NAR Genomics and Bioinformatics, 2023, **5**, 1–9
https://doi.org/10.1093/nargab/lqad107
Advance access publication date: 19 December 2023
Standard Article

Beyond MitoCarta—expanding the list of candidate proteins involved in mitochondrial functions using a biological network approach

Dmitriy Leyfer[1,2] and Jessica L. Fetterman[3,*]

[1]Translational Sciences Department, Mitobridge, division of Astellas, Cambridge, MA 02138, USA
[2]Bioinformatics Program, Boston University, Boston, MA 02215, USA
[3]Evans Department of Medicine and Whitaker Cardiovascular Institute, Boston University Chobanian & Avedisian School of Medicine, Boston, MA 02118, USA

[*]To whom correspondence should be addressed. Tel: +1 617 358 7544; Email: jefetter@bu.edu

Abstract

Mitochondrial diseases are the result of pathogenic variants in genes involved in the diverse functions of the mitochondrion. A comprehensive list of mitochondrial genes is needed to improve gene prioritization in the diagnosis of mitochondrial diseases and development of therapeutics that modulate mitochondrial function. MitoCarta is an experimentally derived catalog of proteins localized to mitochondria. We sought to expand this list of mitochondrial proteins to identify proteins that may not be localized to the mitochondria yet perform important mitochondrial functions. We used a computational approach to assign statistical significance to the overlap between STRING database gene network neighborhoods and MitoCarta proteins. Using a data-driven stringent significance threshold, 2059 proteins that were not located in MitoCarta were identified, which we termed mitochondrial proximal (MitoProximal) proteins. We identified all of the oxidative phosphorylation complex subunits and 90% of 149 genes that contain confirmed oxidative phosphorylation disease causal variants, lending validation to our methodology. Among the MitoProximal proteins, 134 are annotated to be localized to mitochondria but are not in the MitoCarta 3.0 database. We extend MitoCarta nearly 3-fold, generating a more comprehensive list of mitochondrial genes, a resource to facilitate the identification of pathogenic variants in mitochondrial and metabolic diseases.

Introduction

As metabolic hubs, mitochondria house enzymes involved in many metabolic pathways that are tightly regulated through the availability of substrates and cofactors. The availability of substrates and cofactors integrates many cellular activities with the energetic state of the cell. While the mitochondrial genome encodes 13 key catalytic subunits of oxidative phosphorylation complexes I and III–V, along with 22 transfer RNAs (tRNAs) and 2 ribosomal RNAs, the nuclear genome encodes additional subunits and assembly factors of the oxidative phosphorylation complexes, the transcription factors for the mitochondrial genes, replication machinery for the mitochondrial genome, antioxidants and other metabolic enzymes. Communication between the two genomes is essential for mitochondrial function and metabolism.

Pathogenic variants in mitochondrial genes, regulators or metabolic enzymes in either genome, mitochondrial or nuclear, result in human disease, ranging from severe pediatric syndromes to aging-related diseases (1–3). A key challenge in the diagnosis of mitochondrial disease is the heterogeneity in the clinical and biochemical presentation, severity and age of onset, which varies even within families (1–3). The heterogeneity of mitochondrial diseases makes identification of the pathogenic variant difficult and suggests that variants in one or both genomes can alter the penetrance of disease. Although next-generation sequencing is rapidly increasing the rate of identification of previously unknown pathogenic variants in mitochondrial diseases, 40% of patients with complex I deficiencies continue to lack a clear diagnosis of the pathogenic variant (4). A comprehensive list of nuclear-encoded mitochondrial genes that are essential for mitochondrial function is needed to improve the interpretation of patient next-generation sequencing data and systematically evaluate whether variants in mitochondrial genes of either genome alter the penetrance of mitochondrial disease. A list of nuclear-encoded mitochondrial genes is also essential for functional enrichment analysis of transcriptomics and proteomics data.

MitoCarta 3.0 is a catalog of mitochondrial components created through mass spectrometry identification of proteins within mitochondrial isolates and curation of the literature (5). However, MitoCarta 3.0 only contains proteins that are localized to the mitochondrion; hence, MitoCarta 3.0 does not include important regulators of mitochondrial function, such as the master regulator of mitochondrial biogenesis, PGC-1α, that are not localized to the mitochondrion. Consequently, proteins involved in the communication between the mitochondrion and nucleus, as well as other organelles, are missing from the MitoCarta 3.0 catalog.

To facilitate studies on the interplay between mitochondria and other cellular components and pathways, we sought to identify proteins relevant to mitochondrial function that

Received: April 24, 2023. **Revised:** October 25, 2023. **Editorial Decision:** December 4, 2023. **Accepted:** December 6, 2023

may not necessarily be localized to the mitochondria, which we term mitochondrial proximal (MitoProximal) proteins, thereby extending the MitoCarta catalog. To achieve this, a network approach was used to rate the potential mitochondrial involvement of each target in the STRING gene network database ($N = 18\,872$ human proteins). A hypergeometric test was applied to determine a *P*-value of the overlap between each target network neighborhood and MitoCarta 3.0 proteins. Using a data-driven stringent significance threshold, we identified 2059 proteins that were not found in the MitoCarta 3.0 catalog, thereby appending the list of mitochondrial proteins and regulators.

Materials and methods

Datasets

The Gene Ontology (GO) database consists of a catalog of genes organized into biological classes, including mitochondrial genes (6,7). We used the prefix 'mito' to search the C5: GO gene sets, which includes GO term categories BP, MF and CC, in the Molecular Signatures Database (Broad Institute, version 7.4 released on 2021-02-01). We identified 40 gene sets in GO that contained the prefix mito and included the terms mitochondria, mitochondrion or mitochondrial in the GO title or description, excluding gene sets of mitosis genes. After excluding redundant genes and mapping of Entrez IDs to Ensembl protein IDs, a total of 1730 unique protein-encoding genes from the 40 gene sets were identified.

We used the human MitoCarta 3.0 dataset as our list of known mitochondrial proteins (5). The STRING database consists of protein–protein interactions determined from experimental data, curation of the literature, and expression and computational prediction methods, and is regularly updated (8). The STRING database contains ∼24.5 million proteins and their known or predicted interactors (neighbors). We retrieved the protein nearest network neighbors from the entire human STRING database v11.5 and calculated the number of neighbors for each target protein, using the Ensembl protein ID (8). HUGO gene symbols were mapped to the Ensembl protein, gene and transcript IDs, and mouse gene homolog IDs using the db2db function in bioDBnet (9).

