Resource

Cell

Global Survey of Organ and Organelle Protein Expression in Mouse: Combined Proteomic and Transcriptomic Profiling

Thomas Kislinger,[1,6] Brian Cox,[2,3,6] Anitha Kannan,[4,6] Clement Chung,[1] Pingzhao Hu,[1] Alexandr Ignatchenko,[1] Michelle S. Scott,[5] Anthony O. Gramolini,[1] Quaid Morris,[1,4] Michael T. Hallett,[5] Janet Rossant,[3] Timothy R. Hughes,[1,2] Brendan Frey,[1,4] and Andrew Emili[1,2,7,*]

[1] Banting and Best Department of Medical Research, University of Toronto, Toronto, ON M5G 1L6, Canada
[2] Department of Medical Genetics and Microbiology, University of Toronto, Toronto, ON M5S 1A8, Canada
[3] Department of Developmental Biology, The Hospital for Sick Children, Toronto, ON M5G 1L7, Canada
[4] Department of Computer Science, University of Toronto, Toronto, ON M5S 2E4, Canada
[5] Center for Bioinformatics, McGill University, Montreal, QC H3A 2B4, Canada
[6] These authors contributed equally to this work.
[7] Present address: Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, 160 College Street, Room 914, Toronto, ON M5S 3E1, Canada.
*Contact: andrew.emili@utoronto.ca
DOI 10.1016/j.cell.2006.01.044

SUMMARY

Organs and organelles represent core biological systems in mammals, but the diversity in protein composition remains unclear. Here, we combine subcellular fractionation with exhaustive tandem mass spectrometry-based shotgun sequencing to examine the protein content of four major organellar compartments (cytosol, membranes [microsomes], mitochondria, and nuclei) in six organs (brain, heart, kidney, liver, lung, and placenta) of the laboratory mouse, *Mus musculus*. Using rigorous statistical filtering and machine-learning methods, the subcellular localization of 3274 of the 4768 proteins identified was determined with high confidence, including 1503 previously uncharacterized factors, while tissue selectivity was evaluated by comparison to previously reported mRNA expression patterns. This molecular compendium, fully accessible via a searchable web-browser interface, serves as a reliable reference of the expressed tissue and organelle proteomes of a leading model mammal.

INTRODUCTION

Elucidation of gene-product function and regulation is a fundamental objective in human biology. It has become apparent that proper biological activity and cellular homeostasis depend on spatially and temporally restricted partitioning of functionally related sets of gene products. Organ- and organelle-selective protein accumulation represents one basic, conserved mode of biological control. Yet despite a relatively modest number (~25,000) of putative protein-coding genes (Lander et al., 2001; Margulies et al., 2005), much of the human proteome remains poorly annotated in terms of tissue- and organelle-selective expression. Knowledge of the global patterns of protein synthesis and subcellular localization across the major mammalian organ systems should therefore provide insight into the fundamental biological information encrypted in the human genome.

The recent completion of the genomic sequences of human and other mammalian species provides researchers with access to a wealth of relevant sequence information necessary for the functional characterization of gene products in a systematic and comprehensive manner. The use of tractable animal models, such as the laboratory mouse in particular, allows for investigation of the physiological roles, biochemical activities, and disease associations of evolutionarily conserved proteins on a genome-wide scale (Skarnes et al., 2004). Indeed, groundbreaking studies of global mRNA transcript patterns in mouse using DNA microarrays (Pan et al., 2004; Su et al., 2004; Zhang et al., 2004) have uncovered evidence of substantive tissue selectivity in terms of gene expression. Not all transcripts generate protein, however, and alternate translation efficiency and posttranslational turnover may result in differential protein accumulation. Certain proteins may also be transported between tissues, particularly those associated with circulatory or endocrine functions. These differences may underlie at least in part the modest correspondence reported between quantitative measurements of cognate gene transcript and protein levels (Griffin et al., 2002; Gygi et al., 1999), despite an obvious dependency of protein synthesis upon mRNA. Hence, the biological significance of differences in mRNA abundance detected

among tissues remains to be elaborated at the protein level.

One limitation of transcriptional profiling is that little information is gleaned with respect to the subcellular localization of the translated gene products. In contrast, an unbiased "subtractive proteomics" screening approach based on differential detection of proteins in isolated organellar compartments using high-throughput mass spectrometry offers the potential to determine subcellular enrichment directly (Andersen et al., 2002; Beausoleil et al., 2004; Krapfenbauer et al., 2003; Mootha et al., 2003; Nielsen et al., 2005; Schirmer et al., 2003; Wu et al., 2004). Perhaps because the complexity of the mammalian proteome is daunting (Aebersold and Mann, 2003), most proteomic studies published to date have, however, been focused on a single organelle or tissue, with only limited comparisons of the global patterns of protein expression and subcellular localization across tissues in an animal model setting. This contrasts with simpler systems like yeast, where proteomic methods examining protein expression and subcellular localization (Ghaemmaghami et al., 2003; Kumar et al., 2002; Washburn et al., 2001) have been applied successfully on a genome-wide scale.

To address this issue, we have performed an in-depth comparative proteomic analysis of the organelles of six representative mouse organs (adult brain, heart, kidney, liver, lung, and embryonic placenta). Computational and statistical procedures were used in combination with available conventional annotations and the observed proteomic profiles to create a high-quality reference map of the putative subcellular localizations and tissue selectivity of 4768 proteins. Crosscomparisons of the recorded proteomic patterns to the results of two analogous DNA microarray-based studies of global mRNA mouse tissue patterns (Su et al., 2004; Zhang et al., 2004) revealed broad areas of agreement with relatively few (albeit some notable) inconsistencies, confirming the context dependence of mammalian protein function. The entire collection of high-confidence protein profiles, including the primary supporting tandem mass spectra and database search results, is fully accessible through a searchable web-browser interface, allowing for convenient exploration of the biodistributive and colocalization properties of proteins of particular interest.

RESULTS

Proteomic Survey of Mouse Organs and Organelles

To assess tissue and organellar enrichment, we applied a comprehensive comparative proteomic profiling procedure (see Experimental Procedures) based on gel-free multidimensional protein identification technology (MudPIT) (Kislinger et al., 2003; Washburn et al., 2001). We examined the protein composition of four subcellular compartments (cytosol, membrane-derived microsomes, mitochondria, and nuclei) isolated by differential ultracentrifugation from healthy adult laboratory mouse brain, heart, kidney, liver, lung, and embryonic placenta. To compensate for the extreme sample complexity and large dynamic range in protein levels, we performed multiple (between 7 and 9) repeat profiling analyses on each fraction to improve detection coverage. The ~8 million spectra acquired during 203 MudPIT experiments were rigorously searched against a minimally redundant protein sequence database. As more accurate protein quantitation techniques on isotope labeling or extracted peptide ion signal correlation are not well suited for comparative analyses of broadly dissimilar samples and a project of this scope (Ong and Mann, 2005), relative abundance was estimated based on the cumulative number of high-confidence spectral matches recorded for a given protein across each fraction (Liu et al., 2004; Zybailov et al., 2005).

Essential to this screening process were reliable protein identifications. To estimate the rate of incorrect identifications (false positives), the database searches were also performed in parallel against an equivalent number of "decoy" protein sequences presented in inverted amino acid orientation (Kislinger et al., 2003; Peng et al., 2003). A stringent multistep filter was then applied to minimize invalid identifications (i.e., reverse sequences) while maintaining favorable detection of lower-abundance and smaller proteins (see Experimental Procedures). First, we used a rigorous statistical model (Kislinger et al., 2003) to assign a confidence score to each candidate peptide sequence match. Next, given that spurious identifications usually have limited supporting spectral evidence (see Figure S1 in the Supplemental Data available with this article online), we accepted only those proteins detected with a minimum of two or more high-scoring spectra (likelihood p value < 0.05). A final parsimonious interpretation of the combined search results led to a set of 4768 high-confidence protein identifications (Table S1), with an average of approximately 2000 proteins identified per tissue and ~1000 per organelle (Table 1). The vast majority (>85%) of these proteins were assigned probability scores >99% based on at least one unique (unambiguous) peptide sequence (Figure S2). The remaining spectra mapped to clusters of closely related protein isoforms (e.g., splice variants, paralogs, orthologs, or overlapping database entries). After filtering, only ~0.3% of the filtered spectra mapped to decoy proteins and the false positive rate was conservatively estimated to be <5% per tissue.

