MCP Papers in Press. Published on May 21, 2007 as Manuscript M700017-MCP200

Global Survey of Human T Leukemic Cells by Integrating Proteomic and Transcriptomic Profiling

Linfeng Wu[1], Sun-Il Hwang[1#], Karim Rezaul[1#], Long J. Lu[2#], Viveka Mayya[1], Mark Gerstein[2], Jimmy K. Eng[3], Deborah H. Lundgren[1], David K. Han[1]*

[1]Department of Cell Biology and Center for Vascular Biology, School of Medicine, University of Connecticut, Farmington, CT 06030, USA, [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA, [3]Fred Huchinson Cancer Research Center, Seattle, WA 98195

[#] These authors contributed equally to this paper.

[*]Corresponding author
Address correspondence to:
David K. Han
Center for Vascular Biology
Department of Cell Biology
University of Connecticut Health Center
263 Farmington Ave., Farmington, CT 06030
Tel: 860-679-2444
Fax: 860-679-1201
Email: han@nso.uchc.edu

Running Title

Global Survey of Human T Leukemic Cell

Abbreviations

PCNA: proliferating cell nuclear antigen

Lamp-1: Lysosome-associated membrane glycoprotein 1 precursor

GRP78: 78 kDa glucose-regulated protein precursor

F1α: ATP synthase mitochondrial F1 complex, alpha subunit 1

hnRNP C1/C2: Heterogeneous nuclear ribonucleoproteins C1/C2

CDK: Cyclin-dependent kinase

PECAM-1: platelet/endothelial cell adhesion molecule-1

MAPK: Mitogen-activated protein kinase

Bid: BH3-interacting domain death agonist

Lck: Lymphocyte cell-specific protein-tyrosine kinase

TMH: transmembrane helix

SUMMARY

Global protein survey is needed to gain systems-level insights into mammalian cell signaling and information flow. Human Jurkat T leukemic cells are one of the most important model systems for T cell signaling study, but no comprehensive proteomics survey has been carried out in this cell type. Here, we combine subcellular fractionation, multiple protein enrichment methods and replicate tandem mass spectrometry analyses to determine the protein expression pattern in a single Jurkat cell type. The proteome dataset was evaluated by comparison with the genome-wide mRNA expression pattern in the same cell type. A total of 5381 proteins were identified by mass spectrometry with high confidence. Rigorous comparison of RNA and protein expression afforded removal of the false positive identifications and redundant entries, but rescued the proteins identified by a single high-scoring peptide, resulting in the final identification of 6471 unique gene products, among which 98% of the corresponding transcripts detected with high probability. Using hierarchical clustering of the protein expression patterns in five subcellular fractions (cytosol, light membrane, heavy membrane, mitochondria, and nuclei), the primary subcellular localization of 2241 proteins was assigned with high confidence, including 792 previously uncharacterized proteins. This proteome landscape can serve as a useful platform for systems-level understanding of organelle composition and cellular functions in human T cells.

INTRODUCTION

An important goal in functional genomics is to globally profile protein expression and localization in biological systems. Many studies have utilized genome-wide cDNA or oligonucleotide microarrays to measure mRNA expression level, deducing the corresponding protein expression [1,2]. However, despite the obvious dependency of protein synthesis on mRNA, many studies have reported that more than half of the total transcripts are non-coding RNA [3-5]. In addition, quantitative measurements show only moderate or even poor correlation between protein and mRNA expression level due to different translational efficiency and posttranslational turnover [6-8]. Furthermore, subcellular localization of proteins can not be accurately predicted based on mRNA expression. Therefore, biological systems ultimately need to be explained at the level of proteins.

The completed draft sequences of the human genome [9,10] and several other organisms [11,12] combined with mass spectrometry have made large-scale proteomics feasible [13,14]. However, due to the huge diversity and dynamic range of expressed proteins, especially in human cells, identification of all or most of the expressed proteins in cells has remained as one of the greatest challenges [15]. Although many proteome-scale studies on different cells, tissues and subcellular organelles have been reported, no comprehensive analysis of a single human cell type has been carried out to date. In this study, we performed a comprehensive survey of a human Jurkat T

Leukemic cell line by combining proteomics and trasncriptomics profiling. Human Jurkat T leukemic cells are one of the most popular model systems for studying signal transduction since many key advances in the field of T-cell receptor (TCR) signaling were made using Jurkat T cells [16]. Moreover, this cell type is also used for studying other biological phenomena, such as apoptosis and cell engulfment [17,18]. Therefore, a global survey of human Jurkat T cells can serve as a platform for many in-depth characterizations of cellular function and signaling transduction. Moreover, in contrast to the recent survey of organ and organelle protein expression in mouse [8,19], our study was carried out in a single cell type, which makes it more suitable for studying protein network and signaling flow within cells.

EXPERIMENTAL PROCEDURES

Whole Cell Lysate Preparation

Human Jurkat A3 T leukemic cells from American Type Culture Collection (Bethesda, MD) were used for this study. Jurkat T cells were grown to a maximal density of 0.5-0.8 × 10^6 cells/ml, and then collected by centrifugation at 400 × g for 10 min at 4°C. The cell pellets were washed twice with ice-cold PBS. To obtain whole cell lysates, cells were re-suspended in lysis buffer (50 mM HEPES, pH 7.5, 100 mM NaCl, 1 mM EDTA, 1% Tween-20, and a cocktail of protease inhibitor [Roche Diagnosis GmbH, Mannheim, Germany]) on ice for 30 min, and then centrifuged at 12,000 × g for 20 min at 4°C. The supernatant (designated whole cell lysate) was collected and stored at -80°C for later analysis.

Subcelluar Fractionation

Subcelluar fractionation was carried out as described below. One volume of Jurkat cell pellet was incubated in five volumes of hypotonic buffer (buffer A: 20 mM HEPES, pH 7.5, 10 mM KCl, 1.5 mM $MgCl_2$, 1 mM EDTA, 1mM EGTA, 1 mM DTT and a cocktail of protease inhibitor) for 2 minutes, and then mixed with equal volume buffer A supplemented with 0.5 M sucrose resulting in a 0.25 M sucrose isolation buffer. After 10 minute incubation on ice, cells were homogenized with a glass Dounce homogenizer until ~50% of the cells became trypan blue positive. The homogenates were centrifuged at 650 × g for 10 minutes at 4°C. The post-nuclear supernatants were incubated on ice for other subcellular fraction preparation. The pellets were further homogenized in buffer A until ~95% of the cells became trypan blue positive. The nuclear pellets were isolated by centrifugation at 650 × g for 10 min at 4°C, and then rinsed with isolation buffer once. The nuclear pellets were gently re-suspended in 2.5 ml buffer A supplemented with 1.28 M sucrose, layered over 5 ml buffer A supplemented with 2.3 M sucrose, and then centrifuged at 60,000 × g for 90 minutes at 4°C. The purified nuclear pellets were rinsed with buffer A, centrifuged at 12,000 × g for 10 min and stored at -80°C for subsequent experiments.

