

Table of Content

Supplemental Methods	2-4
References - supplemental methods	4
Supplemental_Fig_S1	5
Supplemental_Fig_S2	6
Supplemental_Fig_S3	7
Supplemental_Fig_S4	8
Supplemental_Fig_S5	9-10
Supplemental_Fig_S6	11-12
Supplemental_Fig_S7	13
Supplemental_Fig_S8	14-15
Supplemental Datasets Legends and links	16

Supplemental Methods

NUMT exclusion from RNA-seq data: One source of contamination of our analyzed mtDNA gene expression data is sequencing reads of primitive mtDNA fragments that were transferred to the nDNA during the course of evolution (Hazzkani-Covo et al. 2003; Mishmar et al. 2004). To control for such contamination, GTEx used nuclear mitochondrial pseudogene transcript sequences as their reference transcriptome (appears as mtXPN in the GTEx data, where X represents the mitochondrial gene and N stands for the number of the pseudogene). Therefore, reads that had higher similarity scores to any given pseudogene, as compared to the reference mtDNA gene, were excluded from the reads count for mtDNA-encoded genes. To further control for NUMTs, we used the normalized read counts (reads per kilo-base per million reads - RPKM) available in the GTEx consortium data files as our raw data. According to GTEx documentation, the read alignment distance was ≤ 6 (i.e., sequence alignments must not contain more than six mismatches), thus notably reducing possible contamination by NUMT reads.

Control of RNA-seq data for age, gender and cause-of-death: The age variable was obtained as a continuous factor, whereas gender and cause-of-death were categorical factors. Dummy variables (1 for gender and 5 for cause-of-death) were used to correct for the effect of these factors on gene expression, while correcting for multiple testing. Finally, the residual expression for each gene in each sample was calculated as the difference between the observed expression and the expression value predicted for this gene by the linear regression, considering phenotypic parameters (i.e. age, gender and cause-of-death).

Clustered heat maps construction: The median Spearman's rank correlation coefficient between the pairs of genes/transcripts under consideration were used to create a distance matrix for hierarchical clustering. Such clustering was obtained according to the similarity of the correlation of gene/transcript expression with other genes/transcripts under consideration. The correlation values were color-coded: red – positive correlation, blue – negative correlation. Non-significant or zero correlations are marked in white.

To construct heat maps for each of the analyzed tissues, we used the gene/transcript order that was obtained from the hierarchical clustering generated while analyzing the total expression pattern across all tissues. Additionally, while constructing tissue-specific correlation distance matrixes, we considered all samples available for the tested tissue.

Brain cell-type determination in single cell RNA-seq analysis: Cells were considered excitatory-neurons if they displayed higher maximal expression of the pan-glutamatergic marker-genes (*Slc17a6* and *Slc17a7*) than both the maximal expression of pan-GABAergic marker-genes (*Gad1*, *Gad2* and *Slc32a1*) and the non-neuronal marker genes (*Olig1*, *Gja1*, *Xdh*, *Ctss* and *Myl9*). Cells were considered inhibitory-neurons if they displayed higher maximal expression of the pan-GABAergic marker-genes as compared to both the maximal expression for pan-glutamatergic and non-neuronal marker-genes. Other cells were considered non-neuronal (Sugino et al. 2006; Tasic et al. 2016).

Statement - usage of human genome sequence versions (GRCh37 and GRCh38): The GTEx dataset is based on mapping of RNA-seq reads against the GRCh37/hg19; we used mapped reads which are publicly available, and do not have access to the restricted area of the GTEx data. As our analyses focus on gene expression correlation among OXPHOS genes, involving more than 80 genes that showed consistent and significant correlation values across and within tissues, it is safe to state that possible small changes in some of the genes' sequences, among different versions of the human genome, would not significantly alter our conclusions.

Structural predictions: Swiss-model server (<https://swissmodel.expasy.org/>) (default parameters) was used to predict the 3D structures of the discussed OXPHOS subunit alternatively spliced variants. In addition, we used Jpred 4 (<http://www.compbio.dundee.ac.uk/jpred/>), to predict secondary structures (default parameters) and assess their conservation between the canonical isoforms and the alternative ones.

