We have developed a systematic analytical approach, termed PRISM (Proteomic Investigation Strategy for Mammals), that permits routine, large scale protein expression profiling of mammalian cells and tissues. PRISM combines subcellular fractionation, multidimensional liquid chromatography-tandem mass spectrometry-based protein shotgun sequencing, and two newly developed computer algorithms, STATQUEST and GOClust, as a means to rapidly identify, annotate, and categorize thousands of expressed mammalian proteins. The application of PRISM to adult mouse lung and liver resulted in the high confidence identification of over 2,100 unique proteins including more than 100 integral membrane proteins, 400 nuclear proteins, and 500 uncharacterized proteins, the largest proteome study carried out to date on this important model organism. Automated clustering of the identified proteins into Gene Ontology annotation groups allowed for streamlined analysis of the large data set, revealing interesting and physiologically relevant patterns of tissue and organelle specificity. PRISM therefore offers an effective platform for in-depth investigation of complex mammalian proteomes.

The laboratory mouse is a powerful model organism for investigating fundamental aspects of mammalian cell physiology, development, and disease (1), and it is currently the focus of systematic efforts aimed at large scale gene prediction and functional annotation (2–4). The use of oligonucleotide- and cDNA-based microarrays, in particular, is providing unprecedented insight into the regulation of global patterns of gene expression (5). Nevertheless, since protein abundance does not always correlate with transcript levels (6) and since the subcellular localization and turnover rate of biologically active protein can only be determined directly, the development of sensitive, accurate methods for comprehensive analysis of cellular protein expression patterns in mouse and other mammals is broadly needed.

Two-dimensional polyacrylamide gel electrophoresis has been the traditional method of choice for high resolution proteome analysis (7, 8). Despite recent advances, this approach is biased against membrane-associated proteins, low abundance proteins, or proteins with extremes in isoelectric point or molecular weight (9, 10). The identification of gel-separated proteins by mass spectrometry (MS) is also tedious due to the need to extract, digest, and analyze individual gel spots. Consequently techniques for gel-free chromatographic separation of protein or peptide mixtures coupled to on-line MS detection are currently in development. One promising method, based on multidimensional capillary-scale liquid chromatography-electrospray ionization tandem MS (LC-MS) protein identification technology (MudPIT) pioneered by Yates and colleagues (11–13), permits shotgun sequencing of large numbers of proteins present in cell extracts. MudPIT has been applied successfully to several model organisms, leading to the identification of 1,484 proteins in yeast (12), 2,363 proteins in rice (14), and, most recently, 2,415 proteins in Plasmodium (15). Other powerful gel-free approaches, such as the use of accurate mass tag detection by Fourier transform ion cyclotron resonance MS (16), isotope-coded affinity tags (17), and high accuracy quadrupole MS (18), also allow for significant proteome coverage. Nevertheless, the mouse proteome is predicted to be very complex and highly regulated (19), involving many thousands of proteins regulated by means of differential synthesis and selective subcellular localization. Furthermore, current high throughput experimental proteomic approaches do not allow for ready transformation of raw data into meaningful, easy to interpret output.

Here we describe the development and application of PRISM, a generic Proteomic Investigation Strategy for Mammals that allows for systematic, efficient, and unbiased detection and simplified follow-up analysis of large numbers of proteins expressed in mammalian cells and tissues. PRISM consists of a series of integrated experimental and analytical
steps, starting with subcellular fractionation and high throughput protein shotgun sequencing using an optimized MudPIT procedure followed by automated statistical validation, annotation, and categorization of the identified proteins based on universal Gene Ontology (GO) annotation terms (20). PRISM was evaluated on healthy adult mouse lung and liver, and physiologically significant differences in the tissue specificity and subcellular localization of hundreds of proteins were readily detected, confirming the utility of the approach for global analysis of complex mammalian proteomes.

**EXPERIMENTAL PROCEDURES**

**Materials**—All solid chemicals were from Sigma, while HPLC grade acetonitrile, methanol, and water were purchased from Fisher Scientific, and heptfluorobutyric acid was obtained from BioLynx (Brockville, Ontario, Canada). Bulk Porosyme immobilized trypsin was obtained from Applied Biosystems (Streetsville, Ontario, Canada), and endoproteinase Lys-C was from Roche Diagnostics.