The post-nuclear supernatants were used to isolate other subcellular fractions, including cytosol, heavy membrane, light membrane and mitochondria as described previously [20].

In order to isolate plasma membrane fraction, one volume of cells (~ 2 × 10^9) were re-suspended in five volumes of hypotonic buffer A, homogenized by a glass Dounce homogenizer 15 times. The homogenates were centrifuged at 650 × g for 10 minutes at 4°C and the pellet was removed. The supernatant was further centrifuged at 100,000 × g for one hour at 4°C. The pellet was re-suspended in buffer A supplemented with 0.25 M sucrose, dounced 10-15 times, layered on 6.5 ml sucrose buffer (buffer A supplemented with 38% sucrose), and then centrifuged at 200,000 × g for two hours. The membranes were collected from the phase between the 0.25 M and 38% sucrose, diluted in 10 ml 0.25 M sucrose buffer, and centrifuged at 100,000 × g for 1 hour. The pellet was designated as plasma membrane.

In order to isolate lipid raft fraction, the PBS-washed cell pellets were re-suspended in 1 ml lysis buffer (25 mM MES, pH 6.5, 150 mM NaCl, 0.1% Triton X-100, 1 mM sodium vanadate, 5 mM EDTA and a cocktail of protease inhibitor). Cells were dounced 20-30 times, and then mixed with 1 ml 80% sucrose in lysis buffer without Triton X-100. The lysates were placed in the bottom of a 14 X 89 mm clear centrifuge tube (Beckman), gently overlaid with 6.5 ml of 30% sucrose and 3.5 ml of 5% sucrose in lysis buffer without Triton X-100, and then centrifuged at 200,000 × g at 4°C for 18 hours (the machine was set at NO BRAKE condition). The low density membrane raft in the 5% sucrose fraction was collected and designated as lipid raft. The details of cell culture and other enrichment approaches are shown in supplemental materials.

In-gel Digestion and Nano-LC-MS/MS Analysis

Proteins were digested with trypsin and analyzed as described previously [20] with minor modifications as outlined below. Digested proteins were analyzed using a linear ion trap mass spectrometer (Finnigan LTQ; Thermo Finnigan, San Jose, CA). Samples were loaded onto a 10-cm × 100-μm capillary C18 reversed-phase column by a micro-autosampler (Famos, Dionex, Sunnyvale, CA) followed by LC-MS/MS analysis on the LTQ. For stable isotope-free peptide samples, each full MS scan was followed by five MS/MS scans of the five most intense peaks in the MS spectrum with dynamic exclusion enabled. The m/z scan range was either 300-1700 or 400-1700 for full mass range. For stable isotope-labeled peptide samples, which mainly come from subcellular fractions and phosphoprotein enrichment, each full MS scan was followed by one MS/MS scan of the most intense peak in the MS spectrum with dynamic exclusion enabled.

Database Searching and Data Processing

All the mass spectrometry raw files were converted to .dat files using Xcalibur software (version 1.4 SR1), which were then converted to mzXML using the conversion software dat2xml (http://regis-web.systemsbiology.net/rawfiles/converter/linux/dat2xml). Peak lists were generated automatically without smoothing and de-isotoping, and charge states assigned based on the MS and MS/MS scans as previously described [21]. The minimum signal count for full MS is 1000. All the mzXML files were searched against a local copy of the non-redundant human protein database (56,709 entries, Nov. 30th 2004 release version) from NCI's Advanced Biomedical Computing Center using the SEQUEST algorithm [SEQUEST-PVM v.27 (rev. 0)] [21]. SEQUEST parameters were outlined as follows: all the filtering thresholds were off; mass tolerance of 1.0 Da for precursor ions and 0.5 Da for fragment ions, full tryptic constraint allowing one missed cleavage, allowing oxidization (+16 Da) of methionine. If the peptides contain heavy isotope-labeled amino acids, the corresponding amino acid modification was also allowed. The detail description of labeled amino acids used in each experiment is shown in Table S1. The database search results were processed using INTERACT program [22] and filtered with the following criteria: *Xcorr* cutoff values are 1.9, 2.2, 3.7 for 1+, 2+, 3+ peptides, respectively; *ΔCn* cutoff value is 0.1; partially isotope-labeled peptides were excluded. Proteins identified by at least two distinct peptides in the same experimental fraction were collected for indept analysis. To estimate the false-positive rate, the datasets were searched against a forward and reversed concatenated human protein database [23]. To address redundancy issue in the list of identified proteins, the peptides filtered by the above *Xcorr* and *ΔC* values were used to compute peptide probability and protein probability using PeptideProphet (version 1.0) [24] and ProteinProphet (version 2.0) softwares [25], which combines the redundant proteins into a unique protein group and indicates whether the peptide sequences are unique to the corresponding protein group. The searching parameter was set as minimum probability 0.0 to include all the results for the in-dept analysis. In addition, a genome-wide transcript profiling of the same human Jurkat cell type was performed and compared to our proteomic dataset.

Transcript Profiling

The Sentrix Human-6 Expression BeadChip (Illumina, San Diego, CA) that contain 50-mer gene-specific oligonucleotide probes corresponding to >46,000 human transcript variants were used in this study. There are on average 30x redundancy for each transcript per array. Total RNA was isolated from Jurkat cells at log phase using guanidine thiocyanate (GTC) method. All the solution and materials were RNAase-free if necessary. 1x10^8 Cells were washed twice with PBS, lysed with 4ml GTC (4 M guanidine thiocyanate, 30 mM sodium Acetate, 1% 2-mercapotaethanol, pH 7.0), and then homogenized using 20 gauge syringe needle 20 times. The cell lysate was gently layered on the top of 3 ml CsCl solution (5.7 M CsCl, 30 mM sodium Acetate). Then the sample was centrifuged for 20 hours at 27,000 rpm in SW 41 rotor. The supernatant was removed carefully and the pellet was dissolved in 200 µl H_2O. Total RNA was further purified with phenol:chloroform:isoamyl alcohol (25:24:1 v/v/v) and precipitate with 3M sodium acetate and ethanol. The final total RNA pellet was lyopholised in speed vacuum and stored in -80°C. Poly-A enriched mRNA was isolated from total RNA using Absolute mRNA Purification Kit (Stratagene). The qualities of total RNA and poly-A enriched mRNA were examined by electrophoresis on a formaldehyde 1% agarose gel and northern blot hybridization (Fig. S1).