We obtained additional annotation information for the MitoProximal proteins from National Center for Biotechnology Information (NCBI) Gene, GeneCards 5.12 and the Human Protein Atlas version 23.0 (Supplementary Data) (10,11). Enrichment analysis of the top 200 most significant genes in the MitoProximal dataset was performed using the Molecular Signatures Database 3.0 (12).

Hypergeometric test

We performed a hypergeometric test using the hypergeom function imported from scipy.stats to determine the *P*-value of the intersection of known mitochondrial proteins (MitoCarta 3.0 proteins) with the STRING neighbors for each target protein using the Ensembl protein IDs in Python 3.8.10 (Figure 1). The hypergeometric test utilizes the following equation:

$$p\left(k,\ M,\ n,\ N\right) = \frac{\binom{n}{k}\binom{M-n}{N-k}}{\binom{M}{N}}.$$

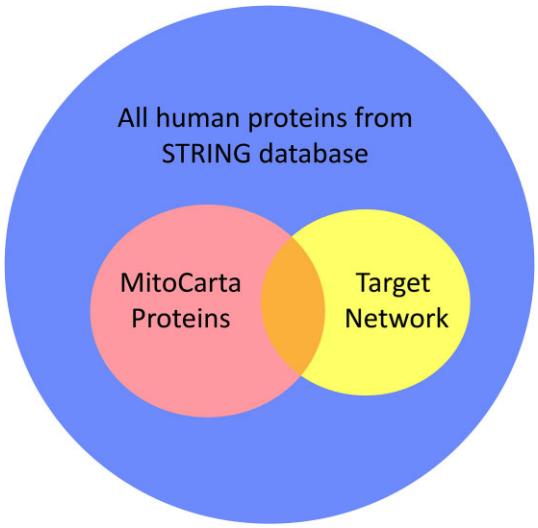

Figure 1. Hypergeometric test identifies protein networks that are enriched in mitochondrial proteins. A hypergeometric test determines the significance of the intersection of known MitoCarta mitochondrial proteins with the network of each protein in the STRING database. A *P*-value that reaches statistical significance indicates that known mitochondrial proteins (from MitoCarta 3.0) are overrepresented in the target network (STRING). The targets that were not present in the MitoCarta 3.0 protein dataset, but having networks enriched in MitoCarta 3.0 proteins, were classified as MitoProximal proteins.

M was defined as the total number of human proteins in the STRING database ($N = 18\,872$). n was defined as the number of proteins in the nearest neighbors for each individual protein from the STRING database. N was defined as the total number of MitoCarta proteins ($N = 11\ 36$) and k was defined as the number of MitoCarta proteins that were found in the target neighbors.

Defining the *P*-value threshold of significance

We used the elbow method to determine the hypergeometric *P*-value threshold of significance to define the MitoProximal proteins (13). When the percent of rediscovered MitoCarta mitochondrial proteins was plotted against a protein's *P*-value rank (Figure 2A), a clear separation was observed between the region where the percentage of the rediscovered MitoCarta proteins was rapidly increasing up to ∼90% of the MitoCarta protein list and a region of the 'law of diminishing returns' where a further decrease in *P*-value did not return the proteins from the original dataset as rapidly (Figure 2B).

We identified the cutoff for the region enriched in known mitochondrial proteins using the elbow method. The 'elbow' method is a heuristic often used in bioinformatics data analysis to determine the optimal number of clusters in a k-means algorithm or to choose a number of principal components in dimensionality reduction algorithms to determine a point where separating the dataset further results in overfitting (13). Plotting the percentage of discovered proteins from the original dataset versus the base 2 logarithm of the *P*-value ranks revealed a sigmoidal curve with three regions: (i) the first from rank 1 to rank ∼64, where *P*-values fall very rapidly by almost 100 orders of magnitude, (ii) then a region where *P*-values fall exponentially up to rank ∼3000 and (iii) finally a plateau after rank ∼3000 (Figure 2C).

