Published in final edited form as:
J Proteome Res. 2012 March 2; 11(3): 1521–1536. doi:10.1021/pr200460s.

Annotator: Post-processing Software for generating function-based signatures from quantitative mass spectrometry

Juliesta E. Sylvester[*,1,4], **Tyler S. Bray**[*,2], and **Stephen J. Kron**[3,4]

[1]Department of Biochemistry and Molecular Biology, The University of Chicago, Chicago, Illinois 60637

[2]Department of Computer Science, The University of Chicago, Chicago, Illinois 60637

[3]Department of Molecular Genetics and Cell Biology, The University of Chicago, Chicago, Illinois 60637

[4]Ludwig Center for Metastasis Research, The University of Chicago, Chicago, Illinois 60637

Abstract

Mass spectrometry is used to investigate global changes in protein abundance in cell lysates. Increasingly powerful methods of data collection have emerged over the past decade, but this has left researchers with the task of sifting through mountains of data for biologically significant results. Often, the end result is a list of proteins with no obvious quantitative relationships to define the larger context of changes in cell behavior. Researchers are often forced to perform a manual analysis from this list or to fall back on a range of disparate tools, which can hinder the communication of results and their reproducibility. To address these methodological problems we developed Annotator, an application that filters validated mass spectrometry data and applies a battery of standardized heuristic and statistical tests to determine significance. To address systems-level interpretations we incorporated UniProt and Gene Ontology keywords as statistical units of analysis, yielding quantitative information about changes in abundance for an entire functional category. This provides a consistent and quantitative method for formulating conclusions about cellular behavior, independent of network models or standard enrichment analyses. Annotator allows for "bottom-up" annotations that are based on experimental data and not inferred by comparison to external or hypothetical models. Annotator was developed as an independent post-processing platform that runs on all common operating systems, thereby providing a useful tool for establishing the inherently dynamic nature of functional annotations, which depend on results from on-going proteomic experiments. Annotator is available for download at http://people.cs.uchicago.edu/~tyler/annotator/annotator_desktop_0.1.tar.gz.

Keywords

Quantitative Analysis; Software; Significance; Biological Context; Stable Isotope Labeling; Phosphoproteome

Correspondence should be addressed to Stephen J. Kron. Telephone: (773) 834-0250. Fax: (773) 702-4394. skron@uchicago.edu..
*Authors contributed equally.

Supporting Information Available: This material is available free of charge via the Internet at http://pubs.acs.org.

INTRODUCTION

Over the past decade, there has been a significant effort to describe differences among cellular states by changes in the proteome. Toward this end, technological advances, especially in the field of mass spectrometry, have allowed increasingly efficient, parallel and quantitative analyses of protein abundance and modifications. These techniques generate large amounts of data that require, as a practical necessity, specialized software for organization, management, and analysis (reviewed by [1]). Although existing tools have met many of the challenges associated with processing large LC-MS/MS data sets,[2-11] many of the tasks involved with assigning biological significance are still performed using an ad hoc assortment of tools, often requiring manual validation, that leave conclusions subject to the judgment of the individual researcher (reviewed by [12]). Accordingly, the major bottleneck currently facing proteomic analyses is not the rate at which data is generated but the time that it takes to interpret data in a biological context.[13]

Current approaches for deriving biological meaning from proteomic data depend either on the use of network-based models or enrichment analyses. Network-based approaches use libraries of protein interaction profiles, provided either by high-throughput monitoring strategies[14] or compiled from pairwise interactions referenced in the literature.[15] Software is available to overlay experimentally identified proteins onto these curated interaction networks. For example, provided with a short list of proteins, MetaCore from GeneGo, Inc. will display a hypothetical network built from previously observed interactions.[16] The longer the list of proteins, however, the harder it can be to match the data to a compiled network. In fact, these interaction networks have not been shown to be universally applicable and each interaction may not be physiologically significant.[17] Finally, the extensive degree of crosstalk and functional interdependence between signaling pathways can make it very difficult to extract signatures that are easily associated with observable cell functions.[18]

Nevertheless, interaction networks provide a framework for inferring protein dependencies and have been used successfully to profile essential gene expression[19] and to describe regulatory architectures (reviewed by [20]). The challenge arises in generalizing curated interaction profiles to describe consequences resulting from the differences in protein abundance that are observed by LC-MS/MS. Ubiquitous signaling mechanisms such as activation and inhibition, with feed-forward and feedback loops, do not depend on protein abundance for the modulation of network activity. Therefore, systems approaches that use circuitry to represent activity are not useful for determining how changes in protein abundance represent changes in cellular function.