References - supplemental methods

- Hashimshony T, Wagner F, Sher N, Yanai I. 2012. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell reports* **2**(3): 666-673.
- Hazkani-Covo E, Sorek R, Graur D. 2003. Evolutionary dynamics of large numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications. *Journal of molecular evolution* **56**(2): 169-174.
- Mishmar D, Ruiz-Pesini E, Brandon M, Wallace DC. 2004. Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. *Hum Mutat* **23**(2): 125-133.
- Sugino K, Hempel CM, Miller MN, Hattox AM, Shapiro P, Wu C, Huang ZJ, Nelson SB. 2006. Molecular taxonomy of major neuronal classes in the adult mouse forebrain. *Nature neuroscience* **9**(1): 99-107.
- Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, Levi B, Gray LT, Sorensen SA, Dolbeare T. 2016. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature neuroscience* **19**(2): 335-346.

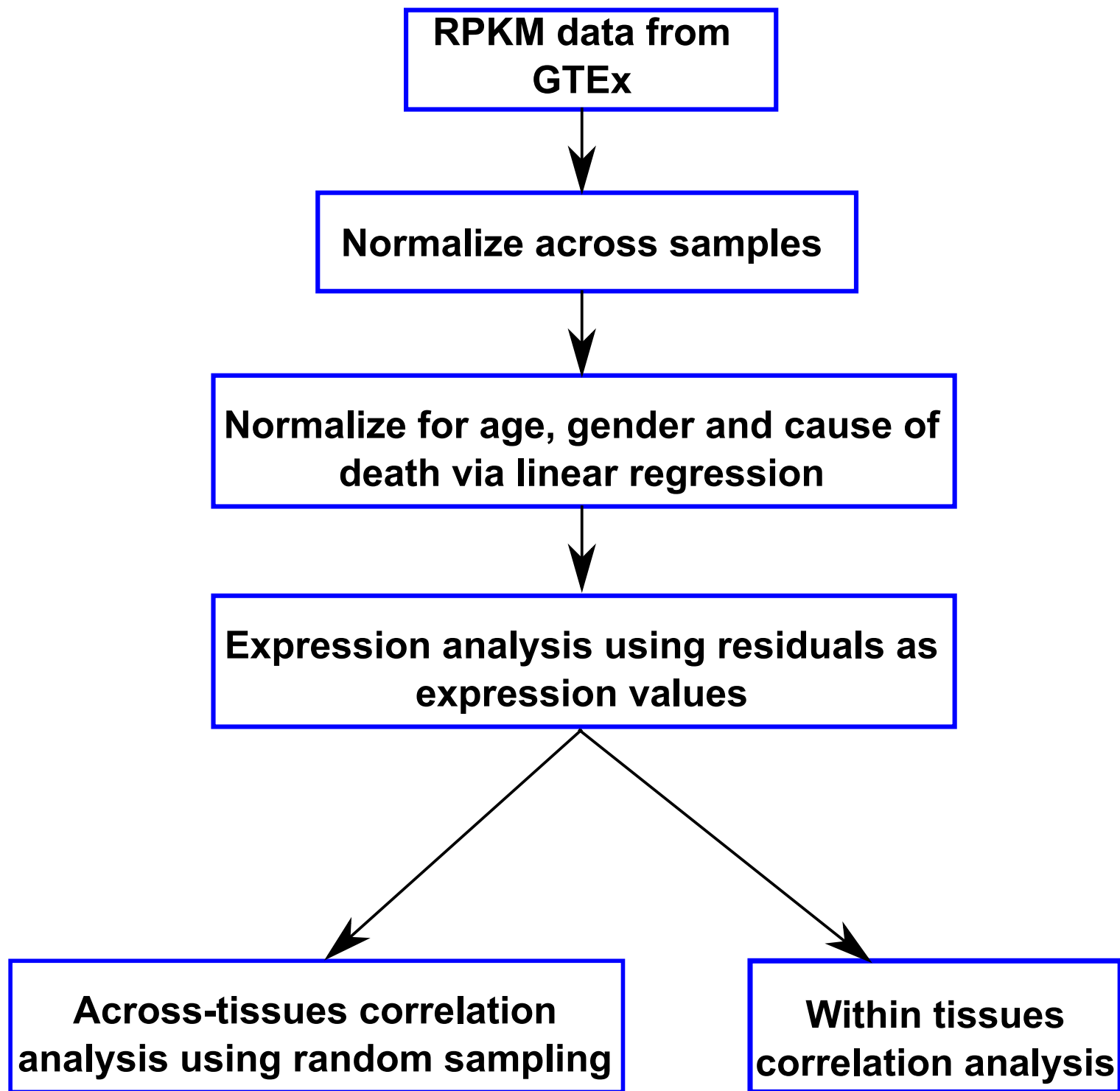


Figure S1. Workflow for RNA-seq data analysis of OXPHOS mito-nuclear gene expression correlation.