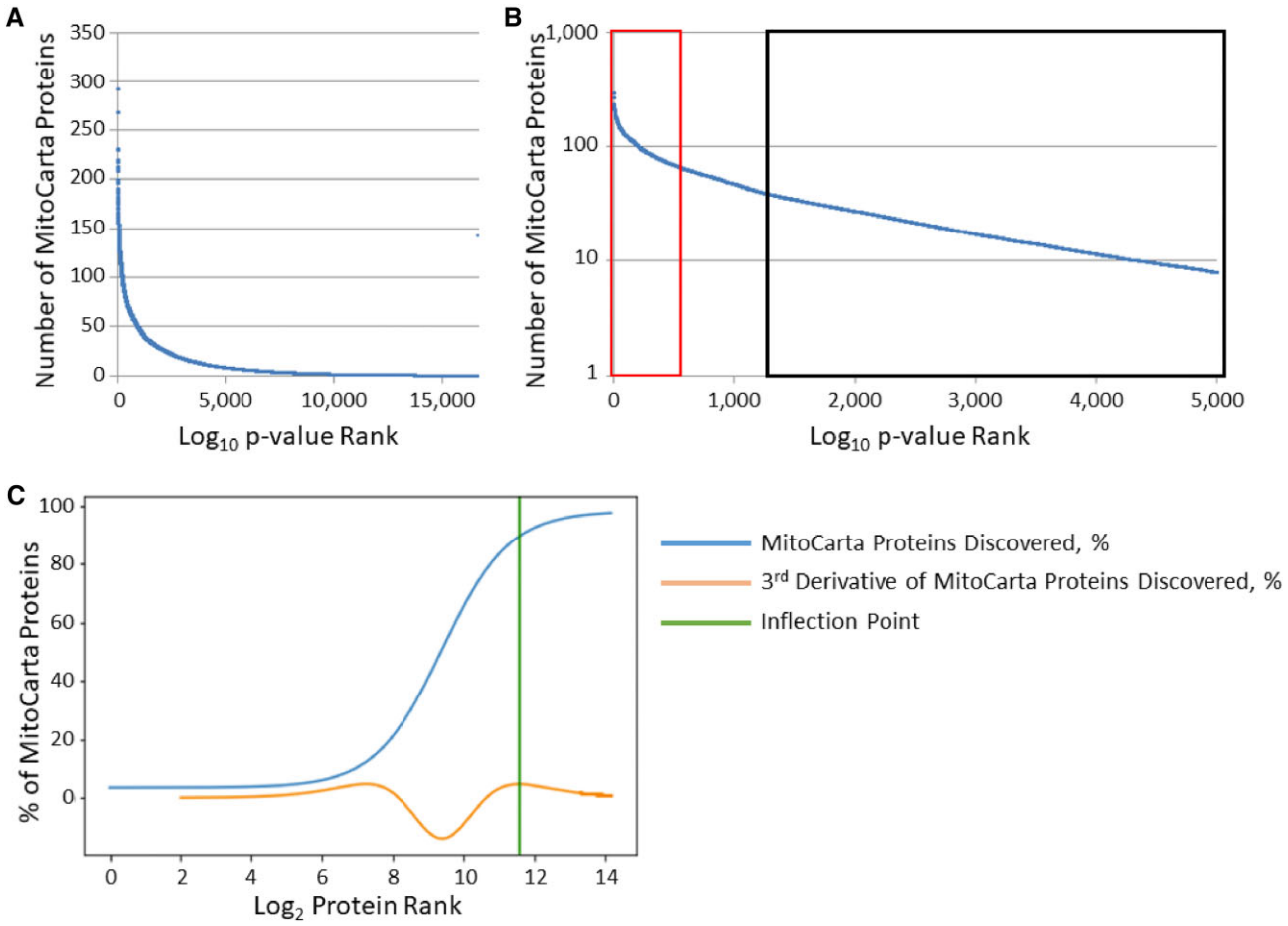

Figure 2. Defining the *P*-value threshold of significance. The number of rediscovered MitoCarta 3.0 proteins was plotted against the $\log_{10}$ *P*-value rank (**A**). A rapid super-exponential decrease in the number of MitoCarta proteins rediscovered with a decreasing *P*-value rank (red box) was noted, after which the *P*-values decreased exponentially (black box, **B**). To determine a data-driven *P*-value threshold, we used the elbow method and took the third derivative (orange line) of the percentage of MitoCarta genes discovered (blue line) to determine the inflection point (vertical green line, **C**), where a further decrease in *P*-value did not return the proteins from the original dataset as rapidly. *X*-axis is the $\log_2$ *P*-value rank.

We used the inflection point between the second and the third region as the cutoff for inclusion of the proteins in the list of MitoProximal proteins. The inflection point was identified by taking a third derivative of the sigmoid curve, corresponding to the *P*-value rank 3096 and *P*-value of 4.49×10^{-3}. MitoProximal proteins were thus identified as those proteins identified in the hypergeometric test with a *P*-value $<4.49 \times 10^{-3}$ and not present in the MitoCarta 3.0 dataset.

Results

GO (6,7), a catalog of genes organized into biological classes, including mitochondrial genes, had limited intersections with MitoCarta 3.0 (Supplementary Figure S1). Thus, experimentally confirmed mitochondrial proteins from MitoCarta are missing from GO, while genes identified as mitochondrial in GO are not in MitoCarta.

To create a more comprehensive list of mitochondrial-related proteins that extends beyond the mitochondrial localized proteins in MitoCarta 3.0, we performed a hypergeometric test to calculate the significance of overlap between target network neighbors and MitoCarta proteins, on the background of all proteins in the STRING database. The test returned 3094 proteins that reached a *P*-value of significance after calculating a data-driven threshold (*P*-value $<4.49 \times 10^{-3}$), of which 1035 proteins were in the MitoCarta 3.0 dataset (Figure 3A and Supplementary Table S1). We identified 2059 MitoProximal proteins (Supplementary Table S2) with a *P*-value $<4.49 \times 10^{-3}$ that were not already present in MitoCarta 3.0 (Figure 3A).

To validate our methodology, we systematically analyzed the resulting output to confirm the inclusion of several additional protein/gene sets relevant to mitochondrial function and disease in the MitoProximal protein list. All of the 97 oxidative phosphorylation subunits were identified in the MitoProximal gene list. A total of 149 protein-encoding genes are considered to contain confirmed pathogenic variants that cause oxidative phosphorylation diseases, or primary mitochondrial disease (14). Of the 149 confirmed oxidative phosphorylation genes involved in mitochondrial diseases, 134 genes (90%) were encompassed in our MitoProximal list (Figure 3B). Of the 15 OXPHOS disease-causing proteins not found in the MitoProximal list, none were found in MitoCarta and only 2 were found in the GO dataset. Further, a genome-wide CRISPR/Cas9 death screen identified 191 genes as essential for oxidative phosphorylation (14). Of the 191 high-confidence (false discovery rate <0.1) genes identified in the screen, 74% were found in our MitoProximal dataset

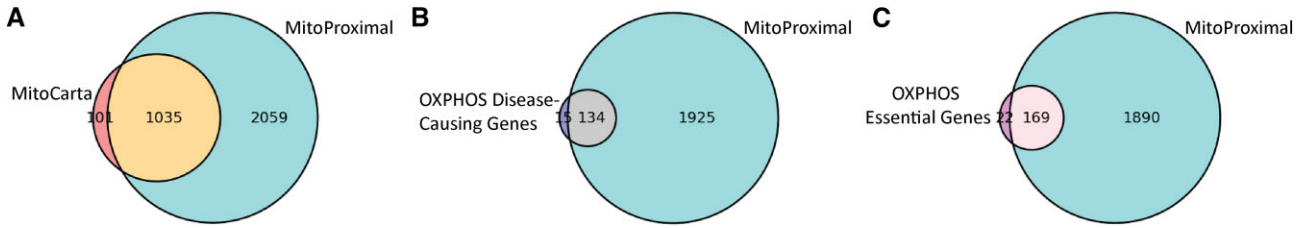

Figure 3. Overlap of the MitoProximal dataset with existing mitochondrial protein and gene datasets. Most of the MitoCarta 3.0 catalog (91%) was rediscovered among the significant proteins identified in the hypergeometric test, along with an additional 2059 novel candidate mitochondrial-related proteins identified (MitoProximal proteins, **A**). Of the protein-encoding genes containing confirmed pathogenic variants in oxidative phosphorylation diseases (primary mitochondrial diseases), 90% were identified in the MitoProximal dataset (**B**). Among genes identified as required for oxidative phosphorylation through a CRISPR/Cas9 genome-wide death screen (14), 88% were in the MitoProximal dataset (**C**). OXPHOS, oxidative phosphorylation.

(Figure 3C). Of the 22 OXPHOS essential proteins not in the MitoProximal list, all but one (SERCA1) was in MitoCarta. In contrast, only 1 of the 22 OXPHOS essential proteins not in MitoProximal list was found in the GO dataset.

