Quantitative Organelle Proteomics of MCF-7 Breast Cancer Cells Reveals Multiple Subcellular Locations for Proteins in Cellular Functional Processes

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · NOVEMBER 2009

Impact Factor: 4.25 · DOI: 10.1021/pr9008332 · Source: PubMed

Claire Mulvey
University of Cambridge
16 PUBLICATIONS 133 CITATIONS

Jasminka Godovac Zimmermann
University College London
124 PUBLICATIONS 3,334 CITATIONS

research articles Journal of proteome research

Quantitative Organelle Proteomics of MCF-7 Breast Cancer Cells Reveals Multiple Subcellular Locations for Proteins in Cellular Functional Processes

Amal T. Qattan,[†] Claire Mulvey,[†] Mark Crawford,[†] Darren A. Natale,[‡] and
Jasminka Godovac-Zimmermann*,[†]

*Division of Medicine, University College London, 5 University Street, London WC1E 6JF, United Kingdom, and
Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center,
3300 Whitehaven Street, NW, Washington, D.C. 20007*

Received September 16, 2009

We have combined sucrose density gradient subcellular fractionation with quantitative, tandem-mass-spectrometry-based shotgun proteomics to investigate spatial distributions of proteins in MCF-7 breast cancer cells. Emphasis was placed on four major organellar compartments: cytosol, plasma membrane, endoplasmic reticulum, and mitochondrion. Two-thousand one-hundred eighty-four proteins were securely identified. Four-hundred eighty-one proteins (22.0% of total proteins identified) were found in unique sucrose gradient fractions, suggesting they may have unique subcellular locations. 454 proteins (20.8%) were found to be ubiquitously distributed. The remaining 1249 proteins (57.2%) were consistent with intermediate distribution over multiple, but not all, subcellular locations. Ninety-four proteins implicated in breast cancer and 478 other proteins which share the same five major cellular biological processes with a majority of the breast cancer proteins were observed in 334 and 1223 subcellular locations, respectively. The data obtained is used to evaluate the possibility of defining more exact sets of subcellular organelles, the completeness of current descriptions of spatial distribution of cellular proteins, the importance of multiple subcellular locations for proteins in functional processes, the subcellular distribution of proteins related to breast cancer, and the possibility of using these methods for dynamic spatio/temporal studies of function/regulation in MCF-7 breast cancer cells.

Keywords: quantitative proteomics • mass spectrometry • breast cancer • MCF-7 cells • subcellular organelles • subcellular location

Introduction

Although a century of extensive research generated 199 368 scientific publications in the period 1894−2009, breast cancer remains the second leading cause of cancer deaths in women today (after lung cancer). According to the American Cancer Society, about 1.3 million women will be diagnosed with breast cancer annually worldwide and about 465 000 will die from the disease.[1,2] The completion of human genome sequencing and the development of DNA-microarray technology led to extensive investigation of gene expression associated with breast cancer. Large-scale screening of mRNA levels showed that multiple and extensive changes in mRNA levels are commonly seen in breast cancer.[3–5] On the other hand, eukaryotic cell proliferation is known to involve complex molecular choreography of mitogens that stimulate cell growth, receptors, their signaling pathways, and downstream effectors of cell division,[6,7] indicating a need for complementary, highly parallel studies at the protein level.

The spatial and temporal distribution of proteins within cells is a very complex, but essential, feature of cellular function. The analysis of such distributions is complicated by the facts that a given protein may have multiple subcellular locations, may exist in multiple transcriptional or post-translational isoforms within the same cell and that the different isoforms may have different spatial and temporal distributions as well as different functional roles.[8,9] Common highly parallel methods such as analysis of mRNA abundance can give information on inputs to cellular protein abundance. However, the mRNA methods do not always correlate well with direct measurements of protein abundance,[10] require additional complexity to measure transcriptional isoforms, do not detect post-translational isoforms, and do not give information on spatial location. Conversely, direct measurements of spatial location by methods such as fluorescence microscopy usually do not distinguish isoforms, are usually semiquantitative, and are difficult to achieve in highly parallel formats.

One goal of the presently reported work was to establish high throughput proteomics methods that are capable of analyzing dynamically at least some of the complexity involved in subcellular protein distribution. The estrogen-dependent MCF-7 malignant breast epithelial cell line was selected due to the

* To whom correspondence should be addressed. Prof. J. Godovac-Zimmermann, e-mail: j.godovac-zimmermann@ucl.ac.uk.
[†] University College London.
[‡] Georgetown University Medical Center.

10.1021/pr9008332
Published on Web 11/16/2009

wealth of information available in the literature and its relevance to breast cancer.[11,12] Proteomics methods based on mass spectrometry are only suitable for indirect measurements of spatial location and we have therefore concentrated on the distribution of proteins between different subcellular organelles. To avoid the need for multiple purification procedures for many different organelles, we have used partial purification based on sucrose gradient centrifugation followed by high throughput proteomics analysis of the protein content of different fractions from the sucrose gradient. Distribution analysis is used to show that there are many proteins which can be reliably shown to be present in more than one subcellular organelle. On the basis of these initial results, we assess the prospects that the present "low resolution" distribution between cytosol, plasma membrane, endoplasmic reticulum and mitochondria can be expanded to a more complete set of subcellular organelles. We show evidence that the majority of observed proteins have multiple subcellular locations, that current annotations of protein subcellular location are still sparse, and that multiple locations can be monitored for large numbers of proteins implicated in breast cancer. We consider the prospects for analyzing dynamic spatio/temporal changes in protein distribution between different subcellular locations as a consequence of cellular functional state.

Experimental Procedures

Cell Culture Conditions. The adenocarcinoma mammary epithelial breast MCF-7 cell line (ATCC HTB-22, Manassas, VA) was cultured at 37 °C with 5% CO_2 in Dulbecco's Modified Eagle Media DMEM/F-12 (Gibco, Invitrogen, Paisley, UK) with L-Glutamine, 15 mM HEPES supplemented with 10% defined FBS, 100 U/ml penicillin and 100 μg/mL streptomycin.

Preparation of Subcellular Organelles by Sucrose Gradient Density Centrifugation. All procedures were performed at 4 °C in the presence of protease and phosphatase inhibitor cocktails (Roche Diagnostics, Mannheim, Germany). Cells were lysed in Break Buffer (0.3 M Sucrose, 1 mM EDTA, Heparin 5 U/mL, 10 mM HEPES, 5 mM $MgCl_2$, pH 7.4) and homogenized gently by liquid shear methods with 40–50 strokes of a tight-fitting Dounce homogenizer (0.05–0.08 mm clearance). Phase-contrast microscopy was used to confirm that the organelle membranes remained intact. The cell suspension was centrifuged for 5 min at 800*g* to pellet cellular debris and the supernatant was collected for subcellular fractionation (Eppendorf centrifuge 5415R, Hamburg, Germany). A 10 mL discontinuous sucrose density gradient was formed by carefully layering equal volumes of decreasing concentrations of sucrose buffer (from the bottom of the gradient: 1.46, 1.3, 1.16, 1.02, 0.87, 0.73, 0.58, and 0.43 M sucrose in 1 mM EDTA, Heparin 5 U/mL, 10 mM HEPES, and 5 mM $MgCl_2$, pH 7.4). The cell suspension was carefully overlaid onto the sucrose gradient and ultracentrifugation was performed for 18 h at 14 440*g* in a swing-bucket rotor (TST41 rotor, Optima LE-80K centrifuge, Beckman, MN). Following ultracentrifugation, 24 × 500 μL fractions were collected by careful aspiration at the meniscus of the gradient. Prior to further analysis, the refractive index (Rf) of each fraction was assessed (Refractometer, Sun Instruments, Torrence, CA) and enzyme activity assays were performed at this point. Each gradient fraction was then diluted 1:1 with Dilution Buffer (1 mM EDTA, Heparin 5 U/mL, 10 mM HEPES, 5 mM $MgCl_2$, pH 7.4) and centrifuged for 60 min at 22 000*g* in a TLA-100.4 fixed rotor (TLX Ultracentrifuge, Beckman, Chaska MN) in order to precipitate proteins from the sucrose suspension. For each fraction, at this step the pellet was retained for proteomic analysis and the supernatant was subjected to acetone precipitation in order to obtain any additional solubilized proteins from the sucrose solution. The acetone-precipitate was combined with the above pellet for each of the 24 fractions and resuspended in 1× Solubilisation Buffer for final proteomic analysis (Solubilisation Buffer (2×): 20 mM PIPES pH 7.3, 300 mM NaCl, 2% Triton X-100, 0.2% SDS, 2% deoxycholic acid).

Enzyme Activity Measurements, Gel Electrophoresis, and Immunoblotting. For the determination of enzyme activities, cytochrome c oxidase assays (Sigma-Aldrich, Poole, Dorset, UK) and lactate dehydrogenase LDH enzymatic assays (Promega, Hampshire, UK) were performed on all subcellular fractions according to the manufacturer's instructions and monitored at the correct wavelength for each substrate. For Western blotting and silver staining analysis the protein content of each subcellular fraction was determined using the BioRad Protein Assay (BioRad, Herts, UK) and 30 μg of each organelle fraction was electrophoretically separated by 12% (w/v) SDS-PAGE with a Mini-Protean III system (BioRad, Herts, UK) using standard techniques according to Laemmli.[13] Gels were either silver-stained (ProteoSilver Plus kit, Sigma-Aldrich, Poole, Dorset, UK) or alternatively electro-transferred onto nitrocellulose membranes by a semi dry transfer apparatus (BioRad, Herts, UK). The nitrocellulose membranes were blocked with 5% milk-TBS-Tween buffer and probed for 3 h to overnight with appropriate dilutions of one of the following primary antibodies: anti-GADPH, anti-E-cadherin, anti-Lys-Asp-Glu-Leu (KDEL) or anti-VDAC. Primary antibodies and peroxidise-conjugated secondary antibodies were obtained from Cell Signaling, (New England Biolabs, HERTS, UK) with the exception of anti-KDEL (Stressgen, Canada). The chemiluminescence reagents ECL Plus and Hyper Film were supplied by GE Healthcare (Bucks, UK).