Figure S2. nDNA-encoded OXPHOS genes show significantly higher co-expression with mtDNA-encoded OXPHOS genes. A box-plot showing the values of the correlation coefficients between pairs of OXPHOS genes (intra-OXPHOS) and between OXPHOS genes and randomly chosen protein-coding genes (OXPHOS-Genome). *** $p < 1 \times 10^{-100}$.

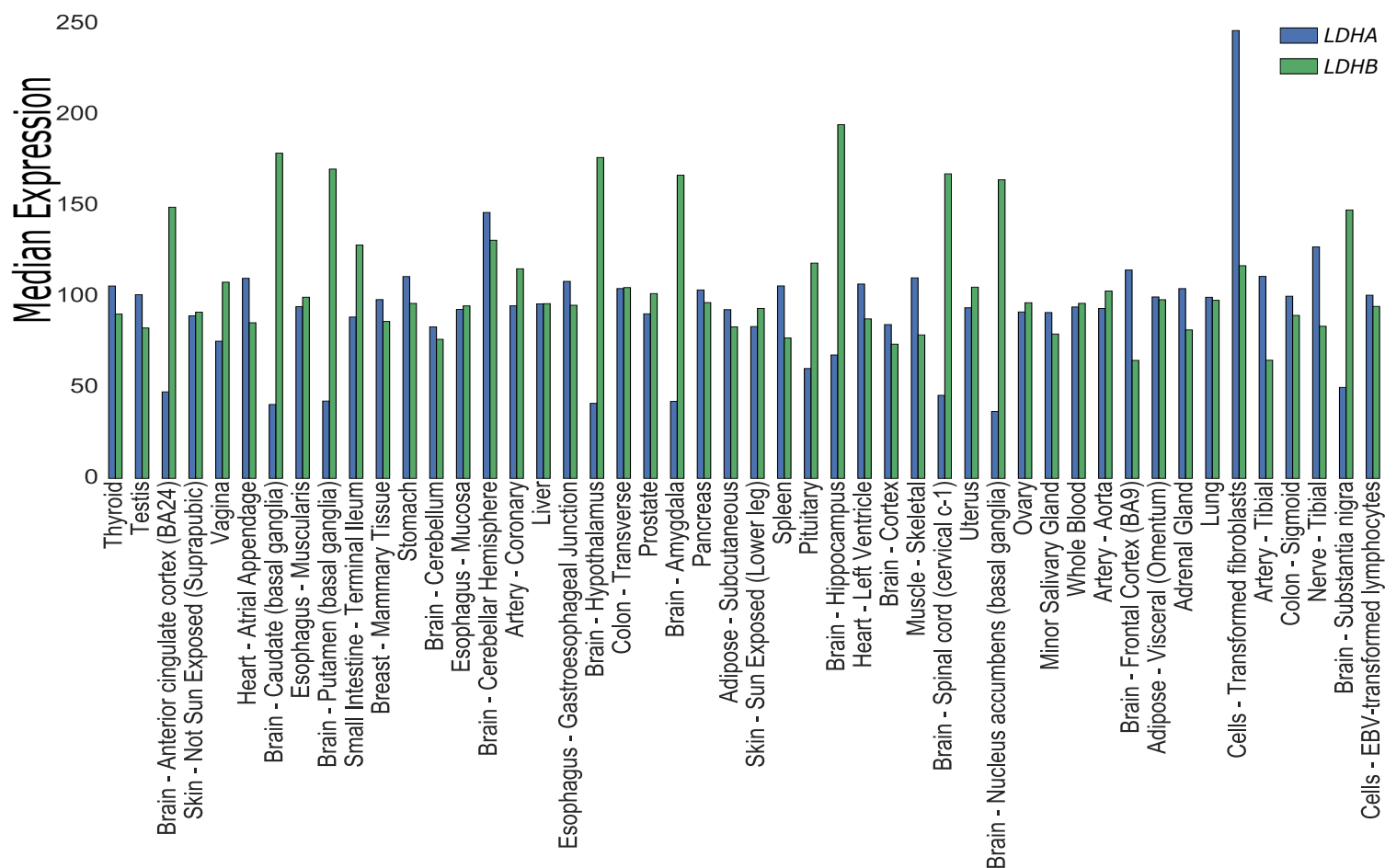


Figure S3. Expression of LDH subunits across human tissues. A bar-plot showing the median expression values of the two LDH subunit-coding genes (*LDHA* and *LDHB*) across all human tissues inspected in this study.