We performed a gene enrichment analysis to determine the pathways or biological processes that were enriched within the MitoProximal dataset. Enrichment analysis of the top 200 most significant genes in the MitoProximal dataset revealed that the top 10 cellular processes were those involved in metabolism (Table 2), which is consistent with the role of mitochondria as a metabolic hub of the cell (15). Reactome Metabolism of Amino Acids and Derivatives and GOBP Cellular Amino Acid Metabolism Process were among the top biological processes identified. Branched chain amino acid catabolism occurs within mitochondria to support ATP generation with several enzymes involved, including BCAT1 and ILVBL among the top 200 most significant genes in the MitoProximal dataset (16). Additional enzymes involved in amino acid metabolism that were among the 200 most significant genes identified as MitoProximal proteins include PYCR3, PHGDH, ASS1 and GPT.

Further, MitoProximal proteins were enriched in the GOBP Carbohydrate Derivative Metabolic Process, which encompasses proteins involved in glycolysis, one-carbon metabolism and the pentose phosphate pathway. Although glycolysis occurs within the cytosol, the NADH generated by glycolysis is transported into mitochondria for use by complex I of the respiratory chain in ATP generation. One-carbon metabolism serves to generate the one-carbon units used to synthesize purines, for methylation reactions and to sustain glutathione levels (17). The folate cycle, a component of one-carbon metabolism, takes place within mitochondria with the amino acids glycine and serine used to ultimately generate purines (17). Relatedly, the top 200 MitoProximal proteins were enriched in GOBP Nucleobase-Containing Small Molecule Metabolic Processes with key enzymes in *de novo* purine synthesis ADSL, PPAT, GART, ADSL and GMPS among the most significant proteins identified in the hypergeometric test (Supplementary Table S2).

The MitoProximal dataset contains a number of proteins involved in mitochondrial functions and regulation, as well as enzymes involved in multiple metabolic pathways (Figure 4). Proteins involved in the regulation of mitochondrial biogenesis, including PPARGC1A, PPARGC1B, PPRC1 and NRF1, were identified in the MitoProximal dataset. These mitochondrial biogenesis proteins are not localized within the mitochondrion, and hence not in MitoCarta 3.0, but are nonetheless essential for mitochondrial function. Similarly, CLUH is a cytosolic RNA-binding protein localized to RNA granules where it preserves the translation of mRNAs encoding proteins involved in oxidative phosphorylation, citric acid cycle, fatty acid oxidation and amino acid catabolism under conditions of stress (18–20). CLUH is not localized to mitochondria, and hence not found in MitoCarta; however, CLUH was identified using the hypergeometric test ($P = 1.5 \times 10^{-26}$).

Within the MitoProximal dataset, many ion and metabolite carriers and transporters were identified, which would be expected to regulate metabolism within the cell and consequently mitochondrial activities. Additionally, a number of proteins (e.g. ATG proteins, BECN1, ULK1) involved in autophagy were identified, which is consistent with the importance of autophagy for the clearance and turnover of dysfunctional mitochondria (21).

Among the MitoProximal proteins, 134 are localized to mitochondria based upon annotation in the Human Protein Atlas or NCBI Gene but are not found in MitoCarta 3.0 (Figure 4 and Table 1). Interestingly, NDUFA4L2 was identified as a MitoProximal protein ($P = 1.41 \times 10^{-101}$). NDUFA4L2 is localized to mitochondria and inhibits complex I of the electron transport chain under conditions of hypoxia, limiting oxidant generation (22–24). Hence, identification of such proteins that play essential roles in response to cellular stress may be missing from the existing mitochondrial catalogs as such genes are only expressed under specific conditions. Several MitoProximal proteins ($N = 16$) are localized to mitochondria in the Human Protein Atlas but have not yet been assigned a function.

Discussion

To date, studies cataloging mitochondrial proteins were primarily generated using proteomics approaches of isolated mitochondria or green fluorescent protein (GFP) fusion microscopy to identify proteins localized to mitochondria (5,25–31). Such methods capture proteins localized to mitochondria but miss proteins essential for mitochondrial function that are not localized to mitochondria, are lost through the mitochondrial isolation process or are only expressed under specific cellular stress conditions. We used a hypergeometric test to identify the overlap between MitoCarta 3.0 proteins and network interactions within the STRING database to identify proteins important for mitochondrial function, which we termed MitoProximal proteins. We identified a total of 2059 proteins that were not located in MitoCarta and that met our data-driven *P*-value threshold of significance.