**Tissue Preparation and Organelle Fractionation**—Healthy adult female mice (ICR) were CO2-asphyxiated and sacrificed. The organs of interest were perfused with cold phosphate-buffered saline, rapidly removed, rinsed, and homogenized for 2 min in ice-cold lysis buffer containing 250 mM sucrose, 50 mM Tris-HCl (pH 7.4), 5 mM MgCl2, 1 mM DTT, and 1 mM phenylmethylsulfonyl fluoride, filtered through cheesecloth to remove debris, layered onto 4 ml of cushion buffer, and pelleted in an ultracentrifuge at 80,000 g for 15 min; the supernatant served as source of cytosol, mitochondrial, and microsomal fractions. The supernatant was cleared of nuclei by centrifugation at 21,000 g for 15 min; the supernatant served as source of cytosol, mitochondrial, and microsomal fractions. Mitochondrial proteins were isolated by 100,000 g ultracentrifugation for 1 h (Beckman SW41 rotor). The mitochondrial pellet was washed with 20 mM Tris-HCl, pH 7.8, and treated with 5 mM iodoacetamide for 1 h at room temperature followed by carboxymidomethylation with 5 mM iodoacetamide for 1 h at 37 °C. The samples were then diluted to 4 mM urea with 50 mM ammonium bicarbonate, pH 8.5 and digested with a 1:150 molar ratio of endoproteinase Lys-C at 37 °C overnight. The next day the mixtures were further diluted to 2 mM urea with 50 mM ammonium bicarbonate, pH 8.5, supplemented with CaCl2 to a final concentration of 1 mM, and incubated overnight with Porosyme trypsin beads at 30 °C with rotating. The resulting peptide mixtures were solid phase-extracted with SPEC-Plus PT C18 cartridges (Ansys Diagnostics, Lake Forest, CA) according to the manufacturer's instructions and stored at −80 °C until further use.

**MudPIT Analysis**—A fully automated 15-cycle, 30-h MudPIT chromatographic procedure was set up essentially as described previously (12, 13). Briefly, an HPLC quaternary pump was interfaced with an LCO DECA XP ion trap tandem mass spectrometer (ThermoFinnigan, San Jose, CA). A 150-μm-inner diameter fused silica capillary microcolumn (Polymicro Technologies, Phoenix, AZ) was pulled to a fine tip using a P-2000 laser puller (Sutter Instruments, Novato, CA) and packed with 10 cm of 5-μm Zorbax Eclipse XDB-C18 resin (Agilent Technologies, Mississauga, Ontario, Canada) and then with 6 cm of 3-μm Partisphere strong cation exchange resin (Whatman). Samples were loaded manually onto separate columns using a pressure vessel. The chromatography was carried out as described by Wolters et al. (13).

**Protein Identification and Validation**—The SEQUEST program (a kind gift from Jimmy Eng and John Yates III) was used to search peptide spectra essentially as described previously (21). The database was populated with non-redundant mammalian Swiss-Prot and TrEMBL protein sequences in both a normal and inverted amino acid orientation (22). Statistical analysis (error modeling) was performed on the SEQUEST scores obtained for over 30,000 peptide matches. Formally the output of the analysis, Yi, was given as: $Y_i = 0$ (spectrum is incorrectly matched to an inverted peptide sequence; $Y_i = 1$ (spectrum is matched to a normal peptide sequence, possibly incorrect); $Y_i = 2$ (the spectrum matches the correct peptide sequence). We estimated a function $F(x,y,z)$ that characterizes the likelihood that a peptide match with score $X_i = (x,y,z)$ is correct as

$$F(x,y,z) = P(Y_i = 2|X_i = (x,y,z))$$

(Eq. 1)

For a protein with multiple peptide matches, $(x_1, x_2, \ldots, x_m)$, one can then estimate the probability of correct identification by

$$P(x_1, x_2, \ldots, x_m) = \max_{1 \leq i \leq m} P_i$$

(Eq. 2)

To compute the detection sensitivity (coverage), an estimate of the number of proteins actually present in the sample was made using

$$NPepN = \lambda PepN/(AProl/AProt)$$

(Eq. 3)

where $PepN$ is the number of observed peptides, $AProl$ is the average amino acid length of proteins in the database, $AProt$ is the average length of a peptide in the database, and $\lambda \approx 1$ is a positive constant proportional to the number of matches to an actual peptide. $F$ was approximated by first carving the $X$ regions and then fitting a smooth function. Monotonicity implies that for every $\beta$, there is a rectangular region, $R$, for which the function $F$ will have a value of at least $\beta$. Assumption (i) implies that for a large $\beta$ and a rectangle $R$ with a large number of observations ($K > 100$) the following applies.
Large Scale Proteomic Investigation Strategy for Mammals

\[ \beta = F(\tilde{X}) = P(Y = 2 \mid \tilde{X} \in R) = 2\alpha - 1 \]  
(Eq. 4)

where \( \alpha \) represents the proportion of 1's in region \( R \).

