Nucleic Acids Research Advance Access published October 7, 2015

Nucleic Acids Research, 2015 **1**
doi: 10.1093/nar/gkv1003

MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins

Sarah E. Calvo[1,2,3,*], Karl R. Clauser[2] and Vamsi K. Mootha[1,2,3,*]

[1]Howard Hughes Medical Institute and Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA, [2]Broad Institute, Cambridge, MA 02141, USA and [3]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

Received August 04, 2015; Revised September 21, 2015; Accepted September 23, 2015

ABSTRACT

Mitochondria are complex organelles that house essential pathways involved in energy metabolism, ion homeostasis, signalling and apoptosis. To understand mitochondrial pathways in health and disease, it is crucial to have an accurate inventory of the organelle's protein components. In 2008, we made substantial progress toward this goal by performing in-depth mass spectrometry of mitochondria from 14 organs, epitope tagging/microscopy and Bayesian integration to assemble MitoCarta (www.broadinstitute.org/pubs/MitoCarta): an inventory of genes encoding mitochondrial-localized proteins and their expression across 14 mouse tissues. Using the same strategy we have now reconstructed this inventory separately for human and for mouse based on (i) improved gene transcript models, (ii) updated literature curation, including results from proteomic analyses of mitochondrial subcompartments, (iii) improved homology mapping and (iv) updated versions of all seven original data sets. The updated human MitoCarta2.0 consists of 1158 human genes, including 918 genes in the original inventory as well as 240 additional genes. The updated mouse MitoCarta2.0 consists of 1158 genes, including 967 genes in the original inventory plus 191 additional genes. The improved MitoCarta 2.0 inventory provides a molecular framework for system-level analysis of mammalian mitochondria.

INTRODUCTION

There is increasing appreciation for the essential roles that mitochondria play not only in oxidative phosphorylation and energy metabolism, but also in small molecule metabolism, ion homeostasis, immune signalling and cell death. Mitochondria originally descended from an endosymbiotic bacterium, predicted to resemble modern-day α-proteobacteria, early in eukaryotic evolution (1). Mammalian mitochondria contain their own genome (mtDNA), which encodes a total of 13 proteins that are all core components of oxidative phosphorylation. However, all of its remaining >1000 proteins (2) are nuclear encoded and imported into the organelle. Mutations in either the mtDNA or the nuclear genome underlie the largest collection of inborn errors of metabolism (3), and there is growing evidence that a gradual decline in mitochondrial activity is associated with aging and age-associated disorders.

To fully understand the molecular basis of mitochondrial physiology and the organelle's role in disease, it is very useful to have a complete protein parts list for this organelle. In 2008, we constructed the MitoCarta1.0 inventory of mitochondrial proteins using multiple experimental and computational approaches (4). At that time, we purified mitochondria from 14 mouse tissues and performed in-depth tandem mass spectrometry (MS/MS) to identify mitochondrial proteins. We then compiled complementary clues of mitochondrial localization from homology to yeast and *Rickettsia prowazekii* proteins, presence of mitochondrial targeting signals and protein domains, and RNA coexpression across tissues and during mitochondrial biogenesis (Figure 1). Using a naïve Bayes integration (5), every mouse gene was assigned a combined score of mitochondrial localization from the seven data sources, each weighted by its accuracy based on large training sets of known mitochondrial and non-mitochondrial mouse genes. The resulting MitoCarta1.0 inventory of 1098 mouse genes contained 591 curated mitochondrial components used for training, 131 proteins validated using GFP/microscopy, and 376 proteins assigned to the organelle at a 10% false discovery rate (FDR). MitoCarta1.0 has been widely used to elucidate the function of uncharacterized genes and pathways (6–9), including reverse genetic screens (8,10) and forward genetic approaches to identify genes underlying rare mitochondrial disorders (11–14).

Here we present an updated MitoCarta2.0 inventory using the same overall strategy (Figure 1). The seven un-

*To whom correspondence should be addressed. Tel: +617 714 7687; Email: scalvo@broadinstitute.org
Correspondence may also be addressed to Vamsi K. Mootha. Tel: +1 617 643 3059; Email: vamsi@hms.harvard.edu