RNA sample was amplified using TotalPrep RNA amplification kit (Ambion, Foster City, CA), followed by hybridization, labeling and scanning of the chips according to the Ilumina's protocol. The data was extracted, normalized and analyzed using the Illumina BeadStudio software.

Quantitative Analysis

The semi-quantitation of protein abundance was calculated by normalizing the spectral counts of each protein in one fraction relative to the total spectral counts in the corresponding fraction. The normalized profiles were hierarchically clustered based on uncentered correlation with centroid linkage using Cluster 3.0 [26], and visualized using Java TreeView [27].

RESULTS

Saturated Protein Identification

To begin to identify the total proteome of human Jurkat T cells, we separated 33 µg of whole cell lysates by one-dimensional gel electrophoresis and cut the whole gel lane into 18 gel slices. The proteins contained in the gel slices were trypsinized and peptides were extracted as previously described [20]. The peptides were then analyzed by liquid chromatography and tandem mass spectrometry (LC-MS/MS) using LTQ iontrap mass spectrometer. This process was repeated fourteen times using the same amount (33 µg) of proteins from the same whole cell lysate sample. Each replicate analysis identified approximately 2200 to 2700 proteins (Fig. 1A and Table S1). The overlap between each two replicates is about 82% with only 1.8% standard deviation. Differences in protein identification among replicates can be attributed to the complexity of the peptide sample and random sampling during data acquisition by LC-MS/MS. When the data for cumulative total number of unique proteins was analyzed, we found that 3620 proteins were identified from the 14 replicates. We observed that the proteome coverage was enhanced by about 40% when the sample was sequenced fourteen times compared to a single analysis (Fig. 1B). Moreover, the first five analyses reached about 94% of the total proteins identified by 14 replicate analyses. The protein accumulation curve approaches a slope of 0 after nine repeats, suggesting that fourteen replicate analyses are more than enough to achieve saturation.

Global Proteome Survey of Human Jurkat T Leukemic Cells

Although we saturate protein identification in whole cell lysate by replicate analyses, it is obvious that we did not identify all the proteins. Therefore, we next attempted to detect proteins that were refractory to replicate analysis. We aim to increase the proteome coverage by reducing the sample complexity using the well-established and validated fractionation methods [20,28]. First, we sub-fractionated the cell into seven fractions, including cytosolic, light membrane, heavy membrane, mitochondrial, nuclear, lipid raft, and plasma membrane fractions. Western blot assessments confirmed the appropriate partitioning of several organellar markers across these subcellular fractions, providing a basic confirmation of fraction purity (Fig. S2). Second, whole cell lysates from Jurkat cells were used to enrich phosphorylated proteins and validated by western blotting (Fig. S3). Third, we enriched glycosylated proteins by using the lectin wheat germ agglutinin (WGA), which preferentially binds N-acetyl glucosamine (GlcNAC), terminal GlcNAC structures and sialic acid. The enrichment of the glycosylated proteins from the whole cell lysates was validated by western analysis (Fig. S4). Fourth, we tried to detect previously masked subsets of proteins by differential protein depletion. Since proteins have different binding rates to a particular medium, we hypothesized that the abundance level of proteins in a complex mixture might shift during the process of binding, resulting in enrichment of previously masked proteins. To investigate this hypothesis, whole cell lysates of Jurkat cells were incubated with CNBr-activated sepharose beads which covalently bind to free amine groups. Un-coupled proteins were collected at different time points. We found that the protein-abundance pattern shifted after depletion (Fig. S5).

Next we repeatedly analyzed the proteomes from all of the above sub-fractions using one dimensional gel electrophoresis combined with LC-MS/MS, and a combined total of 1707 LC-MS/MS runs were performed. The flow diagram of our experimental strategy is shown in Figure 2A. All the mass spectrometry data were then searched against a non-redundant human protein database using the SEQUEST algorithm [21] followed by stringent filtering, resulting in the identification of 9611 unique proteins. Further selection of proteins identified by at least two high-scoring unique peptides and exclusion of several apparent contaminants introduced during sample handling (*e.g.,* trypsin, keratins), lead to a total identification of 5381 proteins with high-confidence, with ~1000 to ~3600 proteins identified in each fraction (Fig. 2B, Table 1 and S2). The peptide false positive rate in each fraction was lower than 0.7% when the entire dataset was searched against the concatenated forward and reversed database (Table 1). These data indicated that the criteria we used to filter our spectra are stringent and the final protein list contained very low false positive identifications.

One essential issue in shotgun proteomics is that peptide sequences can be present in multiple protein entries due to closely related proteins (e.g. splice variants, homologs, paralogs, orthologs, or redundant entries in protein database), which leads to an overestimation of protein identification number. To address this issue, we computed protein probability and observed that the vast majority of our identified proteins (80%) were assigned unambiguously, that is each unique protein groups have only one representative protein in our identified protein list and were distinguished by at least one high-scoring (peptide probability ≥0.9) unambiguous peptide. The remaining proteins could not be identified unambiguously as the same peptide sequences mapped to more than one protein. In these cases, it is not reliable to claim which proteins are truly expressed in the cells simply based on bioinformatics prediction. Therefore a straight bioinformatics method based on mass spectrometry data is inadequate to completely address the redundancy issues, which indicates that an alternative method is needed to support large-scale shotgun proteomics data.

Proteomic and Transcriptomic Profiles Comparison

To further address the redundancy issue, we next performed a genome-wide mRNA analysis in Jurkat cells as an independent approach to support our proteome profiling. Both total RNA and mRNA were prepared from Jurkat cells, and then examined with commercialized human oligonucleotide microarray which contains >46,000 transcript-specific probe sequences per array. Duplicate and triplicate arrays were analyzed in parallel using total RNA and poly-A enriched mRNA respectively. Genes with detection *p* value less than 0.05 were regarded as positive identification. A total of 15,592 and 15,286 unique gene targets were detected in total RNA and mRNA, respectively. 13,973 gene targets were jointly detected in both total RNA and mRNA (Fig. 3A).