Detection Coverage

The majority of the identified proteins were highly enriched in a particular organelle and tissue (~75% and ~50%, respectively). Hierarchical clustering of the proteomic profiles (Figure 1) revealed distinct expression patterns, including broadly expressed (Figure 1A) and tissue-specific (Figure 1B) groupings. Protein membership within these clusters was enriched for select functional annotations and phenotypic associations. For example, a significant fraction of ubiquitously detected nuclear proteins were crossreferenced to the Gene Ontology (GO) annotation terms "DNA binding," "transcription," and/or "nucleus."

Table 1. Numerical Summary of the Proteomics Data

Organ	Organelle	Proteins	Spectra
Brain	total	2,243	90,456
	cytosol	1,366	37,813
	membrane	1,040	15,800
	mitochondrion	1,075	19,259
	nuclei	907	17,584
Heart	total	1,652	79,197
	cytosol	806	25,915
	membrane	702	16,162
	mitochondrion	667	15,621
	nuclei	1,044	21,499
Kidney	total	1,699	60,768
	cytosol	731	19,471
	membrane	608	11,182
	mitochondrion	789	14,019
	nuclei	796	16,096
Liver	total	1,728	71,172
	cytosol	739	23,411
	membrane	567	13,130
	mitochondrion	776	17,653
	nuclei	824	16,978
Lung	total	2,686	90,339
	cytosol	1,310	26,945
	membrane	1,669	30,414
	mitochondrion	1,072	15,566
	nuclei	1,452	17,414
Placenta	total	2,464	97,158
	cytosol	1,170	25,029
	membrane	1,162	20,204
	mitochondrion	901	25,273
	nuclei	1,135	26,652
All	total	4,768	489,090

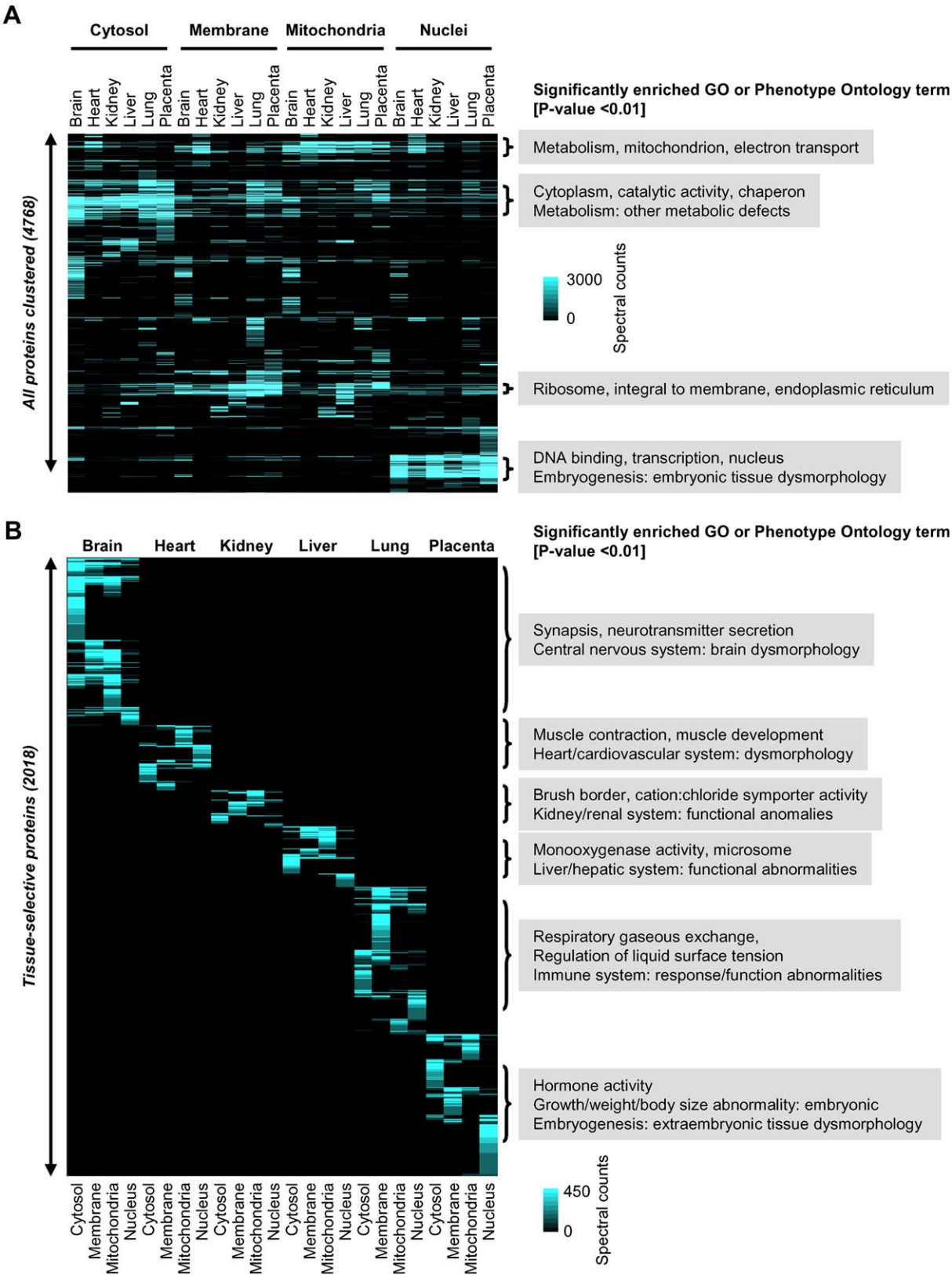
Cumulative number of high-confidence proteins and their associated spectral counts, identified in each of the six tissues and four organelles analyzed in this study.

Proteomic screening methods are known to preferentially detect higher-abundance proteins (Ghaemmaghami et al., 2003; Washburn et al., 2001). Although the MudPIT experiments were highly reproducible (Figure S3), repeat analysis of each fraction largely overcame the undersampling bias associated with the stochastic process of precursor peptide ion selection (Liu et al., 2004). Indeed, nearly saturating detection was evident (i.e., an asymptote or plateau was seen in plots of the cumulative number of proteins detected per fraction, as seen in Figure S4). Nevertheless, coverage was incomplete in that not all of the subunits of well-established multimeric protein complexes (e.g., RNA polymerase II), whose levels might be expected to be stoichiometric, were observed (Table S1), presumably due to a fundamental limitation in instrument sensitivity. Nevertheless, only a modest overall bias was evident in terms of sampling of different functional categories (i.e., GO terms) as assessed using the hypergeometric distribution (Table S2). The most notable exception was that proportionally fewer plasma-membrane proteins were identified than were expected relative to the predicted proteome. This bias may stem in part from overrepresentation of certain membrane-protein classes (e.g., odorant receptors) in the reference sequence database, as well as from inefficient recovery and/or ionization of hydrophobic, lower-abundance integral outer-membrane proteins such as transporters (Washburn et al., 2001).

To better evaluate the coverage achieved with membrane proteins, we deduced the occurrence of putative transmembrane helices (TMH) in the identified proteins (see Experimental Procedures). A total of 668 proteins had at least one well-defined TMH, while 244 proteins were predicted to contain two or more TMHs (Table S3 and Figure S5). Although more vigorous membrane-extraction protocols can improve global proteomic detection of integral membrane proteins (Wu et al., 2004), we concluded that reasonable coverage of membrane-associated proteins, especially internal vesicle bound factors, was achieved.

To more rigorously assess the overall detection coverage obtained by our profiling procedure, we compared our entire dataset of proteomic tissue patterns to the results of two recently published genome-scale surveys of mRNA transcript levels in mouse tissues. The gene expression study by Zhang et al. (2004) used high-density inkjet-synthesized long oligonucleotide microarrays, whereas the report by Su et al. (2004) was based on custom short oligonucleotide Affymetrix gene chips. Of the ~9000 highly correlated transcripts detected in the six organs by both microarray studies (Q.M., T.R.H., and B.F., unpublished data), 1758 gene products were likewise detected in common across all three platforms in a three-way crossmapping (Table S4). Although it appears highly unlikely that these microarrays detected every transcript expressed in these six tissues (Bertone et al., 2004), these data imply that substantive (albeit incomplete) proteomic sensitivity was indeed achieved.