Protein Separation and In-Gel Enzymatic Digestion. For mass spectrometric analysis of the selected subcellular organelle regions, the fraction of interest was separated on SDS-PAGE gels as described above, proteins were visualized by silver-staining and the gel lane was divided into approximately 40 equally sized pieces which were excised from the gel and destained (30 mM $K_3Fe(CN)_6$; 100 mM $Na_2S_2O_3$) prior to further processing. Gel processing was conducted with a Progest Investigator Instrument (DigiLab, Genomics Solutions, Cambs, UK) according to established protocols.[14] Briefly, the gel pieces were washed with three cycles of 25 mM NH_4HCO_3 pH 8.0 and acetonitrile, (followed by reduction (10 mM DTT; 50 mM NH_4HCO_3, 15 min, 60 °C) and alkylation (100 mM iodoacetamide; 50 mM $NH_4HCO_{3,}$ 45 min, RT). The gel pieces were washed with three further cycles of 25 mM NH_4HCO_3 pH 8.0 and acetonitrile. Finally, gel plugs were rehydrated in 20 μg/mL sequencing grade modified trypsin (Promega, Hamps, UK) and incubated overnight at 37 °C. Tryptic peptides were eluted, vacuum-dried, and resuspended in 0.1% formic acid.

Mass Spectrometric Analysis. Peptide samples were loaded via an autosampler (Surveyor MS Pump Plus and Micro AS) onto a Michrom C18 Captrap to initially desalt samples and from there were introduced directly into a LTQ-Orbitrap MS (Thermo Fisher Scientific, Surrey, UK) via a fused silica C18 capillary column (Nikkyo Technos CO, Tokyo, Japan) and a nanoelectrospray ion source. Separation was achieved by a linear gradient of 5–60% Buffer B for 100 min at a flow-rate of 250 μL/min. (Buffer A = 0.1% formic acid; Buffer B = 100% acetonitrile, 0.1% formic acid.) Measurements were performed in the positive ion mode. The FTMS full scan MS spectra (from 450 to 1600 m/z) were acquired with a resolution of $r = 60\,000$. This was followed by a data dependent MS/MS fragmentation of the most intense ion from the survey scan using collision induced dissociation (CID) in the linear ion trap (normalized collision energy 35%, activation Q 0.25; electrospray voltage 1.4 kV; capillary temperature 200 °C: isolation width 2.00). This MS/MS scan event was repeated for the top 3 peaks in the MS survey scan. Target ions already selected for MS/MS were dynamically excluded for 40 s. Singly charged ions were excluded from the MS/MS analysis. The acquired tandem mass spectra were evaluated and searched against an NCBInr database and its reversed database (implemented in BioWorks 3.3.1, Thermo Fisher Scientific, UK) using the SEQUEST algorithm.[15] The following SEQUEST search parameters were used: peptide mass tolerance of 20 ppm; fragment tolerance of 0.5 Da; 2 max allowed missed cleavages; dynamic/variable modifications = oxidation (methionine); static/fixed modifications = carboxyamidomethylation and duplicate peptide matches were not considered (deselect). Protein and peptide identifications were accepted if they contained at least two peptides and could be established at greater than 95.0% probability as specified by the ProteinProphet and PeptideProphet algorithms using Scaffold software (Version 2.1.03, Proteome Software Inc., Portland, OR).[16–18]

Normalization and Quantification Based on Label-Free Methods. Selected search results files (SRF) from the BioWorks 3.3.1. analysis were submitted to Scaffold software (Version 2.1.03, Proteome Software Inc., Portland, OR) to calculate spectral counts. Protein and peptide identifications were accepted if they contained at least two peptides and could be established at greater than 95.0% probability as specified by the ProteinProphet and PeptideProphet algorithms.[16–18] Analysis of the presence/absence of the proteins in different gradient fractions was mostly performed with unweighted spectral counts. Normalization of protein abundance was carried out in several ways for different comparisons. First, normalization using the Scaffold software, which entails averaging the spectral counts for all the samples and then multiplying the spectral count in each sample by the average divided by the individual sample's sum to give weighted spectral counts. Second, to counterbalance the tendency of larger proteins to contribute more peptides, the unweighted spectral count from a protein was divided by the protein's length (number of amino acids) to define the Spectral Abundance Factor (SAF). The normalized spectral abundance factor (NSAF) was calculated by dividing the SAF for a particular protein by the sum of the SAF for all N proteins.[15,19–21] Finally, for some purposes the SAF were used as input to the GeneSpring MS analysis platform 1.2.0 (Agilent Technologies Inc., Santa Clara, CA), and the data was normalized using the "Per Mass Normalized to the Median" algorithm.

Clustering was performed to identify the similarity profile for protein distribution across the sucrose gradient fractions. The abundance factor was submitted to GeneSpring MS analysis platform 1.2.0 (Agilent Technologies Inc., Santa Clara, CA). A distance-based measurement (metric) was used to find the relationship between all the possible pairs in the data set using the Pearson correlation (PE) coefficient (cosine-angle distance). A "bottom-up" agglomerative clustering algorithm, using average-linkage as an aggregation procedure, was used to construct the final dendrogram.

Bioinformatic Analysis of MS Data Sets. The protein NCBI GI numbers based on the GenBank NCBInr database and its reversed database obtained from the MS data were converted to UniProtKB accession numbers using the following methods. The ID mapping table pertaining to human protein entries from UniProtKB release 15.3 was downloaded from the Protein Information Resource (PIR) ftp site (ftp://ftp.pir.georgetown.edu/databases/idmapping/mapping_by_sp/h_sapiens.tb). The file was parsed to look for the UniProtKB accession corresponding to each GI number. It was further possible to obtain the accession numbers used by the LOCATE subcellular localization database (http://locate.imb.uq.edu.au). Because the nature of the entries in GenBank NCBInr differs from those in UniProtKB, it is possible to get not just one-to-one correspondences, but also one-to-many and many-to-many mappings. Entries in GenBank NCBInr are sequence-based, meaning that proteins from any species could be represented if the sequences are identical. Entries in UniProtKB are gene-based, meaning that all proteins from a given gene in a given species could be represented regardless of sequence, thus all isoforms and sequence variants from a single gene are described within one entry. In some cases, where the ID mapping table gave ambiguous UniProtKB identifiers, the peptides obtained by MS were used to supplement the list of the GI numbers to identify the protein corresponding to each ambiguous/missing/TrEMBL identifier. GI numbers were used to search against a database of human sequences that contained not only the canonical sequences, but also the known splice variants. A positive mapping in this procedure occurred when all peptides used to identify the original GenBank entry were present in the same UniProtKB entry. Preference was given to hits to UniProtKB/Swiss-Prot entries over UniProtKB/TrEMBL entries. The file containing UniProtKB/Swiss-Prot entries for release 15.3 was downloaded from the UniProt ftp site (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/uniprot_sprot_human.dat.gz). FASTA files for UniProtKB/Swiss-Prot canonical sequences and splice variants, and for UniProtKB/TrEMBL sequences (for release 15.3) were processed to retain only the human sequences and were downloaded from the UniProt ftp site (ftp://ftp.uniprot.org/pub/databases/uniprot/ current_release/knowledgebase/complete/).

The locations for the proteins as noted in the literature can be obtained from UniProtKB records using both "Subcellular Location" keywords and GO cellular component terms, and from LOCATE records. Human protein localization predictions and annotations were downloaded from LOCATE subcellular localization database (http://locate.imb.uq.edu.au/info_files/LOCATE_human_v6_20081121.xml.zip). In all cases, the provided term was mapped to a "GO slim" term to provide a uniform vocabulary for comparing and merging the information. Proteins known or suspected to be implicated in breast cancer were examined by UniProtKB, Reactome Pathway and BioBase Biological Databases BIOBASE Knowledge Library (BKL) and ExPlain 2.3 (BIOBASE GmbH, Germany) for common ontology, biological process, molecular function terms and for common Reactome Pathway terms as recorded in the relevant UniProtKB records. These terms were then used as lures to obtain the subset of proteins found in this study that share the same terms.

Results

The basic experimental methodology used in the present experiments is summarized in Figure 1. Our use of subfractionation of cellular organelles by sucrose gradient centrifuga-

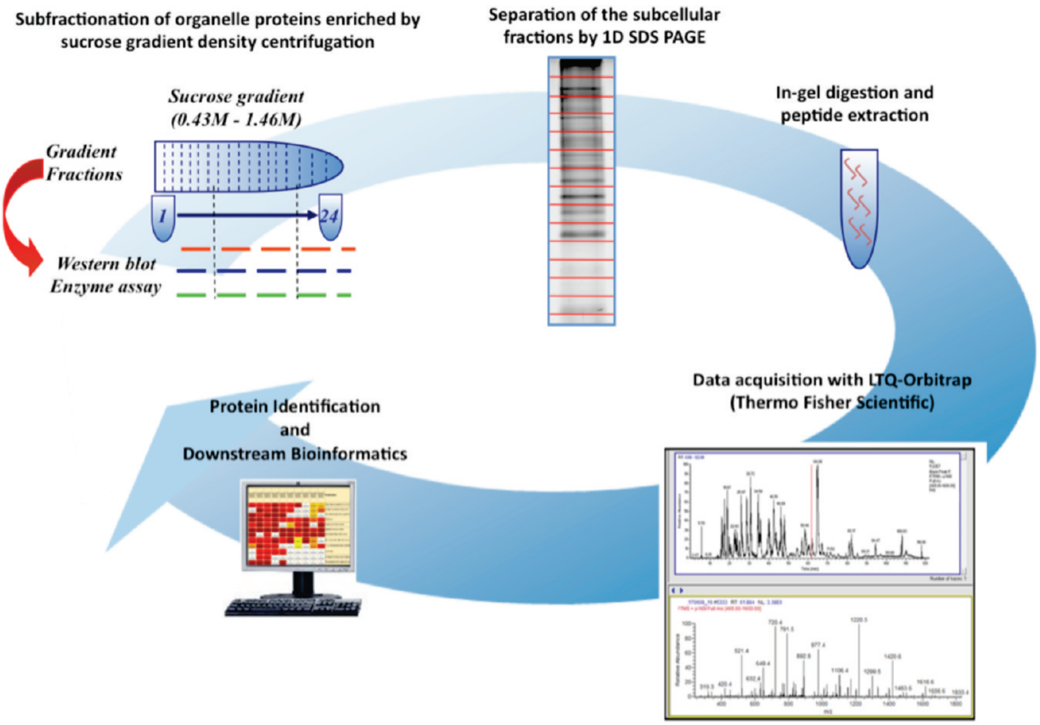
Subfractionation of organelle proteins enriched by sucrose gradient density centrifugation
Sucrose gradient (0.43M - 1.46M)
Gradient Fractions
1
24
Western blot
Enzyme assay
Separation of the subcellular fractions by 1D SDS PAGE
In-gel digestion and peptide extraction
Data acquisition with LTQ-Orbitrap (Thermo Fisher Scientific)
Protein Identification and Downstream Bioinformatics

Figure 1. Subcellular organelle proteomics workflow.