Bridging gene symbols with protein accession numbers resulted in about 7000 gene/protein pairs for the expression comparison study. According to their different detections using mass spectrometry and microarray tools, we categorized these gene/protein pairs into four groups (A, B, C, D) (Table 2 & S3). Group A includes gene products detected by at least two unique peptides and high-confidence mRNA expression. Majority of group A genes (4270 out of 4522 genes) were matched to a single protein accession number, which indicates that there is no redundancy among these protein identifications. A small set of group A genes (252 genes/540 proteins) were matched to multiple protein accession number, which might be due to two reasons. First, there are redundant protein entries in our list. Second, some of the oligonucleotide probes on the chip may not be splice isoform-specific as claims. To address this issue, we investigate the ProteinProphet results and observed that at least 332 proteins were identified in this subgroup based on at least one unambiguous peptide sequences. Therefore a total of 4602 proteins (4270 + 332)were accepted as unambiguous identification in group A. Group B includes 360 proteins (350 genes) identified with at least two peptides, while their transcripts were not detected with high confidence. This group might be due to either false positive identification or their mRNA level was too low, or even degraded after translation. Among them 74 proteins were identified with more than four unique peptides including at least one unambiguous peptide sequence, which were accepted as positive identification. The remaining proteins were considered false positive identification. In group C, the identified proteins have no corresponding gene target on the microarray. In these cases, we accepted 62 proteins which were identified with more than four unique peptides including at least one unambiguous peptide sequence. Group D comprises proteins identified with a single peptide and their mRNA expressions were jointly detected in both purified total RNA and mRNA with high-confidence. In this study, we identified more than 4000 proteins with single peptide which have a relatively high false positive rate. However, transcript expression confirmation effectively reduces the number to 1733 proteins which we accepted in the final protein count and list these proteins separately in table S4. Thus, using a these stringent criteria, we were able to accept a total of 6471 unique proteins (4602 + 74 + 62 + 1733).

It is known that membrane-associated proteins are more difficult to be detected with multiple peptides. Therefore we compared the distribution of integral membrane proteins identified with single peptide (group D) and multiple peptides (combined group A, B, C). As expected, we observed that proteins in group D have a higher coverage of integral membrane proteins than those identified by multiple peptides (Fig. 3B).

In total, excluding the redundant protein entries and potential false positive identification, we identified 6471 unique gene products by mass spectrometry, among which 98% are verified by high-confidence transcript expression (Table S2, S3 & S4). We also investigated the presence of membrane-associated proteins as a measure of proteome detection coverage. Using TMHMM Server 2.0 [29] to predict protein transmembrane helix (TMH), we found that a total of 998 proteins in our accepted proteome dataset had at least one putative TMH, while 492 proteins had two or more TMHs (Table S2). These results are better than a recent study where similar scale proteome was characterized from the mouse organs and organelles [8]. Therefore, we concluded that our dataset provides high coverage of membrane-associated proteins.

Subcellular Localization

One major advantage of proteomic measurements over mRNA profiling is the ability to deduce protein subcellular localization, providing insights into the organelle functions. However, one limitation associated with proteomic study is the difficulty to isolate pure subcelluar organelles. Although we have attempted to purify sub-fractions efficiently and optimized the purification methods [20,28], it may be inaccurate to assign protein localization solely based on the protein identification in the enriched fractions due to the possible cross-contamination during sample preparation. The specificity of protein assignment to particular organelles requires appropriate comparison and data analysis.

Since it has been reported that spectral count of proteins is a semi-quantitative measure of protein abundance [30], the proteomic data reported here might be useful for protein subcellular localization prediction. In this study, we focus on analyzing subcellular localization of proteins identified by at least two unique peptides but not having redundant matches to the same gene target (cytosol, light membrane, heavy membrane, mitochondrion and nucleus fractions, Table S5) based on their spectral count. In addition to the subcellular localization analysis performed in these five fractions, we also provide detailed protein identification list for all of the fractions that were analyzed (Supplementary Table S2). This information can be used to predict the global localization studies of all of the identified proteins.

First, we compared the normalized spectral counts obtained in this study with western blotting results (Figure S2). It was observed that the normalized spectral counts of several biomarker proteins agree with their distribution in different organelles. We therefore assigned a primary localization to each of the proteins based on the normalized spectral count followed by hierarchical clustering (Fig. 4A, Table S5). The clustering result shows significant pattern differences between the five subcellular fractions, which confirms that our samples enriched distinct proteins belonging to different functional categories.

Next we use a “gold standard” protein list to assess the accuracy of our subcellular assignment. This gold-standard list is constructed based on the Gene Ontology (GO) terms and comprises four specific subcellular compartments, i.e., cytosol, non-mitochondrial ribosome, mitochondria, and nuclei. The accuracy of the GO terms assignments were further confirmed by comparison with human protein reference database (HPRD) which only contains information manually extracted from the literature by expert biologists [31]. Only those proteins in agreement with the primary localization annotation in HPRD were accepted as the gold-standard proteins. Our list and the gold-standard list have 616 proteins in common. By plotting the distribution of these gold standard proteins, we observed that cytosolic proteins are mainly enriched in the cytosolic fraction, non-mitochondrial ribosomal proteins in the light membrane fraction, mitochondrial proteins in the mitochondrial and heavy membrane fraction, and nuclear proteins in the nuclear fraction (Fig. 4B). We hierarchically cluster these gold standard proteins as previously to assign a primary localization (Table S6). If our assignment to a protein agrees with the annotations, we consider it as a correct assignment. We quantify the degree of correct assignment in the four clusters that we are able to evaluate (cytosol, nuclei, mitochondria and light membrane) using an enrichment score, which is defined as the ratio of the number of correct assignment to the number of total assignments in this cluster (Table 3). A high enrichment score was observed for the nucleus and mitochondrion (0.93 and 1.00 respectively). The cytosol cluster contains a moderate (0.73) enrichment score. We analyzed the potentially incorrect assignment in the cytosol, most of which (42 out of 44) are nuclear proteins based on the annotation. This may be due to protein shuttling between organelles or incorrect protein annotation in the literature. For example, proliferating cell nuclear antigen (PCNA) is known to be present in the cytosol, but also shuttles to the nucleus during cell proliferation [32]. Therefore, incomplete assignment of many proteins in the literature penalized the actual enrichment score in the cytosol. For the light membrane cluster, although most non-mitochondrial ribosomal proteins are assigned to this cluster, many proteins from other organelles are also included, resulting in a relatively low enrichment score (0.32). This is because light membrane fraction comprises multiple membrane compartments, such as lysosome, smooth/rough endoplasmic reticulum, Golgi apparatus, and even the debris of the nucleus. As for heavy membrane fraction, also called crude mitochondrial fractions in some studies, most proteins (22 out of 25) clustered in this fraction are mitochondrial proteins, which indicates some proteins in this fraction localize in the mitochondrion. However, based on the distribution of several biomarkers, heavy membrane also enriched with other membrane structures, such as endoplasmic reticulum and plasma membrane. Therefore, it is also difficult to specifically assign a primary localization to proteins clustered in heavy membrane. Therefore we concluded that the overall protein assignments in the three clusters (cytosol, nuclei and mitochondria) are highly reliable, and caution should be paid to the light membrane and heavy membrane assignments.

