doi:10.1093/nar/gkm950

LOCATE: a mammalian protein subcellular localization database

Josefine Sprenger, J. Lynn Fink, Seetha Karunaratne, Kelly Hanson, Nicholas A. Hamilton and Rohan D. Teasdale*

ARC Centre of Excellence in Bioinformatics, Institute for Molecular Bioscience, The University of Queensland, St Lucia, Queensland 4072, Australia

Received September 11, 2007; Revised October 15, 2007; Accepted October 16, 2007

ABSTRACT

LOCATE is a curated, web-accessible database that houses data describing the membrane organization and subcellular localization of mouse and human proteins. Over the past 2 years, the data in LOCATE have grown substantially. The database now contains high-quality localization data for 20% of the mouse proteome and general localization annotation for nearly 36% of the mouse proteome. The proteome annotated in LOCATE is from the RIKEN FANTOM Consortium Isoform Protein Sequence sets which contains 58 128 mouse and 64 637 human protein isoforms. Other additions include computational subcellular localization predictions, automated computational classification of experimental localization image data, prediction of protein sorting signals and third party submission of literature data. Collectively, this database provides localization proteome for individual subcellular compartments that will underpin future systematic investigations of these regions. It is available at http://locate.imb.uq.edu.au/

INTRODUCTION

A cell is divided into different cellular compartments and each compartment is associated with a different range of biochemical processes; by localizing a protein to a specific compartment, or set of compartments, the cellular role of the protein can be inferred. Also critical is determining the membrane organization of individual proteins namely their topology relative to the membrane or if they are embedded in the lipid bilayer. Without this knowledge the function a protein has within the cell cannot be fully elucidated. This information provides insight into understanding hypothetical or novel proteins and can provide a more specific organellar context in which to investigate a particular protein. Historically, these data have been difficult to produce on a large scale for higher eukaryotic organisms. However, recent advances in membrane organization prediction methods and high-throughput subcellular localization assays have made it possible to generate these datasets. We used high-throughput methods to predict the membrane organization for the entire proteome and to determine the subcellular localization of a subset of the proteome. We then developed a database, LOCATE, to organize and warehouse these data.

GROWTH OF DATABASE CONTENT

The original mouse LOCATE database (1) has been updated and extended to include a human proteome. The original database content is described in detail (1) and updated features are outlined below.

Dataset

The mouse and human proteome FANTOM3 Isoform Protein Sequence set (IPS8) were generated by the RIKEN FANTOM Consortium (2). This dataset is comprised of protein sequences based on transcript sequences generated from direct sequencing of full-length transcripts. The sequenced transcripts were clustered into transcriptional units (TU) where a TU is a grouping of transcripts that arise from a single genomic locus. The mouse proteome contains 58 128 unique protein isoforms encoded by 29 682 TUs, while the human proteome contains 64 637 unique protein isoforms encoded by 26 583 TUs.

Membrane organization

Protein orientation with respect to the membrane was predicted by MemO, a high-throughput, automated pipeline, which combines publicly available feature predictors with empirically determined annotation rules (3).

*To whom correspondence should be addressed. Tel: +61 7 3346 2056; Fax: +61 7 3346 2101; Email: r.teasdale@imb.uq.edu.au

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Table 1. Distribution of membrane organization classes and high quality localization data in LOCATE

	MemO data	Subcellular localization data		
Membrane organization class	IPS proteins in class (TUs/Isoforms)	Isoforms with experimental data (mouse only)	TUs with literature-mined data	Total represented (TUs/Isoforms)
Soluble, intracellular protein	M: 20487/39809 H: 20061/45611	1566	M: 1948 H: 1250	M: 6492/14448 H: 6169/14094
Soluble, secreted protein	M: 2882/4231 H: 2487/4418	11	M: 464 H: 290	M: 850/1562 H: 983/1866
Type I membrane protein	M: 1308/2199 H: 1287/2531	16	M: 442 H: 287	M: 437/1112 H: 538/1272
Type II membrane protein	M: 3132/4526 H: 3126/5040	242	M: 568 H: 350	M: 830/1766 H: 689/1630
Multi-pass membrane protein	M: 4998/7363 H: 3595/7037	233	M: 583 H: 378	M: 994/1878 H: 1220/2487
Total proteins analyzed	M: 29682/58128 H: 26583/64637	2068	M: 4005 H: 1963	M: 9603/20766 H: 9599/21349

