Accounting for protein subcellular localization

Jadot, Michel; Boonen, Marielle; Thirion, Jacqueline; Wang, Nan; Xing, Jinchuan; Zhao, Caifeng; Tannous, Abla; Qian, Meiqian; Zheng, Haiyan; Everett, John K; Moore, Dirk F; Sleat, David E; Lobel, Peter

Published in:
Molecular & cellular proteomics : MCP

DOI:
10.1074/mcp.M116.064527

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

MCP Papers in Press. Published on December 6, 2016 as Manuscript M116.064527

Accounting for protein subcellular localization: a compartmental map of the rat liver proteome

Michel Jadot[1,#,*], Marielle Boonen[1,#], Jaqueline Thirion[1], Nan Wang[2], Jinchuan Xing[2], Caifeng Zhao[3], Abla Tannous[3], Meiqian Qian[3], Haiyan Zheng[3], John K. Everett[3], Dirk F. Moore[4], David E. Sleat[3,*], Peter Lobel[3,*]

[1]URPhyM-Laboratoire de Chimie Physiologique, Université de Namur, 61 rue de Bruxelles, Namur 5000, Belgium

[2]Department of Genetics, Human Genetics Institute of New Jersey, Rutgers, The State University of New Jersey, Piscataway, NJ 08854 USA

[3]Center for Advanced Biotechnology and Medicine, Rutgers Biomedical and Health Sciences, 679 Hoes Lane West, Piscataway, New Jersey 08854, USA.

[4]Department of Biostatistics, School of Public Health, Rutgers Biomedical and Health Sciences, 683 Hoes Lane West, Piscataway, New Jersey 08854, USA.

*Corresponding authors

[#]These authors contributed equally to this work

Summary: Accurate knowledge of the intracellular location of proteins is important for numerous areas of biomedical research including assessing fidelity of putative protein-protein interactions, modeling cellular processes at a system-wide level and investigating metabolic and disease pathways. Many proteins have not been localized, or have been incompletely localized, partly because most studies do not account for entire subcellular distribution. Thus, proteins are frequently assigned to one organelle while a significant fraction may reside elsewhere. As a step towards a comprehensive cellular map, we used subcellular fractionation with classic balance sheet analysis and isobaric labeling/quantitative mass spectrometry to assign locations to >6000 rat liver proteins. We provide quantitative data and error estimates describing the distribution of each protein among the eight major cellular compartments: nucleus, mitochondria, lysosomes, peroxisomes, endoplasmic reticulum, Golgi, plasma membrane and cytosol. Accounting for total intracellular distribution improves quality of organelle assignments and assigns proteins with multiple locations. Protein assignments and supporting data are available online through the Prolocate website (http://prolocate.cabm.rutgers.edu). As an example of the utility of this dataset, we have used organelle assignments to help analyze whole exome sequencing data from an infant dying at six months of age from a suspected neurodegenerative lysosomal storage disorder of unknown etiology. Sequencing data was prioritized using lists of lysosomal proteins comprising well-established residents of this organelle as well as novel candidates identified in this study. The latter included copper transporter 1, encoded by *SLC31A1*, which we localized to both the plasma membrane and lysosome. The patient harbors two predicted loss of function mutations in *SLC31A1*, suggesting that this may represent a heretofore undescribed recessive lysosomal storage disease gene.

Introduction:

An outstanding challenge in cell biology is to understand how proteins are organized and interact in biological pathways. Analogous to the organs found in multicellular organisms, the cell contains organelles, which are macromolecular assemblies whose components orchestrate specialized functions. Knowledge of the organellar residence of a given protein can provide valuable clues to its physiological role while conversely, understanding the protein composition of a given organelle helps describe its functional capabilities. In addition, accurate knowledge of the location of proteins is critical in assessing the biological significance of a variety of big data initiatives including protein network and pathway analysis (1, 2) as well as facilitating studies on the genetic basis of disease (3).

Microscopy and subcellular fractionation are typically used to ascertain protein localization. Fluorescence and electron-microscopy-based approaches (4) can localize proteins to morphologically recognizable cellular structures and to each other, potentially with extremely high spatial and temporal resolution but experimental parameters may result in incorrect assignments. For example, when expressing exogenous proteins, the presence of a tag and/or non-physiological steady-state levels can disrupt their normal trafficking and localization. For immunolocalization of endogenous proteins, there is increasing awareness that antibody specificity is a serious area of concern (5). In addition, even with high quality reagents, it may be difficult to determine whether a signal represents all or a subset of the protein of interest due to a variety of factors including destruction of a tag after targeting of a fusion protein to a particular organelle, quenching of fluorescent labels under certain cellular environments or relative accessibility of antibodies to different cellular structures. It is also worth noting that unless conducted with rigorous quantitation and appropriate controls, there is a tendency with morphological approaches to focus on fluorescence that is high intensity and punctate, versus lower intensity but diffuse.

Subcellular fractionation is a biochemical approach where cells are disrupted and cellular structures separated, typically by centrifugation based on their sedimentation coefficients (6). Proteins in fractions are measured using immunoassays, activity assays, or mass spectrometry (MS), the latter being amenable to high-throughput analysis. Analytical fractionation can be extremely quantitative, but there are limitations. While it is generally assumed that the protein composition of the vesicles represents the parent structures, some degree of mixing or relocalization during the fractionation process is possible. In addition, different types of structures may comigrate in a given fractionation scheme, potentially limiting resolution. Nonetheless, if one remains aware of these caveats when evaluating data, fractionation is a powerful tool for understanding the organization of the cell (7-9).

Many cellular localization studies employing subcellular fractionation have focused on cataloging proteins associated with individual organelles and therefore cannot employ balance sheet analysis ("bookkeeping") to account for the distribution of each protein throughout the cell. A failure to account for total distribution is a significant problem because, even if the bulk of a particular protein is distributed elsewhere within the cell, identification of some portion of this protein within a given organelle frequently can become dogma *as its sole site of residence*. Not only is this incorrect *per se*, this can lead to erroneous conclusions regarding function of a protein if it is considered only in the context of its organelle of minor residence. In contrast, bookkeeping is a central tenet of analytical subcellular fractionation and is used to tally recoveries, providing confidence that the distribution reflects the total cellular population of each protein.

An early study by Mann and coworkers used "protein correlation profiling" to follow distribution of multiple proteins in mouse liver subcellular fractions (10), and, more recently, high-quality maps have been published for mouse pluripotent stem cells (11) and HeLa cells (12). While these studies do list multiple residencies for some proteins, with the exception of estimating relative distribution among

nuclear, cytosol, and organellar fractions (12), relative distributions among different organelles were not estimated.

We have combined multiple orthogonal subcellular fractionation schemes with high-resolution mass-spectrometry as a platform for an approach that applies the principles of bookkeeping to estimate the distribution of rat liver proteins associated with the eight major compartments that together account for most of the cell: nucleus, mitochondria, lysosomes, peroxisomes, endoplasmic reticulum (ER), Golgi, plasma membrane (PM) and cytosol. This has allowed us to compile a draft map of the mammalian liver proteome, providing information regarding cellular locations for approximately a third of the genes predicted to be expressed in liver based on transcriptome analysis (13). This database (Prolocate) is accessible through a web portal (prolocate.cabm.rutgers.edu) with the ability to query the subcellular distribution of individual proteins and examine underlying data, compile lists of candidate residents of different organelles, and identify proteins with similar multi-compartmental distributions.

Experimental Procedures

Subcellular fractionation. All experiments and procedures involving live animals (adult male Wistar rats) were conducted in compliance with approved Institutional Animal Care and Use Committee protocols. In Experiments (Expts) A and C, animals were fasted overnight in solid bottom cages to reduce glycogen stores prior to fractionation, while animals used in Expts B and D were fed *ad libitum*. Step-by-step subcellular fractionation protocols are provided in Supplemental Methods and the overall scheme is shown in Fig. 1A. To summarize, liver homogenization and classical differential centrifugation were performed essentially as described to produce N, M, L, P and S fractions (14). Note that in this protocol, pellets are washed by a resuspension-recentrifugation procedure, with the combined supernatants being used for subsequent centrifugation steps. The L fraction was subfractionated to yield an L1 pellet and L2 supernatant as shown in Fig. 1A and as described in Supplemental Methods. Further fractionation of the L1 fraction in a Nycodenz step gradient was conducted essentially as described (15) to prepare a fraction highly enriched in lysosomes over peroxisomes ("Nyc2") (Fig. 1A and Supplemental Methods). In Expts C and D, L1 fractions prepared from control and Triton WR 1339-("Triton") injected rats (16) were adjusted to a density of 1.18 g/mL, centrifuged, and the tube divided into top and bottom fractions (Fig. 1A and Supplemental Methods). Marker enzyme activity and protein assays were conducted as described previously (17).

Experimental Design and Statistical Rationale. This study consists of four experiments using rat liver, with livers from two animals used for each subcellular fractionation. Since feeding status does not appreciably affect the overall distributions of the vast majority of liver proteins (see Results), there were two biological replicates for each type of experiment. Expts A and B were biological replicates that included analysis of all differential centrifugation fractions. In addition, Expt A contained a single fraction from a Nycodenz gradient performed on the L1 differential centrifugation fraction to help distinguish lysosomal and peroxisomal proteins. Expts C and D represent biological replicates of isopycnic centrifugation experiments on lysosomal/peroxisome-enriched fractions from control and Triton-treated animals. Given the overall concordance from each set of experiments involving biological replicates, sample sizes were considered acceptable

Mass Spectrometry. Samples indicated in Table 1 were reduced, alkylated and processed for MS by in-gel trypsin digestion as described (18). Note that the homogenate (H), used for bookkeeping purposes, was reconstituted by mixing appropriate ratios of the initial low speed pellet (N fraction) and the pooled post nuclear supernatant (E) shortly before sample processing to minimize agglutination and sampling

error. Two purified histidine-tagged bacterial proteins (Northeast Structural Genomics Consortium target identification numbers DrR57 and GmR40, kindly provided by Dr. Guy Montelione) were added before digestion as internal standards as described (17). GmR40 was added at a constant protein weight:weight ratio of 1 part per 300 parts experimental sample. DrR57 was added at different amounts to yield weight:weight ratios of 0:1, 1:600, 1:300 or 1:100 of DrR57:experimental sample. Tryptic peptides were extracted and 100 µg of each sample labeled with isobaric reagents (Expts A, B and D, iTRAQ-8plex, Expt C, TMT-6plex) using manufacturer's protocols. All individual labeled samples were combined into one and desalted. To reduce sample complexity, the resulting pooled sample was further fractionnated by strong cation ion exchange (SCX) and/or alkaline reverse phase ("high pH", or HP) chromatography prior to LC-MS. Fractions resulting from off-line chromatography fractionnation were then each analyzed by LC-MS using a Thermo Orbitrap Velos mass spectrometer. The number of individual LC-MS runs conducted for each experiment were as follows: Expt A, 288 (10 different SCX fractions and 254 different 2D (SCX x HP) fractions); Expt B, 252 (unfractionated, 16 different SCX fractions and 189 different 2D fractions); Expt C, 272 (unfractionated, 26 different SCX fractions, 40 different HP fractions, and 163 different 2D fractions); and Expt D, 199 (14 different SCX fractions and 153 different 2D fractions). (Note that some fractions were run in replicate, thus the number of fractions analyzed is less than the total number of LC-MS/MS runs). Chromatography details for prefractionation and LC-MS have been deposited in MassIVE in the methods folder for each experiment, as have LC-MS raw files, peak lists, search results, and processed data. GPM .xml search result files containing supporting information for spectral assignments can be viewed using the public installation of the GPM (http://h.thegpm.org/tandem/thegpm_upview.html) as well as using local installations (ftp://ftp.thegpm.org/projects/gpm/gpm-xe-installer/).

MS peak list data files (.mgf) were created using Proteome Discoverer vs. 1.4 (Thermo Fisher Scientific). Individual reporter ion intensities were abstracted from mgf files by summing intensities of ions within ±20 ppm of each reporter ion theoretical m/z. The mgf files were further preprocessed to include only spectra where at least two of the reporter ions had intensities ≥1000. Spectra were matched to peptide sequences using a local implementation of the Global Proteome Machine (GPM) (14, 15) with X!Tandem version SLEDGEHAMMER (2013.09.01). All mgf files within a given experiment were searched together to produce a merged "MudPIT" output file. Target proteins consisted of 24989 unique rat sequences (ENSEMBL Rnor_5.0.75 processed to retain only the first entry for any sequence that was assigned to two or more protein accession numbers), the two bacterial internal standards and 41 potential contaminants (dust/contact proteins and trypsin from the GPM cRAP database). All search parameters are listed in the GPM xml files. Briefly, searches allowed for up to one missed trypsin cleavage and used precursor and product mass tolerances of 10 and 20 ppm, respectively. Initial searching was done using complete carboxamidomethylation at cysteine and selenocysteine, complete isobaric labeling at amino terminal peptide residues and lysine, and potential methionine oxidation. During model refinement, the following potential modifications were allowed: round 1, methionine and tryptophan oxidation, asparagine and glutamine deamidation; round 2, methionine and tryptophan dioxidation; round 3, isobaric labeling at tyrosine and loss of label at the amino terminus and lysine. The spectrum synthesis option was set to "no". Peptide false positive rates (FPR) were calculated by GPM and are as follows: ExptA, 0.58%; Expt B, 0.63%; ExptC, 0.56% and ExptD, 0.85%. Peptide assignments for all spectra with peptide expectation values of ≤ 0.1 were exported to excel files and merged with processed reporter ion intensities (Workbooks S2A-D, GPM… worksheets). For the latter, fragment ion m/z were first linearly recalibrated so that the most intense reporter ion had zero mass error. Intensities associated with each reporter ion (m/z ±10 ppm) were adjusted for spillover into neighboring channels using vendor-supplied correction factors.