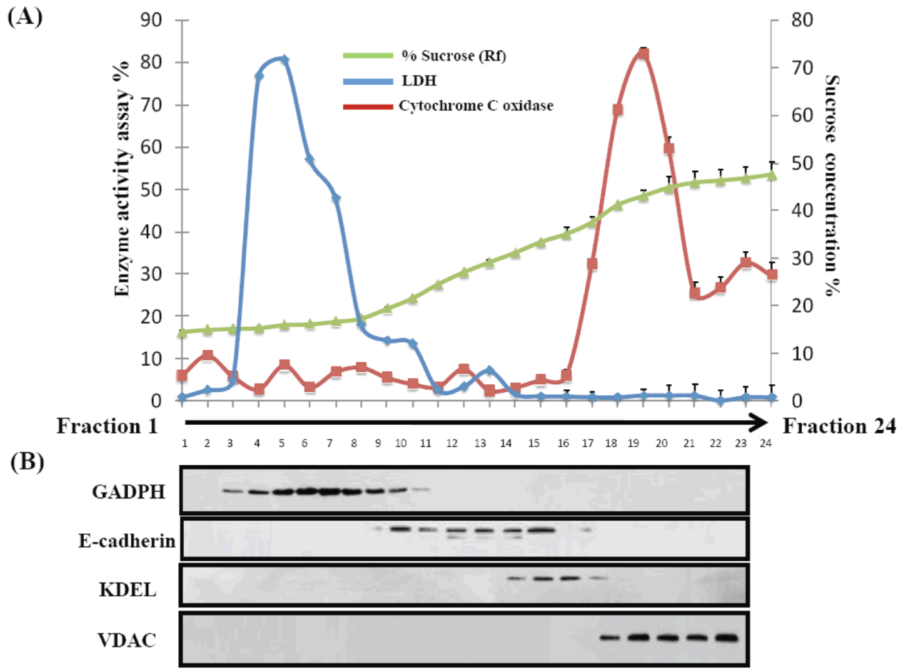

Figure 2. Subcellular verification of protein markers by enzyme assay and Western blot. A total of 24 fractions was collected from a sucrose-density gradient (0.43−1.46 M) in which fractions 1 and 24 represent the top and the bottom, respectively. (a) Typical density profile of sucrose fractions calculated by refractive indices. Also shown is the distribution of activity across the gradient for the subcellular enzyme markers LDH and Cytochrome C oxidase. (b) Immunoblot analysis of the distribution of known marker proteins. The cytosolic marker GADPH was detected mainly in fractions 3−10, plasma membrane marker E-cadherin was in fractions 10−15, the endoplasmic reticulum marker KDEL was in fractions 15−18, and the mitochondrial marker VDAC was found in fractions 18−23.

tion is based on previous proteomics work indicating that fractionation of cytosol, plasma membrane, endoplasmic reticulum, Golgi and mitochondria is readily obtained[22] and that with more sophisticated analysis of protein distribution along the gradient, as many as 10 subcellular locations can be distinguished.[23] The subsequent steps in identifying the proteins present in different fractions of the sucrose gradient (1D SDS gels, gel slicing, proteolytic production of peptides, and identification of peptides by MS) are standard proteomics techniques, and our use of them is described in detail in the Materials and Methods section. We note that the MUDPIT approach[15] has been used in connection with a fused silica C18 capillary column for elution and a nanoelectrospray ion source.

The basic functioning of the sucrose gradient fractionation was controlled by biochemical assays (Figure 2). Enzymatic assays showed maximum activity for lactate dehydrogenase in fractions 3−10 and of cytochrome oxidase in fractions 17−21 (Figure 2A), which is consistent with their localization to cytosol and mitochondria respectively. Similarly, Western blot detection of glyceraldehyde 3-phosphate dehydrogenase, E-cadherin, KDEL and Voltage-dependent anion channel (Figure 2B), indicated sucrose gradient fractions enriched in cytosol, plasma membrane, endoplasmic reticulum, and mitochondrial proteins, respectively. On the basis of this data obtained in two replicate experiments, fractions 9, 13, 16, and 20 from the sucrose gradient fraction were subjected to detailed analysis of protein content by MS methods.

Two aspects of the MS analysis are important in the context of the goals of the present work: secure identification of as many proteins as possible within any given gradient fraction (see below) and accurate measurement of the (relative) amount of any specific protein across the different fractions. We have used direct spectral counts from MS/MS runs for quantitative measurements of the peptides.[15] As seen previously by others,[20] we observed strong linear correlation between spectral counts and the relative abundance of characteristic proteins. This is a good indication that relative amounts of the same protein in different gradient fractions can be measured with considerable confidence using spectral counts.

Table 1 shows a summary of the experimental MS data. A total of 15 527 different peptides were used to identify 2184 proteins in fractions 9, 13, 16, and 20 of the sucrose gradient. At least 2 peptides were sequenced for each protein identified. The raw distribution of these proteins over the four fractions is shown in Table 1A. The MS data for the proteins is given in Supplementary Table 2 (Supporting Information).

The initial set of MS data contained 5514 (protein, fraction, abundance) data points for 2184 proteins, there was an average of 2.5 locations per protein (Table 1B). This initial data set contained a substantial number of (protein, fraction, abundance) data points for which in a particular fraction only a single peptide with a small number of spectral counts was observed for some proteins. The assignment of these proteins is less certain for these fractions. Removal of 876 data points for fractions where only a single peptide and 1 or 2 spectral counts were observed gave the “normal” data set in Table 1. 106 (protein, fraction, abundance) data points with only a single peptide in a fraction, but with 3 to 74 spectral counts, were retained to give a total of 4638 data points. The normal data set, which was used for many of the analyses below, corresponds to an average of 2.1 locations per protein. For some of the analyses, we have also removed from the normal data set those (protein, fraction, abundance) data points where less than 4% of the total amount of a given protein was observed in a specific fraction. This “trimmed” data set reduced the number of data points to 4576, that is, an average of 2.1 locations per protein. In the following we will refer to the three sets of (protein, fraction, abundance) data points used for further analysis as the initial, normal and trimmed data sets (Table 1B), all of which contain a total of 2184 proteins. For individual proteins that were detected in multiple fractions, we will also use the term “primary location” to refer to the (protein, fraction) pair with the highest abundance and the term “secondary location” to refer to other (protein, fraction) pairs with lesser abundances for the same protein.

Table 1. Summary of MS Data

A. Distribution Over Sucrose Gradient Fractions

	fraction			
total number of	F9	F13	F16	F20
Initial Data[a]				
Unique MS Spectra	4393	11435	8628	9654
Unique Peptides	3969	9876	7553	8588
Total Proteins per Fraction	852	1611	1441	1610
Unique Proteins per Fraction[b]	129	116	27	209
Normal Data[c]				
Unique MS Spectra	4233	11233	8341	9427
Unique Peptides	3810	9674	7267	8359
Proteins per Fraction	692	1409	1154	1383
Unique Proteins per Fraction[b]	189	239	69	347
Trimmed Data[d]				
Unique MS Spectra	4092	11223	8320	9375
Unique Peptides	3669	9664	7246	8311
Proteins per Fraction	657	1405	1145	1369
Unique Proteins per Fraction[b]	189	239	69	350

B. Data Sets

data set	number of (protein, fraction, abundance) data points	number of proteins
Initial[a]	5514	2184
Normal[c]	4638	2184
Trimmed[d]	4576	2184

[a] Includes all (protein, fraction, spectral counts) data points verified by Scaffold. [b] Number of proteins found only in one fraction. [c] Excludes (protein, fraction, spectral counts) data points where only a single peptide with 1 or 2 spectral counts was observed in a specific fraction. [d] After removal from the normal data set of (protein, fraction, abundance) data points for which the proportion of the protein in a specific fraction was less than 4% of the total protein abundance in all four fractions.

With the normal data set, many of the proteins were observed in more than one sucrose gradient fraction and hierarchical clustering[24] was used to analyze their distribution over the gradient (Figure 3). This indicated that in many cases the observation of the same protein in multiple fractions was not due to “tailing” of the proteins in the sucrose gradient. The data shows numerous examples of bimodal distribution of proteins over two fractions that are not adjacent in the gradient (e.g., fractions 9 and 20 in Figure 3C), as well as examples of proteins with more complicated bimodal distributions over three of the four fractions (Figure 3B) that are highly unlikely to arise from tailing.

A Venn diagram (Figure 4) has been used to summarize the observed distribution of the proteins over the four sucrose gradient fractions as determined by the hierarchical clustering. A notable characteristic for the normal data set (Figure 4A) is that only 844 of the 2184 proteins (38.6%) were uniquely found in a single fraction. A further 296 proteins (13.6%) were found to be ubiquitously distributed over all fractions. The remaining 1044 proteins (47.8%) were consistent with intermediate distribution over multiple, but not all, subcellular locations. Of these 1044 proteins, 248 (11.4% of total proteins) were distributed over two fractions (e.g., 9 and 20, Figure 3C) or over three fractions (e.g., 9, 13 and 20, Figure 3B) in a “bimodal” manner that is inconsistent with inclusion in a single subcellular organelle and “tailing” over the sucrose gradient.