The MemO data columns show the absolute numbers of proteins classified by MemO into each membrane organization class. The subcellular localization data columns show the number of protein isoforms that have an experimentally determined subcellular localization and the number of transcriptional units (TUs) that have a literature-mined subcellular localization as well as the total numbers of TUs and isoforms that have any subcellular localization data. Individual TU may contain protein isoforms from more than one membrane organization class (4).

This allowed us to categorize proteins into five membrane organization classes based on the presence or absence of a transmembrane domain and the presence or absence of a signal peptide (Table 1). Previously we have documented that an individual TU may contain protein isoforms representing more than one membrane organization class (4). The percentage of TU with variable membrane organization within these mouse and human proteome are 9.3 and 12.6%, respectively.

Subcellular localization

Proteins with an N-terminal myc tag were expressed in HeLa cells and their subcellular localization was detected by indirect immunofluorescence (5). Representative images were collected and analyzed to determine the protein's subcellular localization. The annotations were reviewed using automatic image classification techniques (6). To date within the mouse proteome, experimental subcellular localization data originating from our group have been generated for 2068 protein isoforms representing a five-fold increase since the initial report. In addition, we have continued to generate independent subcellular localization annotations based on primary literature review (1) for 9245 proteins (3232 TUs) that represents a 1.9-fold increase. While we consider these sources of annotations to be of a high quality they are not yet comprehensive. To provide a localization description as complete as possible for any given protein, we also therefore include localization data mined from other online databases including LIFEdb (7), Mouse Genome Informatics (8), UniProt (9), ENSEMBL (10), and others. For mouse, 14 659 protein isoforms (7506 TUs) are annotated with subcellular localisation data from these sources.

In addition, we have included subcellular localization predictions for the mouse proteome from five prediction programs as reported in Sprenger *et al.* (11). These predictors were selected because they can be easily applied to proteome-scale datasets and they predict localization to at least nine major subcellular locations. Although we do not place high confidence in these predictions, we believe they are worth reporting to enable individuals to consider them in combination with other localization data.

In total, we have high-quality localization data for 4786 mouse TUs and 10 883 mouse protein isoforms representing 16 and 19% of the IPS8 set, respectively. Including the data of unknown quality retrieved from external sources, we report localization data for 9603 TUs and 20 766 isoforms representing nearly 36% of the mouse proteome. Table 1 shows a breakdown of the new data by membrane organization class, source, and quality.