After mapping spectra to peptides, we conducted a series of steps to improve data quality. All steps are documented in spreadsheets for each experiment (WorkbooksS2A-D), downloadable through the Prolocate website (prolocate.cabm.rutgers.edu) and also deposited in MassIVE.

1. The first step was to filter spectra based on peptide assignment. Peptides were not used for quantification unless they met all the following criteria: fully tryptic with no missed cleavages; complete isobaric labeling of amino terminal residues and lysines; and absence of adventitious modifications deemed to increase variability (asparagine or glutamine deamidation, methionine di-oxidation, tryptophan mono- and di-oxidation, and isobaric labeling of tyrosine at positions other than the amino terminus). Spectra meeting these criteria are flagged "Y" in the "acceptable peptide" columns in Workbooks S2A-D.
2. We used the internal standard GmR40 added in equal amounts to all samples to help correct for potential differences in efficiencies of digestion, labeling and recovery prior to pooling the samples in each experiment. Here, the intensity of each reporter ion from a given spectrum was divided by the corresponding GmR40 reporter ion intensity (average of all spectra from acceptable GmR40 peptides). Corrected normalized reporter ion ratios were then calculated.
3. We conducted a balance sheet analysis for each spectra. Here, we first scaled the normalized reporter ion ratio (based on labeling of same amount of protein in each sample) to that which would be present in the entire fraction derived from one gram of liver. We then estimated recoveries by summing fractions of interest and dividing by the starting material (i.e., Expts A and B, (N+M+L1+L2+P+S)/H; Expts C and D, (top sucrose + bottom sucrose)/L1 for both the control and triton-treated samples). In theory, the amount of peptide in fractionated material should be the same as that of the starting material, and this ratio can be used to check data quality. Spectra with recoveries less than 2/3 or greater than 3/2 were eliminated from further analysis.

Gene product assignments. A single gene can be associated with multiple protein accession numbers due to alternative splicing, protein processing, and other considerations such as database annotation. Protein accession numbers are mapped to the gene name (obtained using the ENSEMBL Biomart tool) which is used as the primary identifier in this study. In some cases, peptides could be equally well assigned to multiple, genetically-distinct proteins. Here, we assigned the peptide to the protein with the highest GPM expectation value which was generally the one with the greatest sequence coverage in the search. In addition to this "best" protein assignment, Workbooks S2A-S2D lists all alternative protein accession numbers, gene accession numbers, and gene names associated with a given peptide. Organelle assignments were generated for gene products represented by at least one rat protein with a GPM expectation score of $\leq 10^{-3}$.

Multiple assignments using constrained proportionate assignment (CPA). We developed this procedure for analysis of Expts A and B. CPA uses enrichment factors (EF) (levels of a protein of interest in different centrifugation fractions compared to its level in starting material), expressed as profiles, to assign proteins to different subcellular compartments. To do this, we select several proteins that have well-established residencies for each of the compartments (single-compartment marker or reference proteins, Supplemental Table 2) and use these to produce an average profile for each compartment (eight compartments for Expt A, seven compartments for Expt B). We then fit the profile of each individual protein in our data set to an optimal combination of the reference profiles, thereby obtaining estimates of its distribution among all compartments.

The detailed CPA procedure conducted on Expt A is as follows:

1) We first calculate seven EFs (N, M, L1, L2, P, S and Nyc2) for each spectrum. EFs are equivalent to relative specific activities (RSAs) used in classical subcellular fractionation (14) and the steps for these calculations are detailed in Workbook2A. EFs are log2 transformed after setting values >16 and <1/16 to 16 and 1/16, respectively. Spectra with unacceptable recoveries or assigned to unacceptable peptides are dropped from further consideration (see above).

2) We then perform an outlier screen for proteins with at least three spectra. For each EF for a given protein, we use all spectra to compute "normal scores", given by z = (x – mean(x))/std(x), where x is the log2 transformed EF. We discard any spectrum for which any z lies in the outer 1% tails of the normal distribution. This process is repeated for all proteins.

3) Once outlier spectra have been removed, the next step is to determine a mean profile for each protein. If a protein has at least 4 spectra and at least 3 different sequences (peptides), we ordinarily have enough data to fit a random effects model. This model computes a weighted average and standard error of the log2 EFs from all spectra associated with a given protein, accounting for the fact that spectra are nested within sequences. This computation is carried out using the "lmer" function in the "lme4" R package (19). The result is essentially the mean and standard error of the observations, with an adjustment for the nested structure of the data. This procedure prevents a sequence with a very large number of spectra from dominating the estimates of the mean and standard error. Occasionally the "lmer" program will fail to converge for a particular EF. If that happens or if the condition of having at least 4 spectra and 3 different sequences is not met, we compute the simple mean and standard error of the log2 transformed data for that EF. If there are fewer than 3 spectra, we cannot compute a standard error, and thus only report the mean. When this process has been completed, every protein has a mean profile consisting of mean EFs for each of the seven fractions, denoted $\underset{\sim}{x} = (x_1, x_2, \ldots, x_7)$.

4) Using the reference proteins listed in Supplemental Table 2, we average all spectra containing acceptable log2 transformed EFs to form profiles for each of the eight compartments (Cytosol, ER, Golgi, Lysosome, Mitochondria, Nucleus, Peroxisome, and PM). The eight compartmental profiles are denoted as $\underset{\sim}{s}_1, \underset{\sim}{s}_2, \ldots, \underset{\sim}{s}_8$, where each compartmental profile $\underset{\sim}{s}_j$ is a list of its seven mean EFs.

5) For each protein, we find estimated proportions $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_8$ so that weighted profile $\underset{\sim}{y} = \hat{p}_1\underset{\sim}{s}_1 + \cdots + \hat{p}_8\underset{\sim}{s}_8$ is as close as possible to the observed profile $\underset{\sim}{x}$, subject to the constraints $0 \le \hat{p}_j \le 1$ for all j, and $\sum_{j=1}^{8} \hat{p}_j = 1$ where "close" is defined by the minimizing the sum of squares of the differences $Q = \sum_{j=1}^{7} (y_i - x_i)^2$. In other words, we may view the p_j as proportional allocations of the eight standard profiles to form $\underset{\sim}{y}$, which is as close as possible to the observed $\underset{\sim}{x}$ for this particular protein. This constrained optimization is carried out using the "spg" function in the R package "BB (20) (also manuscript in preparation, Varadhan R "Projection of infeasible iterates in constrained nonlinear optimization using BB::spg"). These estimated proportions are referred to as classification coefficients and are used to estimate the relative amount of a given protein that is associated with each compartment.

6) For proteins with ≥ 3 acceptable spectra, we also compute 95% confidence limits for the $\hat{p}_j$ by creating 1000 bootstrap samples of the original data, and fitting the CPA method to each of them. The 95% confidence limits were obtained by ordering the estimates, and selecting the 2.5% and 97.5% percentiles of the estimates.

A similar procedure is used for Expt B (Workbook2B), except there is no Nyc2 fraction, and we estimate distribution among seven compartments, merging the lysosomal and peroxisomal compartments into a Lyso/Perox compartment.

Classification of lysosomal and non-lysosomal proteins using Gaussian clustering. This procedure used for analysis of Expts C and D is a variant of classical Fischer discriminant analysis. The purpose of the Triton density shift experiments is to use the differential separation of lysosomal proteins from non-lysosomal proteins when animals are injected with Triton WR-1339, as described in "Results". The ratio of a particular protein in the top (t) to bottom (b) sucrose fractions from the Triton-treated animals is denoted Tt/Tb, and the corresponding ratio for uninjected control animals is denoted Ct/Cb. For non-lysosomal proteins, Triton has little or no effect on these ratios, while for lysosomal proteins, Tt/Tb > Ct/Cb. Our classification procedure is as follows:

1) The ratios are log2 transformed after setting values of >16 and <1/16 to 16 and 1/16, respectively. Spectra with unacceptable recoveries or assigned to unacceptable peptides are dropped from further consideration (see above).

2) Next we remove outliers similarly to that described in Step 2 for the CPA procedure (see above) i.e., for each protein we remove any spectrum for which either Log2(Tt/Tb) or Log2(Ct/Cb) lies in the outer 1% tail of a normal distribution fitted to all the spectra of that protein.

3) Having obtained a clean set of spectra, we compute the mean values of Log2(Tt/Tb) and Log2(Ct/Cb) for each protein and also the covariance matrix for these quantities. We use reference protein means and covariances for Log2(Tt/Tb) and Log2(Ct/Cb) as a "training" set for our classification procedure, which is a generalization of Fisher quadratic discriminant analysis. Our procedure uses the training set to define a mixture of two bivariate normal distributions with the aforementioned means and covariances, one for the lysosomal reference proteins and one for the non-lysosomal ("other") reference proteins. We initially evaluated the proteins used as markers in Expts A and B as well as those whose assignments were consistent in Expts A and B with various thresholds for classification coefficient point estimates and lower confidence limits. All showed a similar pattern, but given the complete concordance of proteins with a lower confidence limit of ≥ 0.7 in both Expts A and B (Table 2), we used these for purposes of classification, excluding GZMA, EPX and MPO from the lysosomal classification set as these are found in specialized lysosome-like granules but do not undergo the triton shift. We use the R procedure "Mclust" from the "Mclust" package (21, 22) to define these two normal distributions and to predict, for each protein, the probability that it belongs to the lysosomal cluster or the other cluster. The predictions consist of probability assignments to these groups.

We generate confidence limits for each protein using a parametric bootstrap procedure. Here, for each protein, using the covariance matrix computed as described above, we generate a bivariate normal parametric bootstrap sample centered at the mean [Log2(Tt/Tb),Log2(Ct/Cb)]. We use these estimates to generate probabilities of the two subcellular locations (lysosomal or other) for all proteins using the procedure described in step 3 above. We repeat this bootstrap procedure 1000 times to obtain 1000

predictions of subcellular locations for each protein. We obtain 95% confidence intervals by selecting the 25th and 750th values of the sorted bootstrap estimates.

Other classification analyses. We performed support vector machine clustering using the "svm" function in the "e1071" R package with the marker proteins (Supplemental Table 2) as a training set. We performed hierarchical clustering using the "hclust" function in the R package. Clustering was based on Euclidean distances between points using the complete linkage method.

Distances between proteins. The distance calculator on the Prolocate website can be used to find proteins with similar distributions, regardless of organelle assignments. Consider two proteins i and j from Expt A with profiles s_i and s_j (each of which is a list of the seven log2-transformed EFs). The distance between them is the Euclidean distance $d_{ij}^{A} = \sqrt{(s_{1i} - s_{1j})^2 + \cdots + (s_{7i} - s_{7j})^2}$. For Expt B, the distance d_{ij}^{B} is defined analogously, with the profiles containing 6 EFs (since there is no Nyc2 column in this experiment). For Expt C, the "profiles" are simply log2(Ct/Cb) and Log2(Tt/Tb), so that $d_{ij}^{C} = \sqrt{(s_{1i} - s_{1j})^2 + (s_{2i} - s_{2j})^2}$, and d_{ij}^{D} is similarly defined for Expt D. To utilize data from Expts A and B (for proteins contained in both), we simply sum the distances, $d_{ij}^{AB} = d_{ij}^{A} + d_{ij}^{B}$. Distances for other combinations are similarly obtained by summing the appropriate individual distances.

Mapping ENSEMBL rat proteins to other databases and/or to mouse and human orthologs. Mapping results are in Workbook S3. A list of 9,871 ENSEMBL protein sequences identified from MS searches with associated accession numbers and gene names were rooted in the ENSEMBL Rnor_5.0.75 database. Corresponding rat UniProt entries were obtained from the top BLASTP hit against the UniProt database (Major release-2015_10). Matches were considered acceptable if they covered ≥ 75% of the ENSEMBL rat protein sequence and had an expectation score ≤ 0.001. Mouse and human orthologs of rat proteins were identified via a BLASTP Reciprocal Best Hits (RBH) approach where each rat protein sequence was BLASTed against the Mouse and Human proteomes (UniProt, major release 2015_10, ENSEMBL GRCm38.p4 (Mus musculus) and GRCh38.p5 (Homo sapiens)). The top three BLAST hits in each ortholog proteome were BLASTed back against the 9,871 rat proteins identified by MS. The data set in WorkbookS3, worksheet "Blasts with statistics" utilizes a y←x nomenclature to denote the results of these analyses where x denotes the BLAST rank of an ortholog candidate protein while y denotes the BLAST rank of the original rat query protein when the ortholog candidate protein is BLASTed back against the rat proteome. For example, the nomenclature 1←1 is used for a RBH identified ortholog where the first BLAST identified protein in the ortholog proteome (x=1), when BLASTed back against the rat proteome, provides a BLAST result where the original query rat protein is the first identified protein (y=1). BLAST hits were only considered where e-values of pair-wise sequence alignments were ≤0.1 and the pair-wise sequence alignments covered ≥75% of the length of the query rat protein. While true RBH identified orthologs (1←1) where identified for the majority of rat proteins, there were instances where the analyses returned proteins not considered true RBH orthologs. For example, when the first ortholog candidate was BLASTed back against the rat proteome, the original rat query protein was the second returned BLAST hit (2←1). Analysis of such results suggests that they primarily arise from multiple isoforms in the ortholog proteome.