The quantitative alternative to visualization with interaction networks is enrichment analysis, which determines the extent to which a sample has been enriched for particular functions. Enrichment algorithms, which are used for example, by DAVID (Database for Annotation, Visualization, and Integrated Discovery)[21] and EASE (Expression Analysis Systematic Explorer),[22] provide a probability that a given sampling of proteins would be chosen at random from a complete proteome. Similarly, the LC-MS/MS-specific software Scaffold[23] and STRAP (Software Tool for Researching Annotations of Proteins)[24] use Gene Ontology information from NCBI and UniProt to generate pie charts visualizing the relative representation of functions derived from a list of proteins that were identified in an experiment. Software for network manipulation, such as Cytoscape[25, 26], can also provide analyses that describe the likelihood that observed interactions have been randomly selected.[27]

Enrichment analyses do not take into account experimentally observed measures of protein abundance. Both network visualization and enrichment analyses use lists of identified proteins to generate biologically relevant hypotheses. The reverse approach, where biological hypotheses generate lists of expected proteins, has been suggested as a targeted approach to proteome monitoring.[28] While this would provide for more efficient data acquisition, it would not change the essential process by which quantitative proteomic data is translated into relevant changes in cell behavior.

Our original motivation was to develop a tool that provided for faster quantitative analysis in a biological context. This final post-processing stage of proteomic analysis required a new approach that relied less on investigator subjectivity, thereby relieving an automation bottleneck and supporting experimental reproducibility and the communication of comprehensive results. It was clear that a common analytical technique for determining biological significance would facilitate direct comparisons within and between experiments. To achieve this standardized method of protein selection and address the issue of small sample sizes inherent in LC-MS/MS data, we investigated the use of common heuristics and robust, simple statistical measures of significance. Novel examples of this are the use of population-based standard deviations instead of arbitrary fold-change thresholds as measures of significance, the use of t-tests to analyze changes in keyword abundance and the inclusion of explicit normality tests to ascertain the effectiveness of t statistics. Cluster analysis and heat map visualization were used to demonstrate significant similarities within and between data sets, an approach that is rarely applied to quantitative proteomic data. We addressed several details particular to quantitative mass spectrometry, including the quality of the quantitative LC-MS data, the presence of stable isotope-labeled standards, and the effect of sample preparations that include separation by gel electrophoresis. The inclusion of optional filters with user specified parameters provide a high degree of control over the exclusion of possible errors introduced by upstream software.

Most importantly, in this new approach biological context became an integral component of the complete quantitative analysis. UniProt and Gene Ontology keywords provide a consistent language for discussing biological trends that is established, accepted and readily accessible. We used these keywords to organize observed data at the peptide level, an approach that differs from existing keyword enrichment analysis algorithms. This provided us with larger sample sizes, giving more power to statistical analyses, and also provided an automated means to reveal quantitative signatures that could not be extracted from lists of protein names. The simplicity of this quantitative approach avoids some of the problems that emerge from attempting to overlay a hypothetical interaction network onto experimental data. By investigating observed keyword overlap we were able to highlight shared protein functions in a biological context.

As an example of our strategy, we monitored changes in the relative abundance of proteins and keywords in neutrophils activated by lipopolysaccharide (LPS) to induce an inflammatory response. There are several experimental methods to assist with the global study of cell response. Selection-based assays, for example using activity-based probes[29] or clonal selection,[30] may be the most successful means of associating proteins with functions and phenotypes. In this work, we used a gallium affinity column to enrich cell lysates for phosphorylated proteins, thereby focusing our analysis on proteins with the greatest likelihood of being involved in a coordinated signaling response.