We also verified the specificity of the clustering results using our protein list. We collected the annotated or predicted cytosolic, mitochondrial and nuclear proteins based on the GO terms and PSORT II prediction [33]. We observed that the majority (259/373) of proteins in mitochondrion cluster are assigned to the mitochondrion, 670 out of the 1100 proteins in the cytosol cluster are assigned to the cytosol, 520 out of the 768 proteins in the nucleus cluster are assigned to the nucleus. Since we obtained a higher enrichment score in these three clusters using the gold standard list, there is a high probability that the rest of the unassigned proteins in these three clusters are predicted correctly. Therefore, our proteome data provide a primary subcellular prediction for 2241 proteins with high confidence, including 792 proteins which are unannotated by the GO terms and unpredicted by PSORT II. For the rest of 1129 proteins in the light membrane and heavy membrane clusters, more dedicated subcellular prediction computation and assessment are required.

DISCUSSION

Eukaryotic cells segregate and organize functionally related proteins into discrete compartments that have distinct structures and functions. Previous organelle proteomics studies have mainly focused on one compartment, providing insights into the biology and functions of these structures. Recently two groups performed magnificent proteomics studies on multiple organelles in mouse organ by combining subcellular fractionation and mass spectrometry technologies. However, no comprehensive characterization of a single human cell type has been carried out to date. In this study, we combine replicate proteomics analyses and extensive subcellular fractionation/enrichment methods in Jurkat cells, identifying 5381 proteins, of which 80% were assigned with at least one unambiguous peptide sequences. Based on comparison between proteomic and transcriptomic profiling in Jurkat cells, we were able to specifically exclude redundant entries and potential false positive identifications, resulting in 4738 protein identifications. Among them, more than 98% were confirmed by high-confidence mRNA expression. Since we use multiple stringent criteria to filter and confirm our proteome dataset, the protein false positive rate was estimated to be closed to zero.

This proteome/transcriptomic coverage is much higher than previous proteomic and transcriptomic comparison studies in mammalian cells [8,34]. It may be because previous studies either did not analyze the global expression of proteome or transcriptome comprehensively, or compared proteomic and transcriptomic data generating from different biological systems (e.g., different mouse strains or cell type).

Although we performed a comprehensive proteome analysis of a single cell type, resulting in the identification of a huge number of proteins with high confidence, many lower-abundance and membrane-associated proteins are still refractory to rigorous identification by mass spectrometry techniques. This incomplete proteome coverage likely arose from the intrinsic limitations in instrument sensitivity and bias of data-dependent acquisition towards high-abundance proteins [30]. Additionally it may also be due to an overly stringent filtering of the mass spectrum search results. Consistent with this, over 4000 proteins were identified by one high-scoring peptide. By simply accepting proteins identified by single high-scoring peptide if the corresponding transcript was jointly detected in both total RNA and mRNA samples, the number of protein identifications could be boosted to 6471. Moreover, the accepted proteins identified with a single peptide have a higher coverage of membrane-associated proteins than the proteins identified with multiple peptides. This result indicates that proteomic and transcriptomic integration is a powerful tool to rescue false negative protein identification. Another benefit of proteomic and transcriptomic integration is to investigate how the mRNA levels reflect protein abundances and the biological mechanisms of the discordance between protein and mRNA expression. This ongoing investigation being conducted in our laboratory will be reported in the near future.

In this study, we use CNBr-activated sepharose beads to covalently couple proteins and identified un-coupled proteins during time course. Using this approach, we were able to detect several hundreds of new unique proteins after previous extensive profiling. We found that this method helps purify specific class of proteins including some hydrophobic proteins after long-time incubation. But the reaction between proteins and the beads are not simply based on protein hydrophobicity (data not shown). The reaction rate between proteins and the beads is likely due to the combinational outcome of multiple protein properties, such as hydrophobicity and the number of accessible free amine groups.

One unique advantage of proteomics profiling over transcriptomics profiling is the ability to provide information on protein post-translational modifications. In this study, we enriched proteins based on two types of post-translational modifications, phosphorylation and glycosylation. Therefore, proteins identified in these two fractions are likely phosphorylated or glycosylated. However, note that we did not specifically detect the modification sites and the proteins were enriched based on affinity purification, some unspecific binding proteins and proteins associated with those phosphoproteins/glycoproteins may also be detected in these two fractions. Therefore large-scale phosphopeptide and glycopeptide identification are needed to complement our dataset.

Another advantage of proteomics profiling is to deduce protein subcellular localization, providing insights into the biological functions of gene products and organelles. Given the difficulty of isolating completely pure organelles, we opted to combined differential protein expression in multiple subcellular fractions with hierarchical clustering to more accurately predict the primary subcellular localization of proteins. Using annotated proteins to assess the prediction accuracy, we were able to provide high-confidence assignment by this method in at least three compartments (the cytosol, nuclei and mitochondria). The primary subcellular localization assignment of 2241 proteins reported here, including 792 previously unassigned proteins by the GO term or PSORT II predictions, add more information into the proteome composition of these organelles in this widely used human cell type. As for proteins clustered in other two compartments (light membrane and heavy membrane), we chose to be more cautious and did not assign a specific primary localization to each of them, because these two compartments comprise multiple subcelluar structures. More defined subcellular fractionation approaches are required to further separate these fractions. However, the information here still provides a clue for protein localization. Together with the proteins detected in lipid raft and plasma membrane fractions, one can deduced much valuable information on those unassigned proteins in this study.

One of the main challenges confronting protein subcellular localization prediction is that many proteins likely shuttle between compartments, having multiple subcellular localizations. In this study we only assign a primary localization to each protein, while most proteins reported here were detected in more than one subcellular fraction. Although some of these cases may stem from crosscontamination during sample preparation, we cannot exclude that it may indeed reflect the real protein localization patterns. Moreover, 1129 proteins clustered in the light and heavy membrane fractions were assigned with a primary localization with low confidence due to the multiple organelle composition of these two fractions. Therefore, it is possible that by applying more advanced machine-learning methods on the proteomics data reported here more accurate subcellular localization assignment can be expected. Despite the caveats in the identification of post-translationally modified protein and the assignment of protein subcellular localization, the proteomics profile reported here provides a global landscape in a single human cell type, precluding the differences among tissues and more suitable for many in-depth characterization of human systems at a cellular level, such as comparison between protein and mRNA expression, integration of protein expression pattern with protein-protein interactions and biological phenotypes. In addition, some of the data in this study were obtained from cells after specific perturbations. More thorough analysis on the dynamic regulation of proteins in these cells were, or will be reported separately [28]. Since many mutant cell lines have been derived from human Jurkat cells and widely used by the biological community [16], the comprehensive proteome survey of this cell type can serve as a useful platform for more extensive experimental characterization and integration studies. Complete summaries of the peptides and proteins identified in this study are accessible in the supplemental data. In addition, all of the raw data generated in this study are being deposited with the MCP data repository. Investigators are encouraged to utilize this rich proteomic resource.