WorkbookS3 also contains accession numbers and assignments from a variety of other studies listing organelle assignments (worksheets "CDbBS" and "AllOtherSudies"). These are mapped to the identifiers in our study in worksheet "Blasts_with_lookups". This is reproduced in a more streamlined form in

WorkbookS1 worksheet "OtherStudies" and assignments compared to our study in WorkbookS1 worksheet "ProteinDistributionExptsA-D". For the mouse liver global subcellular fractionation study of Foster et al (10), only proteins denoted as having a single "Yes" when considering all compartments listed in their Supplemental Table 5 (refined location worksheet) were used for the analysis, allowing "Probably", "Co-migrating", and "No" in other fields (Workbook S3, worksheet AllOtherStudies).

Curation of a benchmark set. For the Compartments database human benchmark set of 12,892 proteins (23), only 7,955 were assigned to a single compartment ("+" for one compartment and either "-" or not listed for all other compartments) used in our classification scheme (Workbook S3, Worksheet CDbBS). Of these, 2256 orthologs were present in our entire data set, and 809 were assigned to our high stringency data set. Creation of the "curated" CDbBS was conducted in two rounds. We initially reviewed 197 proteins, 34 of which the primary compartment identified in our study and in the CDbBS agreed. Here, two investigators each with years of experience in evaluating subcellular location (MB and DES), blinded to the Prolocate and CDbBS assignments and to whether these agreed or disagreed, independently evaluated evidence for subcellular localization based on the validity of existing experimental evidence in published papers, evidence in the Human Protein Atlas (HPA)(24), and, occasionally, common knowledge (e.g., ribosomal proteins are present both in the cytosol and on ER). A third investigator (PL) then reviewed/arbitrated these assignments in a non-blinded manner. There was strong evidence supporting 31 of the 34 proteins that agreed, with inconclusive (but not contradictory) information for the remaining three. Given that the common assignments had excellent agreement we accepted the remaining common assignments. The second round of analysis was conducted on the remaining 118 proteins where the primary compartment identified in our study and in the CDbBS disagreed. Here, the list was split between the first two investigators for initial blind evaluation, and the third investigator reviewed and arbitrated these assignments. Individual investigator assignments, rationale and PubMed references supporting these assignments are documented in Workbook S1, Worksheet OtherStudies).

Exome sequencing analysis. Whole exome sequencing at ~100x coverage was conducted using the Life technologies (ABI) TargetSeq enrichment kit and the SOLiD 5500XL sequencing system at the Center for Targeted Therapy (University of Texas MD Anderson Cancer Center). Variants were called using LifeScope Genomic Analysis Software (ThermoFisher), the CLC Genomics Workbench (Qiagene), and the Genome Analysis Toolkit (25). Based on the ANNOVAR annotation (26), intergenic, intronic, synonymous variants, and variants with > 1% allele frequency in the ExAC dataset and/or > 5% allele frequency in the 1000 Genome Project dataset were removed. Candidate disease causal variants were selected for further analysis by Sanger sequencing.

RESULTS

Enrichment of organelles in subcellular fractions

We conducted four independent experiments to assign protein location. These represent biological replicates of two different experiments, the first of which was to generally assign organellar location while the second was to provide additional data to distinguish between proteins residing in peroxisomes and lysosomes.

The initial fractionation steps were based on the classical differential centrifugation method developed by de Duve and coworkers (6). Here, a rat liver homogenate (H) is separated into five fractions denoted as N (nuclear fraction), M (heavy mitochondrial fraction), L (light mitochondrial fraction), P (microsomal fraction) and S (final supernatant)(14). We modified this procedure by subjecting the L fraction to a

second centrifugation step to produce an L1 and L2 fraction (Fig. 1A). Marker enzyme analysis by colorimetric or fluorometric *in vitro* assays revealed that mitochondria, ER, PM and cytosol have distinct distributions from each other and from lysosomes and peroxisomes, while the latter two organelles have similar distributions (Fig. 1B). Note that it is axiomatic that in subcellular fractionation, organelles are depleted or enriched in the various fractions but are never purified to homogeneity.

Lysosomes and peroxisomes are difficult to distinguish by differential centrifugation thus we used two different approaches to separate these organelles. One exploits intrinsic differences in the densities of organelles in a Nycodenz gradient (15). Here, L1 (the fraction in which peroxisomes and lysosomes are the most enriched) was further fractionated by centrifugation on a bottom-loaded Nycodenz step gradient (Fig. 1A). Marker enzyme analysis revealed that lysosomes and peroxisomes were enriched and depleted, respectively, at the interface between the 1.105 and 1.135 g/cm^3 density solutions (Fig. 1B, Nycodenz Centrifugation, fraction Nyc2). The other approach relies on the selective decrease in the density of liver lysosomes when rats are administered Triton WR 1339 ("Triton") (16). Triton is a lipoprotein lipase inhibitor, inducing a marked increase in circulating triglyceride-rich lipoproteins which are subsequently endocytosed and accumulate in liver cells (27). L1 fractions from control or Triton-treated rats were dispersed into a sucrose solution (final density 1.18 g/cm^3), centrifuged, and the tube divided into top and bottom fractions that were collected separately (Fig. 1A). Marker enzyme analysis revealed that lysosomes are preferentially shifted into the top fraction by the Triton treatment (Fig. 1B, Sucrose Centrifugation).

Protein localization was then assigned based on isobaric labeling/quantitative MS analysis of selected fractions from the four experiments (Table 1). There are intrinsic properties of isobaric labeling experiments that may introduce stochastic and systematic errors and we addressed these in the following manner:

1. Co-isolation of labeled background peptides prior to MS2 introduces background signal leading to reporter ion compression and reduction of dynamic range (28). To address this, for most LC-MS runs, we prefractionated peptide mixtures by strong cation exchange and/or alkaline reverse phase chromatography to reduce sample complexity and thus background peptides. Monitoring of an internal standard added in different amounts to each sample revealed that there was some ratio compression but the response was linear (Supplemental Fig. 1), with r^2 ranging from 0.95 to 0.99 for the 4 experiments.
2. Samples are subjected to multiple experimental procedures before pooling and each step can potentially result in differential digestion and labeling efficiencies as well as different degrees of deamination and oxidation. To address this, we only include spectra assigned to fully tryptic peptides with no missed cleavages, fully labeled at expected residues (amino termini and lysines), and lack of adventitious modifications other than methionine monooxidation.
3. To help eliminate noise and quantification errors, we used classic balance sheet analysis to compare the amount of material represented by each spectrum that was present in all fractions with respect to starting material, and only included those with recoveries ranging from 2/3 to 3/2 (Experimental Procedures and Supplemental Fig. 2).

Of the ~1.8 million spectra assigned to rat peptides in the four experiments, ~900,000 were considered quantifiable based on our inclusion criteria (Supplemental Table 1). These mapped to a total of 9871 different proteins, which was reduced to 6920 when only considering those represented by two or more peptides and at least three spectra within a given experiment (Table 1).

Protein assignment to subcellular organelles

Expts A and B consisted of proteomic analyses of all differential centrifugation fractions and the initial rat liver homogenates and were used to establish primary subcellular localization of different proteins. A single lysosome-enriched fraction from a Nycodenz gradient ("Nyc2") conducted on the L1 fraction was included in Expt A. Note that the Nycodenz fractionation was not performed in Expt B and thus lysosomes and peroxisomes are not distinguishable in this experiment. Expts C and D consisted of a proteomic analysis of a density-based separation of the L1 fraction from control and Triton-treated rats, and were used to help properly distinguish between lysosomal and peroxisomal localizations. It is worth elaborating on the methods used for assignment of protein location to allow researchers to understand and make best use of this information.

We used a custom approach that we term "constrained proportionate assignment" (CPA) adapted from (20) in Expts A and B to estimate the probability that a given protein is located within a given compartment as well as its distribution among all compartments. CPA is described in detail in Experimental Procedures but given its importance in our analysis, it is worth briefly reviewing the overall method.

The basis of CPA is quite simple. If a protein was present in two or more different compartments, its distribution among the centrifugation fractions would reflect the sedimentation properties of the compartments, proportionally weighted to the residencies within these compartments. For our analysis of Expt A, we postulated that each protein resides in one or more of the eight major subcellular compartments: mitochondria, lysosomes, peroxisomes, ER, Golgi, PM, nucleus and cytosol. For each quantifiable spectrum, analogous to the relative specific activity term used in classical subcellular fractionation marker enzyme analysis, we calculated enrichment factors for N, M, L1, L2, P, S and Nyc2. These data were used to create a profile for each protein consisting of its enrichment factor in each fraction. Based on published studies, we selected sets of 4-6 proteins with rigorously established cellular localization as archetypal markers for each of the eight cellular compartments (Supplemental Table 2) and used these to create reference profiles (Fig. 2A, yellow-black dashed lines). In general, the profiles for each compartment were distinct from each other and consistent with respective marker enzyme analyses (Fig. 2). Each protein was then fit to an optimal linear combination of the eight reference profiles, with optimality defined by minimizing the sum of squares of differences between the profiles of the protein and of the references. The eight "classification" coefficients (each constrained to values ranging from 0 to 1 and summing to 1) correspond to a proportionate assignment of residencies of each protein to the eight compartments, with single or multiple residencies possible. Similar analysis was conducted with Expt B except that since we used a seven-compartment distribution, not distinguishing between lysosomal and peroxisomal locations (Fig. 2B). The coefficient of this "Lyso/Perox" classifier would be comparable to the sum of the lysosomal and peroxisomal classification coefficients from Expt A.

It is important to note that accuracy of classification depends both on the quality of the quantitative MS data and the degree of separation of different compartments by subcellular fractionation. To estimate data quality, for proteins represented by at least three spectra, we used variance in the data and a bootstrap procedure to generate confidence intervals for each classification coefficient. For degree of separation by subcellular fractionation, inspection of the marker protein profiles indicates good resolution of most organelles from each other with some exceptions. In particular, ER, Golgi and PM are only subtly different, which may lead to some uncertainty on whether a protein is assigned properly among these three compartments, especially when considering proteins with multi-compartment distributions.

Comparison of the classification coefficients for proteins quantified in both Expts A and B indicates a strong positive correlation (Fig. 3, all, r^2=0.79). Not unexpectedly, Golgi and PM show the greatest scatter. It is likely that this reflects intrinsic difficulties in distinguishing ER, Golgi and PM by differential centrifugation alone rather than differences in underlying protein distributions in the two experiments. If

one sums the values of the ER, Golgi and PM classification coefficients to create a microsomal classification, the overall correlation improves (Fig. 3, All*, r^2=0.88).

There is good agreement when comparing the predominant compartmental classification for proteins quantified in both Expts A and B (Table 2). There is 82.4% concordance when considering all overlapping 2952 proteins. Overall agreement increases to 94.9% when comparing the 2133 proteins where the classification coefficient point estimate is ≥0.5. Concordance improves to 97.3% for the 1877 proteins with classification coefficient lower confidence limits ≥0.5, rises to 99.8% for the 1277 proteins with point estimates ≥0.7, and reaches 100% for the 1094 proteins with lower confidence limits ≥0.7. Thus, the greater the stringency, the more confident the assignment of predominant compartmental localization of a given protein, at the expense of overall proteome coverage.

Fig. 4 illustrates separations achieved in Expts A and B for individual proteins using principal component plots, with symbols indicating predominant compartment assignment for those with point estimates ≥0.7. This also can be illustrated using heatmaps (Supplemental WorkbookS1, worksheets HeatA and HeatB). Note that plots A and B appear different because Expt A contains an additional dimension that contributes to the principal components (the Nyc2 fraction), which also allows lysosomes and peroxisomes to be classified independently. However, in both Expts A and B, proteins with the same predominant classification coefficient from the CPA analysis cluster, although not unexpectedly, ER, Golgi and PM proteins are not well resolved. Support vector machine clustering yielded single compartmental assignments that agreed well with the predominant compartment assigned in our CPA procedure (data not shown). Similar agreement was found using hierarchical clustering. This underscores the utility of the CPA method, which, in addition to agreeing with more conventional procedures in assigning proteins to a major compartment, also provides estimates for proportional residence in multiple compartments.

In Expts C and D (Workbooks S2C and S2D), the Triton shift approach was conducted specifically to distinguish lysosomal from other proteins present in the L1 fraction. Fig. 5A shows the relative distribution of generally accepted lysosomal, peroxisomal and mitochondrial proteins after centrifugation in 1.18 /cm^3 sucrose. All the lysosomal proteins show a dramatic shift in the Triton-treated samples. Triton-treatment had little effect on peroxisomal proteins. In contrast, there was a small effect on mitochondrial proteins. The basis for this is unclear but one possibility is that this is due to the normal process of mitochondrial turnover and degradation within the lysosome. It also is important to note that for multi-compartmental proteins with lysosomal component and a non-lysosomal, non-peroxisomal component, the Triton shift experiments will overestimate the lysosomal proportion of the total population.

