Supplemental Material

Larsen et al. The *Alu* neurodegeneration hypothesis: a primate-specific mechanism for neuronal transcription noise, mitochondrial dysfunction, and manifestation of neurodegenerative disease.

Supplemental Methods and Results

Retrotransposon content analyses.—The Ensembl identifications for 1,145 mitochondrial genes were downloaded from the Broad Institute's MitoCarta2.0 website (https://www.broadinstitute.org/scientific-community/science/programs/metabolic-disease-program/publications/mitocarta/mitocarta-in-0). In R, we used the biomaRt package with the "ENSEMBL_MART_ENSEMBL" biomart and the "hsapiens_gene_ensembl" dataset to extract the 1,145 genes, as well as 5kb of upstream and downstream flanking sequences, corresponding to each MitoCarta2.0 Ensembl ID. Resulting FASTA sequence files are available on GitHub (https://github.com/KelsieHunnicutt/Larsen-etal_Alu_Neurodegeneration).

To quantify retrotransposon content across the extracted sequences, we used RepeatMasker v. 4.0.6 (http://www.repeatmasker.org/) with the latest RepBase database of repeat elements v.20160829 (http://www.girinst.org/repbase/) and the RMBlast search engine 2.2.27 (http://www.repeatmasker.org/RMBlast.html). RepeatMasker was run on all FASTA files using the slow search –s option. RepeatMasker output tables for mitochondrial genes are available in the GitHub repository.

To test for enrichment of *Alu* elements within mitochondrial genes against non-mitochondrial protein coding genes, we generated 10 randomly selected sets of non-mitochondrial genes (each set consisting of 1,145 random genes). Using the biomaRt R package, we produced a list of non-mitochondrial genes by extracting all Ensembl IDs matching the filter biotype=protein_coding then removing those Ensembl IDs belonging to mitochondrial genes within the MitoCarta2.0 database. Nucleotide length for each gene was calculated by extracting the genomic start and end position for each Ensembl ID. To accommodate for length bias, and thereby mobile element content, we randomly selected non-mitochondrial genes (using the R sample function) that were consistent with the length distribution observed in mitochondrial gene set. Based on the length of the genomic sequence, including the 5kb flanks, we partitioned the mitochondrial genes into five length bins (~3kb to ~13kb, ~13kb to ~59kb, ~59kb to ~268kb, ~268kb to ~1.2Mb, and ~1.2Mb to ~5.3Mb). We then selected a total of 1,145 genes of appropriate length from the list of non-mitochondrial genes to fill these bins, thus providing a comparable randomly selected set of non-mitochondrial genes. Genomic and flanking sequences were downloaded for each gene following the methods described above. This process was iterated 9 additional times, resulting in a total number of 8,973 unique genes across all 10 random samples. The gene sets were then analyzed using RepeatMasker as described above to quantify mobile element content. Sequence FASTA files and RepeatMasker output files for the 10 random samples can be found on the GitHub repository.

Bar graphs were generated using ggplot2 in R. Statistical support was calculated using two-tailed t-tests as implemented in R. In both intragenic sequences and flanking regions, mitochondrial genes were found to be significantly enriched for *Alu* elements. Upstream flanks were found to contain a significantly higher number of *Alu* elements (p-value=1.409319e-10) and have a higher percentage of sequence content occupied by *Alu* elements (p-

value=1.151572e-10). Similarly, intra-genic sequences were found to contain a significantly higher number of *Alu* elements (p-value=9.531036e-07) and have a higher percentage of sequence content occupied by *Alu* elements (p-value=2.287497e-09). This pattern was also found in downstream flanks, which had a significantly higher number of *Alu* elements (p-value=4.801572e-08) and a higher percentage of sequence content occupied by *Alus* (p-value= 3.968526e-08).