By continuity, for a not too large \( R \), \( P(Y = k \mid X \in R) \) is approximately equal for all \( X \in R \). Hence we let

\[ p_k = P(Y = k) \quad k = 0, 1, 2 \]  
(Eq. 5)

and

\[ q = p_1 + p_2 = 1 - p_0 \]  
(Eq. 6)

The probability that region \( R \) with \( K \) observations has \( \alpha \times K \) of 1's (meaning either 1 or 2 since 2's cannot be recognized a priori) and \((1 - \alpha) \times K \) of 0's is given by

\[ \left( \frac{K}{0.98} \right) (q)^{0.98} (1 - q)^{0.02} \]  
(Eq. 7)

It is well known that the maximum likelihood estimator of \( q \) is \( \alpha \) (23); we therefore let \( q \approx \alpha \). Assumption (i) implies \( p_1 = p_0 \), resulting in

\[ p_2 = 2\alpha - 1 \]  
(Eq. 8)

Since

\[ p_2 = P(Y = 2 \mid X \in R) \]  
(Eq. 9)

Equation 4 is proven.

Therefore, if \( R \) is a rectangle with at least 1's and the SEQUEST scores \( X \) of \( R \), the probability that a peptide match is correct is approximated by

\[ F(x,y,z) = 2\alpha - 1 \]  
(Eq. 10)

Next rectangular regions are identified for which \( F(x,y,z) \approx \beta \), \( \beta = (0.98, 0.96, 0.9, 0.8, 0.7, 0.6, 0.5) \). To this end, for fixed \( \beta \), we defined a function \( H(x,y,z) = 1 \) if \( (x,y,z) \in R \) and =0 otherwise. For \( \alpha = (\beta - 1)/2 \), we minimized the weighted distance between function \( H \) and the data points. In other words, since the rectangle \( R \) is easily parameterized (example \( R = \{x, y, z\} \) such that \( a < x < b \) and \( c < y < d \) and \( e < z < f \)), one looks for values "a,b,c..." that minimize the following quantity

\[ \sum_i |H(\tilde{X}_i) - Y_i|W_i \]  
(Eq. 11)

To ensure \( \alpha \), the weights were computed as follows

\[ W_i = \begin{cases} 1 & \text{if } Y_i = 0 \\ 1 - \gamma & \text{if } Y_i = 1 \end{cases} \]  
(Eq. 12)

where \( \gamma = 2 - 1/\alpha \). Assumption (ii) implies the existence of a smooth monotone function that can approximate the data. The actual optimization algorithm was an accelerated Random search (24). Computations were run on a desktop computer using FORTRAN.

GOClust—GOClust takes as input a tab-delimited text file of validated proteins. To facilitate comparison across multiple samples, the programs DTASelect and Contrast (a generous gift from Dave Tabb, Scripps Research Institute, La Jolla, CA) were used to arrange the data sets (25). Protein matches to other sequence databases are first mapped to a corresponding Swiss-Prot or TrEMBL entry using the Sequence Retrieval System at the Canadian Bioinformatics Resources (www.cbr.nrc.ca). The GOA flat file (regularly updated) that provides GO annotations for non-redundant Swiss-Prot, TrEMBL, and Ensemble entries was downloaded from the European Bioinformatics Institute (www.ebi.ac.uk). The final output is a series of tables of grouped proteins that share a common annotation to one or more preselected GO terms. The choice of terms is fully flexible to satisfy user interests.

RESULTS

Subcellular Fractionation and High Throughput Protein Identification—The apparent complexity of the mammalian proteome, defined here as the set of proteins produced by cells or tissues, represents a considerable experimental challenge even to high performance LC-MS techniques such as MudPIT (11–13). Hence we chose to use subcellular fractionation, an effective technique for selective enrichment of specific subsets of cellular proteins (26, 27) and organelles (28, 29) as a means of increasing both the proteome coverage and functional insight gained from LC-MS analysis. The entire PRISM methodology is outlined schematically in Fig. 1. As the goal was to develop a simple, generic methodology, we opted for a straightforward procedure based on differential centrifugation (outlined in Fig. 2A; see “Experimental Procedures”), which nonetheless results in respectable enrichment of nuclear, mitochondrial, microsomal, and cytosolic compartments and, importantly, a significant increase in the number of proteins detected by LC-MS (see below).
Proteins are extracted from each of five well defined subcellular fractions and digested with endoproteinase Lys-C and trypsin, and the peptides mixtures are analyzed using a 15-step MudPIT procedure (see “Experimental Procedures”). The MS instrumentation is set to automatically record both the mass-to-charge ratio and the fragmentation pattern of each eluting peptide that undergoes collision-induced dissociation. The fragmentation spectra are then compared with non-redundant human and mouse protein sequences using SEQUEST (30, 31), a database search program that infers amino acid sequence identity by matching fragment ions to translated genomic sequences.

**Protein Validation**—The output of SEQUEST is a series of putative protein matches and associated peptide scores, which include a cross-correlation score based on spectral fit (Xcorr), the normalized difference between the Xcorr of the top and second best matches (ΔCn), and a preliminary ranking based on the number of matched ion peaks (RSp). A subjective combination of these scores as well as other factors such as the charge of the precursor ion, the presence of tryptic termini (relevant in experiments where the peptides are generated by digestion with trypsin), and the number of peptides that map to a given protein is typically used to evaluate the accuracy of a prediction (30). To provide a more rigorous estimate of the accuracy of SEQUEST predictions, we developed a statistical algorithm, STATQUEST, that uses an empirical, probabilistic method for determining the likelihood of each putative peptide match.