SUPPLEMENTAL DATA

Supplemental data include experimental procedures, five figures, six summary tables and the detail information for each identified peptide (peptide atlas file) and their sharing results among different protein entries (ProteinProphet file).

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health Grants RO1 HL 67569, PO1 HL 70694, and RR 13186.

REFERENCE

1 Pan Q., Shai O., Misquitta C., et al. (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol Cell 16, 929-941.

2 Su A. I., Wiltshire T., Batalov S., et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101, 6062-6067.

3 Okazaki Y., Furuno M., Kasukawa T., et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563-573.

4 Carninci P., Kasukawa T., Katayama S., et al. (2005) The transcriptional landscape of the mammalian genome. Science 309, 1559-1563.

5 Cheng J., Kapranov P., Drenkow J., et al. (2005) Transcriptional maps of

10 human chromosomes at 5-nucleotide resolution. Science 308, 1149-1154.

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

7 Chen G., Gharib T. G., Huang C. C., et al. (2002) Discordant protein and mRNA expression in lung adenocarcinomas. Mol Cell Proteomics 1, 304-313.

8 Kislinger T., Cox B., Kannan A., et al. (2006) Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. Cell 125, 173-186.

9 Lander E. S., Linton L. M., Birren B., et al. (2001) Initial sequencing and analysis of the human genome. Nature 409, 860-921.

10 Venter J. C., Adams M. D., Myers E. W., et al. (2001) The sequence of the human genome. Science 291, 1304-1351.

11 Kunst F., Ogasawara N., Moszer I., et al. (1997) The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature 390, 249-256.

12 Adams M. D., Celniker S. E., Holt R. A., et al. (2000) The genome sequence of Drosophila melanogaster. Science 287, 2185-2195.

13 Peng J., Schwartz D., Elias J. E., et al. (2003) A proteomics approach to understanding protein ubiquitination. Nat Biotechnol 21, 921-926.

14 Ptacek J., Devgan G., Michaud G., et al. (2005) Global analysis of protein phosphorylation in yeast. Nature 438, 679-684.

15 Wu L., Han D. K. (2006) Overcoming the dynamic range problem in mass

spectrometry-based shotgun proteomics. Expert Rev Proteomics 3, 611-619.

16 Abraham R. T., Weiss A. (2004) Jurkat T cells and development of the T-cell receptor signalling paradigm. Nat Rev Immunol 4, 301-308.

17 Arur S., Uche U. E., Rezaul K., et al. (2003) Annexin I is an endogenous ligand that mediates apoptotic cell engulfment. Dev Cell 4, 587-598.

18 Wang P., Song J. H., Song D. K., Zhang J., Hao C. (2006) Role of death receptor and mitochondrial pathways in conventional chemotherapy drug induction of apoptosis. Cell Signal 18, 1528-1535.

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

20 Rezaul K., Wu L., Mayya V., Hwang S. I., Han D. (2005) A systematic characterization of mitochondrial proteome from human T leukemia cells. Mol Cell Proteomics 4, 169-181.

21 Eng J., McCormack A. L., Yates J. R., 3rd. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5, 976-989.

22 Han D. K., Eng J., Zhou H., Aebersold R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. Nat Biotechnol 19, 946-951.

23 Peng J., Elias J. E., Thoreen C. C., Licklider L. J., Gygi S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast

proteome. J Proteome Res 2, 43-50.

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

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

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

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

28 Hwang S. I., Lundgren D. H., Mayya V., et al. (2006) Systematic characterization of nuclear proteome during apoptosis: a quantitative proteomic study by differential extraction and stable isotope labeling. Mol Cell Proteomics 5, 1131-1145.

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

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

31 Peri S., Navarro J. D., Amanchy R., et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13, 2363-2371.

32 Vriz S., Lemaitre J. M., Leibovici M., Thierry N., Mechali M. (1992) Comparative analysis of the intracellular localization of c-Myc, c-Fos, and replicative proteins during cell cycle progression. Mol Cell Biol 12, 3548-3555.

33 Nakai K., Horton P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24, 34-36.

34 Hu S., Li Y., Wang J., et al. (2006) Human saliva proteome and transcriptome. J Dent Res 85, 1129-1133.

FIGURE LEGEND

Figure 1: Saturated Protein Identification by Replicate Analyses

(A) The number of unique proteins identified in each replicate analysis of the human Jurkat whole cell lysates is shown.

(B) The cumulative curve of unique proteins identified by replicate analyses of the whole cell lysates from human Jurkat T Leukemic cells is shown. The percentages of proteome coverage after two, three and five repeats are indicated.

Figure 2: Proteome Survey of Human Jurkat T Leukemic Cells

(A) The flow chart of overall experimental strategy for protein identification. Human Jurkat T Leukemic cells were fractionated into ten fractions including cytosolic, light membrane, heavy membrane, mitochondrial, nuclear, lipid raft, plasma membrane, phosphoprotein, glycoprotein, and depletion. Proteins extracted from the Jurkat whole cell lysates and the above subfractions were repeated analyzed by one dimensional gel electropheresis combined with LC-MS/MS (GeLC-MS/MS). A total of 1707 LC-MS/MS runs were performed resulting in the identification of 5381 proteins.

(B)The cumulative curve of total identified proteins following multiple enrichment methods is shown. WCL, whole cell lysates; HM, heavy membrane; LM, light membrane; Mito, mitochondria; Phospho, phosphoproteins; Glyco, glycoproteins; PM, plasma membrane; Depletion, un-coupled proteins by CNBr-activated sepharose beads.

Figure 3: Proteomic and Transcriptomic Profile Integration

(A) High-confidence transcript detection. Using $p < 0.05$ as a cutoff, 15592 and 15286 unique gene targets were detected in purified total RNA and mRNA samples, respectively. 13973 gene targets were jointly detected in both total RNA and mRNA.

(B) Comparison between the distribution of membrane-associated proteins identified by single and multiple peptides. The accepted proteins identified by multiple peptides and a single peptide were applied for transmembrane helix (TMH) prediction by TMHMM server 2.0. The membrane protein distributions in these two categories are shown.

