Readme support for the dataset:

## Data underlying chapter 5 of PhD thesis: Exploring the potential of yeast mitochondria for synthetic cell research

File SI\_1\_WGSdata.xlsx contains all Single-Nucleotide Polymorphisms (SNPs) and Insertions or Deletions (INDELS) within known coding sequences (CDSs) that lead to an amino acid change, that were obtained upon sequencing the evolved *S. cerevisiae* strains IMS1248-IMS1253 compared to a IMX2600 reference genome. Whole Genome Sequencing was performed by Macrogen Europe (Amsterdam, the Netherlands) on a Novaseq 6000 sequencer (Illumina, San Diego, CA) to obtain 151 cycle paired-end libraries with an insert-size of 550 bp using TruSeq Nano DNA library preparation, yielding 2 Gigabases in total per sample. All Illumina sequencing data are available at NCBI (https://www.ncbi.nlm.nih.gov/) under the Bioproject accession number PRJNA923502. Reads were mapped using BWA (version 0.7.15) a IMX2600 reference. Alignments were processed using SAMtools (version 1.3.1), and variants were called by applying Pilon (version 1.18). The first tab, (‘All mutations’) shows a list of all mutations found over the variant calling files (vcf) of the whole dataset and in which strains the mutations were found. The other tabs list all SNPs and INDELS found per strain, with details on the strand orientation (trans\_orient), location in CDS (loc\_in\_cds), position in codon (codon\_pos), the codon substitution observed (codon), if the mutation is a SNP or INDEL (type), the resulting amino acid change in 1-letter amino acid codes (pep), the resulting amino acid change and location in three-letter amino acid codes (peptide) and under “effect” whether the mutation was non-synonymous (NSY), an early stop codon (NON), an extension (RTH). ‘NA’ means non-applicable.

File SI\_2\_Proteomics\_GOterm contains all raw proteomics data and GO-term analysis. Tab “Proteomics raw results” contains raw proteomics data of an unevolved *S. cerevisiae* strain, IMS1251, IMS1252 and IMS1254. 1 µg protein digest was analyzed using a one-dimensional shot-gun proteomics approach in three technical replicate experiment for each biological replicate according to den Ridder et al. 2022. Data were analyzed against the proteome database from *Saccharomyces cerevisiae* (UniProt, strain ATCC 204508 / S288C, Tax ID: 559292, July 2020) using PEAKS Studio X (Bioinformatics Solutions Inc., Waterloo, Canada) as described by den Ridder et al. 2022. The significance score for evaluating the observed abundance changes was calculated using a one-way ANOVA and expressed as the -10∙*log*10(p), where p is the significance testing p-value, which represents the likelihood that the observed change is caused by random chance. Based on the average of the duplicate quantified proteins, the fold change of each protein in a specific condition was calculated relative to the unevolved sample. The average fold changes of the technical replicates were subsequently used to determine the standard deviations of the biological replicates. The dataset has been subsetted in a subset of mitochondrial proteins (tab ‘subset\_mitochondrial\_proteins’), a subset of proteins that were significantly upregulated in all three strains (Upreg = s\_up, tab ‘Significant\_UP’) and a subset of proteins that were significantly downregulated in all three strains (Upreg = s\_down, tab ‘Significant\_DOWN’). These subsets were again subsetted in mitochondrial proteins significantly upregulated (tab ‘MITO\_Sig\_UP’) and significantly downregulated (tab ‘MITO\_Sig\_DOWN’). A functional term enrichment analysis was performed to determine whether specific Gene Ontology (GO [63]) terms or were shared between proteins that were significantly more- or less abundant in two out of three or all three strains. GO-term analysis was performed using the GO::TermFinder [64] accessed through the GO Term Finder webpage hosted by the *Saccharomyces* genome data base (SGD, <https://www.yeastgenome.org>, [65]) using GO version 2023-01-01. A list of genes of interest (e.g. mitochondrial proteins more abundant in all strains), was analyzed against a custom reference list containing all *S. cerevisiae* mitochondrial proteins (SI file 2). The reference list of mitochondrial proteins was obtained by exporting a list of all genes with “mitochondrion” as cellular component (GO:0005739) from SGD. The enrichment strength was calculated by dividing the number of proteins in the entered dataset that associated with a GO-term by the number of proteins associated with the same GO-term in the reference list. Tab ‘GO\_enrichment\_DOWN’ contains a GO-term enrichment of all proteins of MITO\_Sig\_DOWN, Tab ‘GO\_enrichment\_UP’ contains a GO-term enrichment of all proteins of MITO\_Sig\_UP’. Tab ‘mitochondrion annotations’ contains a reference list of mitochondrial proteins. Tab ‘GO\_enrichment\_DOWN\_1251and54’ contains a GO-term enrichment of all proteins of MITO\_Sig\_DOWN of only strains IMS1251 and 1254.