We began our error modeling by considering the criteria mentioned above as a collection of variables and evaluating whether a subset, d, of these might describe a region of d-dimensional space enriched for correctly identified proteins. The goal was to produce a function corresponding to the probability that a given protein with SEQUEST scores \(X_i(x, y, z, \ldots)\) is correctly identified. To this end, we evaluated the distribution of SEQUEST database search scores for doubly charged mouse peptides and the derived probability function \(G(x, y)\).

**Fig. 2.** A, schematic representation of the subcellular fractionation procedure using differential centrifugation. Centrifugal forces and times (’, minutes) are indicated inside circular arrows. The final protein fractions analyzed by LC-MS are highlighted in bold (N, nuclear; C, cytosol; MI, microsomes; MS, soluble mitochondria; MP, insoluble mitochondrial). B, STATQUEST, a statistical algorithm to validate putative protein identifications. Graphical representation of the distribution of SEQUEST database search scores for doubly charged mouse peptides and the derived probability function \(G(x, y)\). C, GOClust, a program for automatic annotation of identified proteins. Matches to proteins within the Protein Information Resource (PIR) and GenPept (GenBank™) databases are first linked to the TrEMBL database. Next Swiss-Prot and TrEMBL linked proteins are annotated using an annotation (GOA) flat file. The annotated proteins are then clustered into user-selected GO subcategories, and a summary spreadsheet is produced.
Since a match to an inverted sequence, or $0$, clearly indicates an incorrect match, we located regions of variable space (i.e. Xcorr, ΔCn, and RSp) where the concentration of $0$’s is low. Since a low concentration of $0$’s relates to a high probability of correct matches, we were able to derive a likelihood function (see “Experimental Procedures”). (To further justify the above approach, we offer Supplemental Figs. F1 and F2 that show that the distribution of Xcorr and ΔCn for matches to normal peptide sequences (or $1$’s; Supplemental Fig. F2) is biased compared with matches to inverted sequences (or $0$’s; Supplemental Fig. F1)).

While the exact form of this function is not known, a good (least squares) fit is achieved with

$$\alpha = G(x,y) = 1 - 1/Q(x,y) \quad (\text{Eq. 13})$$

where $Q(x,y)$ is a polynomial expression of second degree.

Singly, doubly, and triply charged peptides were treated separately, and the three predictors (variables) were fixed as $x = Xcorr$, $y = ΔCn$, and $z = RSp$.

The final function is

$$F(x,y,z) = \begin{cases} 2G(x,y) - 1 & \text{if } z \leq 5 \\ 2G(x,0) - 1 & \text{if } z > 5 \end{cases} \quad (\text{Eq. 14})$$

The output of this function is a probability value for putative matches, which allows for easy assignment of a confidence factor. We found that the dimension of the problem could be reduced by fixing the third variable, $z < 5$, since this results in virtually the same entries; an Accelerated Random Search (24) reduced by fixing the third variable, $z < 5$, since this results in virtually the same entries; an Accelerated Random Search (24) was used as the optimization algorithm. Heuristically the optimization “slides” the upper right corner rectangle until it encounters a region with concentration of 1’s lower than $\alpha$.

For example, if $\alpha$ is set as 0.99, the optimization scheme fits rectangles in such a way that it must add at least 100 of 1’s for every added 0. If there is no such rectangle, the optimization scheme stops. Hence the optimization produces rectangles with maximum area at a concentration of at least $\alpha$. A graphical representation of $G(x,y)$ for doubly charged precursor ions is shown in Fig. 2B. While $G(x,y)$ depends on both the peptide charge and the particular MS instrumentation used, the interpolated functions are virtually indistinguishable for proteins derived from distinct organisms (data not shown). STATQUEST, launched by command line, can be used to rapidly filter large SEQUEST output files based on preselected confidence ($p$ value) cut-off values.

**Functional Annotation and Clustering**—Data analysis poses a significant challenge to large scale proteomic studies. To this end, we developed GOClust, a Perl-based computer program, to automatically annotate and subgroup long lists of validated proteins based on the GO annotation schema, a dynamic controlled vocabulary for describing the known molecular function, subcellular location, and biological role of proteins that changes as knowledge accumulates (20). As outlined in Fig. 2C, GOClust first sorts the proteins based on the database to which each corresponding accession number maps (e.g. Swiss-Prot, Protein Information Resource (PIR), or GenPept). Next the program obtains the GO identification numbers (GOids), and corresponding GO terms, assigned to each protein (or a close homologue) using GOA reference flat files downloaded from the European Bioinformatics Institute (32, 33) and the GO Consortium (20). Lastly the annotated proteins are grouped based on common annotation to one or more subcategories within the GO hierarchy, and the clusters are summarized in a spreadsheet (see “Experimental Procedures”). Users can specify the degree of annotation granularity (sets of returned GO terms) to narrow or broaden the scope of the analysis to areas of interest.