Much of the incomplete coverage of the proteome likely arose from intrinsic limitations in instrument sensitivity, which is biased toward the detection of more abundantly expressed proteins. However, hundreds of presumably lower-abundance proteins, such as sequence-specific transcription factors, protein kinases, and intracellular signaling molecules, were successfully identified (Table S1). The coverage may also have been limited in part due to an overly stringent filtering of the database search results,


A
Cytosol
Membrane
Mitochondria
Nuclei
Brain
Heart
Kidney
Liver
Lung
Placenta
All proteins clustered (4768)
Significantly enriched GO or Phenotype Ontology term [P-value <0.01]
Metabolism, mitochondrion, electron transport
Cytoplasm, catalytic activity, chaperon
Metabolism: other metabolic defects
3000
0
Spectral counts
Ribosome, integral to membrane, endoplasmic reticulum
DNA binding, transcription, nucleus
Embryogenesis: embryonic tissue dysmorphology
B
Brain
Heart
Kidney
Liver
Lung
Placenta
Tissue-selective proteins (2018)
Significantly enriched GO or Phenotype Ontology term [P-value <0.01]
Synapsis, neurotransmitter secretion
Central nervous system: brain dysmorphology
Muscle contraction, muscle development
Heart/cardiovascular system: dysmorphology
Brush border, cation:chloride symporter activity
Kidney/renal system: functional anomalies
Monooxygenase activity, microsome
Liver/hepatic system: functional abnormalities
Respiratory gaseous exchange,
Regulation of liquid surface tension
Immune system: response/function abnormalities
Hormone activity
Growth/weight/body size abnormality: embryonic
Embryogenesis: extraembryonic tissue dysmorphology
450
0
Spectral counts
Cytosol
Membrane
Mitochondria
Nucleus

resulting in a significant false-negative rate. Consistent with this, the number of protein identifications could be boosted by ~12% (to 5373 candidate proteins) simply by accepting tentative database matches with marginal (subthreshold) probabilities (i.e., between 85% and 95% initial likelihood scores) if the corresponding mRNA was likewise jointly detected by the two microarray studies (Figure S6). Alternatively, many lower-level transcripts may not be efficiently translated, or the resulting proteins may be unstable or become secreted or modified in such a way as to make them unrecognizable by spectral searches against a primary sequence database.

Correspondence and Differences between mRNA and Protein Expression Landscapes

The conventional procedure for comparing mRNA and protein abundance has generally been to determine the correlation coefficient (e.g., Pearson or Spearman) between respective expression profiles (Cox et al., 2005). Previous efforts to analyze noisy data with simple correlation metrics have resulted in positive but weak associations (Griffin et al., 2002; Le Roch et al., 2004; Lian et al., 2001), while analyses with more robust statistics have yielded stronger correlations (Gygi et al., 1999). In computing a correlation score, it is generally assumed that gene-product measurements are noise free and follow a normal distribution. These assumptions were not valid in our case; in particular, the spectral counts were discrete and not continuous (as demanded for fitting of a normal distribution) and were markedly skewed in distribution. Therefore, we modeled a Bayesian network to decrease the effect of residual noise and better evaluate the concordance between the mRNA (Zhang et al., 2004) and protein patterns recorded by the two respective platforms (see Experimental Procedures).

Our model was based on the assumption that transcript levels (as measured by probe intensity) are correlated linearly with protein abundance (as measured by filtered spectral counts), as suggested from double log-plots of putative mRNA and protein levels recorded for each tissue (Figure S7). However, unlike traditional correlation analysis, our approach handles measurement uncertainty by modeling noise in the mRNA levels with a Gaussian distribution and in the spectral counts with a Poisson distribution. The model also uses a background distribution to discount unreliable measurements by explicitly explaining mRNA levels independently of the observed spectral counts. The output of the learned Bayesian network is a probability score indicating the strength of the linear relationship between cognate gene-product pairs based on the respective tissue profiles (Table S5). Permutation testing was then performed to determine statistical significance. An important advantage of the model is that arbitrary thresholds are not needed to decide on the closeness of fit. Whereas the Pearson correlation requires a predefined correlation threshold, our model provides a more rigorous statistical cutoff for establishing departure from concordance while at the same time determining the false discovery rate.

Contrary to general expectation (Griffin et al., 2002; Gygi et al., 1999), the overall concordance between the protein and mRNA tissue patterns (Zhang et al., 2004) was found to be conspicuously good (Figure 2). Of the 1758 cross-mapped proteins classified by our approach (Table S5), only 503 pairs of gene product were deemed to be statistically significant "outliers" (not linearly correlated) after permutation testing (Figure 2, bottom panel). The rest were considered to be either highly correlated "inliers" (Figure 2, top panel), wherein the transcript patterns were highly indicative of the corresponding tissue protein levels (409 gene products), or "midliers" (Figure 2, middle panel; 846 gene products), where the gene-product patterns appear similar but did not achieve statistical significance (i.e., did not pass permutation testing). Figure S8 and Table S6 present a comparison of our approach and the results (which are largely in agreement) of traditional (Pearson correlation) methodologies for inferring the relationships between the protein and mRNA tissue profiles.

Several of the outliers were blood-borne factors (e.g., complement), which showed the highest mRNA probe signal intensity in liver (the primary site of synthesis prior to secretion into the circulation), whereas the corresponding proteins were preferentially detected in the lung and placenta (which are rich in blood vessels). Although specious, these classifications confirm the validity of our model. The other uncorrelated gene products were enriched for nuclear and mitochondrial proteins. Mann and colleagues (Mootha et al., 2003) have previously noted an incomplete correspondence between the relative abundance of mitochondrial proteins and the corresponding mRNA transcripts across mouse tissues. Some of the discrepancy may be due to our examination of female mice exclusively, whereas Zhang et al. (2004) and Su et al. (2004) reportedly used both males and females in their microarray analyses.

Although various outliers were detected with low spectral counts and/or weak probe intensities, making the apparent discordance suspect, plausible biologically interesting outliers were also observed. These include cytochrome P450 isoform 4B1 (CP4B1), whose transcript was detected preferentially in kidney and only weakly in lung, whereas the cognate protein was more abundant in pulmonary microsomes and virtually absent in kidney (cf. Table S5), as reported previously (Imaoka et al., 1995).

Figure 1. Mouse Organ and Organelle Protein Expression Patterns

(A) Hierarchical clustering of the proteomic profiles based on the cumulative spectral counts detected in each organelle. A selection of significantly enriched Gene Ontology (GO) and Phenotype Ontology terms are displayed.

(B) Heat-map display of clusters of tissue- and organelle-selective protein expression.

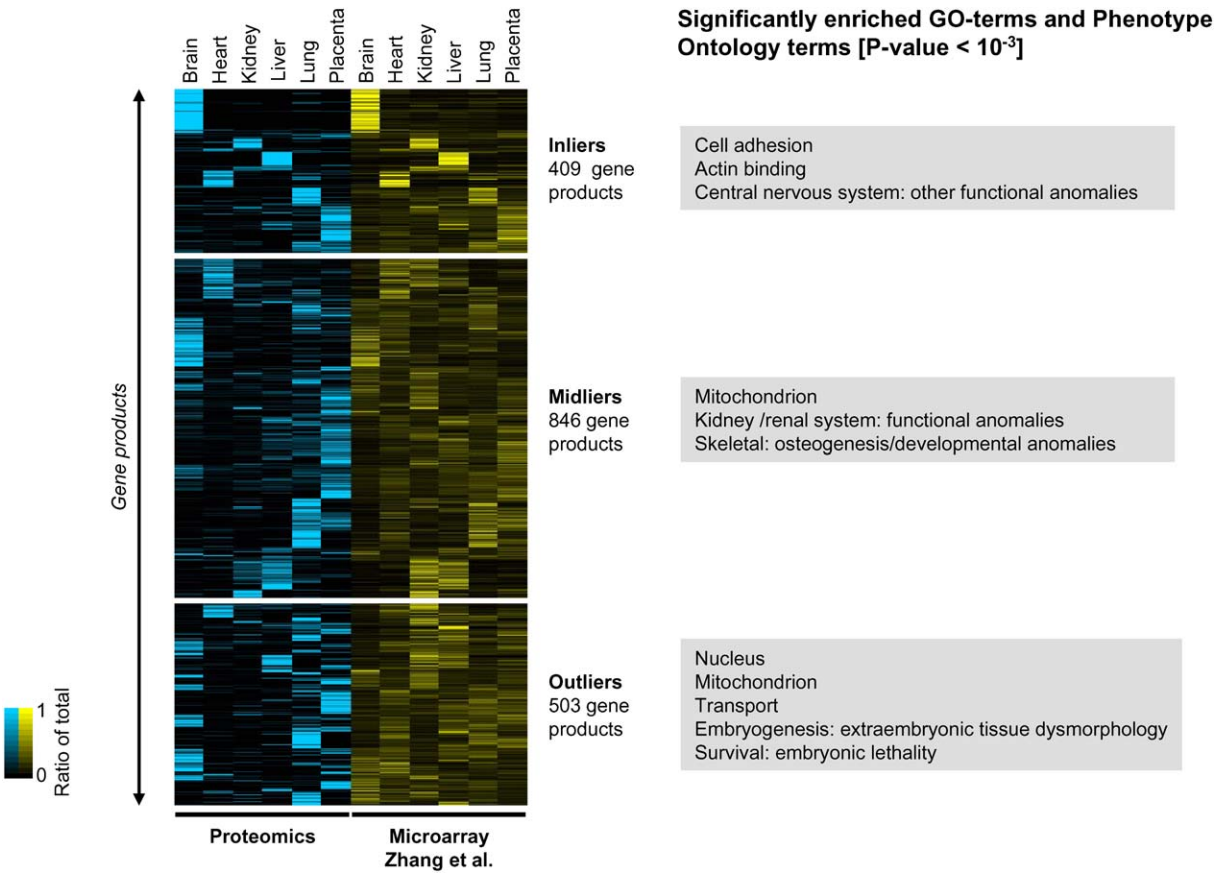


Figure 2. Concordance between mRNA and Protein Tissue Patterns

Comparison of protein (this study) and reproducible mRNA (Zhang et al., 2004) levels recorded for a subset of 1758 gene products detected in a three-dataset crossmapping. The heat maps indicate various model predictions: inliers, highly correlated (linear fit) gene products with a significant p value (409); midliers, ambiguous gene products that do not deviate from the linear model in a statistically significant manner (846); and outliers, uncorrelated gene products (503). A selection of enriched ontology terms mapped to each category is listed.