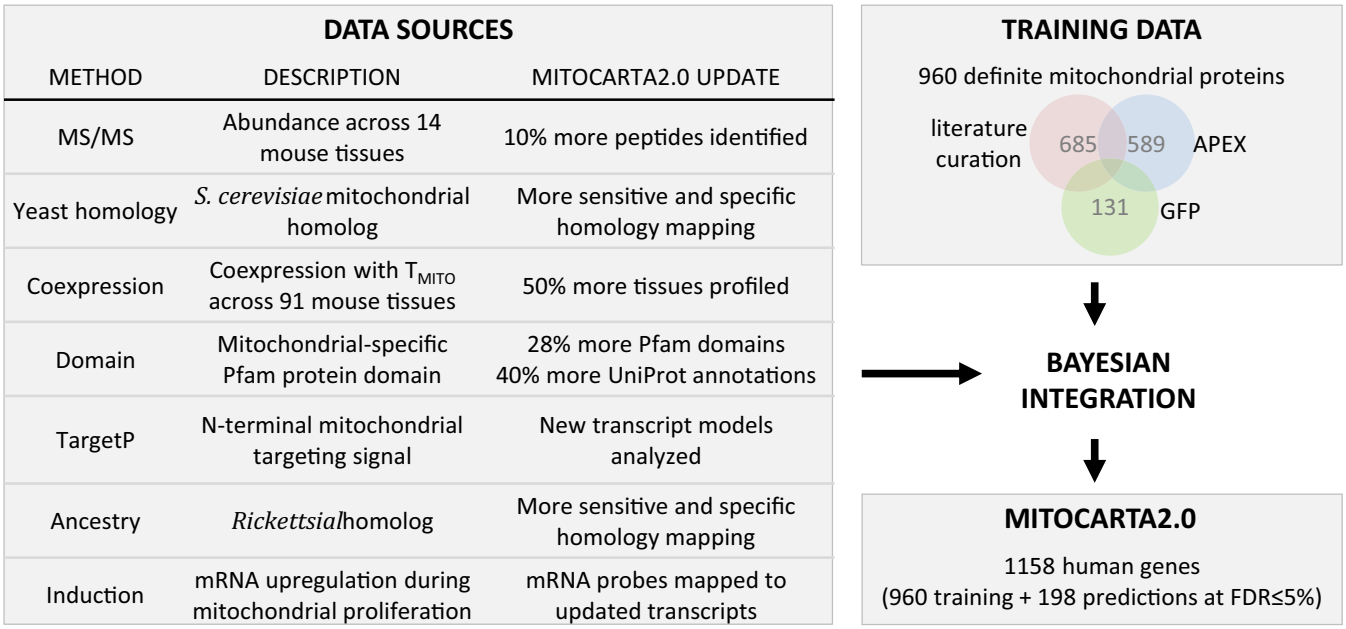

Figure 1. An updated inventory of mitochondrial proteins. MitoCarta2.0 is an inventory of 1158 human genes encoding proteins with strong support of mitochondrial localization. It is built by combining a large training set of 960 mitochondrial genes (based on literature curation, APEX-based mass spectrometry experiments and GFP-tagging/microscopy) with a Bayesian integration of seven genome-scale data sets, updated extensively from the previous MitoCarta1.0 inventory.

derlying data sources have been substantially updated using improved transcript models, MS/MS search algorithms, database versions and homology detection methods. Furthermore, the mitochondrial training set was increased by 60%. The MitoCarta 2.0 inventory consists of 1158 human genes and 1158 mouse genes encoding mitochondrial proteins. The MitoCarta2.0 website www.broadinstitute.org/pubs/MitoCarta freely provides the updated mitochondrial gene identifiers, evidence of mitochondrial localization, protein expression across 14 mouse tissues and protein sequences.

INTEGRATION OF GENOME-SCALE DATA SETS

Method overview

The MitoCarta2.0 inventory of mitochondrial proteins is constructed by first compiling the evidence of mitochondrial localization from seven complementary data sources (Figure 1). In parallel, we compiled large training data of known mitochondrial proteins (T_{mito}) and non-mitochondrial proteins (T_{non_mito}). These training data are used to assess the accuracy of each input data source by computing a likelihood score of mitochondrial localization at a range of input values (Figure 2). The seven individual likelihood scores are combined using a naïve Bayes methodology into an overall score for each gene (15). The resulting naïve Bayes score is far more accurate at scoring the known training data compared to each individual method (Figure 2). The final MitoCarta2.0 inventory is constructed by combining the T_{mito} training data with all genes scoring below a 5% false discovery threshold (Figure 1).

Gene models

The MitoCarta2.0 database is based on human and mouse RefSeq proteins (release 63) (16) that are mapped to NCBI Gene loci (ftp.ncbi.nih.gov/gene/DATA, 02/19/2014) (17).

Training data

All human and mouse genes are partitioned into three sets: T_{mito} (960 human, 961 mouse), $T_{possible\_mito}$ (816 human, 750 mouse) or T_{non_mito} (17468 human, 18918 mouse) as follows.

The T_{mito} set of definite mitochondrial proteins is the union of (i) literature curation of proteins with strong experimental evidence of mitochondrial localization in mammals (see Supplementary Data), (ii) presence in the mitochondrial matrix proteome or intermembrane space (IMS) proteome in HEK 293T cells based on APEX-labeling (18,19) or (iii) confirmed mitochondrial localization by GFP-tagging and microscopy (4). 15 proteins in the previous $T_{mito1.0}$ were excluded based on updated literature curation (*Aadat, Armc4, Eln, Iqce, Mobp, Myl10, Nt5c3, Phyhipl, Pisd, Pla2g15, Pts, Tmem143, Tmem186, Tshz3, Txn1*). We include the APEX-based matrix and IMS proteomes in T_{mito} given the extremely high specificity of APEX-labeling. Human and mouse T_{mito} sets were created using human-mouse orthologs (best reciprocal BlastP hits, Expect < 1e-3) with the addition of species-specific genes with literature evidence.