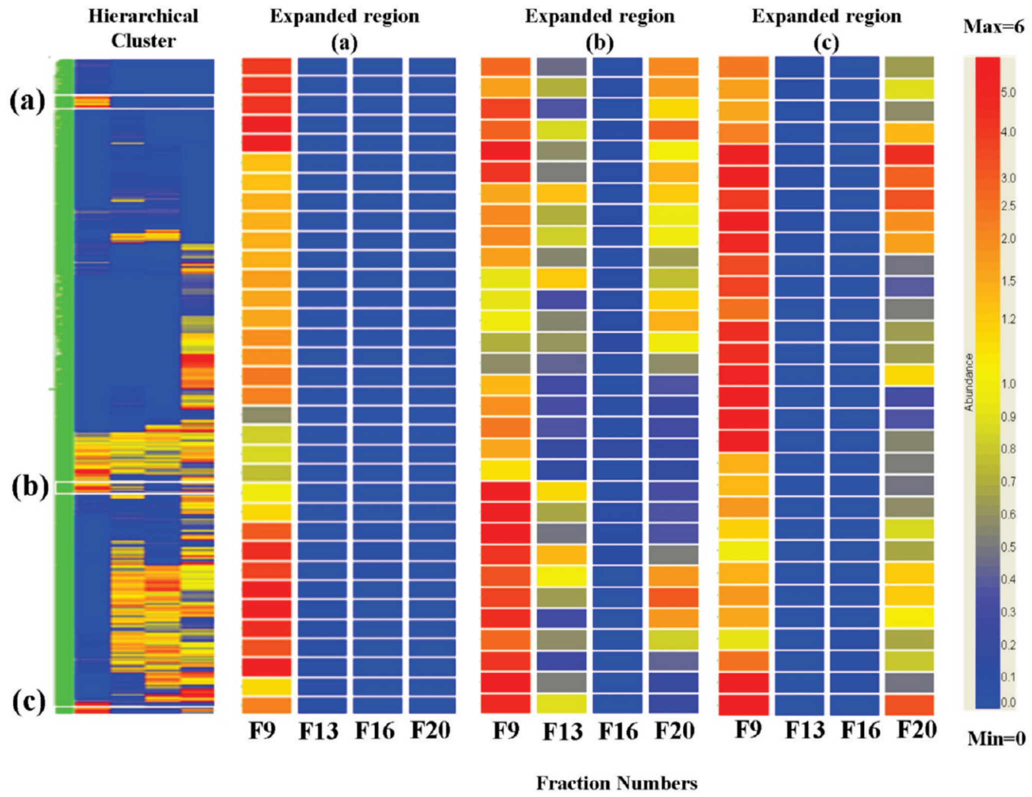

Figure 3. Hierarchical clustering and heat map across the four fractions. Individual proteins are represented by a single row and each fraction is represented by a single column, while each cell represents the abundance of a protein. The color scale is for normalized relative abundance from 6.0 (red) to 1.0 (yellow) to 0.0 (blue, not detected). The expansions show typical regions of the heat map corresponding to: (a) proteins observed uniquely in fraction 9, (b) "bimodal" proteins (see text) observed in fractions 9, 13, 20, and (c) "bimodal" proteins observed in fractions 9 and 20.

Inspection of the distribution of the proteins between primary and secondary locations revealed that they are well dispersed over the regions compatible with a primary location and 1−3 secondary locations (Figure, 5). Thus, for example, proteins for which we detected a primary location and a single secondary location must lie on the line from (0.5, 0.5) to (1.0, 0.0) (green plus signs in Figure 5), but are well dispersed along that line. For 2−3 secondary locations, the initial data set shows better sampling near the edges of the compatible regions, for example, there are more data points at large values of the primary mole fraction and at very small values of the secondary mole fractions. Many of these data points arise from proteins corresponding to sequencing of only one peptide and only 1−2 spectral counts in a specific fraction. This is a consequence of the sampling properties of spectral counting (see below). The dispersion of the data points in Figure 5 over the compatible areas of the plot is a strong indication that the data represent a good sampling of the distribution over multiple subcellular locations for the observed proteins.

We have investigated the robustness of the Venn diagram and the resulting conclusions about distribution of proteins over multiple locations in two ways. First, a more quantitative evaluation of the possibility of tailing in the gradient was obtained by looking for proteins with high abundance in a given gradient fraction, but with no detectable abundance in the adjacent fractions. For the most abundant proteins, the MS detection method was capable of detecting as little as about 0.2% of the protein in an adjacent fraction. Because the proteins may correspond to different subcellular organelles, tailing between two fractions need not be symmetrical, e.g. tailing from F9 to F13 may not be the same as tailing from F13 to F9. This leads to the six tests for the possibility of tailing shown in Table 2. For all the fractions there are many highly abundant proteins which *do not* tail into the adjacent fraction (Table 2). The highly abundant proteins also reveal some characteristics which are common in the data set. Some very abundant proteins were found uniquely in a single fraction (e.g., see hepatoma-derived growth factor and Protein S100-A9 in Table 2). Other proteins were detected in only two fractions, but with a bimodal distribution over the fractions (e.g., see sialic acid synthase and pyridoxal kinase in Table 2). Many proteins were distributed over several fractions, with substantial proportions of the protein present in different fractions (e.g., see ATP-citrate synthase in Table 2). Some proteins were primarily present in a single fraction, but small amounts of the protein were found in other fractions (see e.g. Rho GDP-dissociation inhibitor 1 and nucleophosmin in Table 2). We conclude from the data in Table 2 that spurious tailing of proteins in the sucrose gradient does not make any major contributions to the observed multiplicity of locations.

The second test for robustness of the Venn diagram involved testing whether the multiplicity of locations resulted from "trace" amounts of some proteins observed in some fractions (e.g., see nucleophosmin in Table 2). For this purpose, we calculated for each individual protein the mole fraction of the total protein observed in the different fractions. The (protein, location, mole fraction) data points for all proteins in the normal and initial data sets were then sorted into ascending

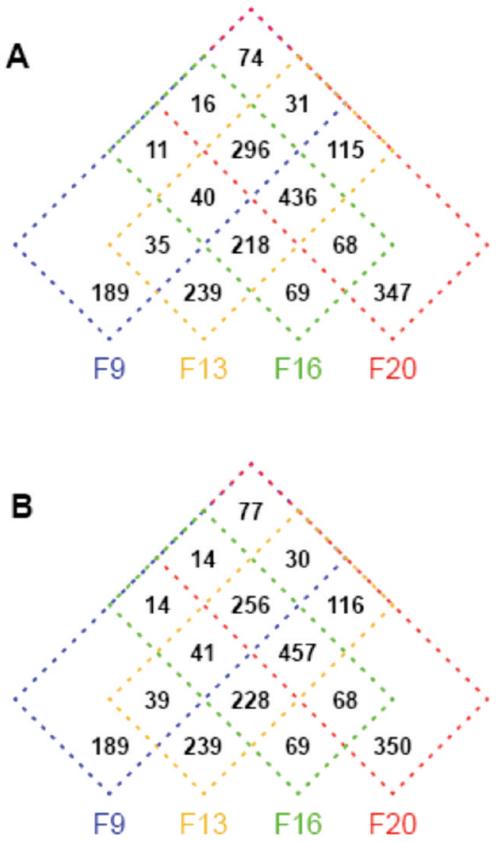

Figure 4. Four-way Venn diagrams summarizing the distribution of the 2184 proteins over different combinations of the sucrose gradient fractions. (A) Normal data set. (B) Trimmed data set.

order of mole fraction and graphed (Figure 6). The proteins which were observed in a single, unique gradient fraction have mole fractions of 1.0 at their location and correspond to the ordered data points at the right side of Figure 6. Two important points are evident from Figure 6. First, for the normal data set the number of (protein, location, abundance) data points that correspond to "trace" proportions of proteins in specific fractions is small. Only 5 of 4638 data points had mole fractions <0.01 and only 62 data points had mole fractions <0.04. Second, 4289 data points correspond to mole fractions >0.10. These latter data points alone already correspond to an average of 1.96 locations per protein for the 2184 proteins. The initial data set shows a higher number of data points corresponding to "trace" proportions of proteins in specific fractions. Twenty-four data points had mole fractions <0.01 and 65 had mole fractions <0.02.

For the normal data set, we recalculated the Venn diagram after eliminating the 62 data points with mole fractions <0.04. We emphasize that the 62 data points were reliably identified and that eliminating them represents a fairly strong test of the robustness of estimations of multiple subcellular locations. There are some changes in the Venn diagram (Figure 4B), but the estimation of multiplicity of locations is changed only moderately. Proteins unique to a single fraction go from 844 (38.6% of all proteins) to 847 (38.8%) and proteins ubiquitously observed in all fractions go from 296 (13.6%) to 256 (11.7%). The other 1081 proteins (49.4%) continue to be consistent with intermediate distribution over multiple, but not all, subcellular locations. Two-hundred forty-eight proteins (11.4%) show bimodal distributions over nonadjacent gradient fractions. We conclude that for the normal data set, neither trace amounts of proteins at secondary locations nor spurious tailing of abundant proteins in the gradient have any major influence on estimations of the proportion of proteins with multiple subcellular locations.

We have used previous subcellular location annotations in the UniProtKB database (in the keyword "subcellular location" field and the ontology "subcellular component" field) and in the Locate Subcellular Location database to compare three aspects of the present work with earlier work: (1) the degree to which the individual sucrose gradient fractions are enriched with proteins corresponding to specific subcellular organelles; (2) the extent to which the multiplicity of subcellular locations observed here is reflected in current annotations of subcellular locations; and, (3) the extent to which there are discrepancies between this work and previous annotations of subcellular locations (see Supplementary Table 1, Supporting Information, for annotation information on the individual proteins).

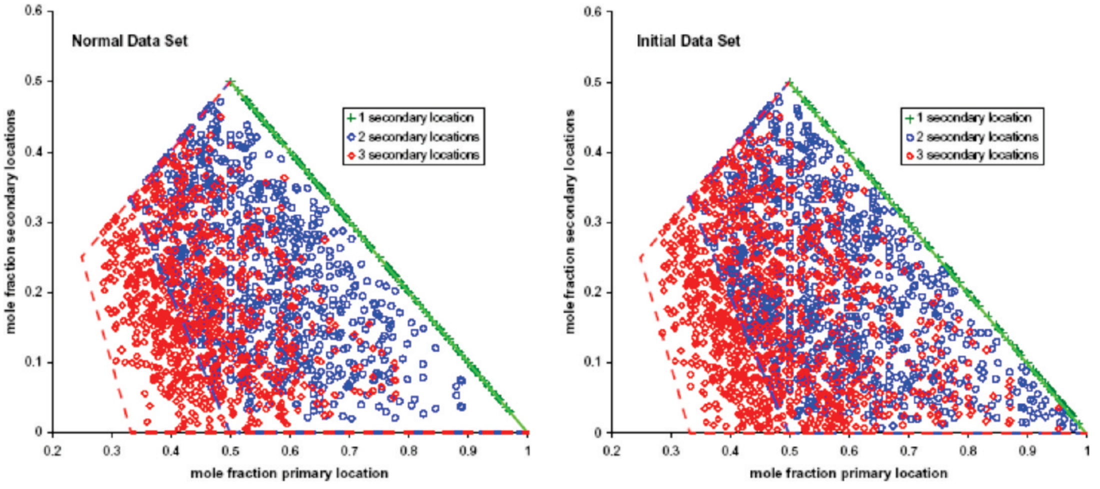

Figure 5. Distribution of proteins with a primary location and 1 (green), 2 (blue), or 3 (red) secondary locations over compatible areas of a plot of primary mole fractions vs secondary mole fractions. For each protein, the spectral counts observed in a specific gradient fraction were expressed as mole fractions of the total number of spectral counts observed in all four gradient fractions. (Left) Normal data set. (Right) Initial data set.