Subcellular Localization

Hierarchical clustering of the protein profiles revealed striking differences between the four organelles (Figure 3). Most (>75%) proteins were preferentially detected in a single compartment, suggesting a specialized biological role. Membership in these clusters was likewise enriched for relevant functional categories. Western blot experiments (Figure S9) confirmed the appropriate partitioning of several well-studied organellar markers across the four fractions, providing a basic confirmation of biochemical purity. Nevertheless, to verify the reliability of these organellar maps, we benchmarked our data against previously reported proteomic analyses of highly purified preparations of mammalian organelles. These included analogous large-scale (albeit less comprehensive) surveys of mouse mitochondria (Mootha et al., 2003), human nuclei and nucleoli (Andersen et al., 2002; Beausoleil et al., 2004), rat cytosol (Krapfenbauer et al., 2003), and discrete mammalian membrane fractions (Nielsen et al., 2005; Schirmer et al., 2003; Wu et al., 2004)(see Table S7 for a complete crosslisting). As expected, the respective subcellular patterns were highly correlated (Figure 3). For example, the majority (~65%) of 357 putative mitochondrial proteins (Mootha et al., 2003) and 446 putative nucleolar proteins (Andersen et al., 2002) identified in our study were preferentially detected in the mitochondrial and nuclear fractions, respectively. These results confirm that reliable inferences regarding subcellular localization can be achieved by differential proteomic comparisons, as previously reported (Schirmer et al., 2003).

Nearly half (2390) of the identified proteins had not been previously assigned to an organelle (based on annotation obtained from the ExPASy web server; Table S8), indicating that our study provides the first experimental evidence for the primary subcellular localization of these proteins in a cell. The modest inconsistencies observed between our proteomic patterns and the literature (Figure 3) could reflect several factors, including inaccurate existing annotations, shuttling of certain proteins between compartments, and residual crosscontamination by higher-abundance proteins (e.g., mitochondrial) (cf. Figure 3), a possibility we address next.

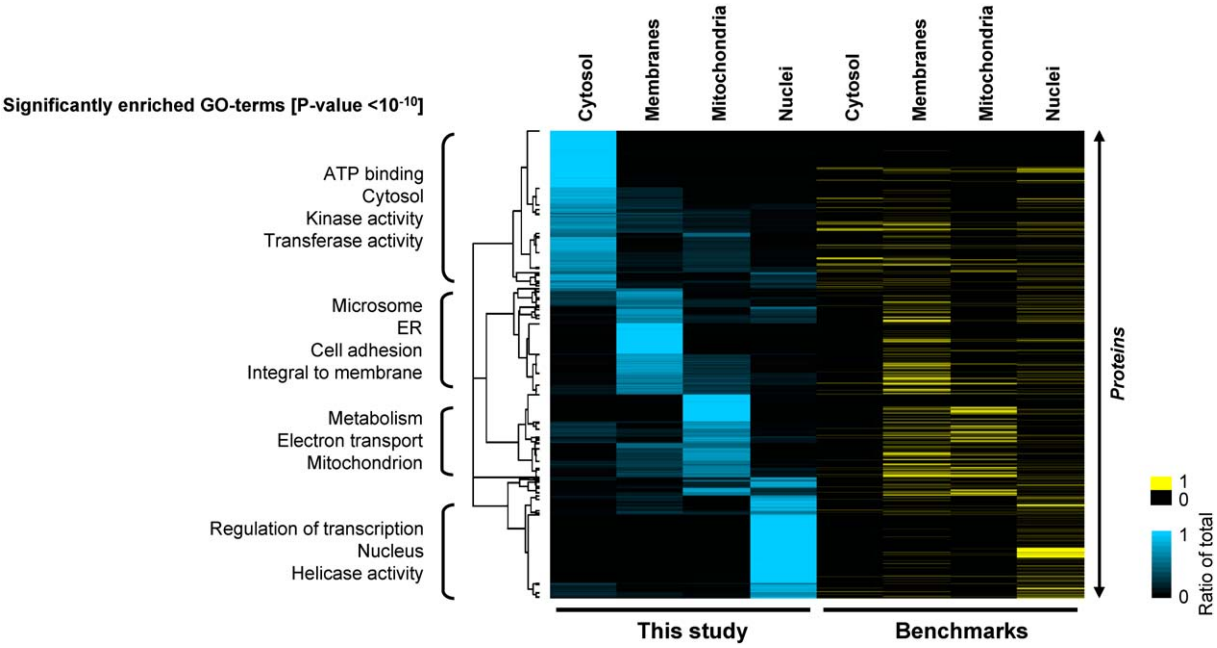


Figure 3. Concordance between Subcellular Location and Benchmark Data Sets

Comparison of the proteomic patterns obtained for the cytosolic, membrane, mitochondrial, and nuclear fractions against benchmark "gold standard" previously reported reference proteomic datasets representing similar compartments: cytosol, rat cytosol (Krapfenbauer et al., 2003); membranes, discrete mammalian membrane fractions (Nielsen et al., 2005; Schirmer et al., 2003; Wu et al., 2004); mitochondria, mouse mitochondria (Mootha et al., 2003); nuclei, human nuclei and nucleoli (Andersen et al., 2002; Beausoleil et al., 2004).

Computational Refinement of Subcellular Localization by Machine-Learning Classifiers

Given the possibility of interorganelle crosscontamination, we used machine-learning techniques to assign a primary subcellular localization and associated confidence score to each of the proteins. Various supervised computational classification approaches, including K-nearest neighbor (KNN) (Cai and Chou, 2004; Huang and Li, 2004), support vector machine (SVM) (Park and Kanehisa, 2003), and Bayesian methods (Lu et al., 2004; Scott et al., 2004), have been used to evaluate protein subcellular localization. We used a weighted variant of the KNN algorithm (WKNN) (see Experimental Procedures), which in our hands generated the most reliable classifications based on a panel of standard statistical performance metrics (P.H., unpublished data). However, there are no firmly established measures for assessing multiplex classifications (i.e., proteins present in multiple compartments) (Chou and Cai, 2005), an issue not fully addressed in previous computational studies of subcellular localization (Cai and Chou, 2004; Huang and Li, 2004; Lu et al., 2004; Park and Kanehisa, 2003). We therefore used a conservative implementation (see Experimental Procedures) to assign a probability to each protein for a given compartment based on a weighted similarity of its proteomic profile to its K-nearest neighbors in a training set of 1558 proteins with known (previously established) localizations (i.e., available SwissProt annotations) (Table S9). We evaluated classifier performance both by 10-fold crossvalidation and by using a separate holdout "gold standard" set of 820 reference proteins previously identified by proteomic screening in a single highly purified organelle (Table S10).

High prediction precision (>77%) and accuracy (>66%) as well as sensitivity and specificity (with the exception of the membrane microsomes) were obtained for both test sets as assessed using receiver operating characteristics (ROC) plots (Figure S10). Based on these classifiers, over two-thirds (3274) of the remaining proteins could be confidently (i.e., with a minimum probability of 80%) assigned to at least one subcellular compartment (Figure 4A and Figure S11). These included 1503 proteins of previously unknown localization, of which 458 were projected to be cytosolic, 553 membrane bound, 60 mitochondrial, and 480 nuclear (bold numbers in Table S8). Only 47 proteins were confidently assigned to more than one compartment, possibly reflecting confounding crosscontamination by higher-abundance mitochondrial factors.