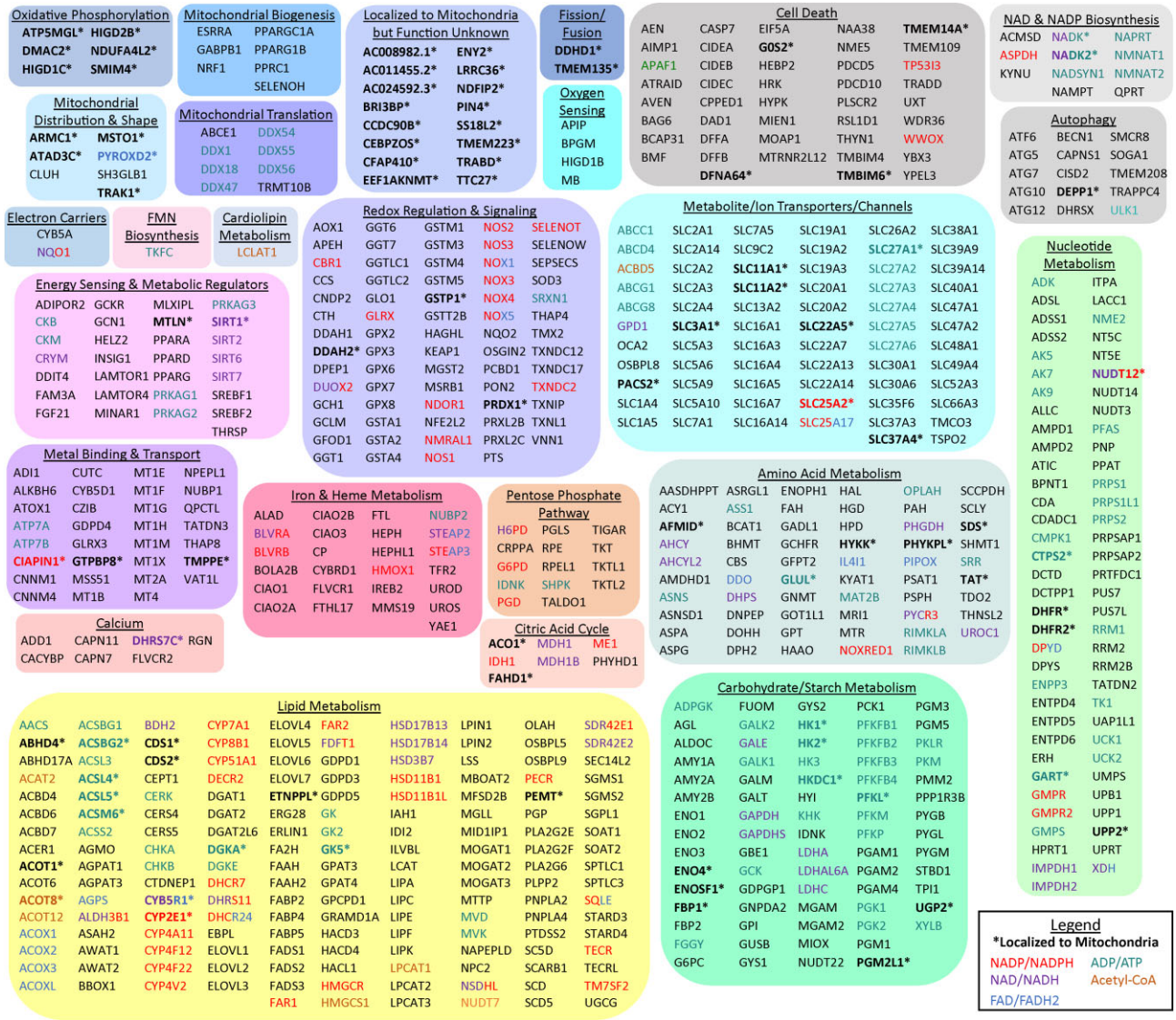

Figure 4. Candidate mitochondrial-related proteins identified by the hypergeometric test. The MitoProximal list consists of proteins spanning many mitochondrial functions and metabolic pathways. Color of font indicates the cofactor(s) required for enzymatic activity (see legend). *Bold font indicates proteins localized to mitochondria. Not all MitoProximal proteins are indicated.

Of the databases of mitochondrial localized proteins, only MitoCarta is maintained and available, while other mitochondrial protein databases, including MitoProteome (32,33), MitoP2 (34,35) and MitoMiner 4.0 (29,36), are no longer available. MitoCarta consists of a catalog of mitochondrial proteins now in its third version (5,28,37). The MitoCarta dataset was initially created by performing proteomics on isolated mitochondria from 14 tissues of the mouse with homologous human proteins identified. GFP tagging and microscopy, computational approaches and curation of the literature were also used to generate and expand the mitochondrial proteome catalog, resulting in the current dataset of 1136 mitochondrial localized proteins in humans (5,28,37). For the majority of the 101 proteins in MitoCarta that were not rediscovered in our method, the proteins did not reach the required level of statistical significance because there were too few or even no interactors (e.g. IQCF5, IQ domain-containing protein F5) in the STRING database. Our study extends the MitoCarta dataset to encompass additional mitochondrial localized proteins and proteins important for mitochondrial function that are not localized to mitochondria.

Our study has several limitations. Despite our stringent, data-driven approach for accounting for multiple testing, it is possible that some of the MitoProximal proteins are false positives. A number of proteins involved in translation and peroxisomes were identified as MitoProximal proteins. Several proteins in the MitoCarta 3.0 list are found in peroxisomes or other membranous, cellular structures, which could have contributed to the identification of additional proteins involved in other organellar processes in the hypergeometric test. Although proteomics approaches have allowed for the identification of the mitochondrial proteome, the close proximity of mitochondria and interactions with other membranous organelles create challenges in obtaining a pure mitochondrial isolate (26). Hence, some of the proteins identified applying proteomics to isolated mitochondria may be false positives, re-

Table 1. MitoProximal proteins localized to mitochondria

Gene symbol	Protein name	*P*-value	Database source	
			HPA	NCBI
ABHD4	Abhydrolase domain containing 4, *N*-acyl phospholipase B	1.2E−09		x
AC008982.1		3.5E−31	x	
AC011455.2		2.2E−58	x	
AC024592.3		3.4E−180	x	
ACO1	Aconitase 1	3.1E−193	x	
ACOT1	Acyl-CoA thioesterase 1	1.4E−25	x	
ACOT8	Acyl-CoA thioesterase 8	3.3E−78	x	
ACSBG2	Acyl-CoA synthetase bubblegum family member 2	2.9E−34		x
ACSL4	Acyl-CoA synthetase long-chain family member 4	5.6E−95	x	
ACSL5	Acyl-CoA synthetase long-chain family member 5	1.8E−52	x	
ACSM6	Acyl-CoA synthetase medium-chain family member 6	5.1E−47		x
ADPRS	ADP-ribosylserine hydrolase	1.1E−09		x
AFMID	Arylformamidase	1.1E−18	x	
ALKBH3	alkB homolog 3, α-ketoglutarate-dependent dioxygenase	2.8E−07	x	
APEX2	Apurinic/apyrimidinic endodeoxyribonuclease 2	2.9E−35		x
ARMC1	Armadillo repeat containing 1	2.0E−25	x	x
AS3MT	Arsenite methyltransferase	5.1E−14	x	
ATAD3C	ATPase family AAA domain containing 3C	3.5E−30	x	x
ATP5MGL	ATP synthase membrane subunit g like	1.1E−137	x	
ATP6V1C2	ATPase H^+ transporting V1 subunit C2	7.5E−23	x	
BRI3BP	BRI3 binding protein	1.0E−03	x	x
C17orf80	Chromosome 17 open reading frame 80	1.3E−73	x	
CCDC90B	Coiled-coil domain containing 90B	2.2E−05	x	x
CDS1	CDP-diacylglycerol synthase 1	5.3E−29		x
CDS2	CDP-diacylglycerol synthase 2	1.8E−29		x
CEBPZOS	CEBPZ opposite strand	7.6E−10		x
CFAP410	Cilia- and flagella-associated protein 410	4.0E−43	x	
CIAPIN1	Cytokine-induced apoptosis inhibitor 1	2.6E−18	x	
CMSS1	cms1 ribosomal small subunit homolog	3.0E−08	x	
CORO7-PAM16	CORO7-PAM16 readthrough	4.4E−47	x	
CTDSP2	CTD small phosphatase 2	2.9E−03	x	
CTPS2	CTP synthase 2	1.0E−50	x	
CTU2	Cytosolic thiouridylase subunit 2	1.0E−05	x	
CYB5R1	Cytochrome b5 reductase 1	6.4E−69	x	x
CYP2E1	Cytochrome P450 family 2 subfamily E member 1	5.1E−09	x	
DDAH2	Dimethylarginine dimethylaminohydrolase 2	4.5E−04	x	x
DDHD1	DDHD domain containing 1	7.5E−06		x
DDT	D-Dopachrome tautomerase	2.2E−14	x	
DFNA64	Diablo IAP-binding mitochondrial protein	1.4E−16	x	
DEPP1	DEPP1 autophagy regulator	6.6E−07	x	
DGKA	Diacylglycerol kinase alpha	5.6E−05	x	
DHFR	Dihydrofolate reductase	3.1E−54	x	
DHFR2	Dihydrofolate reductase 2	3.1E−54	x	x
DHRS3	Dehydrogenase/reductase 3	3.6E−05	x	
DHRS7	Dehydrogenase/reductase 7	1.7E−18	x	
DMAC2	Distal membrane arm assembly component 2	2.1E−06	x	x
DNAJA2	DnaJ heat shock protein family (Hsp40) member A2	7.2E−05		x
DPYSL4	Dihydropyrimidinase like 4	9.9E−04	x	
EEF1AKNMT	EEF1A lysine and N-terminal methyltransferase	9.0E−04		x
EIF2A	Eukaryotic translation initiation factor 2A	2.4E−05	x	
ENO4	Enolase 4	8.7E−89	x	
ENOSF1	Enolase superfamily member 1	9.7E−30		x
ENY2	ENY2 transcription and export complex 2 subunit	6.1E−18	x	
ETNPPL	Ethanolamine-phosphate phospho-lyase	2.9E−09		x
FAHD1	Fumarylacetoacetate hydrolase domain containing 1	4.6E−39	x	
FBP1	Fructose-bisphosphatase 1	1.2E−49	x	
FOCAD	Focadhesin	9.1E−04	x	
G0S2	G0/G1 switch 2	3.3E−06		x
GART	Phosphoribosylglycinamide formyltransferase	4.9E−73	x	
GDAP1	Ganglioside-induced differentiation-associated protein 1	1.9E−20	x	
GK5	Glycerol kinase 5	1.2E−03		x
GLUL	Glutamate-ammonia ligase	5.8E−64	x	
GSTP1	Glutathione *S*-transferase pi 1	8.1E−07	x	
GTPBP8	GTP binding protein 8	7.3E−77	x	
HELB	DNA helicase B	9.7E−09	x	
HIGD1C	HIG1 hypoxia inducible domain family member 1C	4.8E−33		x
HIGD2B	HIG1 hypoxia inducible domain family member 2B	9.9E−40	x	x
HK1	Hexokinase 1	3.3E−63	x	x
HK2	Hexokinase 2	3.6E−46	x	x