Table 2. Tests for Overlap of Proteins Between Sucrose Gradient Fractions[a]

	normalized protein abundance[b]				
accession number	F9	F13	F16	F20	protein
			Overlap from F9 to F13		
	top[c]	ND[c]	all[c]	all[c]	
4758516	15.00	–	–	–	**Hepatoma-derived growth factor**
157829557	11.92	–	–	0.77	Carbonic anhydrase 2
36038	11.76	–	–	0.49	**Rho GDP-dissociation inhibitor 1**
4502105	10.97	–	–	0.31	Annexin A4
4506387	9.53	–	0.24	1.47	UV excision repair protein RAD23 homologue B
7023053	7.79	–	–	1.95	**Sialic acid synthase**
4505701	7.37	–	–	1.28	**Pyridoxal kinase**
			Overlap from F13 to F9		
	ND[c]	top[c]	all[c]	all[c]	
24307879	–	7.25	1.58	0.42	Cytoplasmic dynein 1 intermediate chain 2
68533125	–	6.55	1.28	0.72	**ATP-citrate synthase**
34366439	–	6.33	1.89	0.11	Cytoplasmic dynein 1 light intermediate chain 1
30749633	–	6.00	1.78	0.22	Tyrosine-protein phosphatase nonreceptor type 1
38570062	–	5.81	0.61	0.61	UPF0363 protein C7orf20
620110	–	5.35	1.47	0.21	Coatomer subunit beta
24307879	–	7.25	1.58	0.42	UTP--glucose-1-phosphate uridylyltransferase
			Overlap from F13 to F16		
	all[c]	top[c]	ND[c]	all[c]	
4506773	–	5.26	–	–	**Protein S100-A9**
18655500	–	4.24	–	–	tr|Q6GMX0|Q6GMX0_HUMAN Putative uncharacterized protein
12054072	–	3.03	–	–	Ig gamma-1 chain C region
5454024	–	2.99	–	0.37	Ribonuclease P protein subunit p30
4826659	0.36	2.89	–	0.36	F-actin-capping protein subunit beta
22726189	–	2.65	–	0.38	Proteasome assembly chaperone 2
13876386	–	2.51	–	0.73	Epiplakin
			Overlap from F16 to F13		
	all[c]	ND[c]	top[c]	all[c]	
4506645	–	–	11.43	–	60S ribosomal protein L38
51036603	–	–	4.17	–	Guanine nucleotide-binding protein G(I)/G(S)/G(O) gamma-12
4506761	–	–	4.12	–	Protein S100-A10
4507129	–	–	4.00	2.00	Small nuclear ribonucleoprotein E
5454090	–	–	3.75	2.00	Translocon-associated protein subunit delta
6005860	–	–	3.2	2.00	60S ribosomal protein L35
7661728	–	–	3.2	–	Mitogen-activated protein-binding protein-interacting protein
			Overlap from F16 to F20		
	all[c]	all[c]	top[c]	ND[c]	
4506645	–	–	11.43	–	60S ribosomal protein L38
10190712	–	0.96	8.65	–	Protein S100-A14
150010589	–	0.80	7.20	–	Interferon-induced transmembrane protein 1
17933772	–	2.00	4.80	–	Protein S100-A16
51036603	–	–	4.17	–	Guanine nucleotide-binding protein G(I)/G(S)/G(O) gamma-12
4506761	–	–	4.12	–	Protein S100-A10
3462883	–	1.32	3.51	–	Vesicle transport protein SEC20
			Overlap from F20 to F16		
	all[c]	all[c]	ND[c]	top[c]	
1483131	–	0.34	–	12.24	**Nucleophosmin**
8922331	–	–	–	11.49	Protein mago nashi homologue 2
34201	–	–	–	10.91	60S ribosomal protein L35a
399758	–	–	–	9.52	Heterogeneous nuclear ribonucleoprotein A3
7706425	–	–	–	9.38	U6 snRNA-associated Sm-like protein LSm8
11037094	–	–	–	8.70	tr|Q9HC85|Q9HC85_HUMAN Metastasis related protein
1232077	–	1.44	–	8.19	DNA replication licensing factor MCM2

[a] Proteins where the name is shown in bold correspond to proteins which exemplify general characteristics of the data that are noted in the text. [b] Normalized abundances were calculated from the Spectral Abundance Factor using GeneSpring; that is, the abundances have been normalized using a correction for the differing number of amino acids in the proteins (see Experimental Procedures). For all proteins, the normalized abundances ranged from 0.018 to 22.25. A dash indicates the protein was not detected. [c] Selection criteria. A filter to select nondetected proteins was applied to a chosen fraction (ND). In an adjacent fraction in the sucrose gradient, the proteins were sorted according to abundance and the seven most abundant proteins (top) are shown. For the other two fractions, no filter was applied and all proteins were included (all).

In evaluating these comparisons, it is important to keep in mind that there is not an exact mesh between our experimental strategy and the ontological descriptions of subcellular location used in the databases. The top level of our experimental design

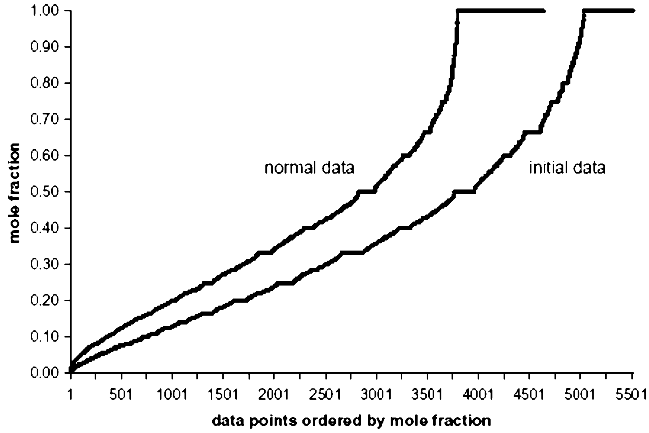

Figure 6. Ordered distribution of the mole fractions observed for all 2184 proteins. The (protein, location, mole fraction) data points observed over all proteins and all sucrose gradient locations were grouped, sorted in ascending order of mole fraction, and graphed (4638 data points for the normal data set, 5514 data points for the initial data set).

matches the levels (extracellular region, plasma membrane, cytoplasm, nucleus) in the GO classification scheme, but the experiment excludes the extracellular region and the nucleus. At a lower level we only tried to obtain an approximate resolution of the cytoplasm as (cytosol, endoplasmic reticulum, mitochondria), while the databases typically use (cytoplasm/cytosol, endoplasmic reticulum, mitochondrion, Golgi apparatus). Overall, relative to the UniProtKB subcellular locations, 271 proteins had no annotations, 1388 had annotations at the top level and 525 had annotations at the lower level.

For the 481 (22.0%) proteins in the initial data set that were observed in only a single fraction, we compared their locations with previous experimental information about subcellular location in the UniProtKB database. Figure 7 summarizes the proportion of these "unique" proteins which were previously assigned to various subcellular locations. This data provides an overview of the enrichment of the four fractions with cytosolic, plasma membrane, endoplasmic reticulum and mitochondrial proteins respectively. First, all four fractions show a substantial proportion of proteins either for which there is no previous annotation of subcellular location, or for which the previous annotation is only nucleus or extracellular region (from 5 (19%) of proteins in F16 to 83 (40%) of proteins in F20). These annotations are compatible with the enrichment of the fractions with their various types of proteins and the present results constitute new annotation information for these proteins. Fraction 9 shows three other major slices: (1) proteins which are fully compatible with cytosolic proteins, (2) proteins which have previously been assigned to cytoplasm, but also to other subcellular locations, and (3) proteins which have been previously assigned to other subcellular locations, but not to cytoplasm or cytosol. There is some ambiguity in the second and third groups since cytosol is not distinguished in many experimental strategies and the assigned locations are daughters of cytoplasm (but not of cytosol) in the GO ontology. Overall for the 127 proteins observed only in fraction 9, 119 (93.7%) have annotations that are compatible with enrichment of this fraction with cytosolic proteins. Only 8 proteins (6.3%) appear to be discrepancies that have other, incompatible locations. Of the 119 compatible proteins, 16 proteins have previous annotations that deviate from observation uniquely in fraction 9. For the other sucrose gradient fractions the cytoplasm/cytosol distinction also leads to some ambiguity, but overall the number/proportion of proteins compatible with enrichment of fraction 13 (plasma membrane), fraction 16 (endoplasmic reticulum) and fraction 20 (mitochondrion) with the respective protein types are 94 (78.3%), 18 (67.0%), and 184 (88.9%) respectively. Because there is some inconsistency between the different subcellular location annotation sources (see below), these numbers vary somewhat if the UniProt subcellular components or the Subcellular Location database are used, but do not change the overall conclusion. Within the limitations of such comparisons, we conclude that the previous annotations are largely consistent with enrichment of the fractions with the expected protein types. Apparent experimental/database annotation discrepancies for all 2184 proteins are considered in more detail below.

Is the apparent multiplicity of protein subcellular locations observed in our experiments captured in current database annotations? To address this question, we used the set of 163 proteins in the initial data set that showed bimodal, nonadjacent distributions over the sucrose gradient fractions (includes proteins observed only in combinations of fractions 9−16, 9−20, 13−20, 9−13−20, and 9−16−20, i.e. proteins that clearly have multiple locations) and which also had at least 8 spectral counts. The latter condition ensures that the classification of these proteins as bimodal is not unduly influenced by the dynamic range limitations of MS/MS spectral counting (see discussion). This set of proteins was compared with (merged) subcellular location annotations from the UniProtKB and LOCATE Subcellular Location databases. Figure 8 shows the distribution over the bimodal combinations of fractions and the annotations of subcellular location for all 163 proteins. As seen above with the proteins identified in only a single fraction, 59 (36.2%) of the bimodal proteins only had annotation at the level (nucleus, extracellular region, no annotation). Furthermore, only 22 (13.5%) of the proteins show multiple locations at the annotation level (cytoplasm/cytosol, plasma membrane, endoplasmic reticulum, Golgi apparatus, mitochondrion). In general these results are consistent with the conclusion that current database annotations of subcellular location are sparse and skewed toward single locations for proteins (see discussion).