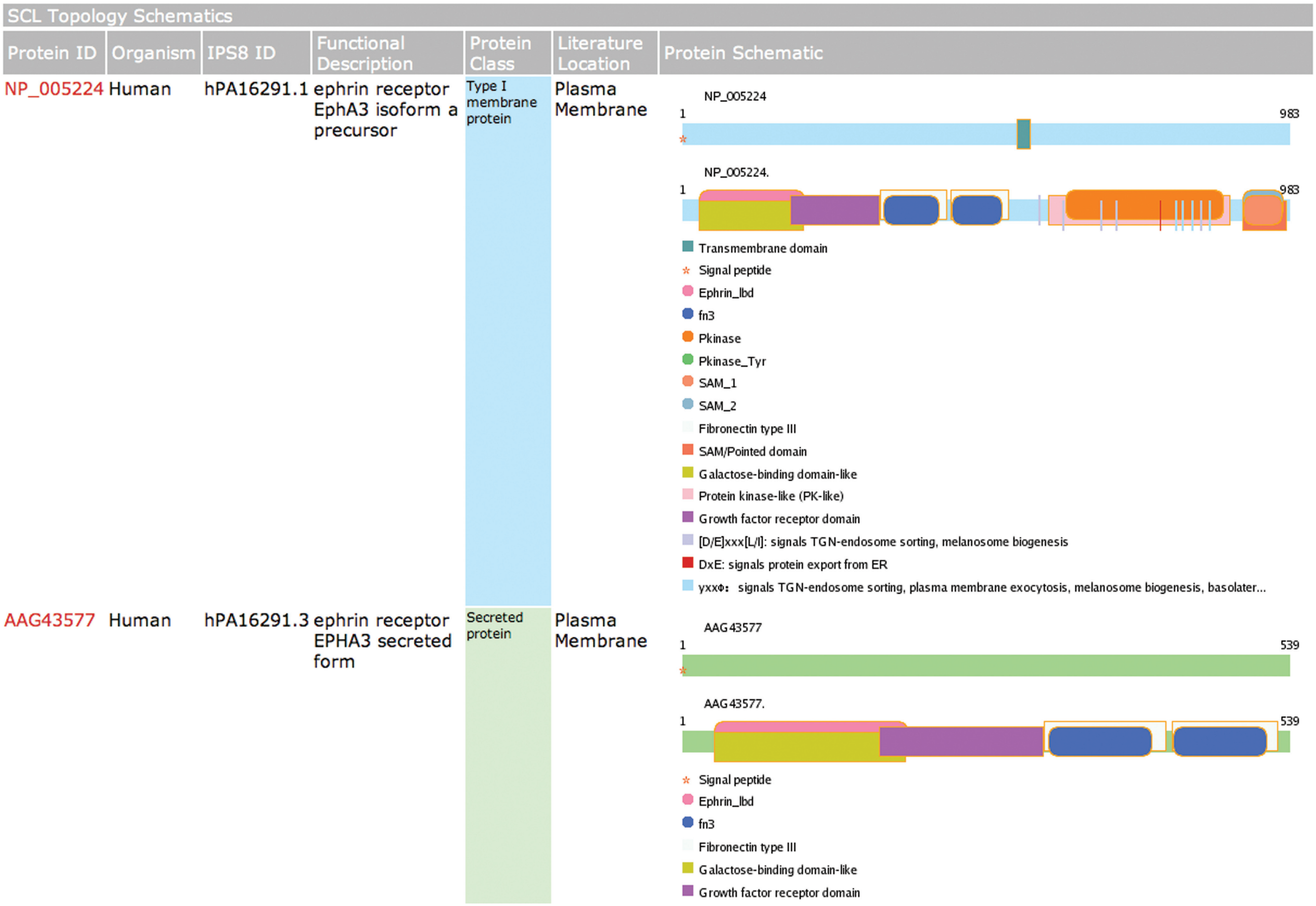
To enable the broader community to contribute information to LOCATE we have developed a submission process to accept subcellular localization annotations based on the published literature from third parties.

IMPROVED DATA PRESENTATION

In order to improve the presentation of the different types of data we have made a number of changes and additions to the existing web pages.

Subcellular localization data

We provide data describing the observed or predicted subcellular localization of a protein from four sources: original experimental data, data mined from the primary literature, data from external databases and data from computational subcellular localization predictors. These localizations are all summarized at the top of the page describing an individual protein so that the data from each of the sources can be compared. We chose not to include predictions from localization predictors in the summary but the top hits for each of the five predictors we used are listed elsewhere on the page along with a link to the detailed output for each predictor.



SCL Topology Schematics

Protein ID	Organism	IPS8 ID	Functional Description	Protein Class	Literature Location	Protein Schematic
NP_005224	Human	hPA16291.1	ephrin receptor EphA3 isoform a precursor	Type I membrane protein	Plasma Membrane	
AAG43577	Human	hPA16291.3	ephrin receptor EPHA3 secreted form	Secreted protein	Plasma Membrane	

NP_005224
1
983
NP_005224.
1
983
Transmembrane domain
Signal peptide
Ephrin_lbd
fn3
Pkinase
Pkinase_Tyr
SAM_1
SAM_2
Fibronectin type III
SAM/Pointed domain
Galactose-binding domain-like
Protein kinase-like (PK-like)
Growth factor receptor domain
[D/E]xxx[L/I]: signals TGN-endosome sorting, melanosome biogenesis
DxE: signals protein export from ER
yxxΦ: signals TGN-endosome sorting, plasma membrane exocytosis, melanosome biogenesis, basolater...
AAG43577
1
539
AAG43577.
1
539
Signal peptide
Ephrin_lbd
fn3
Fibronectin type III
Galactose-binding domain-like
Growth factor receptor domain

Figure 1. Transmembrane topology and predicted motifs and domains display.