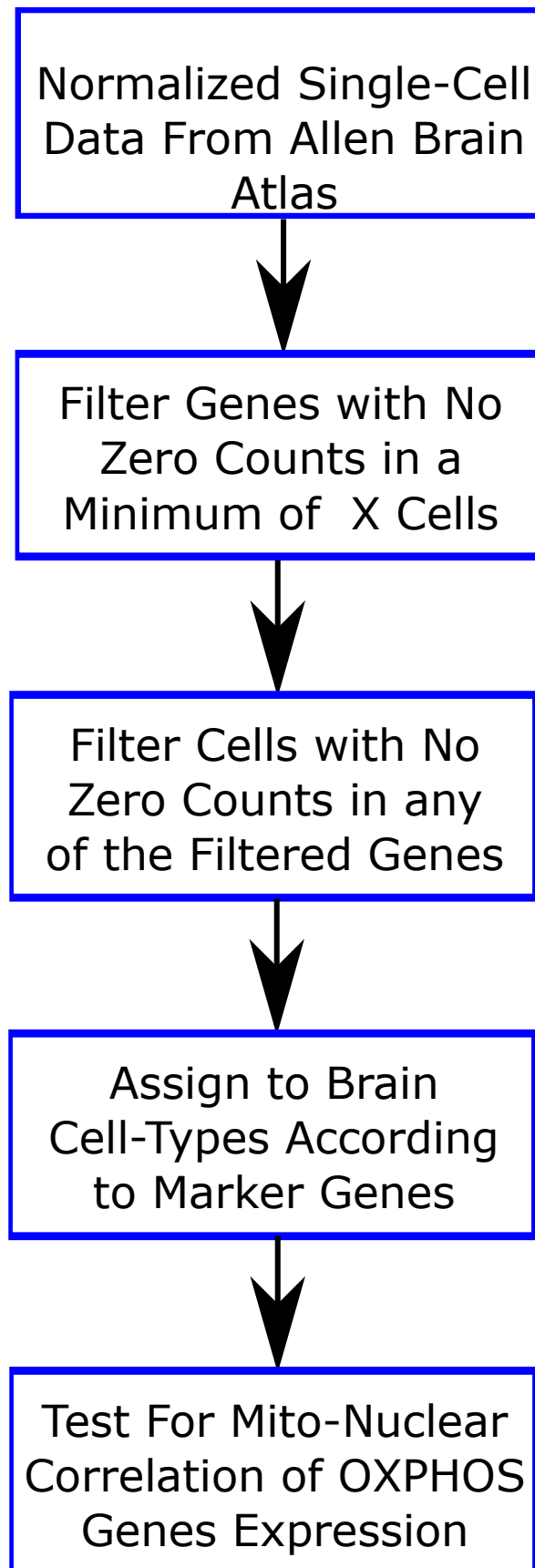