Genes that did not meet our selective criteria for inclusion in T_{mito} but which had some evidence of mitochondrial localization from the MitoP2 database (20) or the NCBI GO database (17,21) (downloaded 2/19/2014) were grouped into a $T_{possible\_mito}$ gene set, not used for training.

A

	Bin	T_{mito}	Sensitivity	cFDR	LogOdds
MS/MS	75-100pure	201	21%	2%	9
	50-75pure	241	25%	4%	8
	75-100ambig	30	3%	22%	6
	25-50pure	69	7%	25%	5
	50-75ambig	71	7%	44%	4
	25-50ambig	90	9%	65%	3
	0-25pure	11	1%	77%	2
	75-100crude	6	1%	88%	1
	50-75crude	17	2%	89%	1
	25-50crude	15	2%	92%	0
	0-25ambig	70	7%	94%	0
	0-25crude	10	1%	97%	-1
	NA	126	13%	99%	-3
Yeast	OrthologMitoHighConf	210	22%	4%	8
	HomologMitoHighConf	148	15%	16%	6
	OrthologMitoLowConf	52	5%	59%	3
	HomologMitoLowConf	75	8%	87%	1
	NoMitoHomolog	472	49%	96%	-1
Coexpression	40-50	160	17%	2%	10
	30-40	76	8%	18%	6
	20-30	78	8%	35%	5
	10-20	130	14%	56%	3
	5-10	137	14%	78%	2
	0-5	369	39%	96%	-1
	NA	7	1%	100%	-4
Domain	MitoDomain	290	30%	15%	6
	SharedDomain	530	55%	86%	1
	NA	88	9%	96%	-1
	NonMitoDomain	49	5%	99%	-3
TargetP	MitoConf1	175	18%	15%	6
	MitoConf2	248	26%	47%	4
	MitoConf3	100	10%	77%	2
	MitoConf4	66	7%	86%	1
	MitoConf5	60	6%	90%	1
	NA	308	32%	97%	-1
Ancestry	Ortholog	123	13%	33%	5
	Homolog	226	24%	70%	2
	NoHomolog	608	64%	95%	-1
Induction	>1.5	119	12%	36%	4
	0.5-1.5	192	20%	57%	3
	<0.5	59	6%	92%	0
	NA	587	61%	95%	-1
	Naive Bayes Integration FDR≤5%	767	80%	5%	8

LogOdds -4 10

B

MitoCarta version	1.0	2.0	**ECHDC1 (ethylmalonyl-CoA decarboxylase 1)**
MS/MS	2	6	Improved MS/MS search algorithm
Yeast	-1	6	Yeast homolog EHD3 newly annotated mitochondrial
Coexpression	-1	3	Coexpression in updated GNFv3 tissue atlas
Domain	1	1	No change
TargetP	-1	-1	No change
Ancestry	2	2	No change
Induction	-1	-1	No change
Naive Bayes Score	1	16	
Naive Bayes FDR	21%	0%	New MitoCarta2.0 prediction FDR<5%

C

MitoCarta version	1.0	2.0	**NFXL1 (nuclear TF, X-box binding-like 1)**
MS/MS	-3	-3	No change
Yeast	-1	-1	No change
Coexpression	0	-1	Updated GNFv3 tissue atlas excludes this gene
Domain	0	-3	Full-length protein has non-mitochondrial ZNF domain
TargetP	9	1	Full-length protein has low-confidence MTS
Ancestry	-1	-1	No change
Induction	-1	-1	No change
Naive Bayes Score	5	-9	
Naive Bayes FDR	10%	88%	MitoCarta1.0 prediction not in MitoCarta2.0