The existence of localization data from each source is also annotated on the results of a BLAST search when a search is performed on the LOCATE database itself. This gives the viewer an overview of the extent of annotation of each isoform and each transcriptional unit.

Transmembrane topology and predicted motifs and domains

The membrane organization of a protein is displayed relative to the other protein domains, using the DomainDraw macromolecular feature drawing program (12). These protein schematic diagrams include Pfam (v21.0) and SCOP (v1.69) predicted domains and subcellular sorting signals based on experimentally defined motifs (Figure 1). The complement of proteins with the individual protein features can be visualized (http://locate.imb.uq.edu.au/list_motifs.shtml).

The topology of a membrane-spanning protein is of interest, especially for the proteins with multiple transmembrane domains (TMDs). We provide the membrane topology as predicted by MemO based on predicted signal peptides and TMDs. However, three of the five TMD predictors generate their own topology prediction without being informed by a signal peptide predictor. We display these topology predictions in addition to the MemO consensus topology.

LOCATION PROTEOMICS—DEFINING A SUBCELLULAR COMPARTMENTS PROTEIN COMPLEMENT

One of the key objectives of this database is to provide the protein content of a particular region of the cell, termed Location Proteome (13). Figure 2 shows the location proteomes of the major cellular compartments. We have compared the data collected from *other* sources with our independently annotated *primary* literature subcellular localization data from LOCATE. The cytoplasm (29.3% *other*; 6.9% *primary*) has been excluded as it contained limited representation in our annotations and proteins remaining at their site of biosynthesis do not represent an active transport event. Within these estimates each TU contributes equally and when multiple subcellular compartments were annotated each annotation was proportionally distributed. The differences between the two subcellular localization datasets have been discussed previously (11). Our *primary* localization annotations are based exclusively on experimental data and aim to represent the predominant subcellular localization. It does not well represent proteins that have multiple cellular localizations in the same cell or across distinct cell types and those induced into trafficking pathways by activation of cellular pathways. In contrast, the *other* subcellular localization dataset captures any subcellular localization without considering the relative distributions across

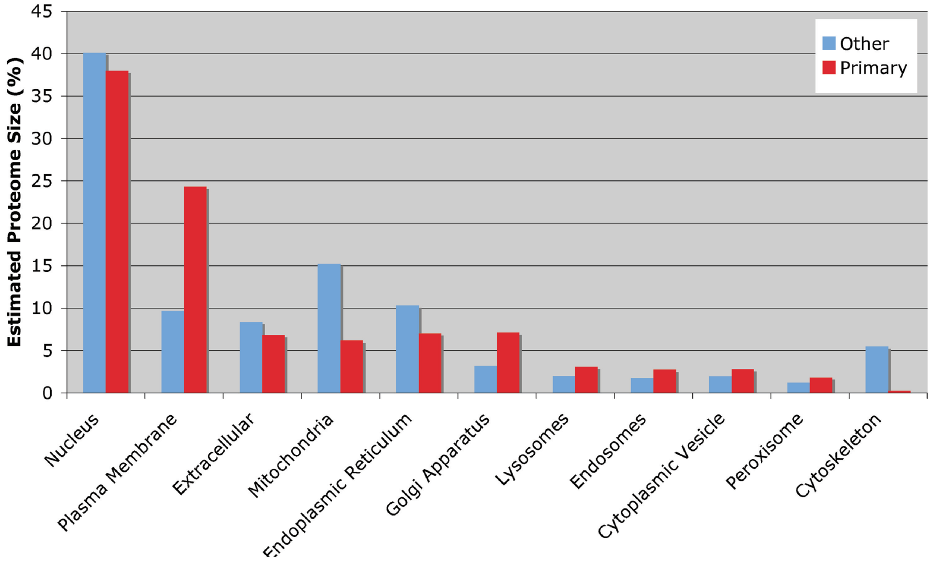


Figure 2. Organelle proteomics—defining the protein complement of individual organelles.

