Published online 6 November 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D1225–D1228***
doi: 10.1093/nar/gky1072

MitoMiner v4.0: an updated database of mitochondrial localization evidence, phenotypes and diseases

Anthony C. Smith and Alan J. Robinson*

MRC Mitochondrial Biology Unit, Wellcome Trust/MRC Building, Cambridge Biomedical Campus, Hills Road, Cambridge CB2 0XY, UK

Received September 13, 2018; Editorial Decision October 12, 2018; Accepted October 19, 2018

ABSTRACT

Increasing numbers of diseases are associated with mitochondrial dysfunction. This is unsurprising given mitochondria have major roles in bioenergy generation, signalling, detoxification, apoptosis and biosynthesis. However, fundamental questions of mitochondrial biology remain, including: which nuclear genes encode mitochondrial proteins; how their expression varies with tissue; and which are associated with disease. But experiments to catalogue the mitochondrial proteome are incomplete and sometimes contradictory. This arises because the mitochondrial proteome has tissue- and stage-specific variability, plus differences among experimental techniques and localization evidence types used. This leads to limitations in each technique's coverage and inevitably conflicting results. To support identification of mitochondrial proteins, we developed MitoMiner (http://mitominer.mrc-mbu.cam.ac.uk/), a database combining evidence of mitochondrial localization with information from public resources. Here we report upgrades to MitoMiner, including its re-engineering to be gene-centric to enable easier sharing of evidence among orthologues and support next generation sequencing, plus new data sources, including expression in different tissues, information on phenotypes and diseases of genetic mutations and a new mitochondrial proteome catalogue. MitoMiner is a powerful platform to investigate mitochondrial localization by providing a unique combination of experimental sub-cellular localization datasets, tissue expression, predictions of mitochondrial targeting sequences, gene annotation and links to phenotype and disease.

INTRODUCTION

Mitochondria are associated with a wide spectrum of metabolic, degenerative and age-related human diseases as well as cancer. This has generated considerable interest in mitochondria from a wide range of researchers. However, the mitochondrial proteome has yet to be conclusively identified—as evidenced by the continued discovery of new mitochondrial disease genes—which hinders investigations into the organelle's role. Many different approaches have been used to address this problem, but each has limitations and no single technique provides full coverage of the mitochondrial proteome. Furthermore, this problem is exacerbated by how the mitochondrial proteome varies with cell type, environment and development.