We explored different methods to assign proteins based on Triton shift experiments. One possibility was to fit the data to the same set of rigorously established reference proteins used for Expts A and B (Fig. 5B). We also considered using proteins that had consistent classification in Expts A and B with various thresholds for inclusion. The proteins assigned with a lower confidence limit of ≥ 0.7 in both Expts A and B (Fig. 5C) have a pattern quite similar to that of the established markers (Fig. 5B). Given that the former are much more numerous, we used these proteins for purposes of classification. Here, we divided the proteins into a lysosomal marker set and a marker set encompassing all other compartments. We then used finite Gaussian mixture clustering (21) (Experimental Procedures) to calculate "Lyso" and "Other" classification coefficients for all proteins in each experiment, generating confidence limits for proteins with at least three spectra.

We also compared the fractionation of all 1923 proteins common to both Expts C and D (Fig. 5, Panel D). There was good correlation in the distribution found in fasted and fed animals regardless of Triton

treatment, indicating minimal overall effects of nutritional status. Because of the greater protein coverage, we rely primarily on Expt C for corroborating lysosomal assignment in Expt A and for distinguishing between lysosomal and peroxisomal residence in Expt B.

Benchmarking protein assignments

While our assignments were largely internally consistent, it was critical to benchmark them against proteins with established locations. For this purpose, we selected a "high stringency set" of 3,034 proteins whose major compartment was consistent in Expts A, B and C, with a minimum of 2 peptides and 3 spectra for Expt A and at least one peptide and spectrum from Expts B and C (Workbook S1). We compared these to the Compartments Database Benchmark Set (23) (CDbBS), which lists locations for 12,892 human proteins. Of these, we found 809 rat orthologs in our high stringency set that were annotated in CDbBS to be associated with only one of our eight compartments. As shown in Fig. 6A, there was moderately good concordance for these proteins. However, given that there is no "gold standard" for compartmental localization, we conducted a blinded analysis to evaluate the published experimental evidence supporting the CDbBS categorizations (see Experimental Procedures). This allowed us to discard questionable entries and create a "curated" database containing 588 proteins. Our major assignments now agree extremely well with those in this high stringency curated CDbBS (Fig. 6B). This not only provides additional confidence in results from analytical subcellular fractionation but also underscores the inaccuracies present in existing subcellular localization databases, of which further examples are described below.

Comparing Prolocate protein assignments with organelle databases

We have compared our data against several commonly used protein localization datasets (see below). However, investigators can readily compare our assignments with any dataset of interest using simple software tools (see Supplemental Materials, "Comparing custom datasets with Prolocate ".

The MitoCarta study utilizes a combination of MS, GFP-tagging and/or bioinformatics criteria to estimate the likelihood in terms of a false discovery rate (FDR) for mitochondrial localization of >20,000 different mammalian gene products (29, 30), 2657 of which are present in our high stringency data set. For these, our Mito classification coefficient decreased with increasing MitoCarta FDR (Fig. 7A). The MitoCarta study also contains a list of proteins that are designated as having strong support of mitochondrial localization, 621 of which are in our high stringency set. Most of these have low MitoCarta FDRs, but some do not, and it is likely that some represent false positives. Our mitochondrial classification coefficients are considerably higher for the designated mitochondrial MitoCarta proteins with low FDRs compared to those with higher FDRs (Fig. 7, Panels B and C, see also Supplemental WorksheetS1, workbook heatMITOCARTA). Use of both databases in tandem should provide increased confidence in mitochondrial assignments, as well as identifying possibly incorrectly assigned proteins.

The MitoMiner database is a searchable database that has compiled an integrated mitochondrial protein index (IMPI) listing putative rat, mouse and human mitochondrial proteins (31). Our high stringency data set contains 998 of the proteins on the rat IMPI list. Only half of these are classified as predominantly mitochondrial in Expt A, while for the others, many have Mito classification coefficients near zero (Fig. 7D, see also Supplemental WorksheetS1, workbook heatIMPI). These data suggest that there are a number of putative mitochondrial proteins on the IMPI list that are worthy of further evaluation.

The Peroxisome Database (32) (peroxisomeDB) is a catalog of the peroxisomal proteome of multiple species, including rat, compiled from experimental literature review, homology and bioinformatic annotation. Fifty-eight rat proteins in the peroxisomeDB were present in the Prolocate high stringency

data set and a majority were assigned a primary peroxisomal localization. However, for a significant proportion of these (19 proteins), we assigned primary localization to compartments other than the peroxisome (Fig. 7E, see also Supplemental WorksheetS1, workbook heatPEROXDB). Most of these were assigned a minor peroxisomal location in Prolocate and this is consistent with published evidence for residence in multiple cellular locations. However, seven had a peroxisomal classification coefficient of <0.01 in Expt A and these warrant further scrutiny regarding intracellular location. Similar results were obtained when our data were compared with a single study (33) focusing on proteomic analysis of peroxisomes (data not shown).

The CLEAR (Coordinated Lysosomal Expression and Regulation) network is a group of genes that are predicted to regulate lysosomal biogenesis and function via the transcription factor EB (TFEB)(34, 35). Members of the CLEAR network are characterized by the presence of binding sites for TFEB within their respective promoters. The Prolocate high stringency data set contains 159 proteins encoded by genes containing CLEAR elements. For 50 of these, we report a primary lysosomal residence but the majority (109 proteins) are assigned to other organelles (Fig. 7F, see also Supplemental WorksheetS1, workbook heatCLEAR) although a minor lysosomal residence was detected for some. A primarily non-lysosomal location for proteins involved in lysosome biogenesis or function is quite conceivable and consistent with this, our data indicate that the presence of a CLEAR element should not be interpreted as an indicator of lysosomal residence.

Other studies have combined MS with subcellular fractionation to determine global protein localization. While these did not estimate proportionate assignments among all the compartments, and also had some differences in definition of compartments, we were able to conduct some comparisons. Note that we compare the Prolocate high confidence data set to all protein orthologs in the other studies (Supplemental Tables 3-5), but in the summary below, we only report final numbers for proteins found in analogous compartments. An early study used rate-zonal centrifugation and protein profile correlation analysis to classify proteins in a mouse liver homogenate (10). Of the 379 mouse liver proteins assigned to a single compartment, 93.6% agreed with our predominant classification (Supplemental Table 3). However, lysosomes and peroxisomes were not part of the mouse liver classification scheme, and the protein correlation profile classifications were heavily biased towards mitochondrial and cytosolic proteins (344 assignments). Two additional studies were published while this manuscript was under initial review. One primarily used isopycnic centrifugation and a ten-channel isobaric labeling MS3 approach (termed "Hyperlopit") to map proteins in mouse pluripotent stem cells (11). Of the 435 overlapping proteins, 93.3% agreed with our assignments (Supplemental Table 4). The latest study analyzed the human HeLa cell line by differential centrifugation and SILAC labeling to map proteins present in an organelle-enriched fraction (essentially corresponding to a combined M+L+P fraction), and also used a label-free approach to estimate proportional residence among the "N", "M+L+P", and "S" equivalents (12). When considering the 687 HeLa proteins assigned to individual organelles with very high, high and medium confidence assignments, 95.6% agreed with the Prolocate assignments (Supplemental Table 5). This increased to 97.3% when using 548 HeLa proteins assigned with very high and high confidence assignments (Supplemental Table 5). This remarkable degree of agreement lends confidence to these resources, not only for the proteins identified in multiple studies, but also for the many proteins unique to each.

Applications of Prolocate

Proteins with complex distributions. One advantage of our approach is that we estimate the entire distribution of a protein among eight different cellular compartments. We have also generated matrices listing distances between all protein pairs identified in each experiment (Supplemental Workbooks 4A-

D), and the Prolocate interface allows integration of the data to find proteins with similar distributions to any protein of interest measured in one or more experiment. Two select examples are as follows:

COG complex. The COG complex comprises 8 individual subunits that together mediate retrograde vesicle transport within the Golgi. We identified all 8 subunits (COG1-8) and, consistent with the known properties of this complex, they exhibited an extremely similar pattern of localization, with ~30% in the Golgi and ~70% in the cytosolic fraction (Fig. 8A). This agrees with morphological (36) and subcellular fractionation (37) studies that localize members of the COG complex to membrane-associated Golgi and cytosol, respectively. Remarkably, when we search for proteins that co-localize with COG1 using the Prolocate distance calculator (Expts A and C), COGs 3, 4, 5, 7 and 8 represent 5 out of the top 6 highest scoring hits.

AP complexes. We also investigated the distribution of individual proteins in AP complexes, which are multi subunit complexes involved in different vesicular transport pathways (38). Based on our data, intracellular distribution of the individual proteins of AP complexes 1, 2, 3 and 5 were remarkably consistent (Fig. 8B). Proteins of the AP1 complex were primarily localized to the Golgi and the cytosol, which is consistent with the role of AP1 in trans-Golgi to endosome trafficking. Proteins of the AP2 complex were restricted to the PM and cytosol, with no lysosomal localization, consistent with the role of AP2 in clathrin-dependent endocytosis. Proteins of the AP3 complex were ~70% localized to the lysosome, with the remainder being cytosolic. AP3 plays a role in the biogenesis of lysosome and related organelles e.g. endosomes. Distribution of the proteins of AP5 was more heterogeneous than that of the other AP complexes but the lysosome was the primary location. The function of AP5 is unclear but it may play a role in endosomal sorting. As with the COG complex, using the Prolocate distance calculator to identify proteins that co-locate with individual AP complex components primarily returns the other complex members as the highest scoring hits.

Candidate disease genes. Knowledge of protein location is useful in investigating disease pathways and we present one example here that arises from our long-term interest in lysosomal diseases of unknown etiology (39-41). In the course of this research, we performed whole exome sequencing on a number of unsolved cases accumulated in the course of our studies but found potential mutations in an unwieldy number (thousands) of genes. We thus used Prolocate to prioritize candidate disease genes based on lysosomal location of the respective encoded proteins.

One of the proteins on our lysosomal candidate list, SLC31A1, also called CTR1, is involved in copper transport (42). The recent HeLa cell study (12) localizes SLC31A1 to the plasma membrane while previous reports localized it to the PM with some evidence for an intracellular pool in different cell lines (43). More recently, evidence has been reported that may or may not support some lysosomal residence (44, 45). We assign SLC31A1 to both the lysosome and to the plasma membrane (Fig. 8C).

Initial whole exome sequencing data highlighted *SLC31A1* as the primary candidate in one of our potential neurodegenerative lysosomal disease subjects, an infant who died in 1988 at age 6 months. Sanger sequencing revealed compound heterozygosity for two potential mutations in *SLC31A1* (Fig. 9). A nonsense mutation at Arg90 results in a severely truncated protein and it is presumably a null allele. A missense Val181Leu mutation is of unclear significance but predicted to be deleterious by Sorting Intolerant From Tolerant (SIFT) analysis software (46).

Several observations support the possibility that SLC31A1 deficits may be the cause of disease in this case. First, neither of these changes was found in the large scale Exome Aggregation Consortium sequence of 60,706 individuals (47) indicating they are not common polymorphisms. Second, a homozygous mouse knockout for *SLC31A1* is embryonic lethal while the heterozygote displays tissue-

specific abnormalities including reduced copper levels in brain (48). Thus, the nonsense mutation, in combination with a hypomorphic allele, could be responsible for disease. Third, a missense mutation (Arg90Gly, coincidentally at the site of the nonsense mutation reported here) in *SLC31A1* has been suggested to be associated with recessive cognitive disease (49). Fourth, defects in other proteins involved in copper transport e.g., a copper-transporting ATPase in Wilson disease, results in neurologic and other abnormalities (50). Thus, while we do not have access to pedigree and detailed case history information, there is intriguing evidence for *SLC31A1* as a potential human disease gene and it warrants further investigation. Use of the Prolocate database to filter the whole exome sequencing data was a key step in identifying this candidate.

DISCUSSION

We report here our progress to date in establishing a global atlas of protein location. Our approach differs from other quantitative MS efforts in the field in several important respects. First, we use balance sheet analysis to filter data for acceptable recovery. Second, we have implemented statistical approaches that allow estimation of the distribution of each protein among multiple compartments. Third, we estimate confidence limits for protein distributions. This is important as there are experimental errors inherent in any quantitative measurement and this is particularly true for data-dependent MS, where there is oversampling of abundant proteins and occasional incorrect assignments of spectra to peptides and of peptides to proteins. Fourth, evidence for location of individual proteins of interest should be scrutinized closely and to facilitate this, the Prolocate website allows inspection of the fractionation profiles of all component peptides assigned to each protein. Profiles of individual spectra, peptides and proteins have been archived on the Prolocate Website and in Supplemental Workbooks 2A-2D.

It is worthwhile noting that our analysis depends on assigning proteins to reference compartments but they do not cover all cellular structures (e.g., different types of endosomes, stress granules, lipid droplets). If a different cellular structure fractionated similarly to one of the reference compartments, proteins in that structure would be assigned to the reference compartment. Alternatively, if a protein resided in a single structure that fractionated distinctly from the eight reference compartments, it would be assigned to multiple reference compartments. In such cases, the organelle assignments may be misleading but identification of proteins with similar fractionation properties (e.g., using the distance calculator) may still reveal potential associations with a common compartment or sets of compartments. Expanding the number of fractionation procedures used as well as use of additional compartment and subcompartment markers should allow incorporation of additional cellular structures into future classification schemes. Nonetheless, concordant results from a series of complementary approaches including analytical fractionation and microscopy are required for the lowest level of uncertainty in assignment.