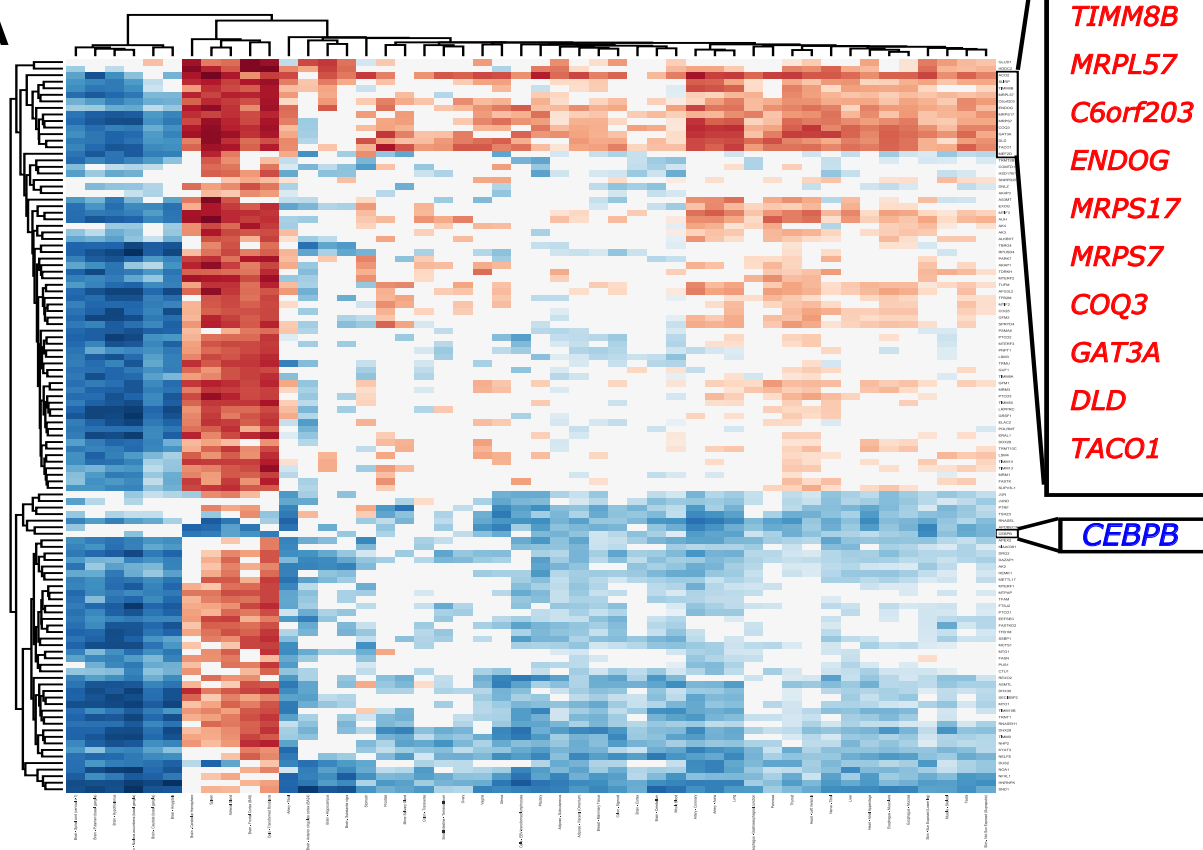