Our assignments compared favorably with predictions produced by the alternate PSLT algorithm (Figure 4A), a Bayesian network predictor that uses orthogonal structural features present in primary protein sequences (i.e., motif occurrence) to forecast subcellular localization (Scott et al., 2004). Moreover, the localizations were largely consistent with biological expectation. For instance, many of the nuclear-specific proteins (Figure 4B) had functional and structural domains consistent with

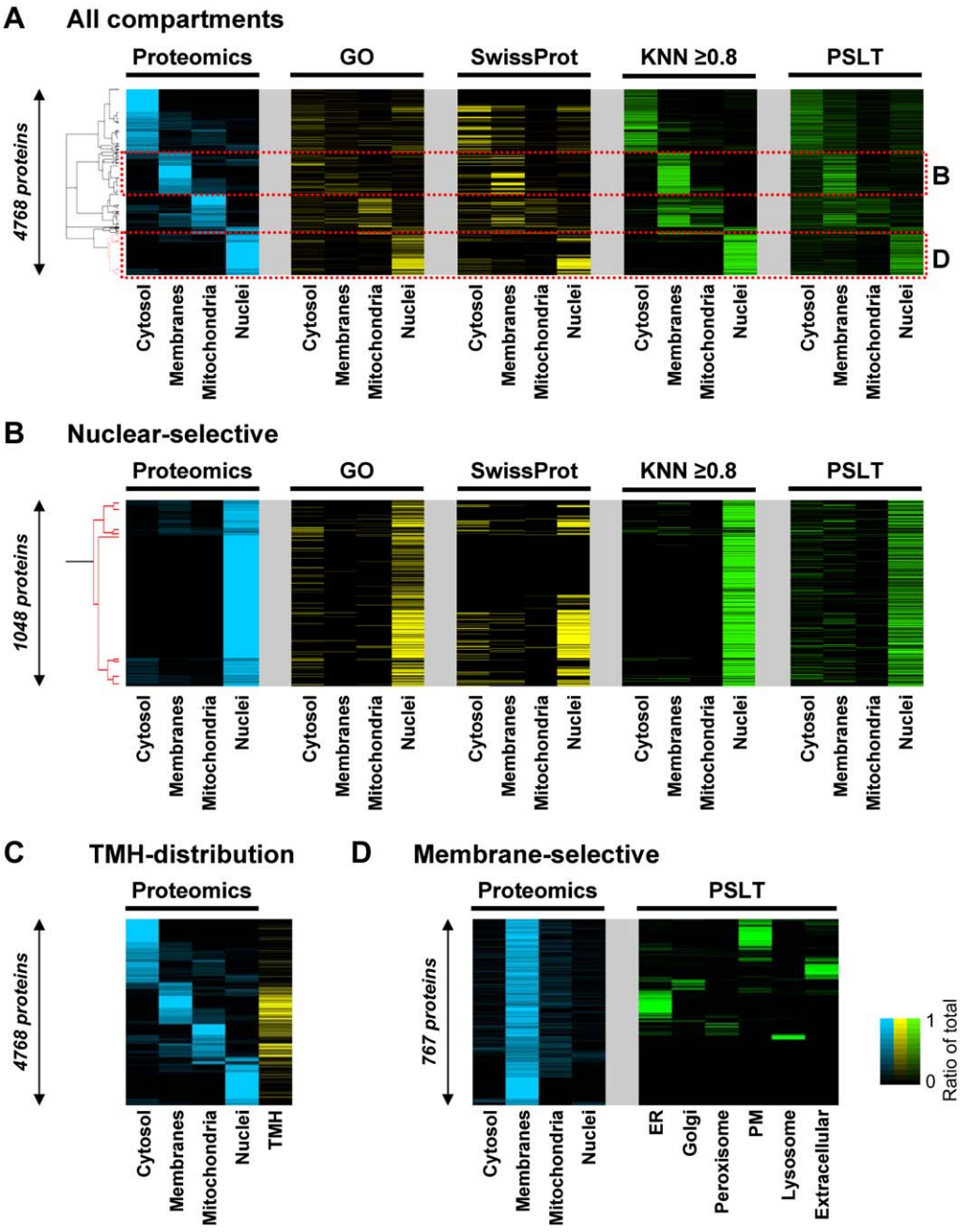


Figure 4. Annotation and Prediction of Subcellular Localization

(A) Comparative clustergrams of the normalized organelle proteomic profiles together with annotations obtained from the Expasy web portal (SwissProt) or GO database and subcellular predictions made using the WKNN (see Experimental Procedures) and PSLT (Scott et al., 2004) machine-learning algorithms.

(B) Zoom-in of a cluster of 1048 putative nuclear-selective proteins extracted from (A) (red highlight).

(C) Distribution of predicted transmembrane helices (TMH) across the organelles.

(D) Subcompartment assignments generated for a panel of 767 putative membrane proteins (highlighted in [A]) based on the application of the PSLT algorithm.

a nuclear-related function (e.g., RNA or DNA binding motifs). Conversely, a sizeable fraction of the proteins preferentially detected in the microsomal fractions had predicted TMH (Figure 4C), suggesting they were indeed membrane bound. Consistent with this, over half (505) of the putative membrane-associated proteins had

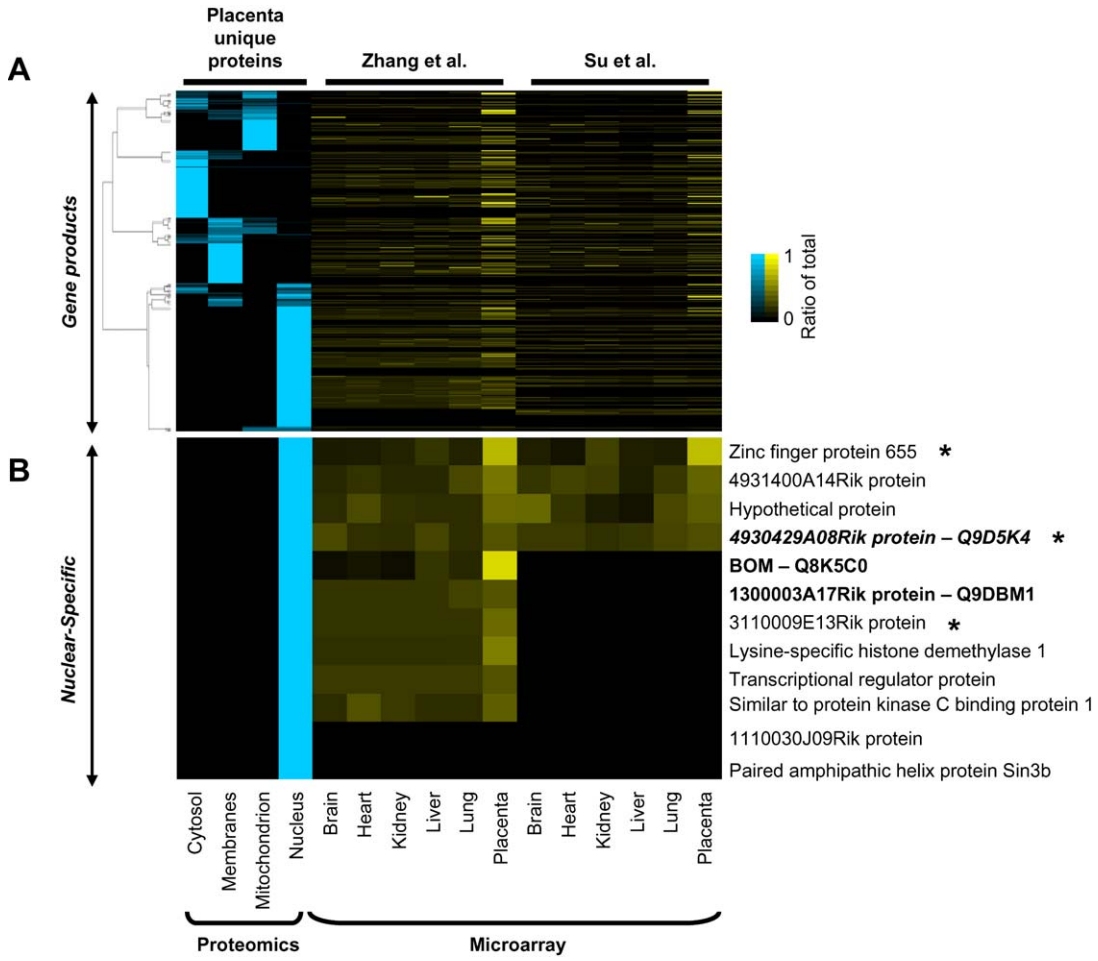


Figure 5. Mining the Proteomics Data for Tissue-Selective Expression Patterns

(A) A cluster of 462 putative placenta-selective proteins together with available microarray-recorded mRNA tissue patterns (Su et al., 2004; Zhang et al., 2004).