Supplemental Figures

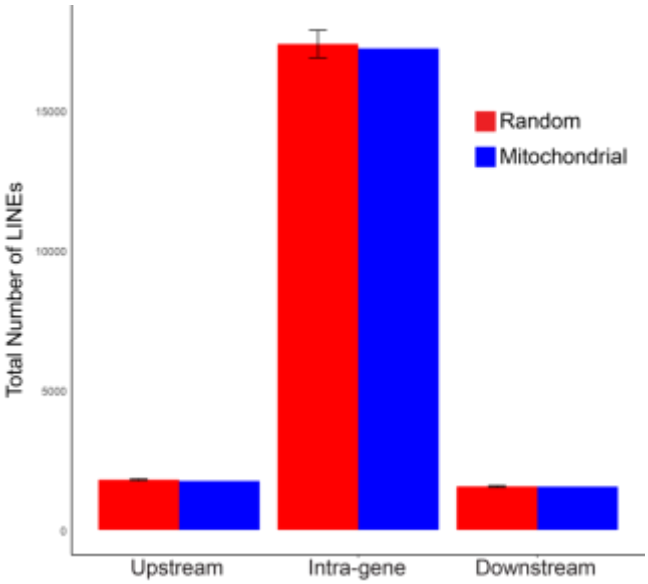

Figure S1. LINE content measured across 1,145 mitochondrial genes and 8,973 randomly selected (non-mitochondrial) genes, as well as flanking genomic regions (5kb), sampled from the human genome (Ensembl build GRCh38).

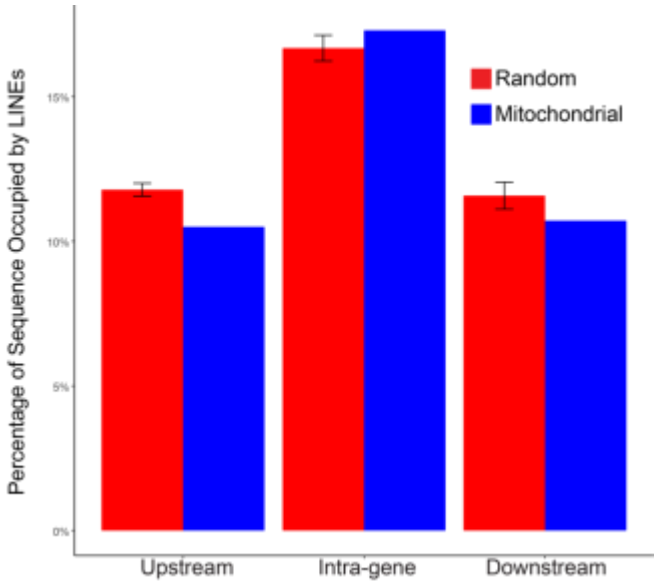

Figure S2. Percent of sequences occupied by LINEs as measured across 1,145 mitochondrial genes and 8,973 randomly selected (non-mitochondrial) genes, as well as flanking genomic regions (5kb), sampled from the human genome (Ensembl build GRCh38).

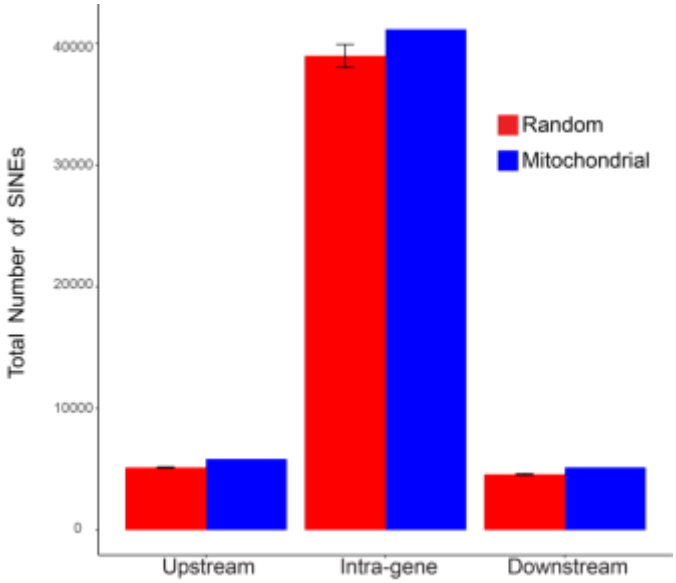

Figure S3. SINE content across 1,145 mitochondrial genes and 8,973 randomly selected (non-mitochondrial) genes, as well as flanking genomic regions (5kb), sampled from the human genome (Ensembl build GRCh38).