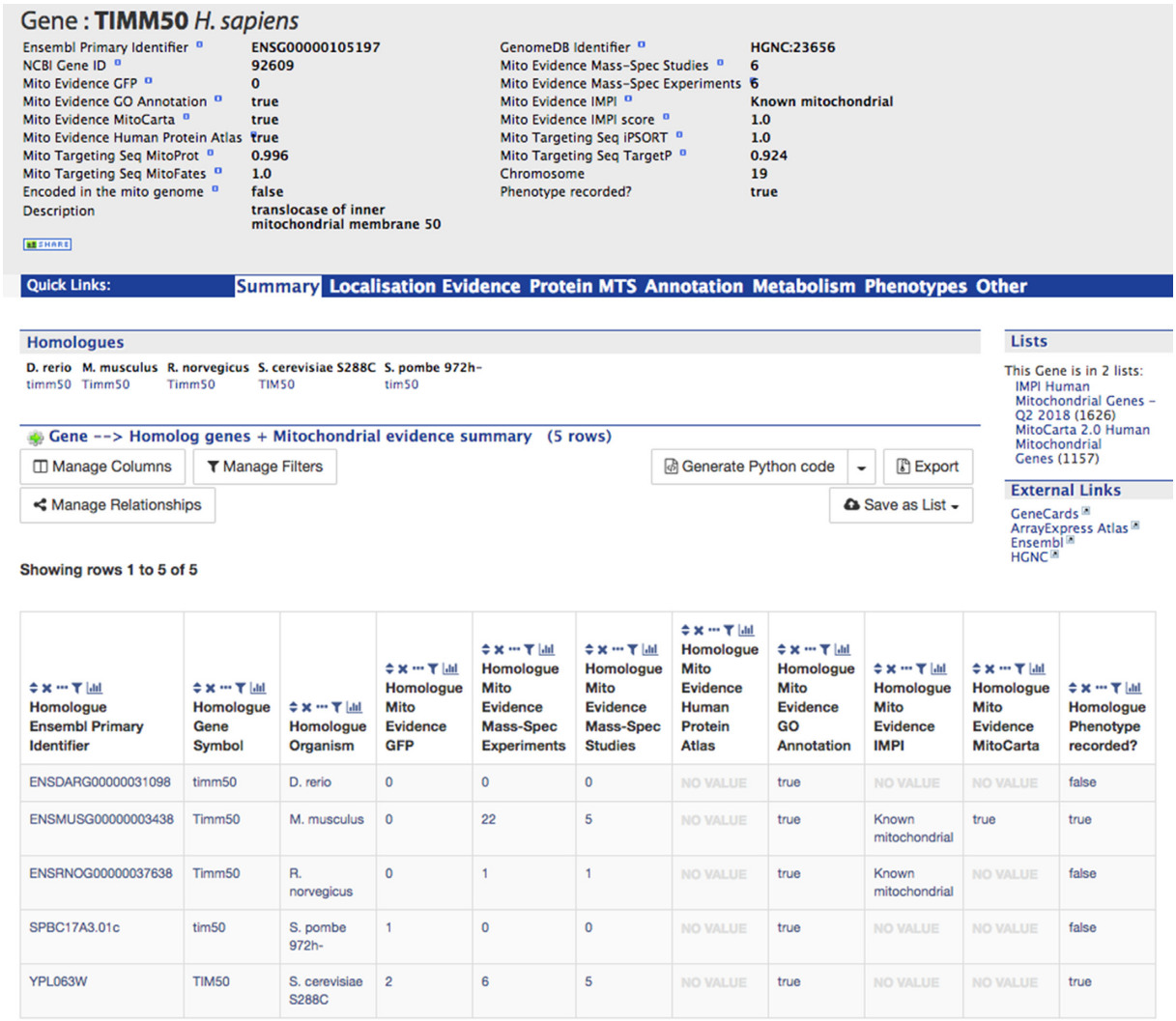
To determine if proteins are mitochondrial, we believe it is necessary to cross-reference between many different evidence types. This has the additional benefit of providing data to apply machine-learning techniques to the problem of identifying novel mitochondrial proteins. We described previously how MitoMiner (1–3) incorporates mitochondrial localization data from a variety of resources—including mass spectrometry of purified cell fractions, GFP-tagging and microscopy, curated annotation in the Gene Ontology (4,5) and computational prediction of mitochondrial targeting sequences. Further, MitoMiner provides a biological context for candidate mitochondrial proteins by integrating information from other public resources, including the Gene Ontology, UniProt (6), OMIM (7) and the Human Protein Atlas (HPA; http://www.proteinatlas.org/) (8,9).

Here we describe a major update to the MitoMiner database including: re-factoring MitoMiner from a protein-centric to gene-centric resource (substantially aiding cross-referencing data and application in analysis of disease gene candidates); new analysis tools and widgets for RNA expression data; new data sources such as a new mitochondrial proteome dataset derived by using machine learning on the data in MitoMiner; and new data types, including human RNA expression data and phenotype descriptions of diseases and genetic mutations.