Figure 4: Protein Subcellular Localization

(A) Hierarchical clustering of protein expression pattern obtained from cytosolic, light membrane, heavy membrane, mitochondrial, and nuclear fractions are shown. The protein expression level was measured by the normalized spectral count, i.e. the spectral count of each protein in one fraction divided by the total spectral counts of the same fraction. The five protein clusters are indicated. This pattern allowed the assignment of primary localization of each of the identified proteins, except in some proteins with multicompartment distributions. Note compartment specific protein distribution pattern as well as the multi-compartmental distribution patterns.

(B) The expression pattern of the “gold standard” proteins in cytosolic, light membrane, heavy membrane, mitochondrial and nuclear fractions are shown. The “gold standard” proteins were selected from the identified proteins in different subcellular compartments using the GO terms.

Figure 1A

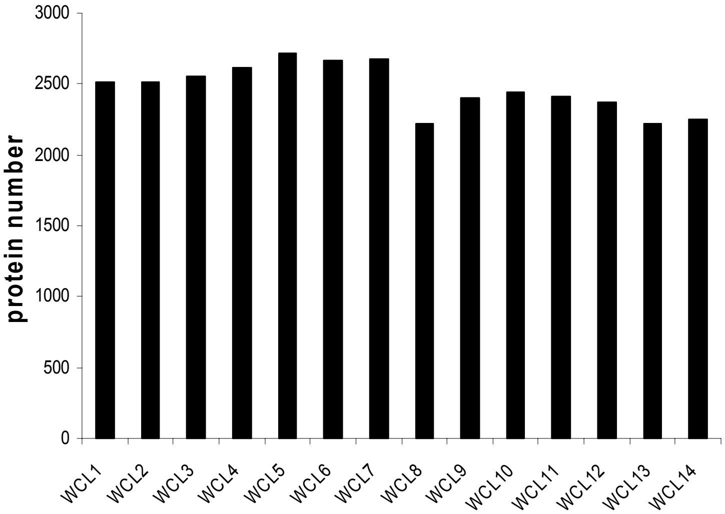

Figure 1B

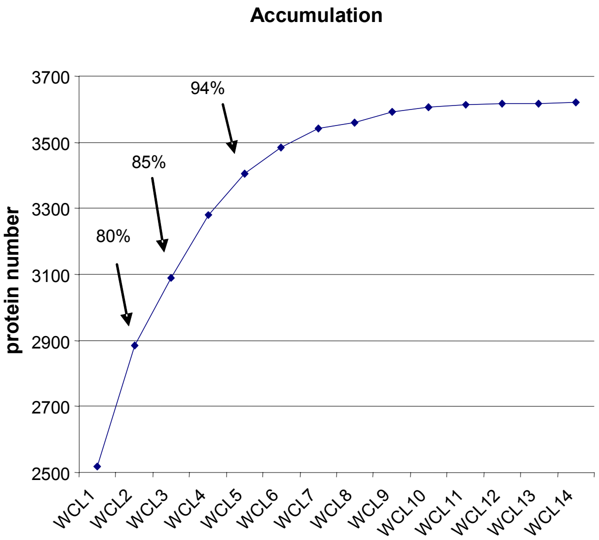

Figure 2A

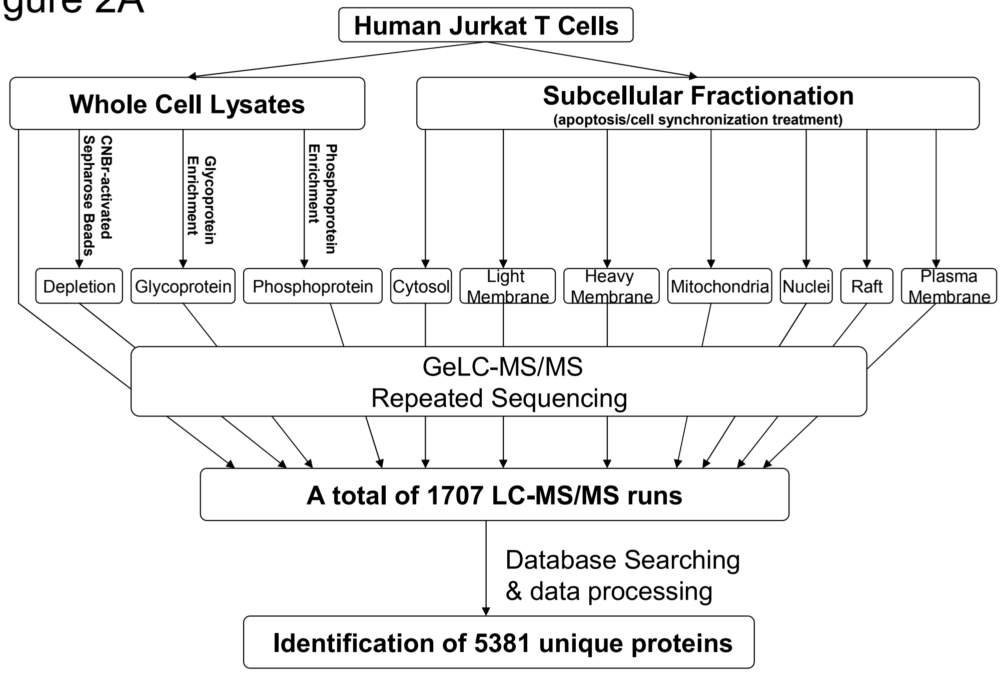

Figure 2B

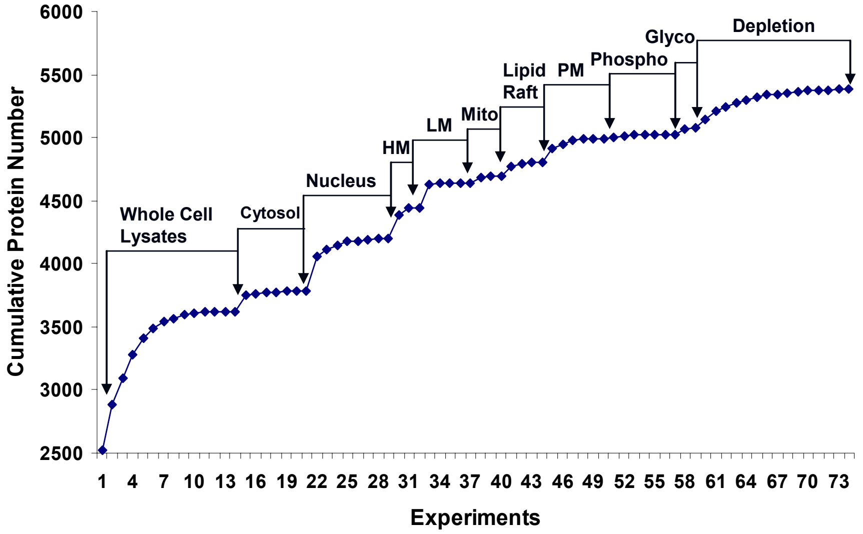

Figure 3A

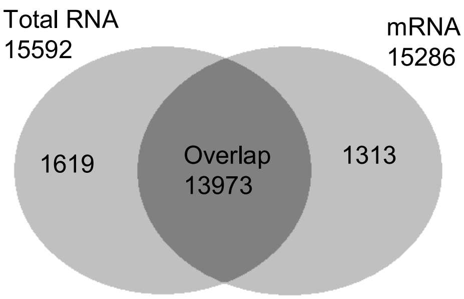

Figure 3B

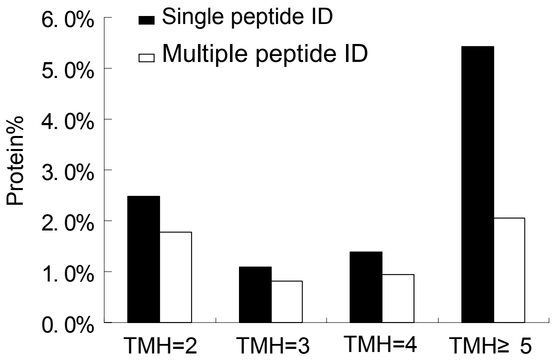

Figure 4A

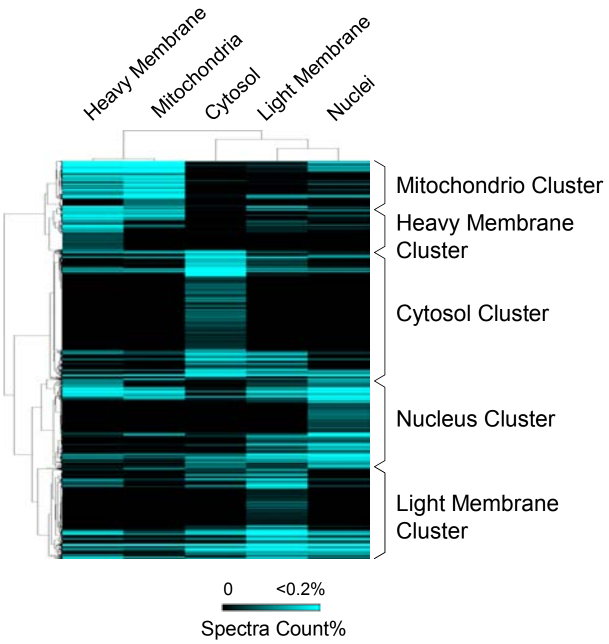

Figure 4B

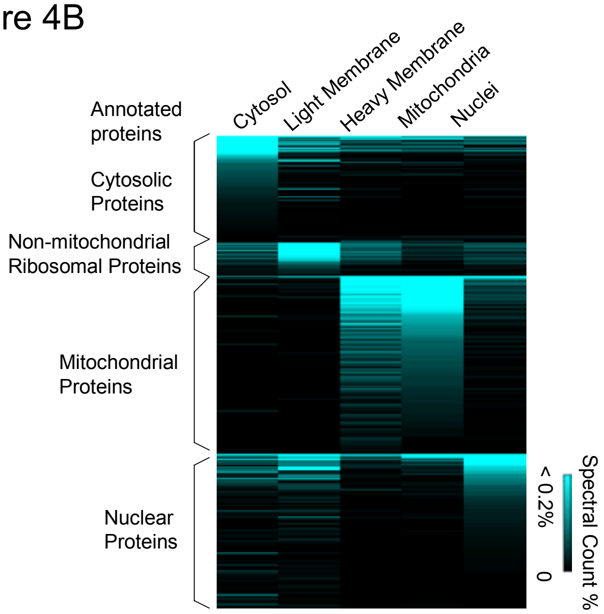

Table 1. Summary of the Proteomics Data. The Number of high-confidence protein identification (more than one high-scoring peptide) and their associated unique peptide counts, spectral counts and peptide false positive rate in each experimental fraction were shown.

Experimental fraction	Proteins	Unique Peptide	Spectra	Peptide False Positive%