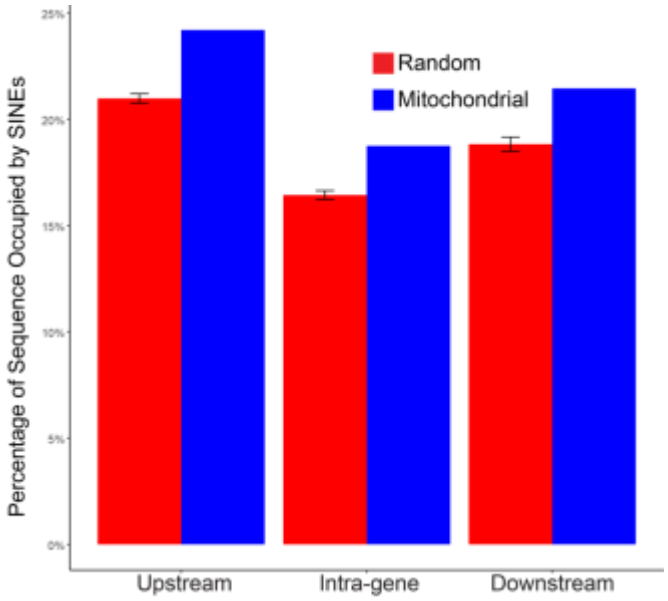

Figure S4. Percent of sequences occupied by SINEs as measured across 1,145 mitochondrial genes and 8,973 randomly selected (non-mitochondrial) genes, as well as flanking genomic regions (5kb), sampled from the human genome (Ensembl build GRCh38).

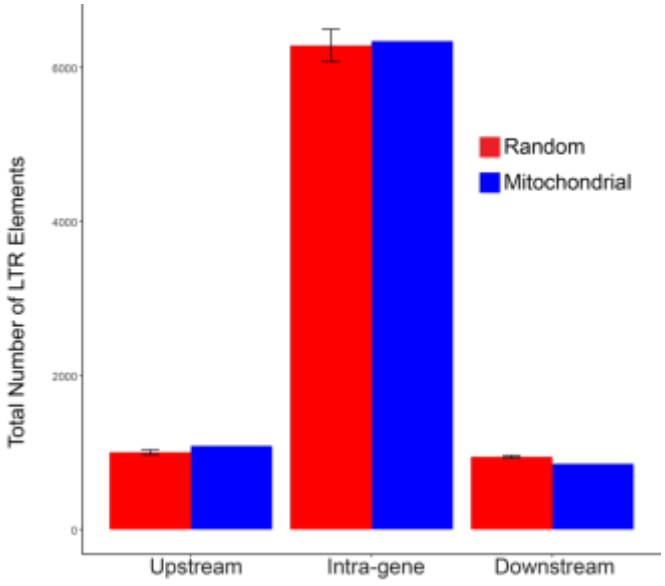

Figure S5. LTR content across 1,145 mitochondrial genes and 8,973 randomly selected (non-mitochondrial) genes, as well as flanking genomic regions (5kb), sampled from the human genome (Ensembl build GRCh38).