Over all of the 2184 proteins, the annotations at the subcellular level in the examined databases tend to be to single locations. Given that many previous proteomics studies were biased against detection of proteins in multiple locations (e.g., studies of purified organelles) and that annotations at subcytoplasmic levels are clearly still very sparse, we consider that the previously available annotations of experimental data are not inconsistent with the proposal that many, probably a sizable majority, of the proteins have multiple subcellular locations.

Using the initial data set of (protein, fraction) pairs, there were a relatively small number of discrepancies between our data and previous annotations of subcellular location in the two databases. Of the 1441 proteins identified in fraction 16, there were a total of 33 proteins previously annotated to endoplasmic reticulum that we did not observe in fraction 16. Similarly for fractions 13 (1611 proteins) and 20 (1610 proteins), there were a total of 58 and 29 proteins previously annotated to plasma membrane and mitochondrion respectively that we did not observe in the corresponding gradient fraction.

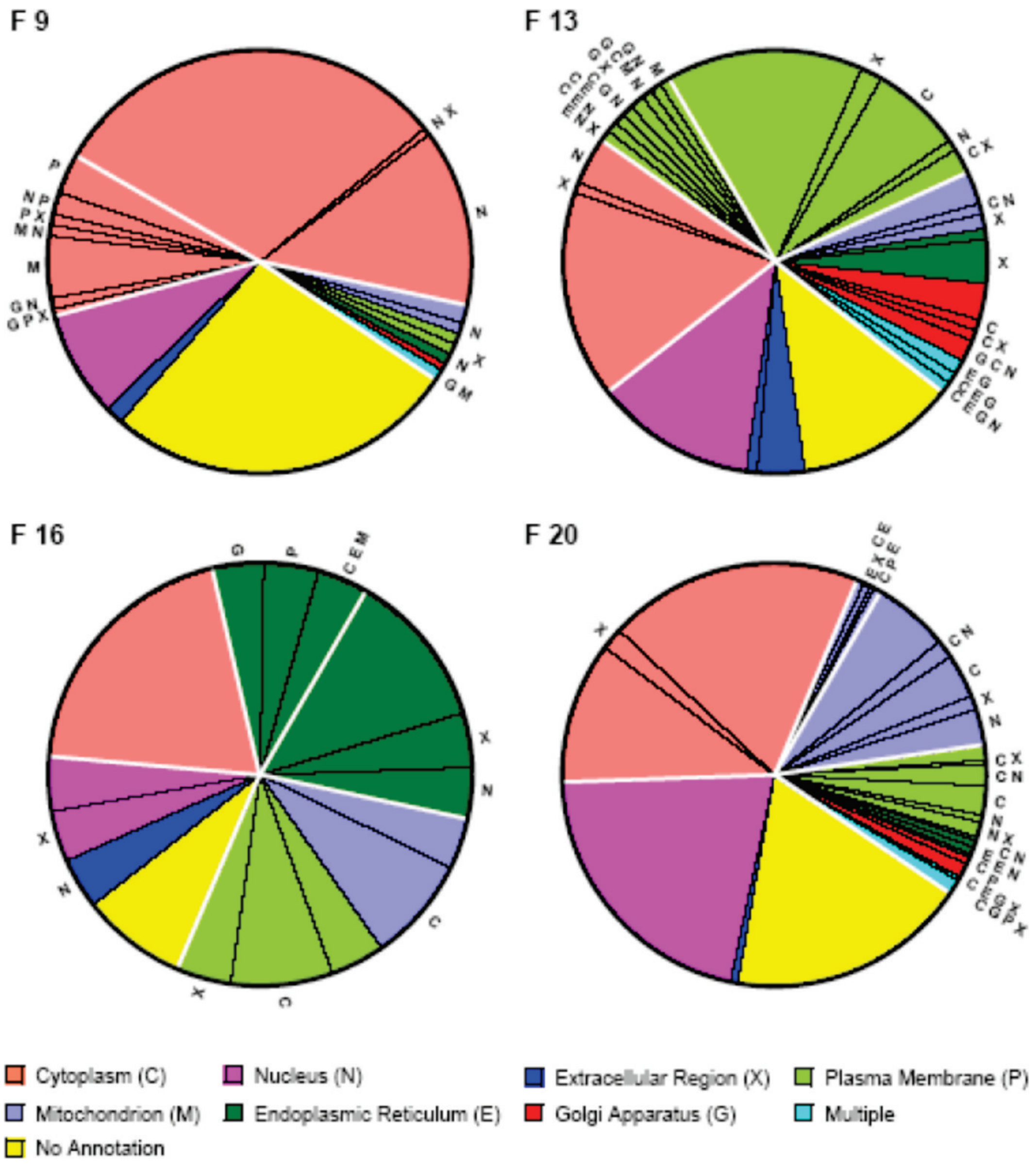

Figure 7. Distribution of current subcellular location annotations in the UniProtKB database over the proteins observed solely in a single sucrose gradient fraction in the initial data set. (F9) 129 proteins in fraction 9. (F13) 116 proteins in fraction 13. (F16) 27 proteins in fraction 16. (F20) 209 proteins in fraction 20. The annotations are color coded (legend) according to the GO classification levels compatible with our experimental strategy (upper: extracellular region, plasma membrane, cytoplasm, nucleus; lower: cytosol/cytoplasm, endoplasmic reticulum, Golgi apparatus, mitochondrion). A small number of proteins had multiple lower level annotations and are shown in the region color coded as multiple. Proteins that had multiple annotations that included other locations different from the color code are indicated by the radial letters. The heavy white lines delineate slice regions that have different compatibility with the experimental data (see text).

Inconsistencies in the databases might contribute to the apparent discrepancies. For the 2184 proteins identified here, Figure 9 shows the status of annotations of plasma membrane (443 proteins), mitochondrion (168) and endoplasmic reticulum (243) proteins. There is rather little concordance between the annotation sets, which presumably must reflect the inclusion of very different experimental data sets. Only 8 of the 443 proteins with annotations of plasma membrane were so annotated in all three data sets! For the proteins annotated to plasma membrane, endoplasmic reticulum and mitochondrion that we did not observe in the corresponding gradient fractions, our data would suggest different primary locations for these 120 proteins, but does not exclude their presence in the annotated subcellular locations as secondary locations which could not be detected at our sensitivity limits (see Discussion).

We believe that some occurrences of apparent discrepancies are almost inevitable for three reasons. First, there is still very little information about whether subcellular distributions of proteins are the same in different cell types or under different cellular conditions. Second, many experiments do not distinguish between different isoforms of the same protein, which may have different subcellular distributions. Indeed, the present data set includes these proteins, which in part show different distributions over subcellular locations for isoforms of the same protein. This data will be analyzed in a separate paper. Third, the databases attempt to aggregate data from experimental

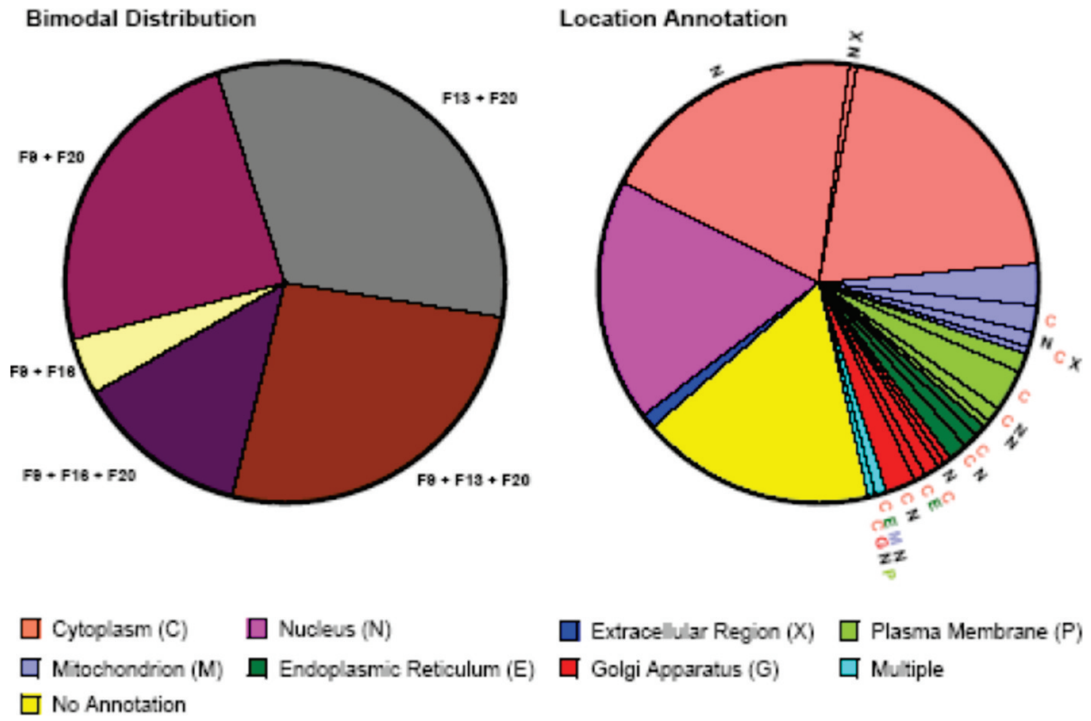

Figure 8. Comparison of the present data on subcellular location of bimodal proteins with (merged) subcellular location annotations in the UniProtKB subcellular location comments, UniProtKB subcellular component GO terms and LOCATE Subcellular Location database. (left) Distribution of the 299 bimodal proteins in the initial data set over different combinations of sucrose gradient fractions. The indicated combinations of fractions can only arise for proteins with at least two different subcellular locations. (right) Summary of the (merged) subcellular location annotations for all 299 bimodal proteins. The annotations are color coded (legend) according to the GO classification levels compatible with our experimental strategy (upper: extracellular region, plasma membrane, cytoplasm, nucleus; lower: cytosol/cytoplasm, endoplasmic reticulum, Golgi apparatus, mitochondrion). A small number of proteins had multiple lower level annotations and are shown in the region color coded as multiple. Proteins that had multiple annotations that included other locations different from the color code are indicated by the radial letters. Slices that have color coded radial lettering are those corresponding to proteins whose annotations indicate multiple subcellular locations within the GO classification levels of our experimental design (see text).

