GCCTTTGCGTAGTTGTATAT
12	2593 - 2612	forward	GTAGCATAATCACTTGTTCC

Supplementary Table 4: Primers list

Supplementary Figure legends

Supplementary Figure 1: MtDNA nucleotide positions with ≥1,000 sequence reads coverage per individual. X axis – nucleotide positions with >1000 sequence reads coverage along the mtDNA in the tested samples. MtDNA positions 1-16,569 are from left to right. Each horizontal dotted line correspond to the indicated tested samples. Each position is indicated by a dot. Only the positions with a minimal coverage of 1,000 sequence reads are marked. Sample IDs are indicated in the Y axis in DNA (A) and RNA (B).

Supplementary Figure 2: Sanger sequencing of the 2617 RDD: Sanger sequence traces encompassing sites 2612 to 2622 in the human mtDNA in several cell types and individuals. The arrows indicate position 2617 in each individual. (a) GM14468 DNA. (b) GM14468 RNA. (c) SHSY5Y DNA. (d) SHSY5Y RNA. (e) ChubS7 DNA. (f) ChubS7 RNA. (g) Colon DNA. (h) Colon RNA. (i) 65080 DNA. (j) 65080 liver RNA. (k) 65080 skeletal muscle RNA. (l) 65080 brain RNA. (m) 64998 DNA. (n) 64998 liver RNA. (o) 64998 skeletal muscle RNA. (p) 64998 brain RNA.

Supplementary Figure 3: Sequence alignment of the top 30 BLAT hits from UCSC to a 100 bp fragment encompassing the RDDs at mtDNA position 2617. The left column lists the BLAT hit loci and exact chromosomal coordinates in the human genome. MtDNA position 2617 is highlighted in yellow. Red highlight – nucleotide positions showing mutational differences from the mtDNA reference sequence. Notably, none of the red marked changes were identified in our DNA and RNA deep sequence data.

Supplementary Figure 4: Positions of primers used for PCR amplification of 16S rRNA and flanking sequences: Amplification of the 16S rRNA and its flanking regions was performed using four primer pairs as described in Materials and Methods. These fragments encompass human mtDNA nucleotide positions 1580-2640, 1580-2655, 2578-3414 and 2593-3414 respectively (rCRS NC_012920). The arrows above the horizontal line (the mtDNA) represent the amplification primers and their orientation.
Numbers indicate reverse and forward amplification primers pairs used for fragments 1-4 of the 16S flanking regions, respectively. Vertical lines indicate gene regions, with the caption underneath representing gene names. MtDNA position 2617 is indicated.

Supplementary Figure 5: Sequence alignment of mitochondrial 16S RNA from 1755 vertebrates. MtDNA positions 2608 to 2634 (human coordinates) are aligned and compared to available vertebrate orthologous sequence. This sequence harbors the stems and loop (Ruiz-Pesini and Wallace 2006) including position 2617 (framed in red). The 1755 vertebrate species include 941 *Actinopterygii* (ray-finned fishes), 109 *Amphibia* (amphibians), 48 *Testudines* (turtles), 109 *Lepidosauria* (sphenodon, lizards and snakes), 177 *Archosauria* (Crocodilains and birds) and 371 *Mammalia* (mammals). Stem, Loop, Stem columns – mtDNA sequence corresponding to the stem-and-loop structure of 16S rRNA around position 2617.

Supplementary Figure 6: Phylogenetic analysis of complete mitochondrial sequences from 334 *Eutheria*. A Maximum Likelihood tree constructed from 334 complete mtDNA *Ehutherian* sequences. Bootstrap scores of 1000 replicates are shown on each branch together with its ancestral state of position 2617. All phylogenetic analyses were performed using MEGA 5.0.

Supplementary Figure 7: Sanger sequence analysis of position 2617 from DNA and RNA samples of mitochondria from various primates: *Pongo pigmaeus*: (a) DNA and (b) RNA; *Gorilla gorilla* (c) DNA and (d) RNA; *Macaca mulatta* (e) DNA and (f) RNA; *Lemur katta* (g) DNA and (h) RNA; *Nycticebus coucang* (i) DNA and (j) RNA. Red arrow indicates nucleotide position corresponding to position 2617 in human mtDNA. Numbers on top of each base indicates positions according to the rCRS.

Supplementary Figure 8: The identified RDDs are not found at the edges of the sequence reads. A representative IGV viewer chart (Robinson et al. 2011) at position 2617 in a representative analyzed sample (GM14447). Upper panel – schematic linear representation of the human mtDNA (rCRS), position 2617 is framed. Numbers at the top part of the figure – nucleotide positions of the rCRS. Lower large panel – schematic representation of the sequence reads encompassing mtDNA position 2617 (thick arrow-like grey bars). Direction of arrow heads are consistent with directions of the reads. Framed red bars – the thymine at position 2617; brown - Guanine, grey - the DNA reference (Adenine).