Whole cell lysates	3620	26865	256375	0.43%
Cytosol	2012	13618	72893	0.64%
Heavy Membrane	1599	9073	34671	0.39%
Light Membrane	2154	13617	58001	0.42%
Mitochondrion	1154	5971	21037	0.46%
Nucleus	1750	10338	44673	0.37%
Raft	1112	5462	20645	0.28%
Plasma Membrane	1529	8239	49381	0.28%
Glycoprotein	936	6175	23668	0.08%
Phosphoprotein	1033	4964	21693	0.50%
Depletion	3035	20156	195167	0.51%
Total	5381	43693	798204	

Table 2. Summary of proteomic and transcriptomic profiles comparison

Protein ID	microarray target	Unique Match no.	Redundant Match no.	Total RNA	mRNA	ID no.	Group
5381 protein identified by more than one peptide	with target	4610	N/A	Yes	Yes	4180	A
				Yes	No	32	
				No	Yes	58	
				No	No	340	B
		N/A	560(262)	Yes	Yes	536 (250)	A
				Yes	No	4(2)	
				No	Yes	0	
				No	No	20 (10)	B
	No target	N/A	N/A	N/A	N/A	211	C
Protein identified by a single peptide	With target	1733	0	Yes	Yes	1733	D

Yes, the corresponding transcripts were detected with $p < 0.05$; No, the corresponding transcripts were not detected with $p < 0.05$.

The numbers outside parentheses represent the matched protein number; the numbers inside parentheses represent the corresponding target gene number.

Table 3: Validation of subcellular localization prediction of proteins based on the normalized spectral count followed by hierarchical clustering.

Protein Group	Cluster	Total Protein no.	Known Correct Assignment no.	Enrichment Score
616 Gold Standard Protein	Cytosol	165	121	0.73
	Nuclei	88	82	0.93
	Mitochondria	208	207	1.00
	Light Membrane	130	42	0.32
	Heavy Membrane	25	N/A	N/A
3370 Identified Proteins	Cytosol	1100	670	0.61
	Nuclei	768	520	0.68
	Mitochondria	373	259	0.69
	Light Membrane	745	N/A	N/A
	Heavy Membrane	384	N/A	N/A