**Mouse Lung and Liver**—To evaluate the effectiveness of PRISM, we used it to investigate differences in the protein expression and subcellular distribution characteristics of adult mouse lung and liver. A combined total of over 300,000 spectra were acquired over the course of 2 weeks from equivalent normalized aliquots of the five respective subcellular fractions (see “Experimental Procedures”). The spectra were searched against non-redundant Swiss-Prot and TrEMBL mouse and human protein sequences downloaded from the European Bioinformatics Institute. Putative matches were filtered by STATQUEST using a $p$ value threshold of 0.02 ($\alpha = 0.99$), which virtually eliminates false positives albeit at the expense of sensitivity (coverage). (We estimate that ~40% of proteins that underwent collision-induced dissociation pass this stringent cut-off (data not shown)). This resulted in the high confidence identification of a total of 2,106 unique proteins (8,606 unique peptides) in the two organs (Table I). Of these, 1,460 proteins (4,242 unique peptides) were detected in lung (Table I), and 1,358 proteins (4,364 unique peptides) were detected in liver (Table I), divided roughly equally among the fractions (the entire expression data set is presented in Supplemental Table I). This analysis also confirmed the expression of over 500 hypothetical proteins predicted by ongoing mouse cDNA sequencing efforts, in particular the RIKEN study (4).
trast, a total of only 413 proteins were identified in a control MudPIT analysis of an unfractionated whole liver extract (data not shown), confirming that the subcellular fractionation procedure provides significantly deeper proteome coverage. GOClust was able to map 1510 (71%), 1186 (56%), and 925 (43%) of the validated proteins to at least one annotation term within the three main GO categories (molecular function, biological process, and cellular component, respectively), which define the known biochemical function, subcellular localization, and physiological role of gene products. This permitted ready examination of the tissue and organelle distribution of proteins across diverse biological processes such as cell metabolism, intracellular signaling, or gene expression (Supplemental Table II). Marked enrichment of specific classes of proteins was evident in the different fractions of both organs, a sampling of which is provided in Fig. 3A. For instance, known nuclear proteins, such as components of the spliceosome and nucleolus, were found as expected almost exclusively in the nuclear fractions (Fig. 3A). Similarly integral membrane proteins known to reside in the Golgi apparatus (Fig. 3A) or the endoplasmic reticulum, lysosome, or peroxisome (Supplemental Table II) were enriched in the microsomal membrane fractions. Some cross-contamination between fractions was evident, likely in part due to proteins bearing annotation to more than one subcellular compartment and to proteins that traffic between cellular compartments such as the signal transducer and activator of transcription STAT3 (see Table III). It should be noted that only those proteins that could be annotated to a GO term within cellular component are shown in Fig. 3A, which does not represent all the proteins detected with known subcellular localization properties, e.g. many more known nucleolar proteins were identified than are presently annotated in the GO schema. This limitation will improve in the near future as the GO database expands.

To collectively examine the samples, we applied a hierarchical clustering algorithm widely used to visualize relationships between genome-scale gene expression data sets (34). As seen in Fig. 3B, certain related pairs of fractions clustered closely together (e.g. the nuclear fractions, lanes 9 and 10) as might be expected for compartments housing similar core biological processes, while other fractions, such as the microsomal membrane fractions, did not, indicative of significant differences between the two organs. Importantly the protein pattern detected in the control liver whole cell extract (lane 2) was most similar to the liver cytosolic fraction (lane 1), consistent with the disproportionate overabundance of metabolic enzymes in the cytoplasm.

Close comparison of the various fractions revealed organelle-specific protein distribution signatures consistent with physiologic expectation. For example, the nuclear fractions were highly enriched for known nuclear proteins, including dozens of proteins involved in transcription, RNA processing, and DNA metabolism (Fig. 3C). Over 170 hypothetical proteins (most predicted by the RIKEN mouse cDNA library sequencing project (4)) were also uniquely detected in the nuclear fractions, strongly suggesting they may have a nuclear related function.