D

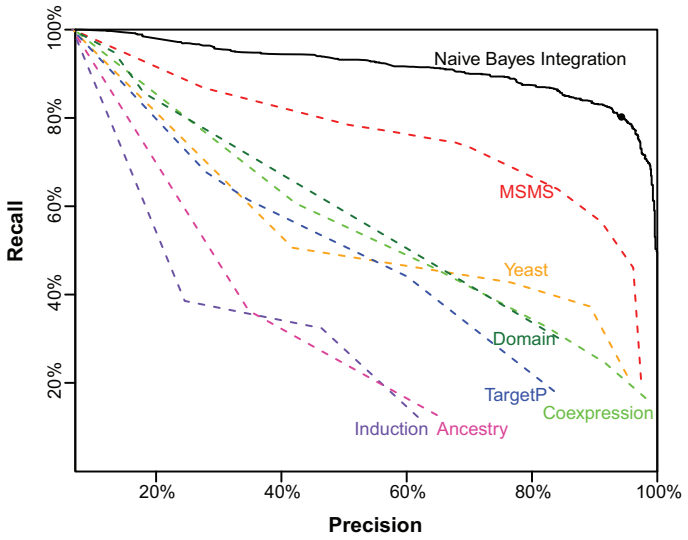

Figure 2. Naïve Bayes integration improves accuracy over individual data sources. **(A)** Accuracy of each of seven data sources and the Maestro naïve Bayes integration, calculated from the human T_{mito} and T_{non_mito} genes. **(B)** LogOdds scores for gene *ECHDC1*, in MitoCarta2.0 but not MitoCarta1.0, highlights improvements in three data sources **(C)** LogOdds scores for gene *NFXL1*, in MitoCarta1.0 but not MitoCarta2.0, highlights improvements to RefSeq gene models—as previous 58aa protein fragment XP_001052092 has been replaced with 918aa full-length protein NP_598682. **(D)** ROC curve shows accuracy of each data source individually as well as the combined naïve Bayes Integration. Black circle indicates 5% FDR threshold.

We define the non-mitochondrial training set T_{non_mito} as all genes not in T_{mito} or $T_{possible\_mito}$. This differs from MitoCarta1.0, where $T_{non\_mito1.0}$ contained 2519 genes whose proteins were reliably localized in non-mitochondrial compartments (e.g. ER, nucleus, lysosome, plasma membrane, vacuole). Thus, $T_{non\_mito2.0}$ now contains thousands of cytoplasmic proteins that were previously underrepresented.

Data integration

As in MitoCarta1.0, seven methods for determining mitochondrial localization were integrated using the Maestro naïve Bayes classifier whereby each method is weighted based on its accuracy (15). Training sets (T_{mito} and T_{non_mito}) were used to create a LogOdds score for each feature F at each predefined bin b (Figure 2A), defined as $\log_2[P(F_b \mid T_{mito}) / P(F_b \mid T_{non\_mito})]$. Assuming conditional independence between the data sets (Supplementary Figure S1), the individual LogOdds scores were summed to create a Maestro score for each gene (Figure 2B, C). For transcript or protein level scores, the gene inherited the highest score of any isoform. The scores for the seven genomic features were calculated at predefined ranges (Figure 2A) as follows (see Supplementary Data for details):

MS/MS: one of 12 categories corresponding to the percent of the protein [0–25%, 25–50%, 50–75%, 75–100%] detected by MS/MS peptides in mitochondria purified from

mouse 14 tissues crossed with a subtractive proteomics enrichment score [crude-enriched, pure-enriched, ambiguous-enrichment], or NA if not detected (4). While the scoring method was identical to MitoCarta1.0, the original MS/MS spectra were searched against new RefSeq transcript models using updated SpectrumMill software that reversed faulty acquisition-time lock mass correction of MS1 scans and precursor masses, which resulted in 75% more spectra identified and 10% more unique peptides identified. The improvements are chiefly due to reversing the faulty lock mass calibration (see Supplementary Data).

Yeast: categorical score [OrthologMitoHighConf, OrthologMitoLowConf, HomologMitoHighConf, HomologMitoLowConf, NoMitoHomolog]. Homology was determined by BlastP (22) top hit (Expect < 1e-3) or jackHMMER (23) reciprocal hit (see Supplementary Data). Orthology was defined as a 1:1 homolog (i.e. the yeast gene had only one homolog in human/mouse). Genes with a yeast homolog/ortholog annotated as mitochondrial in SGD (Saccharomyces Genome Database, 03/06/14) (24) were scored as either MitoHighConf (SGD manual annotation, excluding dual localized proteins) or MitoLowConf (SGD dual localized proteins, or annotated mitochondrial based on high throughput data only). Genes that lacked a yeast homolog or where the yeast homolog was not annotated mitochondrial were categorized as NoMitoHomolog. Compared to MitoCarta1.0, this scoring method was more sensitive due to use of jackHMMER to identify distant homologs, and more specific due to the separate scoring of orthologs and homologs.

Coexpression: N50 score (number of T_{mito} genes found within the gene's 50 nearest transcriptional co-expression neighbours, using Spearman correlation) within the GNF Mouse GeneAtlas V3 survey of gene expression across 91 mouse tissues (GSE10246) (25). In MitoCarta1.0, the GNFv1 atlas surveyed only 61 tissues (26).

Protein domain: categorical score [MitoDomain, NonMitoDomain, SharedDomain or NA] representing presence of a protein domain that is exclusively mitochondrial, exclusively non-mitochondrial, ambiguous or not present in any annotated eukaryotic protein (UniProt Knowledgebase Release 2014_06) (27,28). Protein domains were identified using HMMER (23) based on Pfam version 27 (29). This scoring was identical to MitoCarta1.0, with updated UniProt and Pfam databases.

Targeting sequence: confidence score of mitochondrial targeting signal from TargetP v1.1 (30), as in MitoCarta1.0.