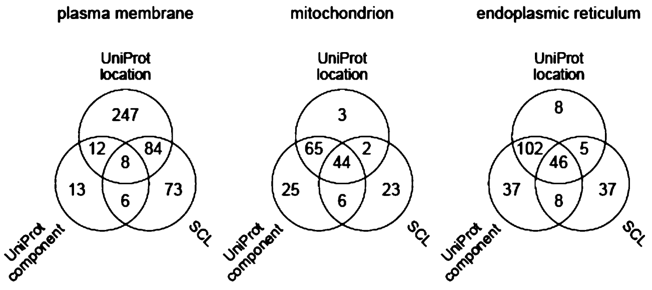

Figure 9. Venn diagrams comparing the status of annotations of subcellular locations in the UniProtKB subcellular location comments, UniProtKB cellular component GO terms and LOCATE Subcellular Location database for proteins observed in this work. (Left) Four-hundred forty-three proteins annotated as plasma membrane. (Middle) One-hundred sixty-eight proteins annotated as mitochondrion. (Right) Two-hundred forty-three proteins annotated as endoplasmic reticulum.

strategies with very different sensitivity, selectivity, dynamic range, and coverage of proteins. Targeted searches for individual proteins in purified subcellular fractions with antibody methods probably have the highest sensitivity for detecting trace amounts of proteins in any specified location, even if the trace is a tiny proportion of the total protein abundance. Conversely, some high throughput methods may have limited resolution for some subcellular locations, for example, distinguishing cytosol from cytoplasm, and may have insufficient sensitivity and dynamic range to detect trace amounts of proteins in specific locations. Aggregating subcellular location information from many cell types and conditions obtained with very different experimental strategies, many of which do not distinguish protein isoforms, then becomes a very tricky task which seems likely to produce some discrepancies with any specific experimental method/data set.

Although only a few of the fractions from the sucrose density gradient have been analyzed, the normal data set provides clear evidence that a minimum of 543 of the 2184 proteins (24.9%) show multiple locations. The minimum estimate is based on those proteins that are either present in all fractions or show bimodal distributions with abundance peaks in nonadjacent fractions of the sucrose gradient (Figure 3B, C). For the 321 proteins (14.7%) that were found only in adjacent fractions of the gradient (i.e., 9−13, 13−16, and 16−20), the present experiments are insufficient to exclude that this might be due to the presence of a single organelle that occupies an intermediate position between the two fractions. On the other hand, we intentionally spaced the analyzed fractions widely in the sucrose gradient and for the 476 proteins (21.8%) that were found in three adjacent fractions (i.e., 9−13−16, or 13−16−20), it is improbable that these proteins have single subcellular locations. Especially since other proteins demonstrated lack of overlap (e.g., proteins in fractions 9−16 or 13−20) and lack of tailing in the sucrose gradient (Table 2). Furthermore, in most cases the relative abundances for the proteins observed in three adjacent fractions were substantial and did not correspond to

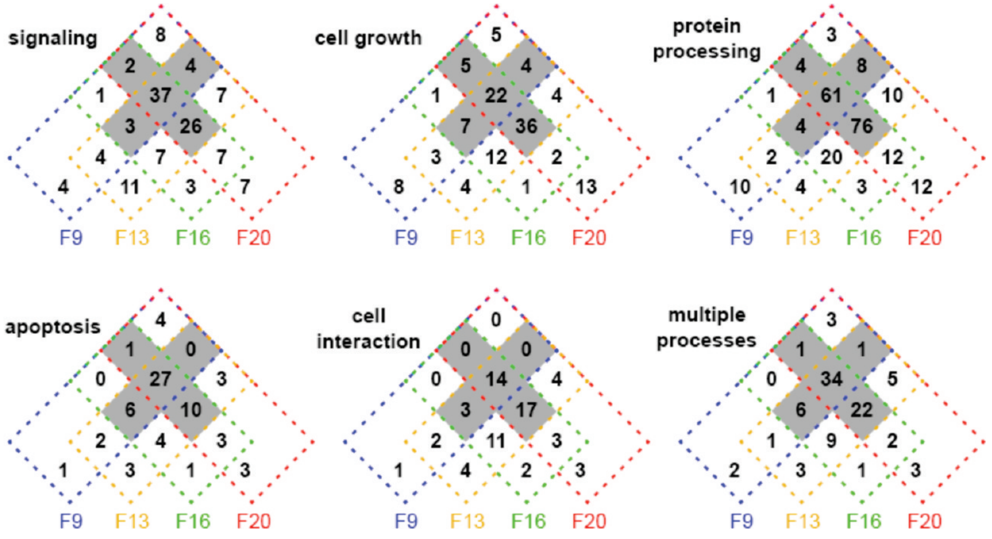

Figure 10. Four-way Venn diagrams summarizing the distribution of the breast-cancer-related set of 519 proteins over the subcellular locations for the cellular processes: signaling (131 proteins), cell growth (127), protein processing (230), apoptosis (68 proteins), and cell interaction (62), as well as for proteins involved in more than one of these cellular processes (93). The shaded regions of the diagrams correspond to proteins with 3 or 4 locations.

trace proportions. Thus, the normal data set provides evidence indicating that 38.6% of the proteins may have unique locations, 24.9% certainly have multiple locations, 21.8% most likely have multiple locations and 14.7% may have either unique or multiple locations.

We have used the observed set of proteins to examine possible connections between subcellular location and function as related to breast cancer. Many of the proteins observed in our experiments have previously been annotated with functional information. Biological process annotations for 1673 proteins, molecular function annotations for 1980 proteins, Reactome Pathway annotations for 176 proteins and post-translational modification annotations for 1653 proteins were available in the UniProt KB database. Supplementary Table 1 (Supporting Information) includes these functional annotations for all 2184 proteins. We used the BioBase Biological Databases, BIOBASE Knowledge Library (BKL) and ExPlainTM 2.3 platform to identify 94 proteins in our data set that are known or suspected to be implicated in breast cancer via disease molecular mechanism, diagnostic marker and therapeutic target association. These proteins were examined for common Gene Ontology (http://www.geneontology.org) biological process and molecular function terms and for common Reactome Pathway (http://www.reactome.org) terms, which were then used as lures to obtain the set of proteins identified in this study that share the same terms.

A majority of the proteins implicated in breast cancer were related to five high level cellular processes that involved a subset of 519 proteins observed in our experiments: apoptosis (68 proteins), cell growth (127), signaling (131), cell interaction (62), and protein processing (230). 93 proteins were involved in more than one of the five processes. Supplementary Table 1 (Supporting Information) includes a more detailed breakdown of these cellular processes to GO daughter terms for the individual proteins. Figure 10 shows how the proteins associated with each cellular process are distributed over the subcellular locations using the initial data set. The striking features are that each process is distributed over all four locations, as might be anticipated for regulated processes, and that for all of the cellular processes there is an appreciable majority of proteins with 3−4 subcellular locations (ranging from 54.8% for cell interaction to 66.5% for protein processing). Furthermore, the latter characteristic was most pronounced for the 93 proteins that were involved in more than one of the high level cellular processes (68.8%). We consider this data further in the discussion.

Discussion

In recent years, considerable effort has been devoted to determining the identities of proteins included in different subcellular organelles by proteomics.[24−28] The most common approach has been purification of individual organelles followed by exhaustive determination of the protein content. The main disadvantages of this approach are that the degree of purification/contamination of the organelle is difficult to ascertain conclusively for lower abundance proteins, that the protein content may be altered by the purification process and that the approach is not very suitable for dynamic studies of protein subcellular location. In a few cases,[22,23] an alternative approach of partial purification of organelles in a sucrose gradient has been employed, but the assignment of proteins to individual organelles has been based on matching gradient profiles of proteins to the profiles of presumptive marker proteins. This is useful for identifying what might be denominated core proteins of an organelle, but is automatically biased against evaluation of proteins in multiple subcellular locations.

There is already substantial evidence that many proteins exist in multiple subcellular locations, for example, in recent years very extensive research on nuclear import/export of proteins has been undertaken.[29] Numerous examples of proteins that can be located in the nucleus, but also in subcellular organelles of the type included in our experiments are already known.[30,31] In the present experiments we detected 268 nuclear proteins and 22 extracellular region proteins that were found in various sucrose gradient fractions, but which had previously only been annotated experimentally to the nucleus and extracellular region respectively. Another 271 proteins that we detected had no prior annotation at either the upper level (cytoplasm, plasma membrane) or lower level (cytosol, endoplasmic reticulum, mitochondria) of our experimental strategy. The present experiments were not designed to obtain specific annotations at the lower level, e.g. to mitochondrion. Hence, observation of a protein in a Fraction 20 that is enriched in mitochondrial proteins should presently only be taken as an indication and not as proof of its presence in mitochondria. Nonetheless, the present experiments gave several hundred new location annotations at the level (plasma membrane, cytoplasm).

There are several ways the limits on MS detection sensitivity may influence the number of locations in which the proteins were observed. In particular, for the highest abundance proteins, the sensitivity and dynamic range of the MS spectral counting methods are such that trace amounts as small as about 0.2% of a protein in a secondary location could be detected. As shown above for the normal data set, trace amounts of abundant proteins in secondary locations do not strongly influence estimates of the proportion of proteins with multiple subcellular locations. Conversely, the proportion of a protein which must be present in a secondary location to be detectable increases as the overall abundance of the proteins decreases, for example, for the lowest abundance proteins, only the highest abundance, primary location falls within the detection limits of the MS methods. Furthermore, for lower abundance proteins or for trace proportions of proteins in specific fractions, the sampling constraints on spectral counting that result from MS/MS sequencing of only the more abundant peptides[20] means that only one peptide may be counted in some fractions. For example, there were 847 (38.8%) proteins classified as "unique" (observed in a single fraction) in the normal data set, but only 481 (22.0%) in the initial data set. This difference corresponds to proteins in various gradient fractions that were only counted with a single peptide and 1 or 2 spectral counts. This means that estimations of multiple locations based on the normal data set are very conservative and certainly underestimate, probably strongly, the proportion of proteins with multiple subcellular locations. Given that estimates based on the normal data set provide evidence for multiple locations of at least 46.7% of the observed proteins, we conclude that a substantial majority of the proteins observed have multiple subcellular locations. Given that only 22% of proteins were seen solely in a single fraction in the initial data set, perhaps as much as 75% of the proteins have multiple locations.