Table 1. Continued

Gene symbol	Protein name	*P*-value	Database source	
			HPA	NCBI
HKDC1	Hexokinase domain containing 1	1.8E−22	x	
HSPB6	Heat shock protein family B (small) member 6	1.7E−05	x	
HYKK	Hydroxylysine kinase	6.2E−09		x
IDNK	IDNK gluconokinase	7.3E−08	x	
ILF3	Interleukin enhancer binding factor 3	9.8E−07	x	
IMPA2	Inositol monophosphatase 2	1.0E−13	x	
JARID2	Jumonji and AT-rich interaction domain containing 2	4.0E−09	x	
LRRC36	Leucine-rich repeat containing 36	8.3E−04	x	
LYSMD2	LysM domain containing 2	1.4E−03	x	
MBTPS2	Membrane-bound transcription factor peptidase, site 2	4.3E−29	x	
MOCOS	Molybdenum cofactor sulfurase	1.3E−36	x	
MSTO1	Misato mitochondrial distribution and morphology regulator 1	5.4E−17		x
MTLN	Mitoregulin	1.4E−20		x
MYBPH	Myosin binding protein H	3.6E−07	x	
NADK2	NAD kinase 2, mitochondrial	7.9E−13	x	x
NDFIP2	Nedd4 family interacting protein 2	2.6E−03		x
NDUFA4L2	NDUFA4 mitochondrial complex associated like 2	1.4E−101	x	
NDUFC2-KCTD14	NDUFC2-KCTD14 readthrough	1.2E−117	x	
NOL7	Nucleolar protein 7	7.9E−05	x	
NUDT1	Nudix hydrolase 1	6.5E−18		x
PACS2	Phosphofurin acidic cluster sorting protein 2	1.2E−10	x	
PAPSS2	3′-Phosphoadenosine 5′-phosphosulfate synthase 2	7.1E−05	x	
PCMTD2	Protein-L-isoaspartate (D-aspartate) O-methyltransferase domain containing 2	6.8E−04	x	
PEMT	Phosphatidylethanolamine *N*-methyltransferase	6.6E−19	x	x
PFKL	Phosphofructokinase, liver type	6.7E−60	x	
PGM2L1	Phosphoglucomutase 2 like 1	4.1E−50	x	
PHYKPL	5-Phosphohydroxy-L-lysine phospho-lyase	1.4E−14	x	x
PIN4	Peptidylprolyl *cis/trans*-isomerase, NIMA-interacting 4	6.3E−44		x
PPCS	Phosphopantothenoylcysteine synthetase	9.9E−35	x	
PRDX1	Peroxiredoxin 1	3.7E−93	x	
PSMB4	Proteasome 20S subunit beta 4	1.4E−25	x	
PYROXD2	Pyridine nucleotide-disulphide oxidoreductase domain 2	4.4E−19	x	x
RAD51	RAD51 recombinase	3.5E−16	x	
RAD51C	RAD51 paralog C	4.9E−23	x	
RPL7L1	Ribosomal protein L7 like	1.0E−41	x	
RRP15	Ribosomal RNA processing 15 homolog	2.6E−04	x	
SDS	Serine dehydratase	6.4E−38	x	
SIRT1	Sirtuin 1	3.1E−10	x	
SLC11A1	Solute carrier family 11 member 1	1.0E−07	x	
SLC11A2	Solute carrier family 11 member 2	1.3E−18	x	
SLC22A5	Solute carrier family 22 member 5	5.4E−07	x	
SLC25A2	Solute carrier family 25 member 2	3.3E−12		x
SLC27A1	Solute carrier family 27 member 1	4.2E−33	x	
SLC37A4	Solute carrier family 37 member 4	7.3E−04	x	
SLC3A1	Solute carrier family 3 member 1	3.5E−03	x	
SMIM4	Small integral membrane protein 4	1.1E−06	x	
SPATA18	Spermatogenesis associated 18	3.5E−03	x	x
SS18L2	SS18 like 2	7.9E−10	x	
TAT	Tyrosine aminotransferase	3.8E−25		x
TMBIM6	Transmembrane BAX inhibitor motif containing 6	1.5E−12		x
TMEM135	Transmembrane protein 135	6.0E−12		x
TMEM14A	Transmembrane protein 14A	3.9E−05		x
TMEM14B	Transmembrane protein 14B	1.2E−33		x
TMEM223	Transmembrane protein 223	1.3E−12	x	
TMPPE	Transmembrane protein with metallophosphoesterase domain	3.4E−25	x	
TRABD	TraB domain containing	1.0E−03	x	
TRAK1	Trafficking kinesin protein 1	2.1E−05		x
TRMT12	tRNA methyltransferase 12 homolog	1.2E−12	x	
TRMT61A	tRNA methyltransferase 61A	3.2E−07	x	
TRPT1	tRNA phosphotransferase 1	1.2E−13	x	
TTC27	Tetratricopeptide repeat domain 27	1.1E−06	x	
UGP2	UDP-glucose pyrophosphorylase 2	4.1E−19	x	
UPP2	Uridine phosphorylase 2	1.1E−09	x	
YJEFN3	YjeF N-terminal domain containing 3	2.6E−06	x	x
ZNHIT3	Zinc finger HIT-type containing 3	2.8E−13	x	