(B) Selection of candidate novel nuclear-localized placental proteins (bold, mentioned in text; *, validated by GFP-fusion imaging).

additional structural properties consistent with a specific membrane-related subcompartment (such as the endoplasmic reticulum or Golgi apparatus) as determined using the PSLT algorithm (Figure 4D).

Mining Tissue- and Organelle-Selective Expression Patterns

Roughly half of all functionally uncharacterized proteins were detected both in a single tissue and organelle (Table S1). Evidence of subcellular and tissue selectivity can be used to generate hypotheses regarding their biological role. For instance, several novel nuclear-localized factors were present in a cluster of 462 proteins identified exclusively in placenta (Figure 5A), suggesting a role in extraembryonic development and/or angiogenesis. Most of these gene products either were unannotated (Figure 5B) or were not previously associated functionally with this tissue (Rossant and Cross, 2001). These include Q9DBM1, an evolutionarily conserved protein composed of D111/G patch domains implicated in nucleic acid binding and mRNA processing (Kawai et al., 2001); Q9D5K4, a member of a small family of mammalian-specific proteins with homologs in human, chimpanzee, dog, and opossum but not other vertebrates (Kawai et al., 2001); and Q8K5C0, a homolog of the *Drosophila* CP2-like transcription factor *Grainyhead/Mindbomb*, whose corresponding transcript was detected exclusively in placenta out of 54 mouse tissues analyzed by microarray (Zhang et al., 2004). These data imply that proteins with specialized functions can be recognized via their proteomic profiles.

Validating Novel Expression Results by GFP Labeling

As an independent validation of our subcellular assignments, we used confocal microscopy to image the localization of several uncharacterized target proteins, similar to previous large-scale proteomic studies (Mootha et al.,

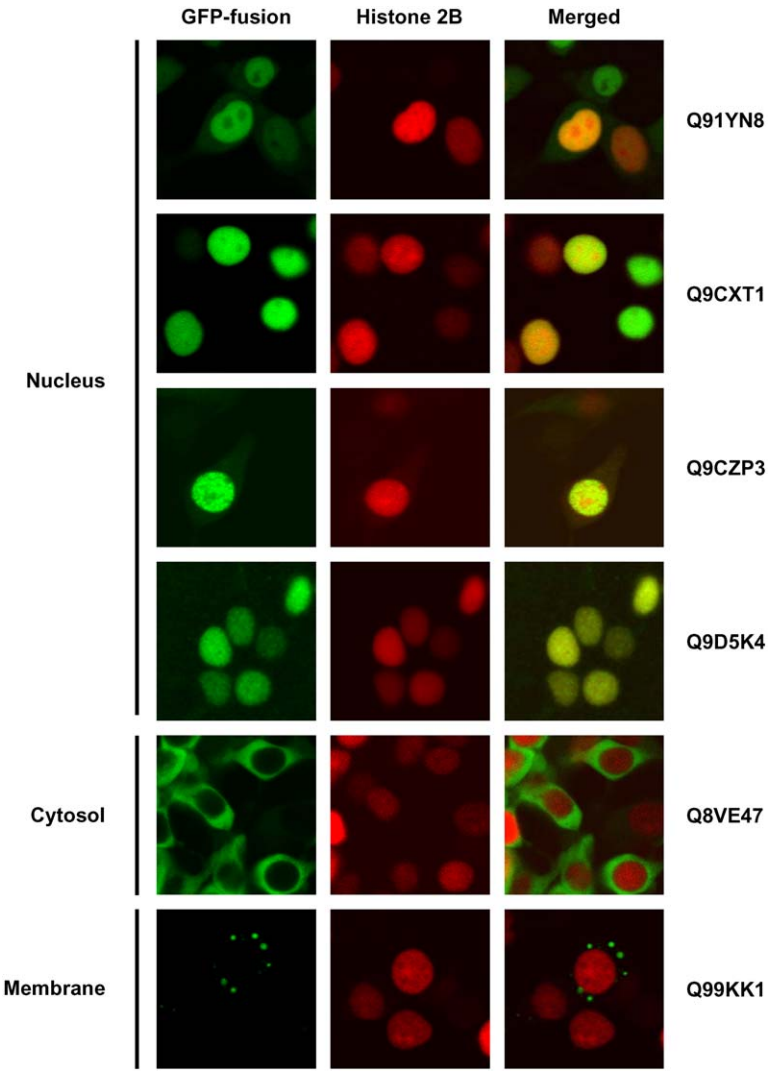


Figure 6. Validating Novel Cellular Localizations by GFP Labeling

Confocal microscopy images of select, previously unannotated proteins expressed as GFP fusions in HEK293T cells (left panel). The constructs were cotransfected along with red-fluorescent-protein-labeled histone 2B as a nuclear marker (middle panel). A merged image is shown on the right-hand panel. Yellow indicates colocalized signals.

2003; Schirmer et al., 2003). For these experiments, we generated sequence-verified clones for ten representative gene products as N-terminal fusions to a green fluorescent marker protein (GFP) and examined their corresponding localization patterns in transient transfections of HEK293T cells. This panel of proteins had strong KNN predictions (five nuclear, two cytoplasmic, two membrane associated, and one mitochondrial; Table S12) but were lacking previous experimentally derived localizations (although several had inferred [electronic] GO annotations, these had not been verified). A histone 2B red fluorescent fusion protein was cotransfected as a nuclear-specific positive control.

Representative images generated for six of these proteins are shown in Figure 6. The results were largely consistent with expectation (although three of the fusions examined could not be unambiguously localized; Table S12). Four of the five putative nuclear proteins were detected exclusively in the nuclei, whereas the other (Q8K335) localized alternately to the transfected nuclei and/or cytoplasm, possibly due to differences in the cell cycle or signaling. In contrast, a putative ubiquitin-activating E1 enzyme homolog (Q8VE47) displayed diffuse cytoplasmic staining, consistent with its ubiquitous detection in the cytosolic fractions of all six tissues, while a putative membrane protein (Q99KK1) detected in lung and placenta was found to localize to discrete intracellular vesicles, possibly secretory granules.

A Community Resource

The entire proteomic dataset reported here is fully accessible to the scientific community via a dedicated web-based database with an easily navigated graphical interface (http://tap.med.utoronto.ca/~mts/). Users can peruse the entire collection of expression profiles obtained for proteins of special interest, accessing complete details of the database search results, the filtered tissue and organelle spectral counts, and the high-confidence

subcellular assignments. All of the primary data can be freely viewed and downloaded, including appropriately labeled visual representations of individual MS/MS spectra together with their respective peptide matches and the corresponding confidence scores. Different search criteria can be used, including browsing by tissue or organelle specificity or based on protein descriptions, GO functional annotation, or similarity to an input protein sequence. Finally, a "bulk" search option is provided for querying longer lists of proteins based on either systematic (SwissProt) names or accession numbers.

DISCUSSION

Eukaryotic cells are generally highly structured, with dedicated subsets of functionally related proteins organized into discrete compartments to provide an optimal context for cellular processes to occur. Systematic elucidation of tissue and organelle expression patterns in a mammalian model therefore provides for a first-pass assessment of the biological roles and molecular functions of evolutionarily conserved proteins on a genome-wide basis. Although this concept has been exploited before in functional genomics studies (Zhang et al., 2004), the consistency of mRNA patterns recorded using different microarray platforms has proven to be less than absolute (Kuo et al., 2002). Hence, we have performed a large-scale proteomic survey of mouse tissue using a rigorous comparative profiling strategy based on a relatively unbiased and sensitive method of detection (i.e., MudPIT) to examine differential protein expression directly (Kislinger et al., 2003; Schirmer et al., 2003; Washburn et al., 2001). Our study builds on a substantive existing body of targeted proteomic studies in mammalian systems (Andersen et al., 2002; Beausoleil et al., 2004; Krapfenbauer et al., 2003; Mootha et al., 2003; Nielsen et al., 2005; Schirmer et al., 2003; Wu et al., 2004) and provides a complementary perspective into the functional organization and regulation of mammalian gene products.