Many tissue-specific proteins were also identified. For instance, members of the P450 family of cytochromes, membrane-bound heme-thiolate monoxygenases involved in NADPH-dependent electron transport pathways active in hepatocytes, were highly enriched in the liver microsomal fraction (Table II). Indeed, of the 27 P450 isoforms detected in this study (25 of which were found in the microsomal fractions), 20 were found exclusively in the liver, 17 of which are known to be liver-specific (35); two were detected in both lung and liver, one of which (C2F2) is known to be expressed in Clara cells present in both organs (36); and three others were detected uniquely in lung, one of which (CP4B) is known to be pulmonary specific (37). In contrast, two-thirds of known mitochondrial proteins detected were found in both organs.
Other notable examples of tissue specificity are listed in Table III, such as the transcriptional regulator STAT3 (STA3), thyroid transcription factor 1 (TTF1), and cAMP-response element-binding protein (CREB)-binding protein (CBP), which were found exclusively in the lung along with several of their known targets, including the pulmonary surfactant-associated proteins, which lower the surface tension at the air-liquid interface in alveoli (38). Conversely hepatocyte nuclear factor 4α (HN4A), a nuclear hormone receptor essential for liver development (39) that drives transcription of liver-specific enzymes such as α1-antitrypsin, apolipoprotein C-III, and transthyretin (40, 41), and hepatocyte nuclear factor 6 (HNF6), a transcriptional activator that binds to cognate promoter consensus sequences upstream of several liver transcribed genes (42), were detected exclusively in the liver nuclear fraction.

**DISCUSSION**

PRISM is an effective approach for systematic global proteome profiling of mammalian cells and tissues since it couples experimental procedures for isolating and identifying hundreds of proteins in select subcellular compartments with automated procedures for validating and organizing the large expression data sets produced. To achieve this, several significant challenges stemming from both the intrinsic complexity of the samples and difficulties in manipulating large proteomic data files had to be addressed. We have validated the utility of PRISM using a well-defined mouse model system.

Although the 2,106 proteins detected in this study encompass only a fraction of the proteins encoded by the mouse genome (≈32,000 proteins are predicted (19)), the overall proteome coverage reported here compares favorably with the number of gene products detected in these same tissues in recent microarray-based expression studies (43). Importantly PRISM revealed physiologically relevant differences in the expression characteristics and subcellular localization of hundreds of proteins with distinct biological and biophysical properties, including membrane-associated proteins and low abundance transcriptional regulators, confirming the unbiased, comprehensive nature of the approach to detect biologically important proteins in extremely complex mixtures.

Examination of the annotated, clustered data set revealed many examples of tissue- or organelle-specific protein expression of notable biological interest. Particularly striking was the detection of low abundance transcription factors linked to liver development (HN4A), liver-specific gene transcription (HNF6), and lung surfactant homeostasis (TTF1 and STA3) that help define the physiological characteristics of (Supplemental Table II), consistent with an organelle expected to be similar between organs.

| ID  | Liver | Lung | P450 Isoform | GO term | Tissue specificitya |
|-----|-------|------|--------------|---------|---------------------|
| C2F2 | ✓     | ✓    | P450 2F2     | Membrane| Clara cells (lung and liver) |
| NCPR | ✓     | ✓    | P450 reductase | Membrane| Liver microsomes     |
| C39A | ✓     | ✓    | P450 39A1    | Membrane| Liver microsomes     |
| CP4B | ✓     | ✓    | P450 4B1     | Membrane| Pulmonary microsomes (not found in liver) |
| CPBA | ✓     | ✓    | P450 2B10    | Membrane| Liver microsomes     |
| CP12 | ✓     | ✓    | P450 1A2     | Membrane| Liver microsomes     |
| CP3B | ✓     | ✓    | P450 3A11    | Membrane| Liver microsomes     |
| CP3G | ✓     | ✓    | P450 3A16    | Membrane| Liver microsomes     |
| CPAC | ✓     | ✓    | P450 2A12    | Membrane| Liver microsomes     |
| CPB9 | ✓     | ✓    | P450 2B9     | Membrane| Liver microsomes     |
| CPCT | ✓     | ✓    | P450 2C29    | Membrane| Liver microsomes     |
| CPD9 | ✓     | ✓    | P450 2D9     | Membrane| Liver microsomes     |
| CPDA | ✓     | ✓    | P450 2D10    | Membrane| Liver microsomes     |
| CPE1 | ✓     | ✓    | P450 2E1     | Membrane| Highest level in liver |
| CPJ5 | ✓     | ✓    | P450 2J5     | Membrane| Wide tissue distribution |
| CPT7 | ✓     | ✓    | P450 17      | Membrane| No information       |
| CPZ2 | ✓     | ✓    | P450 2C37    | Membrane| Liver microsomes     |
| CPZ3 | ✓     | ✓    | P450 2C38    | Membrane| Liver, brain, kidney, and intestine (trace amounts in lung and heart) |
| CPZ4 | ✓     | ✓    | P450 2C39    | Membrane| Liver microsomes     |
| Q8VCX0 | ✓ | ✓ | P450 2D9 | N/A² | Liver microsomes |
| Q99N18 | ✓ | ✓ | P450 CYP4F15 | Monoxygenase | High expression in liver and kidney, lower expression in brain |
| Q5DBW0 | ✓ | ✓ | RIKEN: 1200012E24 | Monoxygenase | No information |
| Q5JKY7 | ✓ | ✓ | P450 CYP2D22 | Monoxygenase | Most abundant in liver |
| Q9JMA7 | ✓ | ✓ | Cytochrome P450 | Monoxygenase | Liver microsomes |

² Annotation obtained from Swiss-Prot.