We use rat liver as a source for analytical centrifugation due to effective fractionation procedures tailored to this tissue (6), and inability to achieve equivalent high-quality fractionation using mouse liver (51). In addition, fresh tissue is highly relevant to *in vivo* physiology. However, while hepatocytes make up the bulk of the liver, there are additional cell types present, and it is possible that this introduces additional heterogeneity in the centrifugation properties of organelles. Cultured cells may provide a more homogeneous source, with the tradeoff that cells undergo changes in culture with accompanying perturbations in protein trafficking, even after short time periods (52). Nonetheless, the remarkable concordance between our assignments on rat liver proteins with those from studies on mouse pluripotent stem cells (11) and HeLa cell (12) proteomes, suggests that such concerns are not a serious obstacle in assigning locations to most proteins. It is worth noting that the use of different cell sources (e.g., rat and

mouse liver, cultured mouse and human cells) and different approaches to fractionation (differential, velocity gradient, isopycnic centrifugation) will results in datasets with some overlap and, thus provide independent validation, but many proteins will be unique to each study. Consolidating these individual studies will be an important step towards establishing a high-quality map of the mammalian subcellular proteome.

In summary, we provide subcellular localization data for over 6000 rat liver proteins, representing approximately a third of the predicted liver proteome (13). This Prolocate database is readily accessible through a web portal. Prolocate can be used to query the location of individual proteins and access the supporting data, and to generate lists of candidate proteins for different cellular compartments. A key feature of Prolocate is the ability to identify sets of proteins that have similar single- or multi-compartmental distributions that may reflect functional interactions within the cell. While future studies will extend coverage of the proteome, explore the relationships between subcellular location and post-translational modification, and address additional cellular compartments, the current map and analytical tools should provide a valuable resource for the biomedical research community.

Acknowledgments.

This work was supported in part by grants from the National Institutes of Health (R01DK5431, R01NS37918, P30NS46593, and S10RR24584). We would like to thank Virginie Tevel, Catherine Lambert de Rouvroit and Marie-France Leruth for assistance with fractionation experiments, Dr. Chang-Gong Liu for exome sequencing, Drs. Robert Donnelly and Dibyendu Kumar for help with genomic sequence analysis, David Fenyo for software used for abstraction of reporter ion intensities and Dr. Marco Sardiello for providing a compilation of CLEAR network proteins. Data are deposited in the MassIVE (http://massive.ucsd.edu) and ProteomeXchange (http://www.proteomexchange.org/) repositories (MS raw data, peak lists, methods and search files: ExptA, MSV000079172 and PXD002418; ExptB, MSV000080231 and PXD005109; ExptC, MSV000080232 and PXD005110; Expt D, MSV000080233 and PXD005112), on the Prolocate website (http://prolocate.cabm.rutgers.edu) (Workbooks) and in Supplemental Materials.

Author contributions.

P.L. and M.J. conceived and designed this study. M.J. and J.T. designed and supervised the subcellular fractionation experiments; M.B., M.J., P.L. and D.E.S. interpreted subcellular fractionation data; N.W., J.X., D.E.S. and P.L. conducted analysis of whole exome sequencing data; D.E.S., C.Z., M.Q., and H.Z. conducted MS analyses; P.L., J.E, D.M. and A.T. designed the Prolocate website; D.M. and P.L. conducted statistical analyses; J.E. conducted bioinformatics analysis to match protein entries in different databases, and; P.L., D.E.S., M.B. and M.J. curated the CDbBS. The initial draft of the manuscript was written by D.E.S., M.J., M.B. and P.L. while all authors contributed towards subsequent revisions prior to submission.

Materials & Correspondence.

Correspondence are requests for materials should be addressed to M.J. (michel.jadot@unamur.be), D.E.S. (sleat@cabm.rutgers.edu) and P.L. (lobel@cabm.rutgers.edu).

Competing financial interests.

The authors declare no competing financial interests.

References

1. Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S. D., Yang, X., Ghamsari, L., Balcha, D., Begg, B. E., Braun, P., Brehme, M., Broly, M. P., Carvunis, A. R., Convery-Zupan, D., Corominas, R., Coulombe-Huntington, J., Dann, E., Dreze, M., Dricot, A., Fan, C., Franzosa, E., Gebreab, F., Gutierrez, B. J., Hardy, M. F., Jin, M., Kang, S., Kiros, R., Lin, G. N., Luck, K., MacWilliams, A., Menche, J., Murray, R. R., Palagi, A., Poulin, M. M., Rambout, X., Rasla, J., Reichert, P., Romero, V., Ruyssinck, E., Sahalie, J. M., Scholz, A., Shah, A. A., Sharma, A., Shen, Y., Spirohn, K., Tam, S., Tejeda, A. O., Trigg, S. A., Twizere, J. C., Vega, K., Walsh, J., Cusick, M. E., Xia, Y., Barabasi, A. L., Iakoucheva, L. M., Aloy, P., De Las Rivas, J., Tavernier, J., Calderwood, M. A., Hill, D. E., Hao, T., Roth, F. P., and Vidal, M. (2014) A proteome-scale map of the human interactome network. *Cell* 159, 1212-1226
2. Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A., and Honig, B. (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* 490, 556-560
3. Lage, K. (2014) Protein-protein interactions and genetic diseases: The interactome. *Biochim Biophys Acta* 1842, 1971-1980
4. Lidke, D. S., and Lidke, K. A. (2012) Advances in high-resolution imaging--techniques for three-dimensional imaging of cellular structures. *Journal of cell science* 125, 2571-2580
5. Bradbury, A., and Pluckthun, A. (2015) Reproducibility: Standardize antibodies used in research. *Nature* 518, 27-29
6. de Duve, C. (1971) Tissue fractionation. Past and present. *The Journal of cell biology* 50, 20d-55d
7. Claude, A. (1975) The coming of age of the cell. *Science* 189, 433-435
8. Duve, C. (1975) Exploring cells with a centrifuge. *Science* 189, 186-194
9. Palade, G. (1975) Intracellular aspects of the process of protein synthesis. *Science* 189, 347-358
10. Foster, L. J., de Hoog, C. L., Zhang, Y., Zhang, Y., Xie, X., Mootha, V. K., and Mann, M. (2006) A mammalian organelle map by protein correlation profiling. *Cell* 125, 187-199
11. Christoforou, A., Mulvey, C. M., Breckels, L. M., Geladaki, A., Hurrell, T., Hayward, P. C., Naake, T., Gatto, L., Viner, R., Arias, A. M., and Lilley, K. S. (2016) A draft map of the mouse pluripotent stem cell spatial proteome. *Nat Commun* 7, 9992
12. Itzhak, D. N., Tyanova, S., Cox, J., and Borner, G. H. (2016) Global, quantitative and dynamic mapping of protein subcellular localization. *Elife* 5
13. Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P. H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Ponten, F. (2015) Proteomics. Tissue-based map of the human proteome. *Science* 347, 1260419
14. de Duve, C., Pressman, B. C., Gianetto, R., Wattiaux, R., and Appelmans, F. (1955) Tissue fractionation studies. 6. Intracellular distribution patterns of enzymes in rat-liver tissue. *Biochem J* 60, 604-617
15. Wattiaux, R., Wattiaux-De Coninck, S., Ronveaux-dupal, M. F., and Dubois, F. (1978) Isolation of rat liver lysosomes by isopycnic centrifugation in a metrizamide gradient. *J. Cell Biol.* 78, 349-368
16. Wattiaux, R., Wibo, M., and Baudhuin, P. (1963) Influence of the injection of Triton WR-1339 on the properties of rat-liver lysosomes. In: de Reuck, A. V. S., and Cameron, M. P., eds. *Ciba Foundation Symposium Lysosomes*, pp. 176-200, Little, Brown, and Company, Boston
17. Della Valle, M. C., Sleat, D. E., Zheng, H., Moore, D. F., Jadot, M., and Lobel, P. (2011) Classification of subcellular location by comparative proteomic analysis of native and density-shifted lysosomes. *Mol Cell Proteomics* 10, M110 006403
18. Sleat, D. E., Della Valle, M. C., Zheng, H., Moore, D. F., and Lobel, P. (2008) The mannose 6-phosphate glycoprotein proteome. *J Proteome Res* 7, 3010-3021

19. Bates, D., Machler, M., Bolker, B. M., and Walker, S. C. (2015) Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* 67, 1-48
20. Varadhan, R., and Gilbert, P. D. (2009) BB: An R Package for Solving a Large System of Nonlinear Equations and for Optimizing a High-Dimensional Nonlinear Objective Function. *J Stat Softw* 32, 1-26
21. Fraley, C., and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* 97, 611-631
22. Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012) mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. . *Technical Report 597*, Department of Statistics, University of Washington
23. 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 (Oxford)* 2014, bau012
24. Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Bjorling, L., and Ponten, F. (2010) Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* 28, 1248-1250
25. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20, 1297-1303
26. Wang, K., Li, M., and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38, e164
27. Huterer, S., Phillips, M. J., and Wherrett, J. R. (1975) Effects of prolonged administration of triton WR-1339 to the rat on morphology and phospholipids of liver. *Laboratory investigation; a journal of technical methods and pathology* 33, 305-310
28. Ting, L., Rad, R., Gygi, S. P., and Haas, W. (2011) MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Methods* 8, 937-940
29. Calvo, S. E., Clauser, K. R., and Mootha, V. K. (2016) MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res* 44, D1251-1257
30. 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., Hill, D. E., Vidal, M., Evans, J. G., Thorburn, D. R., Carr, S. A., and Mootha, V. K. (2008) A mitochondrial protein compendium elucidates complex I disease biology. *Cell* 134, 112-123
31. Smith, A. C., and Robinson, A. J. (2016) MitoMiner v3.1, an update on the mitochondrial proteomics database. *Nucleic Acids Res* 44, D1258-1261
32. Schluter, A., Real-Chicharro, A., Gabaldon, T., Sanchez-Jimenez, F., and Pujol, A. (2010) PeroxisomeDB 2.0: an integrative view of the global peroxisomal metabolome. *Nucleic Acids Res* 38, D800-805
33. Gronemeyer, T., Wiese, S., Ofman, R., Bunse, C., Pawlas, M., Hayen, H., Eisenacher, M., Stephan, C., Meyer, H. E., Waterham, H. R., Erdmann, R., Wanders, R. J., and Warscheid, B. (2013) The proteome of human liver peroxisomes: identification of five new peroxisomal constituents by a label-free quantitative proteomics survey. *PLoS One* 8, e57395
34. Sardiello, M., Palmieri, M., di Ronza, A., Medina, D. L., Valenza, M., Gennarino, V. A., Di Malta, C., Donaudy, F., Embrione, V., Polishchuk, R. S., Banfi, S., Parenti, G., Cattaneo, E., and Ballabio, A. (2009) A gene network regulating lysosomal biogenesis and function. *Science* 325, 473-477
35. Palmieri, M., Impey, S., Kang, H., di Ronza, A., Pelz, C., Sardiello, M., and Ballabio, A. (2011) Characterization of the CLEAR network reveals an integrated control of cellular clearance pathways. *Human molecular genetics* 20, 3852-3866
36. Ungar, D., Oka, T., Brittle, E. E., Vasile, E., Lupashin, V. V., Chatterton, J. E., Heuser, J. E., Krieger, M., and Waters, M. G. (2002) Characterization of a mammalian Golgi-localized protein complex, COG, that is required for normal Golgi morphology and function. *The Journal of cell biology* 157, 405-415