Endosymbiont ancestry: categorical score [Ortholog, Homolog, NoHomolog], where homology was defined by BlastP (Expect<1e-3) or jackHMMER to *Rickettsia prowazekii* (see Supplementary Data), and orthology was defined as a 1:1 homolog (i.e. the *Rickettsia* gene has only one homolog in human/mouse). Compared to MitoCarta1.0, this scoring was more sensitive due to use of jackHMMER and more specific due to separate scoring of orthologs and homologs.

Induction: log2 fold-change of mRNA expression in cellular models of mitochondrial proliferation (overexpression of PGC-1α in mouse myotubes) compared to controls (15,31). Compared to MitoCarta1.0, probes in this data set were re-annotated using new transcript models (32), and data were normalized using gcRMA (33).

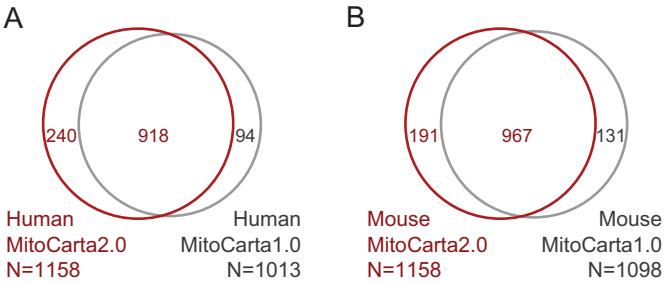

Figure 3. Overlap between MitoCarta versions

In contrast to MitoCarta1.0, these seven features were generated separately for all human genes and all mouse genes. Features that were mouse-specific (MS/MS, Co-expression, Induction) were mapped to human orthologs (BlastP best reciprocal hit, Expect<1e-3). The final MitoCarta2.0 lists for human and mouse were constructed as the union of all T_{mito} genes and all Maestro predictions with FDR≤5%. As in MitoCarta1.0, FDR was defined as (1-SP)/(1-SP + SN x O_{prior}) with specificity SP = TN / (TN + FP); sensitivity SN = TP / (TP + FN); O_{prior} = 1500/21000; TP, true positives; TN, true negatives; FP, false positives; FN, false negatives.

Accuracy of data sets and naïve Bayes integration

We assessed the accuracy of each data set and the combined Maestro naïve Bayes integration using recall and precision at predicting the training sets (Figure 2D). Recall is equivalent to sensitivity and precision is the percent of all predictions expected to be true (TP/(TP+FP) corrected for the size of the training data sets, equivalent to 1-FDR). As shown in Figure 2D, the Maestro naïve Bayes integration has substantially increased accuracy compared with the individual data sets. At the selected 5% FDR threshold on the human data set, the naïve Bayes method has 80% sensitivity and 99.6% specificity—far outstripping any single method. Using ten-fold cross-validation, the naïve Bayes integration showed a similar 79% sensitivity and 99.7% specificity at the same 5% FDR threshold.

HUMAN AND MOUSE MITOCARTA2.0

We separately performed a naïve Bayes integration to create a human-centric and mouse-centric MitoCarta2.0 inventory (Figure 3).

The human MitoCarta2.0 inventory contains 1158 genes, 79% of which overlap MitoCarta1.0 (Figure 3A). Of the 240 genes not in MitoCarta1.0, 100 were detected in the APEX-based matrix or IMS proteomes, 36 have other experimental literature evidence, and 104 achieve high probability of mitochondrial localization at the 5% FDR. For example, *ECHDC1* now has evidence of mitochondrial localization based on updated MS/MS, yeast homology and coexpression (Figure 2B). Of the 94 MitoCarta1.0 human genes that are now retired and not in MitoCarta2.0, 9 were pseudogenes no longer present in the latest RefSeq

database and 85 score below our stringent 5% FDR. For example, *NFXL1* was in MitoCarta1.0 based solely on a high confidence TargetP prediction (Figure 2C), however the more recent RefSeq database replaces the previous protein fragment (XP_001052092) with a full-length protein (NP_598682) that has only a low-confidence TargetP prediction thus it is no longer predicted as resident in the mitochondrion.

The mouse MitoCarta2.0 inventory contains 1158 genes, 83% of which overlap MitoCarta1.0 (Figure 3B). 191 mouse genes were not in MitoCarta1.0, including 84 detected in the APEX-based matrix or IMS proteomes, 31 with other literature experimental evidence and 76 computational predictions (FDR≤5%). The previous version of RefSeq had a larger number of mouse pseudogenes. Of the 131 genes only in MitoCarta1.0, 42 were retired pseudogenes, 15 were $T_{mito1.0}$ genes no longer deemed to have strong evidence and rest were low-confidence computational predictions.

The vast majority of human and mouse MitoCarta2.0 genes are reciprocal top hits (96%). However, the separate inventories contain species-specific genes (e.g. human *ATAD3B,* mouse *Csl*) and predictions that had slightly different species-specific scores and thus exceeded the FDR threshold in only one of the two mammalian species (e.g. human *BOLA3, LDHB* and mouse *Ppm1m*).