SOFTWARE IMPLEMENTATION AND DATA IMPORT

MitoMiner continues to be built on the InterMine data-warehouse (10,11), which provides the infrastructure to keep MitoMiner up-to-date, as well as a powerful and

*To whom correspondence should be addressed. Tel: +44 01223 252860; Fax: +44 01223 252715; Email: ajr@mrc-mbu.cam.ac.uk



Gene : **TIMM50** *H. sapiens*

Ensembl Primary Identifier	ENSG00000105197	GenomeDB Identifier	HGNC:23656
NCBI Gene ID	92609	Mito Evidence Mass-Spec Studies	6
Mito Evidence GFP	0	Mito Evidence Mass-Spec Experiments	6
Mito Evidence GO Annotation	true	Mito Evidence IMPI	Known mitochondrial
Mito Evidence MitoCarta	true	Mito Evidence IMPI score	1.0
Mito Evidence Human Protein Atlas	true	Mito Targeting Seq iPSORT	1.0
Mito Targeting Seq MitoProt	0.996	Mito Targeting Seq TargetP	0.924
Mito Targeting Seq MitoFates	1.0	Chromosome	19
Encoded in the mito genome	false	Phenotype recorded?	true
Description	translocase of inner mitochondrial membrane 50		

Quick Links: Summary Localisation Evidence Protein MTS Annotation Metabolism Phenotypes Other

Homologues

D. rerio	M. musculus	R. norvegicus	S. cerevisiae S288C	S. pombe 972h-
timm50	Timm50	Timm50	TIM50	tim50

Gene --> Homolog genes + Mitochondrial evidence summary (5 rows)

Manage Columns | Manage Filters | Generate Python code | Export | Manage Relationships | Save as List

Showing rows 1 to 5 of 5

Homologue Ensembl Primary Identifier	Homologue Gene Symbol	Homologue Organism	Homologue Mito Evidence GFP	Homologue Mito Evidence Mass-Spec Experiments	Homologue Mito Evidence Mass-Spec Studies	Homologue Mito Evidence Human Protein Atlas	Homologue Mito Evidence GO Annotation	Homologue Mito Evidence IMPI	Homologue Mito Evidence MitoCarta	Homologue Phenotype recorded?
ENSDARG00000031098	timm50	D. rerio	0	0	0	NO VALUE	true	NO VALUE	NO VALUE	false
ENSMUSG00000003438	Timm50	M. musculus	0	22	5	NO VALUE	true	Known mitochondrial	true	true
ENSRNOG00000037638	Timm50	R. norvegicus	0	1	1	NO VALUE	true	Known mitochondrial	NO VALUE	false
SPBC17A3.01c	tim50	S. pombe 972h-	1	0	0	NO VALUE	true	NO VALUE	NO VALUE	false
YPL063W	TIM50	S. cerevisiae S288C	2	6	5	NO VALUE	true	NO VALUE	NO VALUE	true

Lists

This Gene is in 2 lists:
IMPI Human Mitochondrial Genes – Q2 2018 (1626)
MitoCarta 2.0 Human Mitochondrial Genes (1157)

External Links

GeneCards
ArrayExpress Atlas
Ensembl
HGNC

Figure 1. Header of the MitoMiner gene entry for the human TIMM50 gene. MitoMiner's new gene-centric organization presents first an overview of primary mitochondrial localization evidence, then the gene's homologues in other species and a table of their mitochondrial localization evidence, including experimental data, annotation and mutant phenotype. Links to this gene's presence in catalogues of mitochondrial proteomes is provided, as well as links to external resources, e.g. OMIM and Ensembl. The remainder of the gene entry page (not shown) describes other information, including localization evidence, tissue expression data and links to metabolism.

flexible user interface, enabling options from simple text searches to complicated queries with multiple constraints spanning any of the included data types, as well as application programming interfaces for programmatic access with Perl, Python, Ruby and Java.

MitoMiner's database schema is based on the InterMine core schema that specifies different types of biological data. To model MitoMiner-specific data types—e.g. mass spectrometry and GFP tagging datasets, metabolic pathway data and homology mappings—the InterMine core model has been extended with bespoke additions. Data were imported by using either InterMine-provided data loaders, or Perl scripts to convert data files to InterMine compatible XML data files. The MitoMiner data sources and InterMine infrastructure are updated on a 9–12-month cycle.