Figure S4. Workflow for single-cell RNA-seq data analysis for OXPHOS mito-nuclear gene expression correlation. See also 'Single-cell RNA-seq data analysis' under 'Methods'.

Figure S5. Graphical representation of the in vivo identification of candidate promoters in nuclear DNA encoded OXPHOS genes.

- (A) A map of the upstream region of a representative OXPHOS gene (NDUFA8) tested for TF binding in DNaseI Genomic Footprinting (DGF) sites. X-axis represent the positions in the promoter region, relative to the transcription start site (TSS). DGF sites are indicated by blue lines in the upper row and below, all the listed TFs binding to the 1000bp promoter region are shown, with their ChIP-seq peaks shown as black lines. Note that only TFs with a ChIP-seq peak overlapping a minimum of one DGF are indicated.
- (B) Representation of the results for all nDNA-encoded OXPHOS genes. X-axis represent the positions in the promoter region, relative to the transcription start site (TSS). DGFs are shown by blue lines if they do not overlap with any identified TF-binding peak in the ChIP-seq experiments and in black and blue lines if they overlap with at least one such TF-binding peak.

A



B

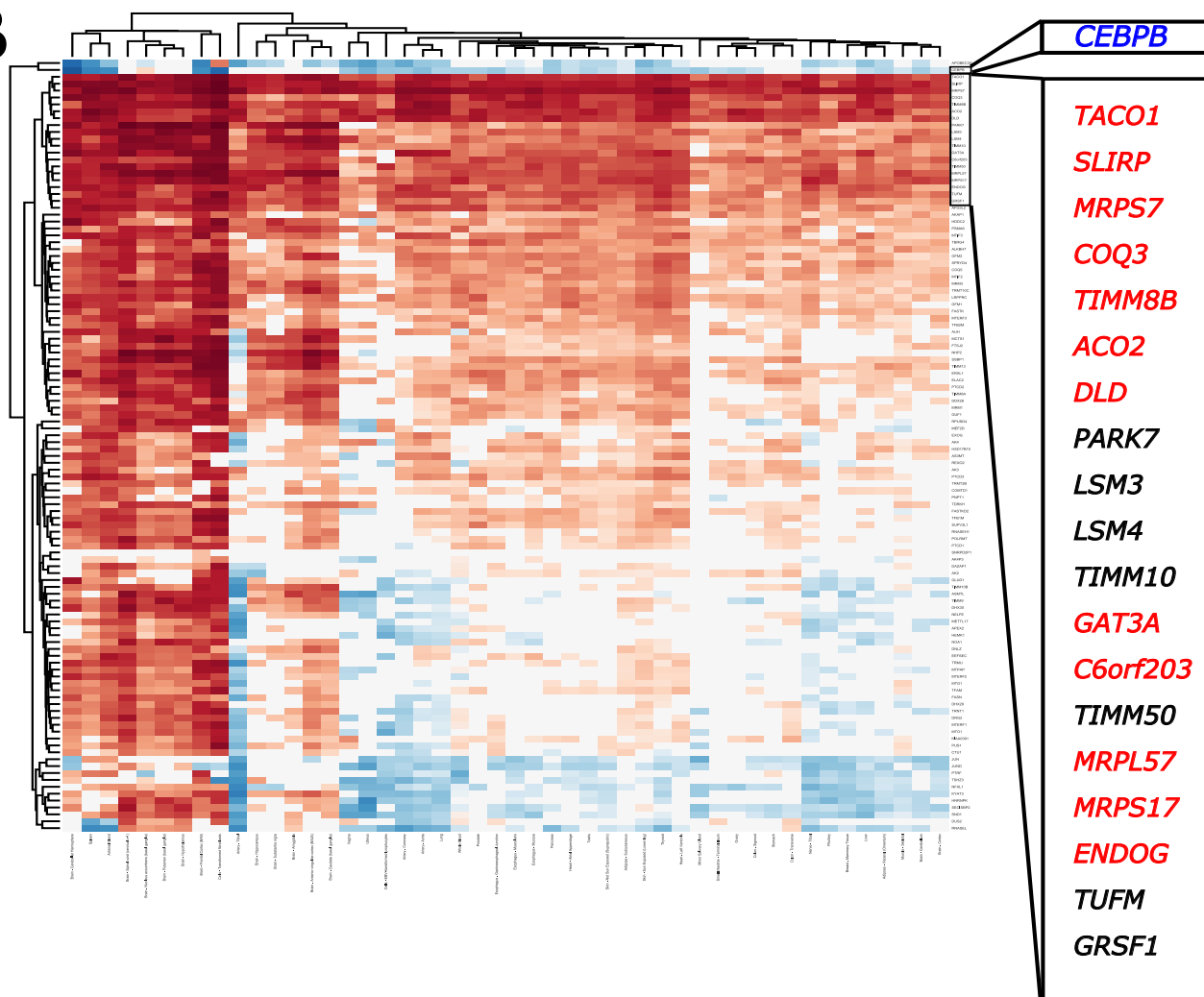


Figure S6. The candidate mito-nuclear OXPHOS expression regulators correlation with OXPHOS genes. Heat maps representing the median correlation values of the different candidate regulators, with the mtDNA-encoded (A) and nDNA-encoded (B) OXPHOS genes.