37. Walter, D. M., Paul, K. S., and Waters, M. G. (1998) Purification and characterization of a novel 13 S hetero-oligomeric protein complex that stimulates in vitro Golgi transport. *J Biol Chem* 273, 29565-29576
38. Park, S. Y., and Guo, X. (2014) Adaptor protein complexes and intracellular transport. *Biosci Rep* 34
39. Tyynela, J., Sohar, I., Sleat, D. E., Gin, R. M., Donnelly, R. J., Baumann, M., Haltia, M., and Lobel, P. (2000) A mutation in the ovine cathepsin D gene causes a congenital lysosomal storage disease with profound neurodegeneration. *EMBO J* 19, 2786-2792
40. Naureckiene, S., Sleat, D. E., Lackland, H., Fensom, A., Vanier, M. T., Wattiaux, R., Jadot, M., and Lobel, P. (2000) Identification of HE1 as the second gene of Niemann-Pick C disease. *Science* 290, 2298-2301
41. Sleat, D. E., Donnelly, R. J., Lackland, H., Liu, C. G., Sohar, I., Pullarkat, R. K., and Lobel, P. (1997) Association of mutations in a lysosomal protein with classical late-infantile neuronal ceroid lipofuscinosis. *Science* 277, 1802-1805
42. Wee, N. K., Weinstein, D. C., Fraser, S. T., and Assinder, S. J. (2013) The mammalian copper transporters CTR1 and CTR2 and their roles in development and disease. *The international journal of biochemistry & cell biology* 45, 960-963
43. Klomp, A. E., Tops, B. B., Van Denberg, I. E., Berger, R., and Klomp, L. W. (2002) Biochemical characterization and subcellular localization of human copper transporter 1 (hCTR1). *Biochem J* 364, 497-505
44. Ohrvik, H., Logeman, B., Turk, B., Reinheckel, T., and Thiele, D. J. (2016) Cathepsin Protease Controls Copper and Cisplatin Accumulation via Cleavage of the Ctr1 Metal-binding Ectodomain. *J Biol Chem* 291, 13905-13916
45. Clifford, R. J., Maryon, E. B., and Kaplan, J. H. (2016) Dynamic internalization and recycling of a metal ion transporter: Cu homeostasis and CTR1, the human Cu(+) uptake system. *Journal of cell science* 129, 1711-1721
46. Ng, P. C., and Henikoff, S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31, 3812-3814
47. Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H. H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., MacArthur, D. G., and Exome Aggregation, C. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285-291
48. Lee, J., Prohaska, J. R., and Thiele, D. J. (2001) Essential role for mammalian copper transporter Ctr1 in copper homeostasis and embryonic development. *Proc Natl Acad Sci U S A* 98, 6842-6847
49. Najmabadi, H., Hu, H., Garshasbi, M., Zemojtel, T., Abedini, S. S., Chen, W., Hosseini, M., Behjati, F., Haas, S., Jamali, P., Zecha, A., Mohseni, M., Puttmann, L., Vahid, L. N., Jensen, C., Moheb, L. A., Bienek, M., Larti, F., Mueller, I., Weissmann, R., Darvish, H., Wrogemann, K., Hadavi, V., Lipkowitz, B., Esmaeeli-Nieh, S., Wieczorek, D., Kariminejad, R., Firouzabadi, S. G., Cohen, M., Fattahi, Z., Rost, I., Mojahedi, F., Hertzberg, C., Dehghan, A., Rajab, A., Banavandi, M. J., Hoffer, J., Falah, M., Musante, L., Kalscheuer, V., Ullmann, R., Kuss, A. W., Tzschach, A., Kahrizi, K., and Ropers, H. H. (2011) Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature* 478, 57-63
50. de Bie, P., Muller, P., Wijmenga, C., and Klomp, L. W. (2007) Molecular pathogenesis of Wilson and Menkes disease: correlation of mutations with molecular defects and disease phenotypes. *Journal of medical genetics* 44, 673-688

51. Dixit, S. S., Jadot, M., Sohar, I., Sleat, D. E., Stock, A. M., and Lobel, P. (2011) Loss of Niemann-Pick C1 or C2 protein results in similar biochemical changes suggesting that these proteins function in a common lysosomal pathway. *PLoS One* 6, e23677

52. Roelofsen, H., Bakker, C. T., Schoemaker, B., Heijn, M., Jansen, P. L., and Elferink, R. P. (1995) Redistribution of canalicular organic anion transport activity in isolated and cultured rat hepatocytes. *Hepatology* 21, 1649-1657

53. Schroder, B. A., Wrocklage, C., Hasilik, A., and Saftig, P. (2010) The proteome of lysosomes. *Proteomics* 10, 4053-4076

54. Cieutat, A. M., Lobel, P., August, J. T., Kjeldsen, L., Sengelov, H., Borregaard, N., and Bainton, D. F. (1998) Azurophilic granules of human neutrophilic leukocytes are deficient in lysosome-associated membrane proteins but retain the mannose 6-phosphate recognition marker. *Blood* 91, 1044-1058

Table 1, Subcellular fractionation/quantitative MS experiments.

Expt	Triton treatment	L1 fractionation	Fractions analyzed by MS	# identified proteins ≥2 peptide & ≥3 spectra (all)	
A	-	Nycodenz	H, N, M, L1,L2, P, S, Nyc2	6075 (8071)	A∩B 2952 (4437) A∩C 4762 (6361) B∩D 1923 (3160) C∩D 2211 (3710) A∩B∩C 2852 (4200) A∩B∩C∩D 1903 (3087) A∪B∪C∪D 6920 (9871)
B	-	none	H, N, M, L1, L2, P, S	3021 (4707)	
C	+ and -	Sucrose	L1, Top & Bottom sucrose	5553 (7859)	
D	+ and -	Sucrose	L1, Top & Bottom sucrose	2245 (3968)	

Table 2. Comparison of major compartment assignments from Expts A and B

Expt A, Highest Classification	# proteins with at least 2 peptides and 3 spectra meeting threshold No threshold ≥0.5 classifier point estimate ≥0.5 classifier lower confidence limit ≥0.7 classifier point estimate ≥0.7 classifier lower confidence limit								(A∩B) /|A|
	Expt B, Highest Classification								
	Mito	**Ly/Per**	**ER**	**Golgi**	**PM**	**Cyto**	**Nuc**	***ALL***	
Mito	463 414 397 303 269	9 1	1 1			2 1		*475* *417* *397* *303* *269*	97.5% 99.3% 100.0% 100.0% 100.0%
Lyso	2 2 1 1	200 171 159 118 105	23 7 4	5 1	2 1	10 3 2		*242* *185* *166* *119* *105*	82.6% 92.4% 95.8% 99.2% 100.0%
Perox	4	75 67 60 31 27	1		1	1		*82* *67* *60* *31* *27*	91.5% 100.0% 100.0% 100.0% 100.0%
ER	7 1	11 1	440 376 320 243 198	49 11 1	25 7 4 1	6	1	*539* *396* *325* *244* *198*	81.6% 94.9% 98.5% 99.6% 100.0%
Golgi		7	18 6 1 1	88 49 36 15 12	9 3 2			*122* *58* *39* *16* *12*	72.1% 84.5% 92.3% 93.8% 100.0%
PM	6	21 3 1	17 6 3	10 1	222 159 126 58 30	7	1	*284* *169* *130* *58* *30*	78.2% 94.1% 96.9% 100.0% 100.0%
Cytosol	3	54 16 11	21 5 3	66 14 9	70 12 5	851 720 676 484 433	12 1 1	*1077* *768* *705* *484* *433*	79.0% 93.8% 95.9% 100.0% 100.0%
Nucleus		1		4	30 4 1	2 1 1	94 68 53 22 20	*131* *73* *55* *22* *20*	71.8% 93.2% 96.4% 100.0% 100.0%
ALL	*485* *417* *398* *304* *269*	*378* *259* *231* *149* *132*	*521* *401* *331* *244* *198*	*222* *76* *46* *15* *12*	*359* *186* *138* *59* *30*	*879* *725* *679* *484* *433*	*108* *69* *54* *22* *20*	*2952* *2133* *1877* *1277* *1094*	82.4% 94.9% 97.3% 99.8% 100.0%
(A∩B)/|B|	95.5% 99.3% 99.7% 99.7% 100.0%	72.8% 91.9% 94.8% 100.0% 100.0%	84.5% 93.8% 96.7% 99.6% 100.0%	39.6% 64.5% 78.3% 100.0% 100.0%	61.8% 85.5% 91.3% 98.3% 100.0%	96.8% 99.3% 99.6% 100.0% 100.0%	87.0% 98.6% 98.1% 100.0% 100.0%	82.4% 94.9% 97.3% 99.8% 100.0%	

Fig. 1. **Subcellular fractionation and organelle assignment.** (**A**) Fractionation scheme showing differential centrifugation (Left) and orthogonal fractionation of L1 (Right). Organelles enriched in given fraction are indicated. Note that pellets are washed by resuspension and recentrifugation, with pooled supernatants used for the next step (Experimental Procedures and Supplemental Materials). (**B**) Marker enzyme analysis. Fractions are labeled on plots showing the lysosomal marker enzyme. Marker enzymes for different compartments are as follows: mitochondria, cytochrome oxidase; lysosome, β-galactosidase; peroxisome, catalase; ER, alkaline α-glucosidase; PM, alkaline phosphodiesterase; cytosol, dipeptidyl-peptidase III. Differential and Nycodenz Centrifugation, data from Expt A; L1 Sucrose Centrifugation, data from Expt C.

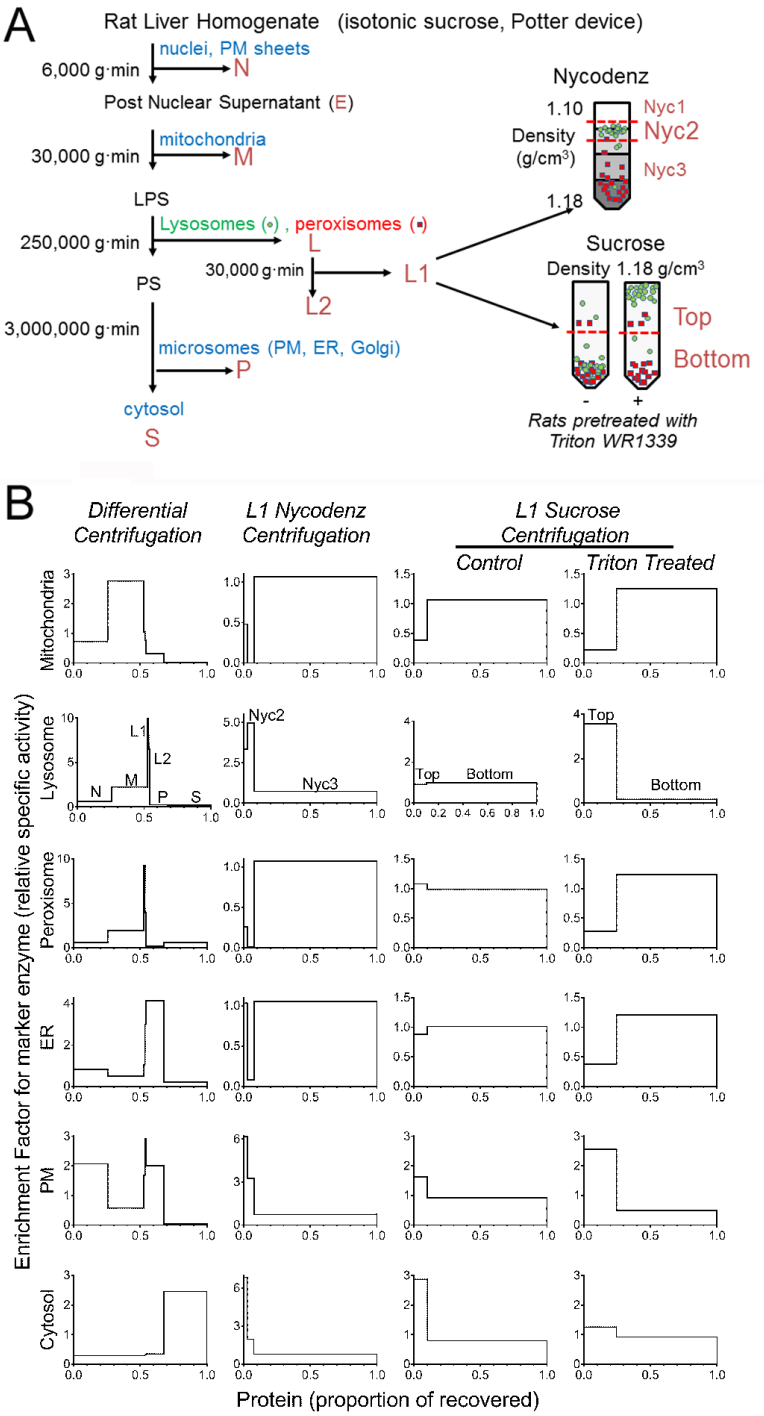

Fig 2. Distribution of marker proteins used for classification of subcellular compartments. (**A**), Expt A. (**B**), Expt B, Red lines, Profiles of individual marker proteins for indicated compartments (Supplemental Table 2) determined from MS data. Black-yellow dashed lines, consensus profiles for indicated compartments. Other lines, data from marker enzyme activity assays. **A**: black solid lines (mitochondria, cytochrome oxidase; lysosome, β-galactosidase; peroxisome, catalase; ER, alkaline α-glucosidase; PM, alkaline phosphodiesterase; cytosol, dipeptidyl-peptidase III) or in black dashed lines (lysosome, TPP1; peroxisome, urate oxidase; ER, NADH cytochrome c reductase). B: as above, except in the lysosome/peroxisome panel, the peroxisomal maker catalase is denoted by a blue solid line and urate oxidase was not quantified.