WEBSITE INTERFACE

MitoCarta2.0 inventory is available at www.broadinstitute.org/pubs/MitoCarta. The human and mouse mitochondrial inventories contain the Maestro naïve Bayes score and FDR, a summary of the evidence supporting mitochondrial localization and protein expression in 14 mouse tissues. Available for download are the naïve Bayes scores and mitochondrial evidence for all human and mouse proteins, BED files of gene coordinates, FASTA files of gene sequences and Excel files of the MS/MS peptides detected across 14 mouse tissues. Images supporting from previous GFP-tagging/microscopy experiments (4) are also available.

COMPARISON TO OTHER MITOCHONDRIAL DATABASES

Multiple research groups have created inventories of mammalian mitochondrial proteins. To our knowledge, these include MitoCarta1.0 (4), MitoP2 (20,34–37), MitoProteome (38,39) and IMPI http://impi.mrc-mbu.cam.ac.uk/), however MitoP2 and MitoProteome are no longer available on the internet. Similar to MitoCarta, IMPI (Integrated Mitochondrial Protein Index) uses machine learning to predict mitochondrial localization in human, mouse, rat and cow based on experimental proteomics data in MitoMiner (40,41), antibody staining from the Human Protein Atlas (42) and mitochondrial targeting sequence prediction tools. The IMPI version Q2 2015 contains 1480 human Ensembl genes with substantial overlap with MitoCarta2.0 (980 in both, 500 IMPI-specific and 178 MitoCarta2.0-specific). Compared to MitoCarta's naïve Bayes methodology, IMPI's machine learning methods (support vector machines and random forests) have the advantage of allowing redundant data sets that are not conditionally independent, however the resulting scores are less readily interpretable and the techniques are more susceptible to overfitting of the training data. Additionally, IMPI does not provide the atlas of protein expression across tissues. Several other mitochondrial-focused web resources (43) aggregate useful mitochondrial data but do not include a reference set of mitochondrial proteins, e.g. MitoMap provides human polymorphism and mutation data (44,45), HMPDb (bioinfo.nist.gov/hmpd) aggregates data from nine knowledge bases and MitoMiner aggregates extensive proteomics data with MitoCarta1.0, UniProt and IMPI (40,41).

There are also many general databases of sub-cellular localization that provide breadth across many species and a hierarchy of subcellular locations. NCBI GO cellular compartments (21) annotates mitochondrial localization based on literature reports (including MitoCarta1.0). It currently includes over 1500 human genes linked to the mitochondrion (of which 1050 are in MitoCarta2.0), however it contains hundreds of genes with annotations electronically inferred from distant species or from single, controversial reports in the literature and furthermore there is no confidence score of mitochondrial localization. Similarly, UniProt (27,28) includes over 1094 human genes linked to mitochondria (of which 76% are in MitoCarta2.0) but lacks confidence scores of localization. COMPARTMENTS (46) lacks a downloadable list of mitochondrial genes, but for any query gene it provides a confidence score of localization to multiple cellular compartments (e.g. mitochondrion, nucleus, ER, cytoplasm) based on aggregating data from knowledge bases (e.g. UniProt), prediction algorithms (e.g. PSORT, yLoc) and text mining.

Overall, MitoCarta2.0 and IMPI provide the most specific inventories of mammalian mitochondrial components to the community, while broader databases such as NCBI GO and UniProt provide more breadth across species and cellular compartments.

CONCLUSION

MitoCarta2.0 represents an easy-to-use inventory of mitochondrial proteins in mouse and human along with the evidence supporting mitochondrial localization for each protein—thereby providing a molecular framework for systematic studies of mitochondrial function and physiology. The MitoCarta database can be tuned to provide more or less stringent predictions of mitochondrial proteins by altering the FDR threshold. For example, when evaluating the results of a high-throughput screen of mitochondrial function, users may want to use a less stringent threshold such as 20% FDR. Similarly, when interpreting whole exome data from patients with mitochondrial disease, a 15% FDR might help interpret recessive mutations in genes with unknown function that may actually underlie mitochondrial dysfunction.

The current database has several important limitations. First, it is static and does not continually incorporate new literature evidence. Second, it is a mitochondrial-centric inventory that does not identify additional cellular localizations for proteins, or those that reside in the mitochondrion only under certain conditions. Third, it a gene-based inven-

tory rather than an isoform-based inventory, because the underlying training data were available only for gene loci. Fourth, the training data were skewed toward proteins that reside within the double-membrane, thus it will be less accurate at predicting proteins of the outer mitochondrial membrane. Additional experimental data sets will be needed to interrogate the outer membrane and additional tissues and conditions not covered in MitoCarta2.0.

Despite these limitations, the updated MitoCarta2.0 mitochondrial inventory provides a valuable research tool to investigate mitochondrial pathways in health and disease. We expect that in the coming years this inventory will help elucidate the function of many specific pathways as well as to interpret many high-throughput data sets in molecular biology and human genetics.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Alexis Jourdain for review of the manuscript and database and Steven Carr for providing resources for proteomic reanalysis.