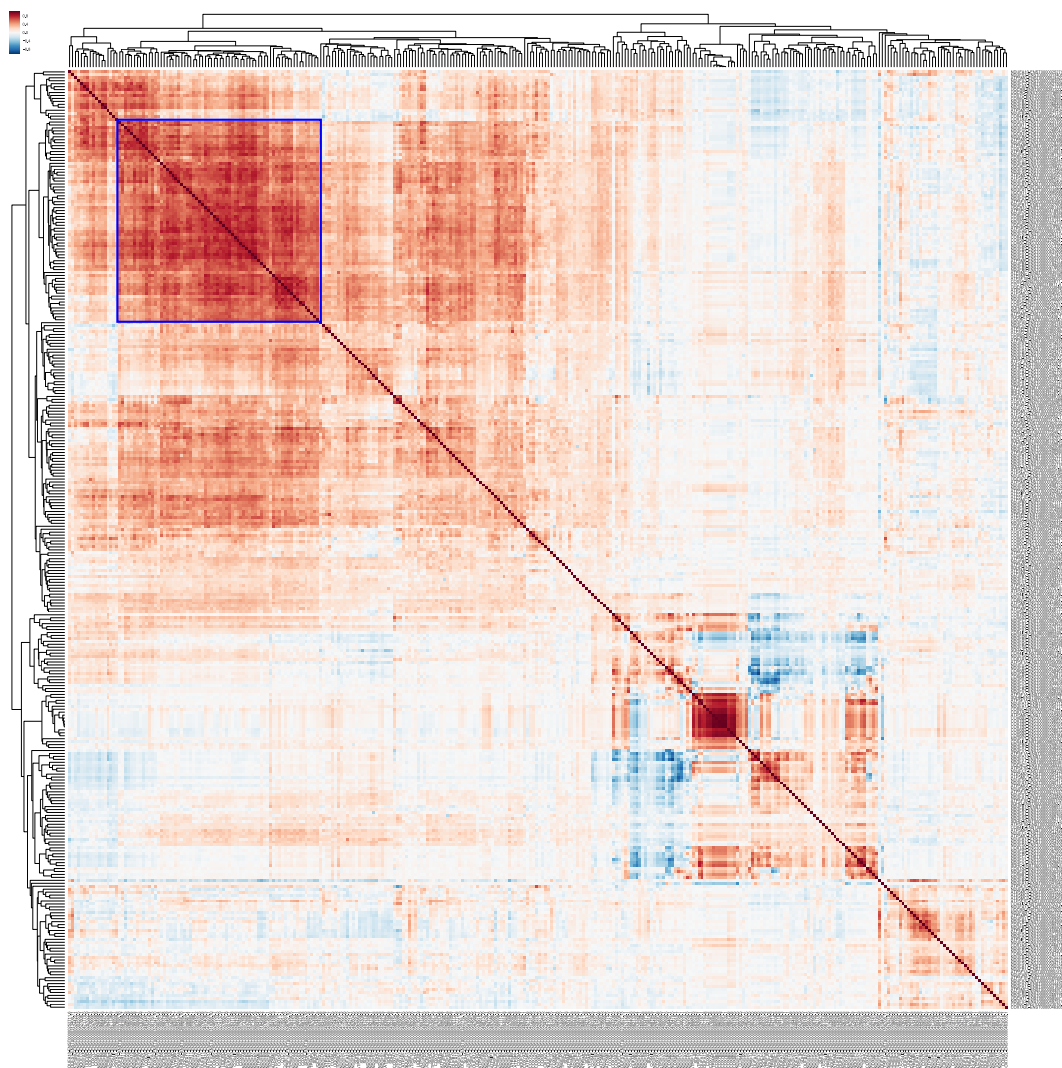
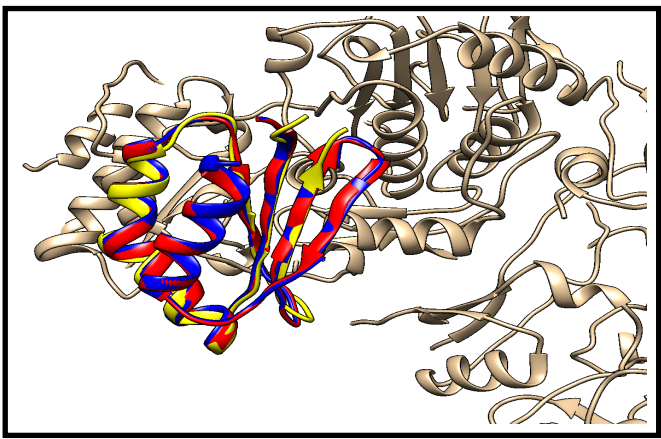