One of the main outstanding questions in expression profiling is how well mRNA levels reflect protein abundance and the biological basis (if any) for any observable differences. Despite the obvious fact that protein synthesis is dependent upon mRNA, earlier studies of the relationships between mRNA and protein profiles have consistently reported a modest correlation between mRNA and protein levels (Griffin et al., 2002; Gygi et al., 1999; Mootha et al., 2003). However, the conclusions drawn from previous reports were generally based on computational methods that may not have fully accounted for systematic or spurious noise. Using a probabilistic framework to better model the relationship between the experimentally recorded protein and mRNA patterns, we have now largely confirmed the overall good concordance of tissue expression patterns of gene products reproducibly detected by microarray-based (Su et al., 2004; Zhang et al., 2004) and proteomic (this study) global profiling procedures. Although our experimental method provided for only a semiquantitative estimate of relative protein abundance (Liu et al., 2004), the overall correspondence between pairs of cognate mRNA and protein profiles was quite impressive, with only ~1/4 of all gene products exhibiting a statistically significant departure from a simple linear relationship at the predicted protein and transcript levels. Some of the remaining discordance likely stems from irrelevant epiphenomena (e.g., different mouse genetic backgrounds) or residual differences in data signal processing (Larkin et al., 2005), but it may also point to interesting posttranscriptional control mechanisms. Nevertheless, incomplete proteome/transcriptome coverage stemming from sheer sample complexity, unknown protein modifications, and poor recovery and detection of lower-abundance and membrane-associated proteins still confounds rigorous definition of the expressed proteome. These problems are also compounded by a dependency on public sequence databases, which are incomplete and often contain errors, for mass spectrometry-based proteomic screening.

Organ-selective gene products can potentially be used as biomarkers to monitor homeostatic perturbations associated with tissue-specific pathologies, such as heart disease, neurological disorders, and cancer. One unique advantage of proteomic measurements over mRNA profiling is the ability to deduce protein subcellular localization directly, providing additional insight into the biological context of uncharacterized gene products that can lead naturally to testable hypotheses regarding function. As the isolation of completely pure organelles is notoriously difficult (Brunet et al., 2003), we opted to combine differential proteomic detection with machine-learning methods to more accurately deduce the primary subcellular localization, benchmarking our results against established (e.g., SwissProt) annotations and alternate hypotheses (Andersen et al., 2002; Beausoleil et al., 2004; Krapfenbauer et al., 2003; Mootha et al., 2003; Nielsen et al., 2005; Schirmer et al., 2003; Wu et al., 2004). The high-confidence assignments for 1503 previously unassigned proteins reported here add substantively to our knowledge of the organization of the organellar proteomes of a leading mammalian model. Nevertheless, ~1/3 of the proteins identified (1494) were assigned to organelles with confidence scores below our threshold cutoff (likelihood < 80%). Much of this ambiguity stems from proteins identified with low spectral counts, ubiquitous organellar distributions, or differences in the organellar patterns among the six tissues.

Despite the fact that many proteins likely shuttle between compartments or have multiple (i.e., pleiotropic) roles in the cell, relatively few proteins could be unambiguously assigned to more than one compartment (aside from cases of probable crosscontamination). These results highlight the ongoing challenges of rigorously defining subcellular localization (Phizicky et al., 2003). While such patterns may indeed be reflected in the raw proteomic datasets, we chose to be cautious in our current interpretation of the data.

These statistics might, however, be enhanced by applying improved forms of pattern recognition. Computational methods for predicting subcellular localization generally fall into one of three categories based on either amino acid composition (Nakashima and Nishikawa, 1994), sequence-derived parameters integrating literature-derived rules (e.g., PSORT; Nakai and Kanehisa, 1992), or sequence homology (Chou and Cai, 2005; Lu et al., 2004; Mott et al., 2002). It is possible that integrating aspects of these alternate approaches together with the proteomic profiles reported here might allow for more complete and accurate classifications, ideally without the bias toward monocompartment predictions.

Despite these caveats, the protein patterns reported here should serve as a useful bridgehead for more extensive experimental characterization of core mammalian biological systems, including relatively poorly defined organs like the placenta and the mechanisms controlling protein expression, stability, and organellar trafficking. By providing unfettered access to the data via a web portal, investigators are encouraged to navigate and contemplate this proteomic landscape.

EXPERIMENTAL PROCEDURES

Tissue Fractionation and Organelle Isolation

The preparation of mouse tissue organelle fractions was as previously described (Kislinger et al., 2003). For detailed protocols, see Supplemental Experimental Procedures.

Mass Spectrometry and Database Searches

The protein fractions were denatured and digested sequentially with endoprotease Lys-C and trypsin and analyzed by data-dependent shotgun (MudPIT) profiling as previously reported (Kislinger et al., 2003). Full details of the entire procedure are provided in Supplemental Experimental Procedures.

Quantitative Analysis

The profiles were clustered based on Spearman correlation ranking with average linkage using Cluster 3.0 (Eisen et al., 1998), as modified by de Hoon and colleagues, and visualized using TreeView (Saldanha, 2004). Protein relative abundance was inferred either using raw spectral counts as a semiquantitative measure (Figures 1A and 1B) or after normalizing the spectral counts per fraction relative to the total recorded per protein (Figures 2–5; "Ratio of total") essentially as previously described (Cox et al., 2005). Functional classification and statistical enrichment were evaluated using an in-house program (MouseSpec; available upon request). Annotations were compiled from the GO and Expasy websites (Table S11). Phenotype Ontology terms were obtained from the Mouse Genome Informatics database (http://www.informatics.jax.org/)

Microarray Dataset Crosscomparison

Global mRNA expression profiles (Su et al., 2004; Zhang et al., 2004) were crossmapped and linked to the proteomics data via SwissProt accession numbers. Only the ~9000 closely correlated transcripts (Q.M., T.R.H., and B.F.; unpublished data) were used for further consideration.

Mathematical Modeling

We took a probabilistic approach to model the relationship between the protein and mRNA tissue patterns. A detailed description is provided in Supplemental Experimental Procedures. Briefly, we used an automatically inferred Bernoulli switch variable that directs toward explaining microarray expression levels of mRNA (probe intensities) either as a linear function of the spectral counts or independently of the counts using a background distribution learned on the mRNA expressions alone using kernel density estimation. The spectral counts were assumed to be Poisson distributed, while the mRNA measurements were modeled as a Gaussian function. The learned model was used to score the strength of relationship between the tissue profiles for each pair of gene products on a gene-by-gene basis. Permutation testing was performed to assign a confidence measure (p value) based on the possibility of observing an extreme probability value with randomized data. Finally, we applied a rigorous probability cutoff of >0.66 to select inliers (positive for a linear relationship), and <0.33 for outliers (negative for a linear relationship), with the remaining gene products (intermediate probabilities $0.33 \leq p \leq 0.66$) classified as ambiguous midliers.

Prediction of Subcellular Localization and Annotation of Organelle Localization

We used a kernel-based variant of the classic KNN algorithm (Hechenbichler and Schliep, 2004) to build the localization classifiers. A detailed description of the training, testing, and prediction process is provided in Supplemental Experimental Procedures. Alternate predictions using the PSLT algorithm based on protein domain architecture (i.e., combinatorial presence of predicted InterPro motifs and putative membrane-spanning domains) were generated following training on mammalian protein sequences in the Hera database as previously described (Scott et al., 2004) using motifs defined in InterPro release 8.0 (Mulder et al., 2005), signal peptides/anchors predicted by SignalP version 3.0 (Bendtsen et al., 2004), and transmembrane domains as deduced by TMHMM version 2.0 (Krogh et al., 2001).

Cloning, Expression, and Imaging

Commercially available plasmids bearing full-length cDNAs of interest were ordered as bacterial glycerol stocks from Open Biosystems (Huntsville, AL, USA). Oligonucleotide primers were designed to amplify the open reading frame from the start codon to the last amino acid, removing the stop codon. Restriction sites were embedded into the primers to facilitate subcloning. PCR was performed with a high-fidelity enzyme (BD Biosciences, Advantage 2), and the end products were TA cloned into TOPO (Invitrogen) or pGEM (Promega; pGEM-T) for dideoxy sequencing. Sequence-verified cDNA clones were subcloned in frame into a vector encoding a C-terminal GFP fusion (Clontech; catalog #6085-1).

Human embryonic kidney 293T cells were plated 24 hr prior to transfection onto gelatin-treated 35 mm glass-bottom culture dishes (MatTek; P35G-0-10-C) to achieve an ~50%–80% confluency. Each GFP fusion plasmid (~0.75 μg) and a histone H2B-RFP control construct (a kind gift from Sean Megason) were cotransfected using FuGENE 6 (Roche). The cells were cultured for a further 24 hr prior to imaging. The cells were live imaged by confocal microscopy using a Zeiss Axiovert 200M inverted microscope fitted with an LSM 510 META confocal system. Channels were sequentially scanned and images collected for each fluorophore using 25× and 40× objectives.

Supplemental Data

Supplemental Data include Supplemental Experimental Procedures, Supplemental References, 12 tables, and 12 figures and can be found with this article online at http://www.cell.com/cgi/content/full/125/1/173/DC1/.