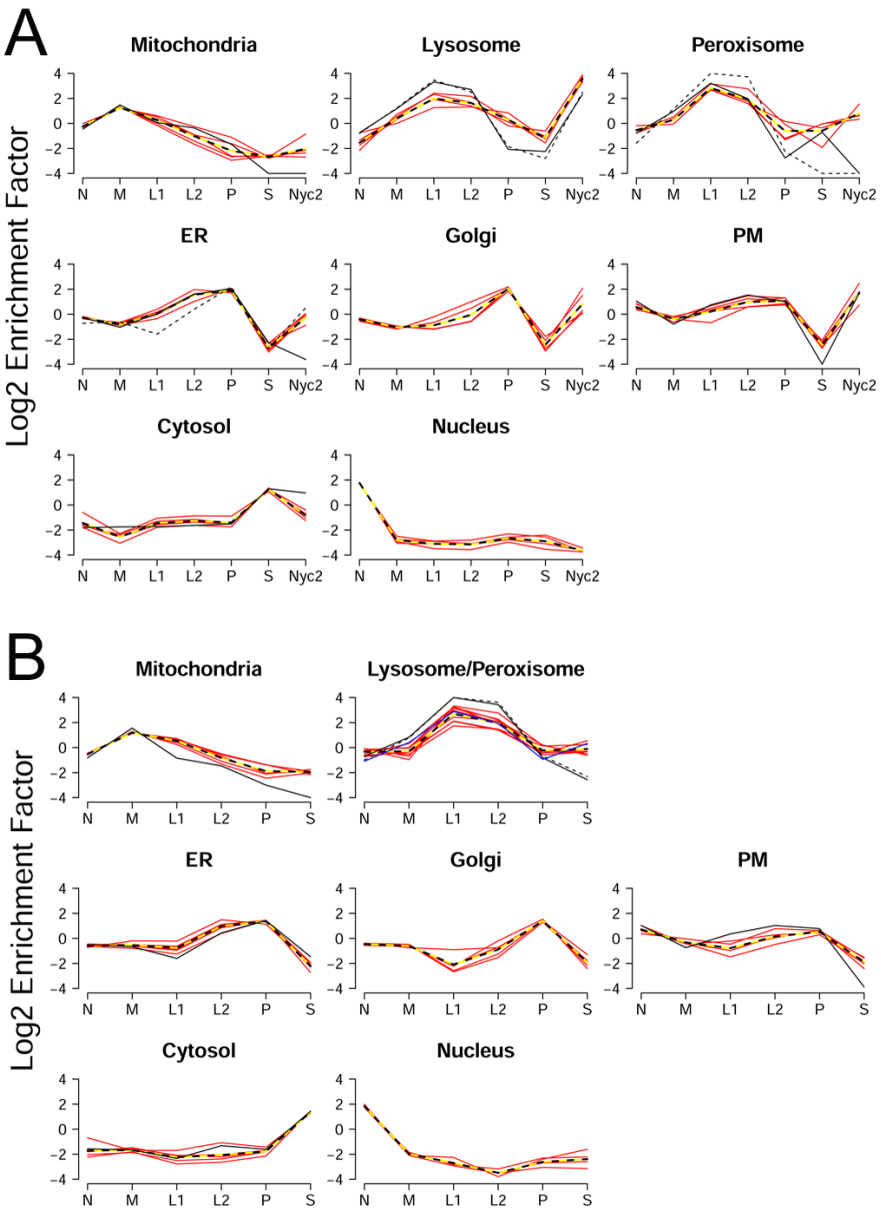

Fig. 3. Comparison of classification coefficient in Expts A and B. Point estimates for the 2952 proteins meeting the criteria of having ≥2 peptides and ≥3 spectra in each of the two experiments (Table 1) are plotted against each other and fit using linear regression. For Lyso+Perox, the sum of the two classification coefficients in Expt A is plotted against the combined lysosome/peroxisome classification coefficient from Expt B. For ER+Golgi+PM (microsomal distribution), the sum of the three classification coefficients is plotted against each other. For All*, the sum of the ER, Golgi and PM coefficients are used instead of the individual coefficients.

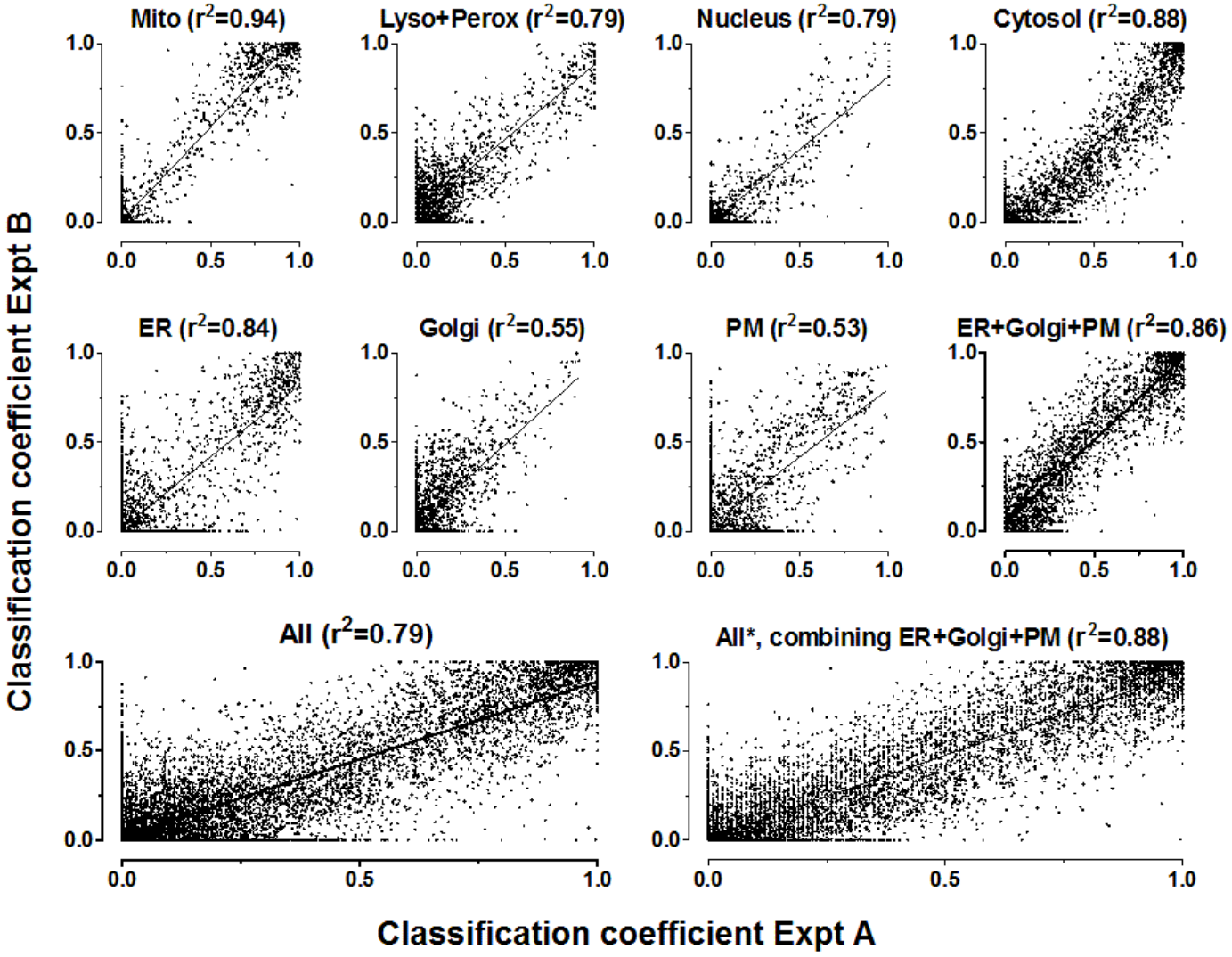

Fig. 4. Principal component analysis. Plots were made for all proteins with at least two peptides and three spectra in Expts A (**A**) and B (**B**). Proteins with point estimates <0.7 are shown in gray, all others are represented by indicated symbol. Plots were generated using the "prcomp" function in R version 3.2 to compute the principal components, with data centered and scaled to have a unit variance prior to the analysis. Explained variances are indicated on the plots.

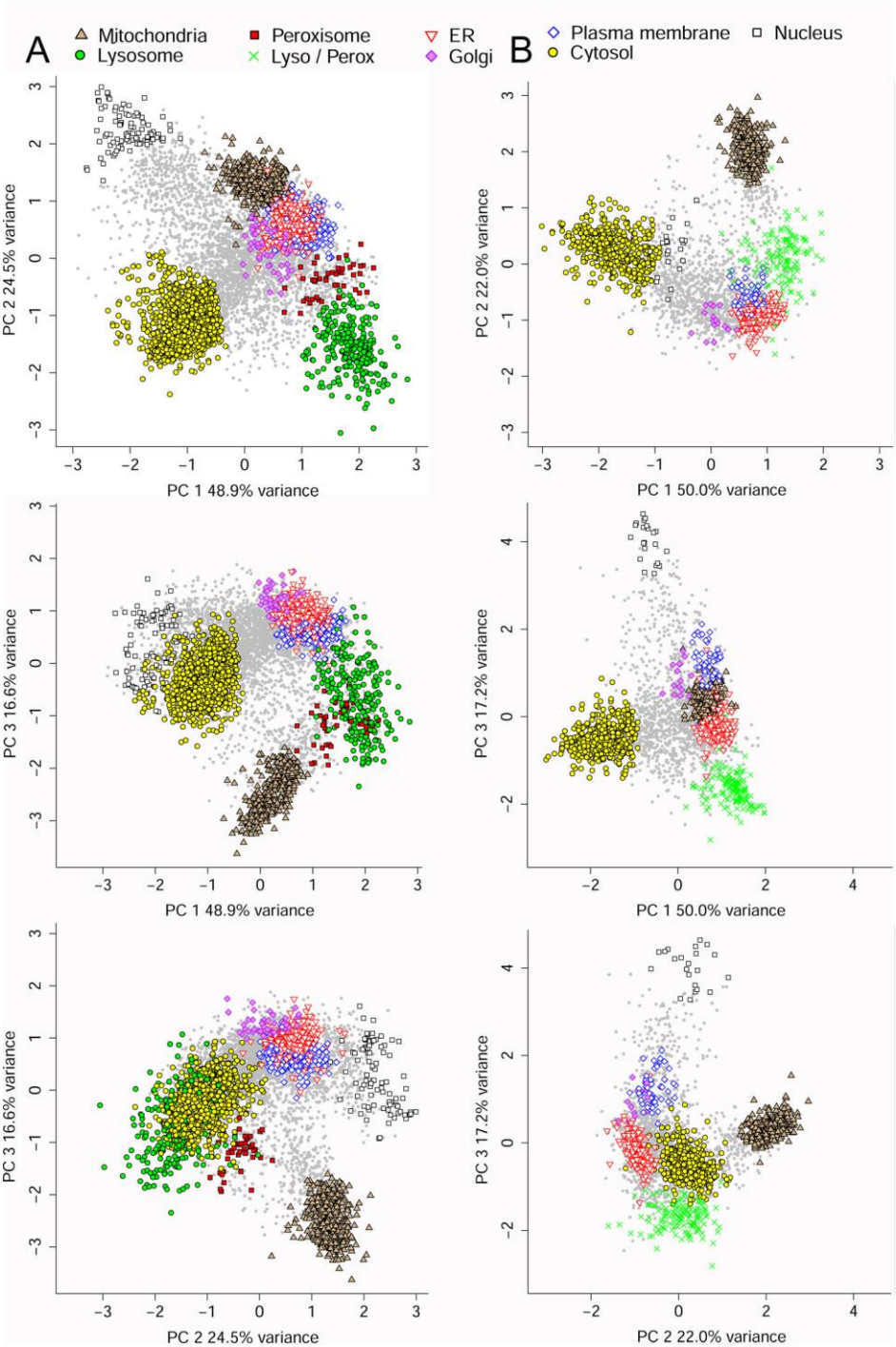

Fig. 5. Analysis of triton-shift experiments. (**A**) Distribution of select lysosomal, peroxisomal and mitochondrial proteins following sucrose density centrifugation of L1 fractions from control (C) and Triton–treated (T) rats. Proteins were selected as having ≥4 peptides in Expts C and D that have been localized to mitochondria (from MitoCarta v2.0 (29) with a FDR ≤0.001); lysosomes (from compendium reviewed in (53) omitting MPO (54)); and peroxisomes (PeroxisomeDB2.0 (32) omitting SOD1(17)). (**B**) Distribution of proteins with at least 2 peptides and 3 spectra (gray points) with other symbols indicating marker proteins from Supplemental Table 2. (**C**) As in Panel B, with other symbols indicating assignments from Expt A for proteins with a predominant compartment lower confidence interval ≥ 0.7 in Expts A and B. Lines indicate boundaries for proportional assignment between lysosomal and other locations (dashed line 1:1; dotted lines, 2:1 and 1:2). (**D**) Comparison of Expts C and D.