FUNDING

National Institutes of Health [GM0077465 to V.K.M.]. V.K.M. is an investigator of the Howard Hughes Medical Institute. Funding for open access charge: National Institutes of Health [GM0077465 to V.K.M.].
Conflict of interest statement. None declared.

REFERENCES

1. Andersson,S.G., Zomorodipour,A., Andersson,J.O., Sicheritz-Ponten,T., Alsmark,U.C., Podowski,R.M., Naslund,A.K., Eriksson,A.S., Winkler,H.H. and Kurland,C.G. (1998) The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature*, **396**, 133–140.
2. Lopez,M.F., Kristal,B.S., Chernokalskaya,E., Lazarev,A., Shestopalov,A.I., Bogdanova,A. and Robinson,M. (2000) High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. *Electrophoresis*, **21**, 3427–3440.
3. Skladal,D., Halliday,J. and Thorburn,D.R. (2003) Minimum birth prevalence of mitochondrial respiratory chain disorders in children. *Brain*, **126**, 1905–1912.
4. Pagliarini,D.J., Calvo,S.E., Chang,B., Sheth,S.A., Vafai,S.B., Ong,S.-E., Walford,G.A., Sugiana,C., Boneh,A. and Chen,W.K. (2008) A mitochondrial protein compendium elucidates complex I disease biology. *Cell*, **134**, 112–123.
5. Calvo,S., Jain,M., Xie,X., Sheth,S.A., Chang,B., Goldberger,O.A., Spinazzola,A., Zeviani,M., Carr,S.A. and Mootha,V.K. (2006) Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat. Genet.*, **38**, 576–582.
6. Baughman,J.M., Perocchi,F., Girgis,H.S., Plovanich,M., Belcher-Timme,C.A., Sancak,Y., Bao,X.R., Strittmatter,L., Goldberger,O., Bogorad,R.L. *et al.* (2011) Integrative genomics identifies MCU as an essential component of the mitochondrial calcium uniporter. *Nature*, **476**, 341–345.
7. Nilsson,R., Schultz,I.J., Pierce,E.L., Soltis,K.A., Naranuntarat,A., Ward,D.M., Baughman,J.M., Paradkar,P.N., Kingsley,P.D., Culotta,V.C. *et al.* (2009) Discovery of genes essential for heme biosynthesis through large-scale gene expression analysis. *Cell Metab.*, **10**, 119–130.
8. Perocchi,F., Gohil,V.M., Girgis,H.S., Bao,X.R., McCombs,J.E., Palmer,A.E. and Mootha,V.K. (2010) MICU1 encodes a mitochondrial EF hand protein required for Ca(2+) uptake. *Nature*, **467**, 291–296.
9. Strittmatter,L., Li,Y., Nakatsuka,N.J., Calvo,S.E., Grabarek,Z. and Mootha,V.K. (2014) CLYBL is a polymorphic human enzyme with malate synthase and beta-methylmalate synthase activity. *Hum. Mol. Genet.*, **23**, 2313–2323.
10. Lanning,N.J., Looyenga,B.D., Kauffman,A.L., Niemi,N.M., Sudderth,J., DeBerardinis,R.J. and MacKeigan,J.P. (2014) A mitochondrial RNAi screen defines cellular bioenergetic determinants and identifies an adenylate kinase as a key regulator of ATP levels. *Cell Rep.*, **7**, 907–917.
11. 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**, 118.
12. Calvo,S.E., Tucker,E.J., Compton,A.G., Kirby,D.M., Crawford,G., Burtt,N.P., Rivas,M., Guiducci,C., Bruno,D.L., Goldberger,O.A. *et al.* (2010) High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. *Nat. Genet.*, **42**, 851–858.
13. Lieber,D.S., Calvo,S.E., Shanahan,K., Slate,N.G., Liu,S., Hershman,S.G., Gold,N.B., Chapman,B.A., Thorburn,D.R., Berry,G.T. *et al.* (2013) Targeted exome sequencing of suspected mitochondrial disorders. *Neurology*, **80**, 1762–1770.
14. Falk,M.J., Pierce,E.A., Consugar,M., Xie,M.H., Guadalupe,M., Hardy,O., Rappaport,E.F., Wallace,D.C., LeProust,E. and Gai,X. (2012) Mitochondrial disease genetic diagnostics: optimized whole-exome analysis for all MitoCarta nuclear genes and the mitochondrial genome. *Disc. Med.*, **14**, 389–399.
15. Calvo,S., Jain,M., Xie,X., Sheth,S.A., Chang,B., Goldberger,O.A., Spinazzola,A., Zeviani,M., Carr,S.A. and Mootha,V.K. (2006) Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat. Genet.*, **38**, 576–582.
16. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
17. Brown,G.R., Hem,V., Katz,K.S., Ovetsky,M., Wallin,C., Ermolaeva,O., Tolstoy,I., Tatusova,T., Pruitt,K.D., Maglott,D.R. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–D42.
18. Hung,V., Zou,P., Rhee,H.W., Udeshi,N.D., Cracan,V., Svinkina,T., Carr,S.A., Mootha,V.K. and Ting,A.Y. (2014) Proteomic mapping of the human mitochondrial intermembrane space in live cells via ratiometric APEX tagging. *Mol. Cell*, **55**, 332–341.
19. 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.
20. Elstner,M., Andreoli,C., Klopstock,T., Meitinger,T. and Prokisch,H. (2009) The mitochondrial proteome database: MitoP2. *Methods Enzymol.*, **457**, 3–20.
21. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
22. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
23. Finn,R.D., Clements,J., Arndt,W., Miller,B.L., Wheeler,T.J., Schreiber,F., Bateman,A. and Eddy,S.R. (2015) HMMER web server: 2015 update. *Nucleic Acids Res.*, **43**, W30–W38.
24. Costanzo,M.C., Engel,S.R., Wong,E.D., Lloyd,P., Karra,K., Chan,E.T., Weng,S., Paskov,K.M., Roe,G.R., Binkley,G. *et al.* (2014) Saccharomyces genome database provides new regulation data. *Nucleic Acids Res.*, **42**, D717–D725.
25. Lattin,J.E., Schroder,K., Su,A.I., Walker,J.R., Zhang,J., Wiltshire,T., Saijo,K., Glass,C.K., Hume,D.A., Kellie,S. *et al.* (2008) Expression analysis of G Protein-Coupled Receptors in mouse macrophages. *Immunome Res.*, **4**, 5.
26. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 6062–6067.

27. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
28. UniProt,C. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
29. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
30. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
31. Mootha,V.K., Handschin,C., Arlow,D., Xie,X., St Pierre,J., Sihag,S., Yang,W., Altshuler,D., Puigserver,P., Patterson,N. *et al.* (2004) Erralpha and Gabpa/b specify PGC-1alpha-dependent oxidative phosphorylation gene expression that is altered in diabetic muscle. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 6570–6575.
32. Pages,H, C,M., Falcon,S. and Li,N. AnnotationDbi: Annotation Database Interface. R package version 1.28.2.
33. Gharaibeh,R.Z., Fodor,A.A. and Gibas,C.J. (2008) Background correction using dinucleotide affinities improves the performance of GCRMA. *BMC Bioinformatics*, **9**, 452.
34. 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.
35. Elstner,M., Andreoli,C., Ahting,U., Tetko,I., Klopstock,T., Meitinger,T. and Prokisch,H. (2008) MitoP2: an integrative tool for the analysis of the mitochondrial proteome. *Mol. Biotechnol.*, **40**, 306–315.
36. Prokisch,H. and Ahting,U. (2007) MitoP2, an integrated database for mitochondrial proteins. *Methods Mol. Biol.*, **372**, 573–586.
37. Andreoli,C., Prokisch,H., Hortnagel,K., Mueller,J.C., Munsterkotter,M., Scharfe,C. and Meitinger,T. (2004) MitoP2, an integrated database on mitochondrial proteins in yeast and man. *Nucleic Acids Res.*, **32**, D459–D462.
38. Guda,P., Subramaniam,S. and Guda,C. (2007) Mitoproteome: human heart mitochondrial protein sequence database. *Methods Mol. Biol.*, **357**, 375–383.
39. Cotter,D., Guda,P., Fahy,E. and Subramaniam,S. (2004) MitoProteome: mitochondrial protein sequence database and annotation system. *Nucleic Acids Res.*, **32**, D463–D467.
40. 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.
41. Smith,A.C., Blackshaw,J.A. and Robinson,A.J. (2012) MitoMiner: a data warehouse for mitochondrial proteomics data. *Nucleic Acids Res.*, **40**, D1160–D1167.
42. 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**, 6220.
43. Calvo,S.E. and Mootha,V.K. (2010) The mitochondrial proteome and human disease. *Annu. Rev. Genom. Hum. Genet.*, **11**, 25–44.
44. Kogelnik,A.M., Lott,M.T., Brown,M.D., Navathe,S.B. and Wallace,D.C. (1996) MITOMAP: a human mitochondrial genome database. *Nucleic Acids Res.*, **24**, 177–179.
45. Lott,M.T., Leipzig,J.N., Derbeneva,O., Xie,H.M., Chalkia,D., Sarmady,M., Procaccio,V. and Wallace,D.C. (2013) mtDNA Variation and Analysis Using MITOMAP and MITOMASTER. *Curr. Protoc. Bioinform.*, **1**, 1.23.1–1.23.26.
46. Binder,J.X., Pletscher-Frankild,S., Tsafou,K., Stolte,C., O'Donoghue,S.I., Schneider,R. and Jensen,L.J. (2014) COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database: J. Biol. Databases Curation*, bau012.