ACKNOWLEDGMENTS

This study was supported in part by funds to A.E. from the McLaughlin Centre for Molecular Medicine, Genome Canada and the Ontario Genomics Institute (OGI), and the National Science and Engineering Council of Canada (NSERC). B.C. was supported by a Canadian

Institutes of Health Research Doctoral Fellowship. We thank Matthew Chow, Rahim Sajoo, and Vincent Fong for expert support with computing and Amy Ralston for assistance with confocal imaging.

Received: September 5, 2005
Revised: December 21, 2005
Accepted: January 26, 2006
Published: April 6, 2006

REFERENCES

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature *422*, 198–207.

Andersen, J.S., Lyon, C.E., Fox, A.H., Leung, A.K., Lam, Y.W., Steen, H., Mann, M., and Lamond, A.I. (2002). Directed proteomic analysis of the human nucleolus. Curr. Biol. *12*, 1–11.

Beausoleil, S.A., Jedrychowski, M., Schwartz, D., Elias, J.E., Villen, J., Li, J., Cohn, M.A., Cantley, L.C., and Gygi, S.P. (2004). Large-scale characterization of HeLa cell nuclear phosphoproteins. Proc. Natl. Acad. Sci. USA *101*, 12130–12135.

Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. *340*, 783–795.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. Science *306*, 2242–2246.

Brunet, S., Thibault, P., Gagnon, E., Kearney, P., Bergeron, J.J., and Desjardins, M. (2003). Organelle proteomics: looking at less to see more. Trends Cell Biol. *13*, 629–638.

Cai, Y.D., and Chou, K.C. (2004). Predicting subcellular localization of proteins in a hybridization space. Bioinformatics *20*, 1151–1156.

Chou, K.C., and Cai, Y.D. (2005). Predicting protein localization in budding yeast. Bioinformatics *21*, 944–950.

Cox, B., Kislinger, T., and Emili, A. (2005). Integrating gene and protein expression data: pattern analysis and profile mining. Methods *35*, 303–314.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*, 14863–14868.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737–741.

Griffin, T.J., Gygi, S.P., Ideker, T., Rist, B., Eng, J., Hood, L., and Aebersold, R. (2002). Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. Mol. Cell. Proteomics *1*, 323–333.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. *19*, 1720–1730.

Hechenbichler, K., and Schliep, K.P. (2004). Weighted k-nearest-neighbor techniques and ordinal classification. Discussion paper *399*, SFB *386*, Ludwig-Maximilians University, Munich. http://www.stat.uni-muenchen.de/sfb386/papers/dsp/paper399.ps.

Huang, Y., and Li, Y. (2004). Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics *20*, 21–28.

Imaoka, S., Hiroi, T., Tamura, Y., Yamazaki, H., Shimada, T., Komori, M., Degawa, M., and Funae, Y. (1995). Mutagenic activation of 3-methoxy-4-aminoazobenzene by mouse renal cytochrome P450 CYP4B1: cloning and characterization of mouse CYP4B1. Arch. Biochem. Biophys. *321*, 255–262.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., et al. (2001). Functional annotation of a full-length mouse cDNA collection. Nature *409*, 685–690.

Kislinger, T., Rahman, K., Radulovic, D., Cox, B., Rossant, J., and Emili, A. (2003). PRISM, a generic large scale proteomic investigation strategy for mammals. Mol. Cell. Proteomics *2*, 96–106.

Krapfenbauer, K., Fountoulakis, M., and Lubec, G. (2003). A rat brain protein expression map including cytosolic and enriched mitochondrial and microsomal fractions. Electrophoresis *24*, 1847–1870.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. *305*, 567–580.

Kumar, A., Agarwal, S., Heyman, J.A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., et al. (2002). Subcellular localization of the yeast proteome. Genes Dev. *16*, 707–719.

Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L., and Kohane, I.S. (2002). Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics *18*, 405–412.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Larkin, J.E., Frank, B.C., Gavras, H., Sultana, R., and Quackenbush, J. (2005). Independence and reproducibility across microarray platforms. Nat. Methods *2*, 337–344.

Le Roch, K.G., Johnson, J.R., Florens, L., Zhou, Y., Santrosyan, A., Grainger, M., Yan, S.F., Williamson, K.C., Holder, A.A., Carucci, D.J., et al. (2004). Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. Genome Res. *14*, 2308–2318.

Lian, Z., Wang, L., Yamaga, S., Bonds, W., Beazer-Barclay, Y., Kluger, Y., Gerstein, M., Newburger, P.E., Berliner, N., and Weissman, S.M. (2001). Genomic and proteomic analysis of the myeloid differentiation program. Blood *98*, 513–524.

Liu, H., Sadygov, R.G., and Yates, J.R., 3rd. (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal. Chem. *76*, 4193–4201.

Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D.S., Poulin, B., Anvik, J., Macdonell, C., and Eisner, R. (2004). Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics *20*, 547–556.

Margulies, E.H., Vinson, J.P., Miller, W., Jaffe, D.B., Lindblad-Toh, K., Chang, J.L., Green, E.D., Lander, E.S., Mullikin, J.C., and Clamp, M. (2005). An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc. Natl. Acad. Sci. USA *102*, 4795–4800.

Mootha, V.K., Bunkenborg, J., Olsen, J.V., Hjerrild, M., Wisniewski, J.R., Stahl, E., Bolouri, M.S., Ray, H.N., Sihag, S., Kamal, M., et al. (2003). Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. Cell *115*, 629–640.

Mott, R., Schultz, J., Bork, P., and Ponting, C.P. (2002). Predicting protein cellular localization using a domain projection method. Genome Res. *12*, 1168–1174.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., et al. (2005). InterPro, progress and status in 2005. Nucleic Acids Res. *33*, D201–D205.

Nakai, K., and Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics *14*, 897–911.

Nakashima, H., and Nishikawa, K. (1994). Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J. Mol. Biol. *238*, 54–61.

Nielsen, P.A., Olsen, J.V., Podtelejnikov, A.V., Andersen, J.R., Mann, M., and Wisniewski, J.R. (2005). Proteomic mapping of brain plasma membrane proteins. Mol. Cell. Proteomics *4*, 402–408.

Ong, S.E., and Mann, M. (2005). Mass spectrometry-based proteomics turns quantitative. Nat. Chem. Biol. *1*, 252–262.

Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A.L., Mohammad, N., Babak, T., Siu, H., Hughes, T.R., Morris, Q.D., et al. (2004). Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol. Cell *16*, 929–941.

Park, K.J., and Kanehisa, M. (2003). Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics *19*, 1656–1663.

Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J., and Gygi, S.P. (2003). Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J. Proteome Res. *2*, 43–50.

Phizicky, E., Bastiaens, P.I., Zhu, H., Snyder, M., and Fields, S. (2003). Protein analysis on a proteomic scale. Nature *422*, 208–215.

Rossant, J., and Cross, J.C. (2001). Placental development: lessons from mouse mutants. Nat. Rev. Genet. *2*, 538–548.

Saldanha, A.J. (2004). Java Treeview–extensible visualization of microarray data. Bioinformatics *20*, 3246–3248.

Schirmer, E.C., Florens, L., Guan, T., Yates, J.R., 3rd, and Gerace, L. (2003). Nuclear membrane proteins with potential disease links found by subtractive proteomics. Science *301*, 1380–1382.

Scott, M.S., Thomas, D.Y., and Hallett, M.T. (2004). Predicting subcellular localization via protein motif co-occurrence. Genome Res. *14*, 1957–1966.

Skarnes, W.C., von Melchner, H., Wurst, W., Hicks, G., Nord, A.S., Cox, T., Young, S.G., Ruiz, P., Soriano, P., Tessier-Lavigne, M., et al. (2004). A public gene trap resource for mouse functional genomics. Nat. Genet. *36*, 543–544.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA *101*, 6062–6067.

Washburn, M.P., Wolters, D., and Yates, J.R., 3rd. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. *19*, 242–247.

Wu, C.C., MacCoss, M.J., Mardones, G., Finnigan, C., Mogelsvang, S., Yates, J.R., 3rd, and Howell, K.E. (2004). Organellar proteomics reveals Golgi arginine dimethylation. Mol. Biol. Cell *15*, 2907–2919.

Zhang, W., Morris, Q.D., Chang, R., Shai, O., Bakowski, M.A., Mitsakakis, N., Mohammad, N., Robinson, M.D., Zirngibl, R., Somogyi, E., et al. (2004). The functional landscape of mouse gene expression. J. Biol. *3*, 21.

Zybailov, B., Coleman, M.K., Florens, L., and Washburn, M.P. (2005). Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. Anal. Chem. *77*, 6218–6224.