* No GO term available.
these respective tissues. Detection of differential expression of a large number of membrane-bound P450 cytochromes in the liver compared with lung also highlights a significant physiological difference between these two tissues with prominent medical relevance. The P450 enzymes represent one of the largest mammalian multigene families with a central role in drug metabolism. As a consequence, the P450 system impacts many of the most important issues of clinical pharmacology, including drug pharmacokinetics, drug metabolism, and undesirable drug-drug interactions (44, 45). The expression characteristics of each of the P450 enzyme isoforms differ in disease states and are an important consideration in drug evaluation studies (46). Of the dozens of known mouse P450 paralogues, PRISM was able to detect most of the clinically important hepatic P450 isoforms, including 1A2 and 2E1 as well as diverse members of the 2C, 2D, and 3A subfamilies. This indicates that PRISM can serve as a platform to investigate the biochemical basis of drug metabolism in a standardized laboratory mouse model setting.

Protein extraction techniques based on differential solubility have previously been shown to significantly increase the number of proteins that can be identified by MudPIT (12). We have

---

**Table III**

Tissue specificity

Representative proteins, linked to a specific GO term, detected uniquely in either liver or lung alone or both organs. The total number of identified proteins bearing annotation to a particular GO term is shown in the second column (Total).

| Annotation (GO term)                  | Total | Examples | Specificity | Functiona |
|--------------------------------------|-------|----------|------------|-----------|
|                                      | Liver | Lung     | ID         |           |
| Cell communication                   | 104   | 186      | UTER       | Inhibitor of phospholipase A2 |
| Cell growth and/or maintenance       | 509   | 555      | MTP        | Transports phospholipids between surfaces |
| Metabolism                           | 424   | 423      | FABL       | Involved in intracellular lipid transport |
| Developmental processes              | 36    | 64       | TBX4       | Transcriptional regulation (mesoderm) |
| Enzyme                               | 304   | 291      | ARG1       | Arginine degradation |
| Ligand binding or carrier            | 348   | 437      | PSPA       | Contributes to lower the surface tension |
| Signal transducer                    | 39    | 62       | TFR2       | Cellular uptake of transferrin-bound iron |
| Ligand                               | 9     | 14       | ATS8       | Antiangiogenic properties |
|                                      |       |          | SY21       | Inhibits hemopoiesis and stimulates chemotaxis |
| Structural molecule                  | 88    | 75       | PRLP       | Anchors basement membranes |
| Transcription regulator              | 36    | 45       | HNF1A      | Transcription factor essential for liver development |
|                                      |       |          | HNF6       | Important for liver gene transcription |
|                                      |       |          | TTF1       | Lung development/surfactant homeostasis |
|                                      |       |          | STA3       | Transcription factor |
|                                      |       |          | GTF1       | Interacts with basal transcription machinery |
|                                      |       |          | TY1        | Multifunctional transcription factor |
|                                      |       |          | UBF1       | Activates RNA polymerase I |
|                                      |       |          | CBP        | Transcriptional regulation of cAMP-responsive genes |
|                                      |       |          |            |           |
| Translation regulator                | 18    | 13       | EF1G       | Anchoring the complex to the other cellular components |
| Transporter                          | 84    | 92       | AB11       |Secretion of bile salts by hepatocytes |
shown that subcellular fractionation can also significantly increase the proteome coverage and depth of information retrieved by LC-MS and have established that differential centrifugation can be an efficient method for isolating fairly pure subcellular compartments. Importantly, of the over 575 proteins detected exclusively in the nuclear fractions, nearly half (265 proteins) were either annotated solely to the nucleus or had a function known to be localized within the nucleus (e.g. transcription). Traditional biochemical techniques, such as Western blotting, immunostaining, or specialized techniques for isolating ultrapure preparations of organelles (47), may be warranted to confirm individual protein distributions. Nonetheless orthologues for approximately ~40% of the mouse nuclear specific proteins reported in this study were identified in two recent proteomic studies carried out on highly purified nucleoli of human HeLa cells (29). Interestingly ~30% of the human nucleolar proteins were encoded by novel or uncharacterized genes (28, 29), roughly the same fraction of hypothetical proteins detected in this study. Considering this enrichment of nuclear related protein functions and the clear segregation of the nuclear fractions during hierarchical clustering, the implication is that a large proportion of the proteins of unknown function detected in the nuclear fractions likely has a biochemical role specific to the nucleus, most likely related to gene expression, RNA processing, or chromosome dynamics. In summary, PRISM not only provides direct evidence for the actual tissue expression of hundreds of hypothetical proteins but can serve as a powerful approach to quickly gain insight into the biological role and/or molecular function of hundreds of novel gene products.