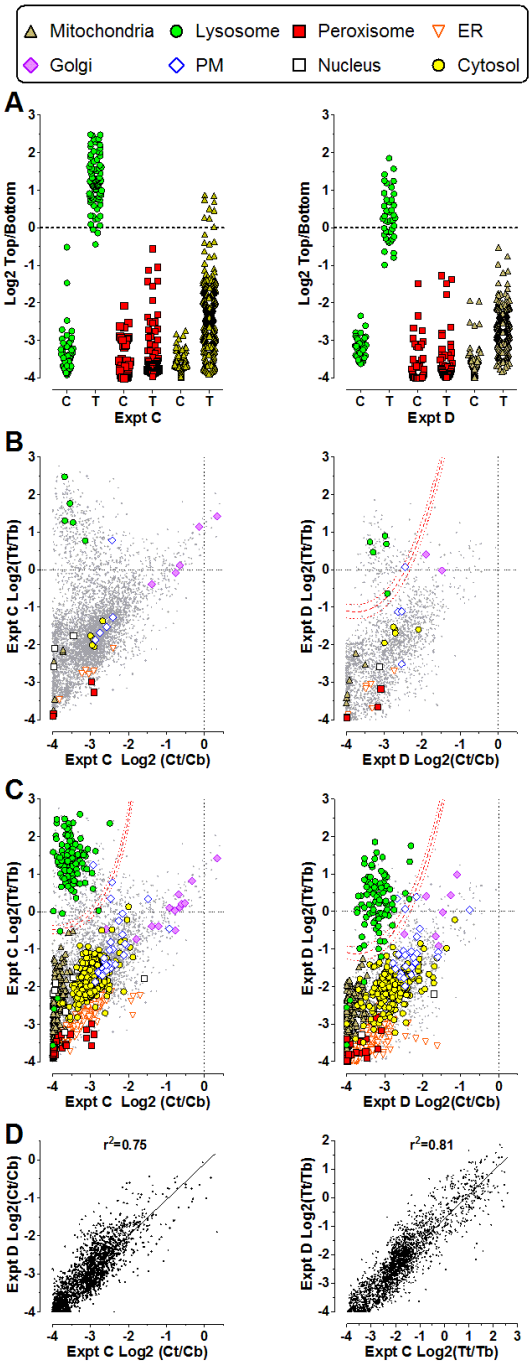

Fig. 6. Benchmarking protein location from the Prolocate high stringency dataset. (**A**) Left Panel, Receiver operating characteristic (ROC) curves showing ability of Expt A classification coefficients (symbols as in **Fig. 5**) to correctly predict location using proteins with single compartment assignment in the CDbBS (23) as reference sets. Sensitivity is true positive rate and 1-specificity is false positive rate. Dashed line indicate ROC curve for an uninformative test. (**B**) As above, using the curated CDbBS (see Text). Area under the curve for data sets shown in Panel A/ Panel B are as follows: Mitochondria, 0.9782 / 0.9994; Lysosome, 0.9930 / 1.000; Peroxisome, 0.9708 / 0.9758; ER, 0.9550 / 0.9978; Golgi, 0.8353 / 0.9995; PM, 0.8191 / 0.9996; Cytosol, 0.7499 / 0.9929; Nucleus, 0.6825 / 0.9964.

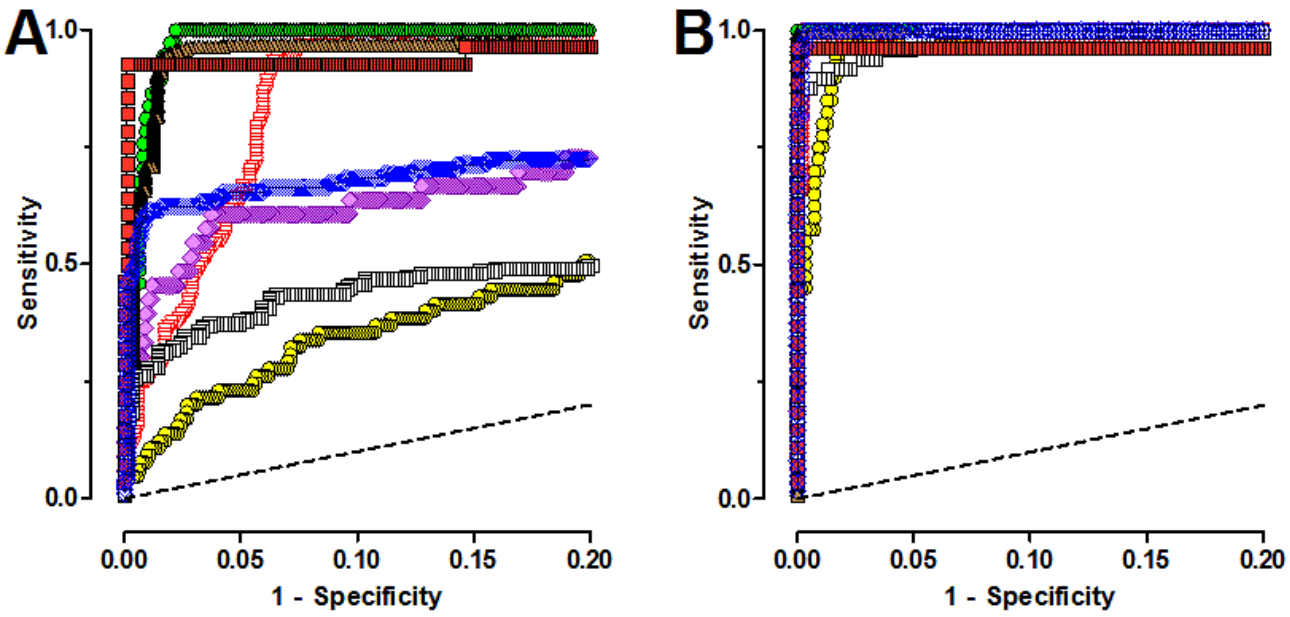

Fig. 7. Comparison of Prolocate high stringency data with other databases. (**A**) Mito classification coefficient for 2640 proteins listed in the MitoCarta 2.0 all mouse genes database (30) that are found in our high stringency classification set, broken down by MitoCarta false discovery rate (FDR); (**B-E**) All classification coefficient values are plotted for each protein found in our high stringency classification set that was found in the following databases: (**B and C**) MitoCarta 2.0 list of designated mitochondrial proteins with (**B**) FDR ≤ 0.1 (394 proteins) and (**C**) FDR > 0.1 (226 proteins); (**D**) MitoMiner, v3.1 - 2015_04 set using the IMPI filter (31) (998 proteins); (**E**) PeroxisomeDB v2.0 (32)(58 proteins); and (**F**) CLEAR network (35) (159 proteins). Red bars represent mean classification coefficient value.

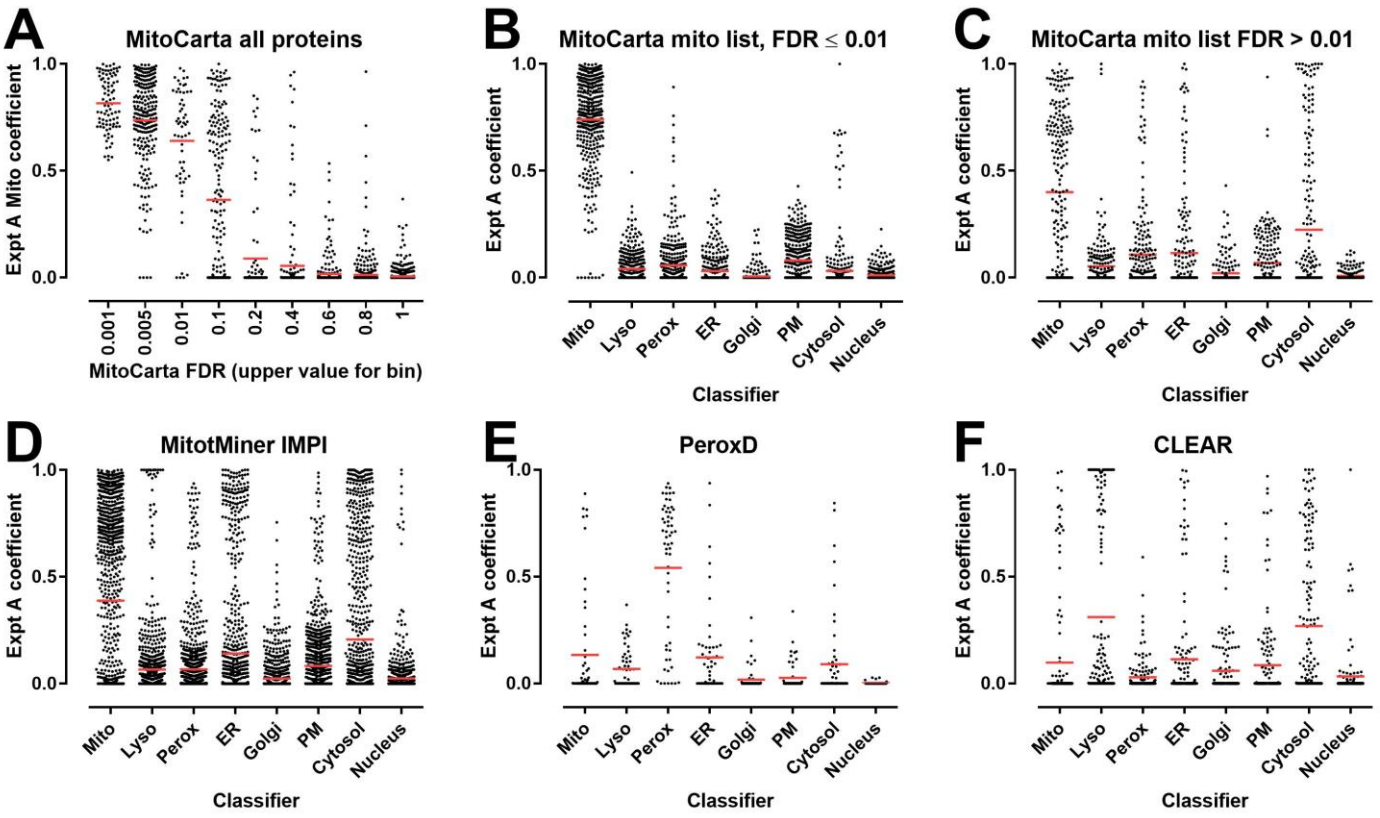

Fig. 8. Localization assignments for dynamic multiprotein complex proteins and a disease gene candidate. (**A**) COG proteins: COG1, open circles; COG2, filled circles; COG3, open squares; COG4, filled squares; COG5, open triangles; COG6, filled triangles; COG7, open diamonds; COG8, filled diamonds. (**B**) AP1 complex: AP1B1, open circles; AP1G2, filled circles; AP1M1, open squares; AP1S1, filled squares. AP2 complex: AP2A1, open circles; AP2A2, filled circles; AP2B1, open squares; AP2M1, filled squares; AP2S1, open triangles. AP3 complex: AP3B1, open circles; AP3B2, filled circles; AP3D1, open squares; AP3M1, filled squares; AP3S2, open triangles. AP5 complex: AP5B1, open circles; AP5M1, filled circles; AP5S1, open squares. (**C**) SLC31A1.

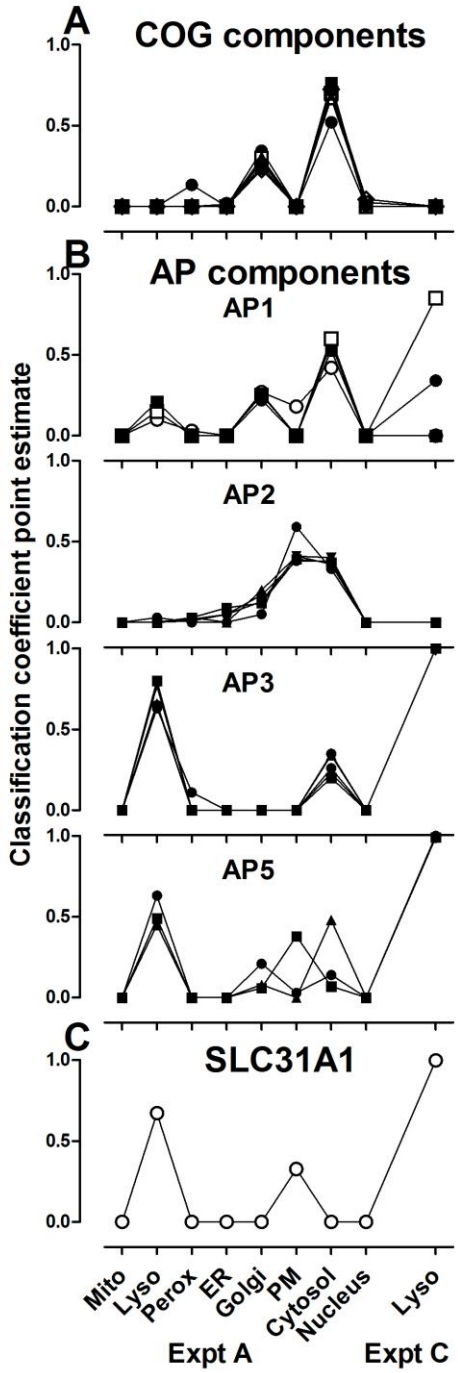

Fig. 9. Compound heterozygosity of *SLC31A1* in a patient with a previously unsolved lysosomal storage disease. Yellow arrows indicate direction of sequencing of plasmids containing PCR-amplified DNA from each mutant allele. Mutant and wild-type nucleotides on electropherograms are boxed in red and black, respectively. The schematic was generated by the UCSC genome browser Custom Tracks tool, with exons shown as thick solid boxes. Exome sequencing and confirmatory PCR/Sanger sequencing revealed that patient 82RD265 was heterozygous at two positions in *SLC31A1* (ENSG00000136868): 9,116021039 C/T, flanking sequence GCC [C/T]GA GAG); 9,116022721 G/T, flanking sequence GCA [G/T]TG GTA. These result in two changes in the protein: Arg90Stop and Val181Leu (numbering based on (ENSP00000363329). To determine if these were on same or different copy of the SLC31A1 gene, long range PCR was conducted to amplify 2.1 kb fragments (forward primer, CAAGCAGTCTGACCAAAAGGT; reverse primer, CTTCAACAACTTCCCACTGCA) containing both regions of interest. After subcloning, sequencing of individual plasmids indicated compound heterozygosity.