HPA, Human Protein Atlas; NCBI, National Center for Biotechnology Information.

Table 2. Top 10 cellular processes enriched for in the MitoProximal dataset

Gene set name	Total number of genes	Number of genes in overlap	*P*-value	*q*-value
GOBP Small Molecule Metabolic Process	1848	115	6.4×10^{-108}	1.7×10^{-103}
GOBP Organic Acid Metabolic Process	966	86	4.1×10^{-90}	5.4×10^{-86}
GOBP Organonitrogen Compound Biosynthetic Process	1821	87	8.1×10^{-68}	7.0×10^{-64}
GOBP Nucleobase-Containing Small Molecule Metabolic Process	678	60	7.8×10^{-61}	5.1×10^{-57}
Reactome Metabolism of Amino Acids and Derivatives	373	45	2.6×10^{-51}	1.3×10^{-47}
GOBP Cellular Amino Acid Metabolic Process	290	41	1.2×10^{-49}	5.3×10^{-46}
Reactome Selenoamino Acid Metabolism	118	32	1.3×10^{-48}	4.7×10^{-45}
GOBP Organophosphate Metabolic Process	1035	59	4.5×10^{-48}	4.8×10^{-45}
GOBP Carbohydrate Derivative Metabolic Process	1113	58	1.7×10^{-45}	5.0×10^{-42}
Reactome Translation	295	37	9.6×10^{-43}	2.5×10^{-39}

sulting from contamination of the mitochondrial isolates with membranes (and hence proteins) from other organelles such as peroxisomes, endoplasmic reticulum or Golgi. Whether such peroxisomal, endoplasmic reticulum or other vesicular-related proteins are truly important for mitochondrial function or bystanders that contaminated the isolation of mitochondria used for the proteomics studies is unclear. Some mitochondrial proteins or regulators of mitochondrial function may not have been identified by our criteria for classification as a MitoProximal protein or may not have reached our *P*-value threshold of significance. Unidentified or poorly understood proteins with few known protein interactions may be missing as well. Experimental validation of the MitoProximal genes will be the focus of future studies.

Using a data-driven approach, we extend the catalog of proteins relevant to mitochondrial function to encompass an additional 2059 MitoProximal proteins beyond those in the MitoCarta 3.0 database. The MitoProximal proteins are involved in multiple metabolic pathways, energy and oxygen sensing, mitochondrial biogenesis and dynamics, cell death, autophagy, metabolite transporters and channels, and transport and binding of metals and heme. The MitoProximal dataset extends the list of mitochondrial-related proteins to facilitate the identification of pathogenic variants and genetic modifiers of human disease.

Data availability

The scripts for performing the hypergeometric test and defining the *P*-value level of significance are available in the GitHub repository (https://github.com/jessicalfetterman/MitoProximal) and Zenodo (DOI: 10.5281/zenodo.10260870). The entire output of the hypergeometric test, annotated MitoProximal dataset and additional datasets used herein are provided in Supplementary Data.

Supplementary data

Supplementary Data are available at NARGAB Online.

Funding

National Heart, Lung, and Blood Institute [K01 HL143142 to J.L.F.]; Astellas Pharma US, Inc. Funding for open access charge: Mitobridge, Inc. (an Astellas Pharma company).

Conflict of interest statement

The authors declare the following financial interests/personal relationships that may be considered potential competing interests: D.L. is a current or former employee of Mitobridge, Inc. (an Astellas Pharma company) and may have or currently own shares of these companies. J.L.F. is a former consultant of Astellas Pharma company.

References

1. Alston,C.L., Rocha,M.C., Lax,N.Z., Turnbull,D.M. and Taylor,R.W. (2017) The genetics and pathology of mitochondrial disease. *J. Pathol.*, **241**, 236–250.
2. Vafai,S.B. and Mootha,V.K. (2012) Mitochondrial disorders as windows into an ancient organelle. *Nature*, **491**, 374–383.
3. Wallace,D.C. (2018) Mitochondrial genetic medicine. *Nat. Genet.*, **50**, 1642–1649.
4. Calvo,S.E., Compton,A.G., Hershman,S.G., Lim,S.C., Lieber,D.S., Tucker,E.J., Laskowski,A., Garone,C., Liu,S., Jaffe,D.B., *et al.* (2012) Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. *Sci. Transl. Med.*, **4**, 118ra110.
5. Rath,S., Sharma,R., Gupta,R., Ast,T., Chan,C., Durham,T.J., Goodman,R.P., Grabarek,Z., Haas,M.E., Hung,W.H.W., *et al.* (2020) MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Res.*, **49**, D1541–D1547.
6. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
7. Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
8. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P., *et al.* (2021) The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.
9. Mudunuri,U., Che,A., Yi,M. and Stephens,R.M. (2009) bioDBnet: the biological database network. *Bioinformatics*, **25**, 555–556.
10. Stelzer,G., Rosen,N., Plaschkes,I., Zimmerman,S., Twik,M., Fishilevich,S., Stein,T.I., Nudel,R., Lieder,I., Mazor,Y., *et al.* (2016) The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics*, **54**, 1.30.1–1.30.33.
11. Uhlen,M., Fagerberg,L., Hallstrom,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,A., Kampf,C., Sjostedt,E., Asplund,A., *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.

12. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdottir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular Signatures Database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
13. Thorndike,R.L. (1953) Who belongs in the family?*Psychometrika*, **18**, 267–276.
14. Arroyo,J.D., Jourdain,A.A., Calvo,S.E., Ballarano,C.A., Doench,J.G., Root,D.E. and Mootha,V.K. (2016) A genome-wide CRISPR death screen identifies genes essential for oxidative phosphorylation. *Cell Metab.*, **24**, 875–885.
15. Spinelli,J.B. and Haigis,M.C. (2018) The multifaceted contributions of mitochondria to cellular metabolism. *Nat. Cell Biol.*, **20**, 745–754.
16. Brosnan,J.T. and Brosnan,M.E. (2006) Branched-chain amino acids: enzyme and substrate regulation. *J. Nutr.*, **136**, 207S–211S.
17. Ducker,G.S. and Rabinowitz,J.D. (2017) One-carbon metabolism in health and disease. *Cell Metab.*, **25**, 27–42.
18. Gao,J., Schatton,D., Martinelli,P., Hansen,H., Pla-Martin,D., Barth,E., Becker,C., Altmueller,J., Frommolt,P., Sardiello,M., *et al.* (2014) CLUH regulates mitochondrial biogenesis by binding mRNAs of nuclear-encoded mitochondrial proteins. *J. Cell Biol.*, **207**, 213–223.
19. Schatton,D., Pla-Martin,D., Marx,M.C., Hansen,H., Mourier,A., Nemazanyy,I., Pessia,A., Zentis,P., Corona,T., Kondylis,V., *et al.* (2017) CLUH regulates mitochondrial metabolism by controlling translation and decay of target mRNAs. *J. Cell Biol.*, **216**, 675–693.
20. Wakim,J., Goudenege,D., Perrot,R., Gueguen,N., Desquiret-Dumas,V., Chao de la Barca,J.M., Dalla Rosa,I., Manero,F., Le Mao,M., Chupin,S., *et al.* (2017) CLUH couples mitochondrial distribution to the energetic and metabolic status. *J. Cell Sci.*, **130**, 1940–1951.
21. Gottlieb,R.A. and Carreira,R.S. (2010) Autophagy in health and disease. 5. Mitophagy as a way of life. *Am. J. Physiol. Cell Physiol.*, **299**, C203–C210.
22. Lai,R.K., Xu,I.M., Chiu,D.K., Tse,A.P., Wei,L.L., Law,C.T., Lee,D., Wong,C.M., Wong,M.P., Ng,I.O., *et al.* (2016) NDUFA4L2 fine-tunes oxidative stress in hepatocellular carcinoma. *Clin. Cancer Res.*, **22**, 3105–3117.
23. Piltti,J., Bygdell,J., Qu,C. and Lammi,M.J. (2018) Effects of long-term low oxygen tension in human chondrosarcoma cells. *J. Cell. Biochem.*, **119**, 2320–2332.
24. Tello,D., Balsa,E., Acosta-Iborra,B., Fuertes-Yebra,E., Elorza,A., Ordonez,A., Corral-Escariz,M., Soro,I., Lopez-Bernardo,E., Perales-Clemente,E., *et al.* (2011) Induction of the mitochondrial NDUFA4L2 protein by HIF-1alpha decreases oxygen consumption by inhibiting complex I activity. *Cell Metab.*, **14**, 768–779.
25. Thul,P.J., Akesson,L., Wiking,M., Mahdessian,D., Geladaki,A., Ait Blal,H., Alm,T., Asplund,A., Bjork,L., Breckels,L.M., *et al.* (2017) A subcellular map of the human proteome. *Science*, **356**, eaal3321.
26. Palmfeldt,J. and Bross,P. (2017) Proteomics of human mitochondria. *Mitochondrion*, **33**, 2–14.
27. Antonicka,H., Lin,Z.Y., Janer,A., Aaltonen,M.J., Weraarpachai,W., Gingras,A.C. and Shoubridge,E.A. (2020) A high-density human mitochondrial proximity interaction network. *Cell Metab.*, **32**, 479–497.
28. Pagliarini,D.J., Calvo,S.E., Chang,B., Sheth,S.A., Vafai,S.B., Ong,S.E., Walford,G.A., Sugiana,C., Boneh,A., Chen,W.K., *et al.* (2008) A mitochondrial protein compendium elucidates complex I disease biology. *Cell*, **134**, 112–123.
29. Smith,A.C. and Robinson,A.J. (2019) MitoMiner v4.0: an updated database of mitochondrial localization evidence, phenotypes and diseases. *Nucleic Acids Res.*, **47**, D1225–D1228.
30. Morgenstern,M., Peikert,C.D., Lubbert,P., Suppanz,I., Klemm,C., Alka,O., Steiert,C., Naumenko,N., Schendzielorz,A., Melchionda,L., *et al.* (2021) Quantitative high-confidence human mitochondrial proteome and its dynamics in cellular context. *Cell Metab.*, **33**, 2464–2483.
31. Rhee,H.W., Zou,P., Udeshi,N.D., Martell,J.D., Mootha,V.K., Carr,S.A. and Ting,A.Y. (2013) Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science*, **339**, 1328–1331.
32. Cotter,D., Guda,P., Fahy,E. and Subramaniam,S. (2004) MitoProteome: mitochondrial protein sequence database and annotation system. *Nucleic Acids Res.*, **32**, D463–D467.
33. Guda,P., Subramaniam,S. and Guda,C. (2007) MitoProteome: human heart mitochondrial protein sequence database. *Methods Mol. Biol.*, **357**, 375–383.
34. Elstner,M., Andreoli,C., Klopstock,T., Meitinger,T. and Prokisch,H. (2009) The mitochondrial proteome database: MitoP2. *Methods Enzymol.*, **457**, 3–20.
35. Prokisch,H., Andreoli,C., Ahting,U., Heiss,K., Ruepp,A., Scharfe,C. and Meitinger,T. (2006) MitoP2: the mitochondrial proteome database—now including mouse data. *Nucleic Acids Res.*, **34**, D705–D711.
36. Smith,A.C. and Robinson,A.J. (2009) MitoMiner, an integrated database for the storage and analysis of mitochondrial proteomics data. *Mol. Cell. Proteomics*, **8**, 1324–1337.
37. Calvo,S.E., Clauser,K.R. and Mootha,V.K. (2016) MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res.*, **44**, D1251–D1257.

Received: April 24, 2023. **Revised:** October 25, 2023. **Editorial Decision:** December 4, 2023. **Accepted:** December 6, 2023