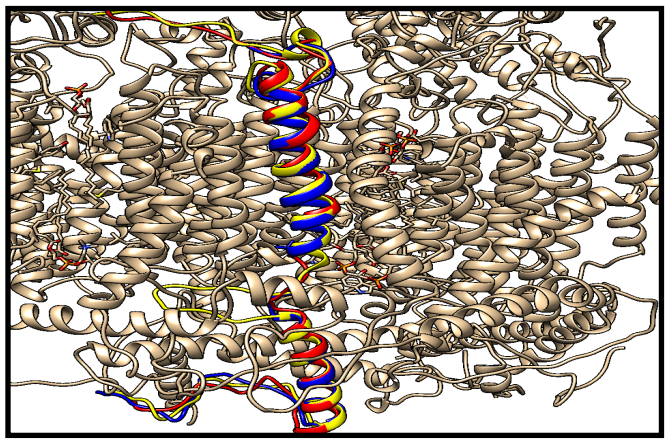
Figure S7. OXPHOS genes splice-variants correlations and the 'main OXPHOS transcript cluster'. A heat map of correlation values for the expression of all OXPHOS transcripts. The 'main OXPHOS transcript cluster' is marked with a blue square. A key for the color code of the Spearman's correlation coefficients is shown in top left corner.

A

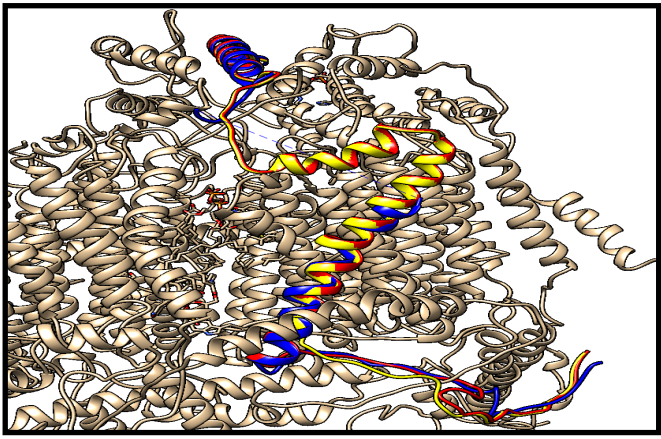
NDUFA2



NDUFB11



NDUFB6



B

NDUFA3

Canonical:

VGAFLKNAWDKEPVLVVSFVVGGLAVILPPLSPYFKYSVMINKATPYNYPVPVVRDDGNMPDVPSPHPQDPQGPSLEWLKKL
 -HHHHHHH-EEEEEHHHHHHHHEEE-HHHHHH-HHHHH-

Alternative:

KCRSRRLLPQEC LGQGASAGRVLRRRGPRCNSAPIEPLLQVLRHDQQGHALQLPSARP
 -HHHH-HHHHHHHH-HHHHHHHHHH-EE-

ATP5MF

Canonical:

MASVGECPAPVPVKDKKLLVKGELPSWILMRDFSPSGIFGAFQRGYYRYNKYINVKKGSISGITMVLACYVLFSSYSFSYKHLKHERLRKYH
 -----EE-HHHHHHHHHHHHHHHH-EEEE-HHHHHHHHHHHHHHHH-----

Alternative:

MASV-----VPVKDKKLLVKGELPSWILMRDFSPSGIFGAFQRGYYRYNKYINVKKGSISGITMVLACYVLFSSYSFSYKHLKHERLRKYH
 ----xxxxxx-HHHHHHHHHHHHHHHH-EEEE-----

Figure S8. Structural comparison between the suggested splice-isoforms and the canonical spliceisoforms. (A) 3D models comparing the structures of the canonical (red), alternative (yellow) and mammalian (sheep) (blue) isoforms of NDUFA2, NDUF6 and NDUF11. (B) Secondary structures prediction for the alternative splice-isoforms of NDUFA3 and ATP5MF, that lacked a homolog with high sequence similarity and a solved 3D structural model, to the canonical isoforms of these subunits.

Supplemental Datasets legends and links

Supplemental Dataset S1: Correlation heat maps for OXPHOS genes and splice variants expression patterns in 48 human tissues. <https://figshare.com/s/cc5cb8e5107cba4da7e3>

Supplemental Dataset S2: Correlation matrices for expression patterns of OXPHOS genes and splice variants of OXPHOS genes in 48 human tissues. <https://figshare.com/s/818279d849951436ee72>

Supplemental Dataset S3: Relative correlation heatmaps for splice-variants of genes with more than one transcript highly OXPHOS correlated across tissues. <https://figshare.com/s/223934fe6e99ab3bf528>

Supplemental Dataset S4: Bash pipeline and associated python scripts and input files for raw data conversion, normalization and co-expression analysis. https://github.com/dmishmar/Barshad_2018