UPDATES TO DATA SOURCES

MitoMiner was conceived as a proteomics resource and annotation attached to protein entries. Thus, in earlier MitoMiner versions, a single gene could be represented by multiple protein entries—each with different data—that needed separate, manual evaluation. However, changing usage patterns—especially with analysis of disease gene candidates from next generation sequencing of mitochondrial disease patients—prompted a major re-factoring of MitoMiner 4.0 to be gene-centric, rather than protein-centric. Thus a new gene entry was created (Figure 1) with data from its protein entries attached directly to it, which can be searched by Ensembl gene identifier (12), Human Gene Organization (HUGO) gene symbol, genome project gene identifier (e.g. from Mouse Genome Informatics (MGI) (13)

and Rat Genome Database (RGD) (14)) and NCBI gene identifier (15). By using this gene entry, protein data are merged and consolidated at the gene level, redundancy has been removed and it is easier to share data amongst homologues of the same gene in different species, as exemplified by the summary table and cross-references (Figure 1). If the proteins of a gene have different values for a property, then the value with the greatest level of evidence is added to the gene entry, e.g. the best mitochondrial targeting sequence prediction.

Next are provided descriptions of data sources new to MitoMiner since version 3.1 (3). MitoMiner retains data sources described in earlier releases (1–3), e.g. metabolic pathways, but these are not described again here.

Mitochondrial protein localization

MitoMiner's focus continues to be collating evidence for the mitochondrial localization of proteins from large-scale studies. New mitochondrial localization datasets have been added—seven from human; one from mouse; one from rat and two from yeast—bringing the total to 56 datasets. The full list of datasets with references is available from the Data Sources page of the MitoMiner website.

New additions to the protein annotation of mitochondrial localization include: mitochondrial targeting sequence predictions from MitoFates (16); cleavage site predictions from MitoFates and MitoProt (17); and confidence notation of prediction reliability from TargetP (18).

Phenotype and disease

A disease gene candidate is more credible if mutation of the homologous gene in a model organism causes a similar phenotype to the human disease. Thus, annotations and links of genes to phenotype and disease have been added to MitoMiner. This includes: human disease (from OMIM (7)) and mitochondrial disease (from the Washington University list of mitochondrial disorders (https://neuromuscular.wustl.edu/mitosyn.html)); phenotypes of spontaneous, induced and engineered mutations in mouse (from MGI); and knock-down and genetic mutant information in zebrafish (from ZFIN (19)) and yeast (from SGD (20)). This information is cross-referenced between homologous genes and available via the 'Summary' section of a gene page (Figure 2).

Mitochondrial proteomes

To decide whether a protein is mitochondrial based upon available data, various definitions of the mitochondrial proteome have been made by evaluating literature, experimental studies and computational predictions. MitoMiner provides two mitochondrial proteome datasets: the well-respected MitoCarta2 (21) and a new Integrated Mitochondrial Protein Index (IMPI), based upon applying machine learning to mitochondrial localization data in MitoMiner (http://impi.mrc-mbu.cam.ac.uk/). Each proteome has its own page (linked from the menu bar) and is provided with a set of template queries to analyse them. Users can evaluate the evidence for a protein being included in a mitochondrial proteome dataset by surveying its gene entry in MitoMiner.

6 Mouse Phenotype

Name	Description
MP:0010025	decreased total body fat amount
MP:0003960	increased lean body mass
MP:0010124	decreased bone mineral content
MP:0002833	increased heart weight
MP:0011100	preweaning lethality, complete penetrance
MP:0003731	abnormal retinal outer nuclear layer morphology

Figure 2. Mutant phenotype data for the mouse TIMM50 gene recorded in the phenotype section of its gene entry. Identifier names and descriptions are from the Mouse Genome Informatics (MGD) database. Cross-references between homologous genes means the existence of these mouse data are recorded in the 'Homologue Phenotype Recorded?' column in the summary table of the human TIMM50 gene entry (Figure 1).