multiple localization or the source of the annotation. Within the *primary* data set the largest compartment proteomes are the nuclear proteome with 38% of the proteins and the extracellular/plasma membrane proteome with 31% of the proteins. The other intracellular organelles proteomes are of a similar size mitochondria proteome 6.2%; endoplasmic reticulum proteome 7.0%; Golgi Apparatus proteome 7.1% and endosome/lysosome 5.8%. Within the *other* subcellular localisation data the mitochondria proteome, endoplasmic reticulum proteome and cytoskeleton proteome have higher estimates. The list of proteins within each region is accessible from the LOCATE homepage.

AVAILABILITY

LOCATE data can be retrieved as individual entries or downloaded as HTML, plain text, or XML files from http://locate.imb.uq.edu.au

ACKNOWLEDGEMENTS

The authors would like to acknowledge Fasheng Zhang for technical support and Emma Redhead for designing the LOCATE XML schema and XML document generator. The work was supported by funds from the Australian Research Council (ARC) and RDT is supported by a National Health and Medical Research Council of Australia R. Douglas Wright Career Development Award. Funding to pay the Open Access publication charges for this article was provided by The University of Queensland.

Conflict of interest statement. None declared.

REFERENCES

1. Fink,J.L., Aturaliya,R.N., Davis,M.J., Zhang,F., Hanson,K., Teasdale,M.S., Kai,C., Kawai,J., Carninci,P. *et al.* (2006) LOCATE: a mouse protein subcellular localization database. *Nucleic Acids Res.*, **34**, D213–D217.
2. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
3. Davis,M.J., Zhang,F., Yuan,Z. and Teasdale,R.D. (2006) MemO: a consensus approach to the annotation of a protein's membrane organization. *In Silico Biol.*, **6**, 387–399.
4. Davis,M.J., Hanson,K.A., Clark,F., Fink,J.L., Zhang,F., Kasukawa,T., Kai,C., Kawai,J., Carninci,P. *et al.* (2006) Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units. *PLoS Genet.*, **2**, e46.
5. Aturaliya,R.N., Fink,J.L., Davis,M.J., Teasdale,M.S., Hanson,K.A., Miranda,K.C., Forrest,A.R., Grimmond,S.M., Suzuki,H. *et al.* (2006) Subcellular localization of mammalian type II membrane proteins. *Traffic*, **7**, 613–625.
6. Hamilton,N.A., Pantelic,R.S., Hanson,K. and Teasdale,R.D. (2007) Fast automated cell phenotype image classification. *BMC Bioinformatics*, **8**, 110.
7. Bannasch,D., Mehrle,A., Glatting,K.H., Pepperkok,R., Poustka,A. and Wiemann,S. (2004) LIFEdb: a database for functional genomics experiments integrating information from external sources, and serving as a sample tracking system. *Nucleic Acids Res.*, **32**, D505–D508.
8. Eppig,J.T., Bult,C.J., Kadin,J.A., Richardson,J.E., Blake,J.A., Anagnostopoulos,A., Baldarelli,R.M., Baya,M., Beal,J.S. *et al.* (2005) The Mouse Genome Database (MGD): from genes to mice–a community resource for mouse biology. *Nucleic Acids Res.*, **33**, D471–D475.
9. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
10. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
11. Sprenger,J., Fink,J.L. and Teasdale,R.D. (2006) Evaluation and comparison of mammalian subcellular localization prediction methods. *BMC Bioinformatics*, **7**(Suppl. 5), S3.
12. Fink,J.L. and Hamilton,N. (2007) DomainDraw: A macromolecular feature drawing program. *In Silico Biol.*, **7**, 0014.
13. Murphy,R.F. (2005) Location proteomics: a systems approach to subcellular location. *Biochem. Soc.Trans.*, **33**, 535–538.