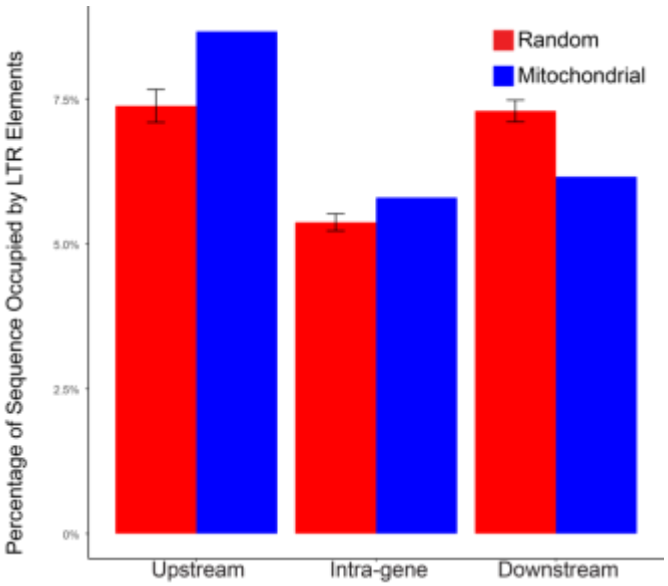

Figure S6. Percent of sequences occupied by LTRs as measured across 1,145 mitochondrial genes and 8,973 randomly selected (non-mitochondrial) genes, as well as flanking genomic regions (5kb), sampled from the human genome (Ensembl build GRCh38).

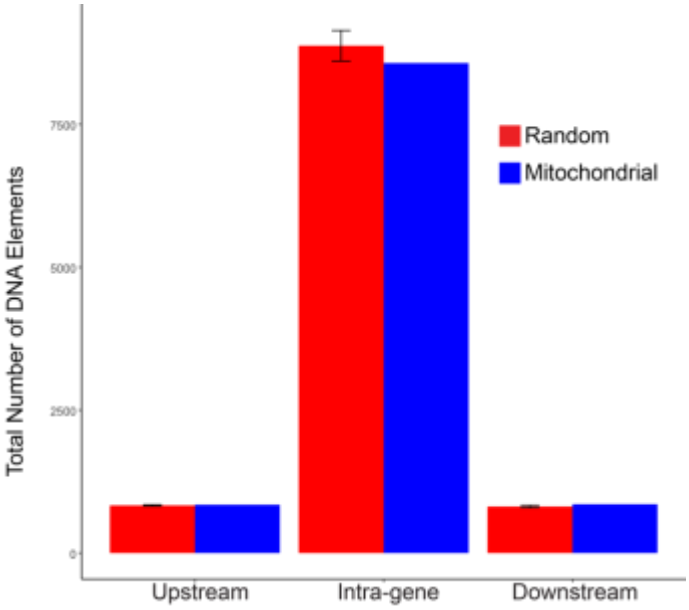

Figure S7. DNA transposon across 1,145 mitochondrial genes and 8,973 randomly selected (non-mitochondrial) genes, as well as flanking genomic regions (5kb), sampled from the human genome (Ensembl build GRCh38).

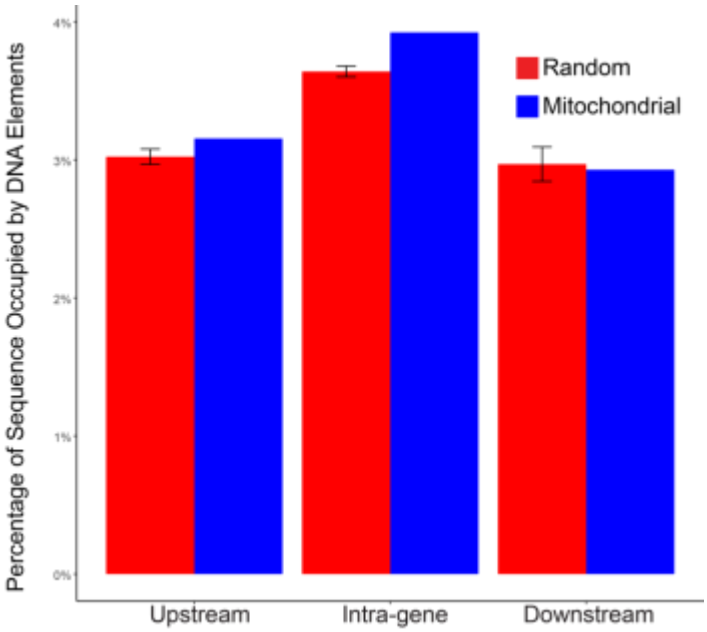

Figure S8. Percent of sequences occupied by DNA transposons as measured across 1,145 mitochondrial genes and 8,973 randomly selected (non-mitochondrial) genes, as well as flanking genomic regions (5kb), sampled from the human genome (Ensembl build GRCh38).