Tissue RNA expression

Genes of mitochondrial proteins tend to be highly expressed in tissues having substantial energy or metabolic demands. For example, when evaluating candidate mitochondrial disease genes it is useful to relate the organ systems affected in the patient with the tissue-specific expression of the candidate gene. Thus, a candidate disease gene identified from a patient suffering from 'mitochondrial encephalopathy, lactic acidosis and stroke-like episodes' (MELAS) is unlikely to be disease-causing if its expression is undetected in brain. To support such queries, MitoMiner now includes RNA tissue expression data in addition to protein level detection from the Human Protein Atlas (9,22), and new display widgets for HPA's tissue and cancer RNA expression were created and are available on gene entry pages.

Tutorials and user guides

The MitoMiner website is accompanied with a full set of support pages including FAQ's, user guides, examples and tutorials (http://mitominer.mrc-mbu.cam.ac.uk/support/). Nine tutorials document new and existing features of MitoMiner, including uploading and analysing data, building bespoke queries by using the Query Builderweb site and storing private lists and queries by using MyMine.

DATA AVAILABILITY

MitoMiner is freely available at the Medical Research Council Mitochondrial Biology Unit website (http://mitominer.mrc-mbu.cam.ac.uk/).

ACKNOWLEDGEMENTS

We would like to thank Julie Sullivan and the rest of the InterMine team for their continued assistance and support of the underlying data-warehouse software.

FUNDING

Medical Research Council, UK [MC_U105674181]. Funding for open access charge: Medical Research Council UK [RG89175].

REFERENCES

1. 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.
2. 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.
3. Smith,A.C. and Robinson,A.J. (2016) MitoMiner v3.1, an update on the mitochondrial proteomics database. *Nucleic Acids Res.*, **44**, D1258–D1261.
4. 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.
5. The Gene Ontology Consortium (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.*, **45**, D331–D338.
6. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
7. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
8. 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.
9. 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.
10. Smith,R.N., Aleksic,J., Butano,D., Carr,A., Contrino,S., Hu,F., Lyne,M., Lyne,R., Kalderimis,A., Rutherford,K. *et al.* (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, **28**, 3163–3165.
11. Kalderimis,A., Lyne,R., Butano,D., Contrino,S., Lyne,M., Heimbach,J., Hu,F., Smith,R., Stepan,R., Sullivan,J. *et al.* (2014) InterMine: extensive web services for modern biology. *Nucleic Acids Res.*, **42**, W468–W472.
12. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Giron,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
13. Smith,C.L., Blake,J.A., Kadin,J.A., Richardson,J.E., Bult,C.J. and Mouse,Genome Database Group. (2018) Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. *Nucleic Acids Res.*, **46**, D836–D842.
14. Shimoyama,M., De Pons,J., Hayman,G.T., Laulederkind,S.J., Liu,W., Nigam,R., Petri,V., Smith,J.R., Tutaj,M., Wang,S.J. *et al.* (2015) The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res.*, **43**, D743–D750.
15. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
16. Fukasawa,Y., Tsuji,J., Fu,S.C., Tomii,K., Horton,P. and Imai,K. (2015) MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. *Mol. Cell. Proteomics*, **14**, 1113–1126.
17. Claros,M.G. and Vincens,P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.*, **241**, 779–786.
18. Emanuelsson,O., Brunak,S., von Heijne,G. and Nielsen,H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.*, **2**, 953–971.
19. Ruzicka,L., Bradford,Y.M., Frazer,K., Howe,D.G., Paddock,H., Ramachandran,S., Singer,A., Toro,S., Van Slyke,C.E., Eagle,A.E. *et al.* (2015) ZFIN, The zebrafish model organism database: Updates and new directions. *Genesis*, **53**, 498–509.
20. Lang,O.W., Nash,R.S., Hellerstedt,S.T., Engel,S.R. and The,S.G.D. Project (2018) An Introduction to the Saccharomyces Genome Database (SGD). *Methods Mol. Biol.*, **1757**, 21–30.
21. 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.
22. Uhlen,M., Zhang,C., Lee,S., Sjostedt,E., Fagerberg,L., Bidkhori,G., Benfeitas,R., Arif,M., Liu,Z., Edfors,F. *et al.* (2017) A pathology atlas of the human cancer transcriptome. *Science*, **357**, eaan2507.