Elegant statistical approaches for eliminating SEQUEST mismatches have been described (22, 48). The use of a statistical algorithm to filter preliminary database sequence matches will allow for standardization in the reporting of protein identifications, permitting comparison between different proteomic studies and providing the basis for rationale estimates of the absolute complexity of a given mammalian proteome. The graded function reported here is particularly powerful since the likelihood that a given identification is correct is well defined, and false positives need not always be eliminated at the expense of true positives. Moreover the necessary statistical assumptions are easy to identify and verify.

Data management, mining, and visualization methods are increasingly a fundamental part of large scale proteomic studies. The bioinformatic tool GOClust described here allows researchers to organize the large numbers of proteins identified by MudPIT, or other large scale proteomic techniques, into smaller, more accessible categories of particular relevance or interest using a standardized nomenclature that permits comparison across multiple experiments. GOClust therefore allows researchers to browse proteomic data sets from a global, system perspective and then drill down to address specific biological questions. The capability of GOClust to reveal insightful patterns of protein expression and point to fruitful new areas for follow-up investigation will increase in concert with the significant annotation efforts ongoing by the GO consortium (20).

In conclusion, PRISM provides a new experimental and analytical framework for systematic, in-depth investigation of the proteomes of mammalian organisms. Although the approach described here is largely qualitative in nature, complementary techniques that allow for the determination of protein relative abundance (21, 49–52) can readily be incorporated. A combination of these and other related genomic scale methodologies should allow unprecedented insight into the complexities and dynamics of the mammalian proteome and their relationship to mammalian physiology, development, and disease.

Acknowledgments—We thank Gerard Cagney, Jimmy Eng, Dave Tabb, and John Yates III for generous use of software; Pete St. Onge, Shiva Amiri, Shaun Ghanny, Ata Gha videl, Thanu ja Premawaraden a, Bianca Tong, Ginny Li, Hassan Qureshi, and Mike Lindo for computing assistance; and Ben Blencowe, Tim Hughes, and Denis Wigle for fruitful suggestions.

* This work was supported in part by a grant (to A.E.) from the National Science and Engineering Research Council of Canada. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[3] The on-line version of this article (available at http://www.mcp online.org) contains Supplemental Figs. F1 and F2 and Tables I and II.

§ These authors contributed equally to this work.
¶ A scholar of the Josef Schormüller Gedächtnisstiftung.
¶¶ Present address: Dept. of Biochemistry and Molecular Biology, University of Calgary, 3330 Hospital Dr., Calgary, Alberta T2N 1N4, Canada.

To whom correspondence should be addressed: CH Best Inst., 112 College St., Rm. 402, Toronto, Ontario M5G 1L6, Canada. Tel.: 416-946-7281; Fax: 416-978-8528; E-mail: andrew.emili@utoronto.ca.

REFERENCES

1. Rossant, J., and McKerlie, C. (2001) Mouse-based phenogenomics for modelling human disease. Trends Mol. Med. 7, 502–507
2. Nadeau, J. H., Balling, R., Barsh, G., Beier, D., Brown, S. D., Bucan, M., Camper, S., Carlson, G., Copeland, N., Eppig, J., Fletcher, C., Frankel, W. N., Ganten, D., Goldowitz, D., Goodnow, C., et al. (2001) Sequence interpretation. Functional annotation of mouse genome sequences. Science 291, 1251–1255
3. Marra, M., Hillier, L., Kucaba, T., Allen, M., Barstead, R., Beck, C., Blistain, A., Bonaldo, M., Bowers, Y., Bowles, L., Cardenas, M., Chamberlain, A., Chappell, J., Clifton, S., Favello, A., et al. (1999) An encyclopedia of mouse genes. Nat. Genet. 21, 191–194
4. Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., Yamanaka, I., Kiyosawa, H., Yagi, K., Tomaru, Y., Hasegawa, Y., et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563–573
5. Schulze, A., and Downward, J. (2001) Navigating gene expression using microarrays—a technology review. Nat. Cell Biol. 3, E190–E195
6. Ogil, 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
7. Hanash, S. M. (2000) Biomedical applications of two-dimensional electro-
Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 6591–6596

50. Smith, R. D., Pasa-Tolic, L., Lipton, M. S., Jensen, P. K., Anderson, G. A., Shen, Y., Conrads, T. P., Udseth, H. R., Harkewicz, R., Belov, M. E., Masselon, C., and Veenstra, T. D. (2001) Rapid quantitative measurements of proteomes by Fourier transform ion cyclotron resonance mass spectrometry. *Electrophoresis* **22**, 1652–1668

51. Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L., Shen, Y., Conrads, T. P., Veenstra, T. D., and Udseth, H. R. (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2**, 513–523

52. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999