We noted above that 120 proteins had annotations to subcellular locations that we did not observe in the corresponding sucrose gradient fractions (33 to endoplasmic reticulum, 58 to plasma membrane and 29 to mitochondrion). We suggested that these discrepancies were not inconsistent with our data if the annotations corresponded to secondary locations. On the basis of the observed spectral counts, there are 39 of these proteins for which our data suggest that the previous annotations correspond to proteins with functional significance in a secondary location, but that >80% of the protein is in a different primary location. This kind of analysis can be extended to many other proteins where the functional activity and the measured mole fractions indicate functional roles at secondary locations. Indeed, some of the proteins that we detected at trace amounts (<3%) in secondary locations already have known functions at those locations. The present experiments thus indicate numerous proteins with primary locations which probably differ from current function/location annotations and for which confirmation of the primary location (and potentially of other functional activities) might be profitably sought.

More generally, the existence of subcellular structures should be expected to lead to proteins with multiple subcellular spatial locations. Cellular regulation will certainly require coordination of functional activities carried out at different spatial locations and it would be surprising if distribution of proteins over multiple locations were not an inherent part of the regulation. Effective regulation, e.g. in response to changed cellular state, suggests that the distribution of proteins over multiple locations is likely to be dynamic. Focusing on the importance of multiple subcellular locations for proteins then suggests new ways of viewing the present data. For example, the localization of individual proteins to two or more locations might be connected to a hierarchy of importance of the protein in cellular regulation. One could anticipate that proteins that distribute between two subcellular locations either are involved in coordination of an individual function/process at each location or are involved in coordinating different cellular processes by participating in different processes at different locations. Those proteins which are located in and can dynamically redistribute among many subcellular locations then might be the most important to cellular regulation. The data for the set of 519 proteins either directly associated with breast cancer or sharing high level cellular processes is consistent with this hypothesis. In all five high level cellular processes, a majority of the proteins were identified in 3−4 locations (Figure 10). If present annotations of proteins to high level cellular processes are reasonably complete, the fact that only 93 of 519 proteins were found to be involved in multiple high level cellular processes would suggest that most multiplicity of location involves regulation of single cellular processes at different locations and then suggest that the 93 proteins involved in multiple processes may have particularly critical roles in cellular regulation related to breast cancer. It seems clear that further information at the (function, location) level is urgently needed. This also serves as a reminder that the databases ought to be organized around (function, location) pairs when possible.

At face value, the present experiments suggest an overwhelming number of (protein, location) pairs for further functional investigation. There is thus a need for high-throughput methods for better defining the locations and for testing the functional relevance of such (protein, location) pairs. Ultimately, we believe that the most productive way for testing whether a given protein may have some regulatory/functional role(s) in a particular subcellular organelle is probably via the analysis of induced dynamic changes in the abundance/form of the protein in the organelle, as opposed to ever better purification of single organelles or ever finer analysis of protein profiles along sucrose gradients. Direct spectral counting of the thousands of available peptides (Table 1) allows definition of reasonable distribution profiles for a limited number of organelles in a sucrose gradient. For individual proteins, where only a few peptides may be observable in fractions corresponding to any given organelle, it will probably be desirable to supplement direct spectral counting with methods that allow more complete detection of dynamic abundance/form changes within the same sucrose gradient fraction. The most efficient approach seems to be encoding dynamic changes with proteomics labeling techniques suitable for detecting abundance changes in the same fraction, while using direct spectral counting to define organelle distribution in a sucrose gradient.

This would allow maximum information to be extracted from the extensive MS measurements needed for a single sucrose gradient (Table 1). Such labeling experiments also offer additional ways to detect any potential artifacts of organelle separations using sucrose gradient fractionation. In the end this amounts to directly analyzing spatial/temporal responses concurrently.

The present experiments suggest 1383 (protein, location, function) data points for 519 proteins involved in five major cellular functional processes for which investigation of functional roles might further elucidate mechanisms involved in breast cancer. This is a very promising situation for experiments aimed at investigating dynamic changes in the spatio/temporal location/form of proteins in MCF-7 cells, their potential roles in regulation and their potential importance in breast cancer. In separate experiments (to be published), we have investigated the time scales which are relevant for studying dynamic changes in subcellular location/protein isoform following stimulation of MCF-7 cells with estrogen.

Finally, in summary, we have found evidence that strongly suggests a majority of the detected proteins have multiple subcellular locations in MCF-7 cells, that even with a fairly simple experiment a wealth of new annotation data can be obtained, that available evidence suggests that for many proteins distribution over multiple subcellular locations can be important to their functional roles, and that large numbers of (protein, location) pairs deserving of further investigation of functional/regulatory roles can be delineated. We are still very far from having good static descriptions of the spatial distributions of cellular proteins, let alone dynamic information on relationships between spatio/temporal distribution and function. However, high-throughput proteomics in combination with other experimental methods seems to offer ways forward.

Acknowledgment. This work was supported by The Wellcome Trust grant to J.G.Z. and King Faisal Foundation scholarship to A.Q.

Supporting Information Available: Supplementary Table 1. Identity Mapping and Annotation Data for the 2184 Proteins. Supplementary Table 2. MS Data for the 2184 Proteins. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) http://www.ncbi.nlm.nih.gov.

(2) http://www.imaginis.com/breasthealth/statistics.asp#1.

(3) Ince, T. A.; Weinberg, R. A. Functional genomics and the breast cancer problem. *Cancer Cell* **2002**, *1*, 15–17.

(4) Van't Veer, L. J.; Dai, H.; van de Vijver, M. J.; He, Y. D.; Hart, A. A.; Mao, M.; Peterse, H. L.; van der, K. K.; Marton, M. J.; Witteveen, A. T.; Schreiber, G. J.; Kerkhoven, R. M.; Roberts, C.; Linsley, P. S.; Bernards, R.; Friend, S. H. A gene-expression signature as a predictor of survival in breast cancer. *Nature* **2002**, *415*, 530–536.

(5) Sorlie, T.; Perou, C. M.; Tibshirani, R.; Aas, T.; Geisler, S.; Johnsen, H.; Hastie, T.; Eisen, M. B.; van de, R. M.; Jeffrey, S. S.; Thorsen, T.; Quist, H.; Matese, J. C.; Brown, P. O.; Botstein, D.; Eystein, L. P.; Borresen-Dale, A. L. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10869–10874.

(6) Sebastian, T.; Johnson, P. F. Stop and go: anti-proliferative and mitogenic functions of the transcription factor C/EBPbeta. *Cell Cycle* **2006**, *5*, 953–957.

(7) Hanahan, D.; Weinberg, R. A. The hallmarks of cancer. *Cell* **2000**, *100*, 57–70.

(8) Roberts, G. C.; Smith, C. W. Alternative splicing: combinatorial output from the genome. *Curr. Opin. Chem. Biol.* **2002**, *6*, 375–383.

(9) Godovac-Zimmermann, J.; Kleiner, O.; Brown, L. R.; Drukier, A. Perspectives in spicing up proteomics with splicing. *Proteomics* **2005**, *5*, 699–709.

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

(11) Lacroix, M.; Leclercq, G. Relevance of breast cancer cell lines as models for breast tumours: an update. *Breast Cancer Res. Treat.* **2004**, *83*, 249–89.

(12) Soule, H. D.; Vazquez, J.; Long, A.; Albert, S.; Brenan, S. A human cell line from a pleural effusion derived from a breast carcinoma. *J. Natl. Cancer Inst.* **1973**, *51*, 1409–1413.

(13) Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **1970**, *227*, 680–685.

(14) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **1996**, *68*, 850–858.

(15) Usaite, R.; Wohlschlegel, J.; Venable, J. D.; Park, S. K.; Nielsen, J.; Olsson, L.; Yates III, J. R. Characterization of global yeast quantitative proteome data generated from the wild-type and glucose repression saccharomyces cerevisiae strains: the comparison of two quantitative methods. *J. Proteome Re.s* **2008**, *7*, 266–275.

(16) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383–5392.

(17) Kulasingam, V.; Diamandis, E. P. Proteomics analysis of conditioned media from three breast cancer cell lines: a mine for biomarkers and therapeutic targets. *Mol. Cell. Proteomics* **2007**, *6*, 1997–2011.

(18) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, 4646–4658.

(19) Florens, L.; Carozza, M. J.; Swanson, S. K.; Fournier, M.; Coleman, M. K.; Workman, J. L.; Washburn, M. P. Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **2006**, *40*, 303–311.

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

(21) Old, W. M.; Meyer-Arendt, K.; Aveline-Wolf, L.; Pierce, K. G.; Mendoza, A.; Sevinsky, J. R.; Resing, K. A.; Ahn, N. G. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* **2005**, *4*, 1487–1502.

(22) Dunkley, T. P.; Watson, R.; Griffin, J. L.; Dupree, P.; Lilley, K. S. Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics* **2004**, *3*, 1128–1134.

(23) Foster, L. J.; de Hoog, C. L.; Zhang, Y.; Zhang, Y.; Xie, X.; Mootha, V. K.; Mann, M. A mammalian organelle map by protein correlation profiling. *Cell* **2006**, *125*, 187–199.

(24) Simpson, J. C.; Pepperkok, R. The subcellular localization of the mammalian proteome comes a fraction closer. *Genome Biol.* **2006**, 222.

(25) Yates, J. R., 3rd.; Gilchrist, A.; Howell, K. E.; Bergeron, J. J. Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell. Biol.* **2005**, *6*, 702–714.

(26) Au, C. E.; Bell, A. W.; Gilchrist, A.; Hiding, J.; Nilsson, T.; Bergeron, J. J. Organellar proteomics to create the cell map. *Curr. Opin. Cell Biol.* **2007**, *19*, 376–385.

(27) Rogers, L. D.; Foster, L. J. The dynamic phagosomal proteome and the contribution of the endoplasmic reticulum. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 18520–18525.

(28) Xu, P.; Crawford, M.; Way, M.; Godovac-Zimmermann, J.; Segal, A.; Radulovic, M. Subproteome analysis of the neutrophil cytoskeleton. *Proteomics* **2009**, *9*, 2037–2049.

(29) Sorokin, A. V.; Kim, E. R.; Ovchinnikov, L. P. Nucleocytoplasmic transport of proteins. *Biochemistry (Mosc).* **2007**, *72*, 1439–1457.

(30) Andersen, J. S.; Lam, Y. W.; Leung, A. K.; Ong, S. E.; Lyon, C. E.; Lamond, A. I.; Mann, M. Nucleolar proteome dynamics. *Nature* **2005**, *433*, 77–83.

(31) Dundr, M.; Misteli, T. Nucleolomics: an inventory of the nucleolus. *Mol. Cell* **2002**, *9*, 5–7.

(32) www.ncbi.nlm.nih.gov, www.uniprot.org, http://locate.imb.uq.edu.au accessed August 2009.

PR9008332