METHODS

Sample Preparation

To provide a source of data that was rich in content for functional annotation, we prepared a complex sample of soluble proteins from neutrophils activated by lipopolysaccharide (LPS). The human promyelocytic HL-60 cell line (ATCC) was differentiated in culture using 1 μM alltrans retinoic acid (Sigma-Aldrich), 6 pM 1α,25-Dihydroxyvitamin D3 (Sigma-Aldrich), and 30 ng/mL granulocyte-colony stimulatory factor (Invitrogen) in Iscove's Modified Dulbecco's medium (Mediatech, Inc.) supplemented with 20% FBS and 4 mM L-glutamine.[31] Cells were activated via treatment with 100 ng/mL of lipopolysaccharide (LPS) from E. coli O111:B4 (List Biological Laboratories) for 30 minutes. The control was treated with an equal volume of double-distilled and autoclaved water. Cells were harvested, washed with 100 mM HEPES, pH 7.4, and lysed in the presence of phosphatase inhibitors. Lysates were prepared and enriched for phosphorylated proteins using the Pro-Q Diamond phospho-enrichment kit (Invitrogen).[32] Fractions were collected, concentrated, and exchanged into 0.25% CHAPS in 25 mM Tris, pH 7.5, by centrifugation at 4 °C using 10 kDa-cutoff concentrators (Millipore) for a final volume near 500 μL.

The total protein content of eluted fractions was determined by Bradford analysis (Pierce) using the average of triplicates. Total protein content was also qualitatively compared by the intensity of Coomassie staining (Pierce) following gel electrophoresis. LPS-treated and control samples were loaded at approximately 10 μg of total protein per lane for separation on 4-12% NuPAGE gradient electrophoresis gels (Invitrogen) using MOPS SDS running buffer. Gels were cut into 11 vertical slices, combining 9 replicate lanes for each vertical slice to increase the amount of protein in each sample. Gel slices were de-stained, reduced, acetylated, and dehydrated.[32] Proteins were digested in-gel by re-hydrating each gel slice with 2 μg of trypsin in 60 mM NH_4HCO_3 with 0.5 mM CaCl for 12 hours at 37 °C. Peptides were extracted from gel slices in two steps, starting with an aqueous extraction with 5% formic acid in water for 1 hour and followed with an organic extraction with 5% formic acid in 50% CH_3CN. Extractions from each step were centrifuged under vacuum separately, combined in water, and lyophilized.

^{18}O Labeling

Isotopic labeling by enzymatic incorporation of ^{18}O was used for relative protein quantitation between the LPS-treated sample and the control.[33] To label peptides at the carboxyl-terminus with ^{18}O, samples were re-suspended in 97% H $^{18}_2O$ (Cambridge Isotope Laboratories, Inc.) and incubated with 30 μL of washed Mag-Trypsin beads (Clontech) for 48 hours at 37 °C. The reaction was monitored by MALDI-TOF MS (4700 Voyager, Applied Biosystems). Beads were removed by magnetic separation, labeled samples were lyophilized and re-suspended in 2% CH_3CN with 0.2 % formic acid in water (mobile phase A), and combined 1:1 (v/v) with the unlabeled sample.

Nanoscale LC-MS/MS

11 LC-MS/MS runs were performed per experiment, corresponding to the number of vertical gel slices. Using an AS1 autosampler and auxiliary isocratic pump (Eksigent Technologies), 10 μL injections were loaded at 10 μL/minute onto a 2.5-μL Opti-Pak pre-column (Optimize Technologies) packed with 5 μm, 200 Å Michrom Magic C_8 solid phase (Michrom BioResources) to remove contaminating salts. Peptides were separated at 350 nL/minute on a 20-cm × 75-μm-inner diameter column packed with 5 μm, 200 Å Michrom Magic C_{18} solid phase (Michrom BioResources). A 90 minute two-step chromatographic gradient was used, starting with a slow separation from 5 - 50 %B over 60 minutes followed

by a rapid increase from 50 - 95 %B over 10 minutes using 80% CH_3CN, 10% n-propyl alcohol, and 0.2 % formic acid in water (mobile phase B).

Samples were analyzed on an LTQ-Orbitrap Hybrid FT mass spectrometer (Thermo Scientific). Data were collected in full profile mode from m/z 375 to 1,950 at 60,000 resolving power with internal calibrant lock masses. The five most abundant double- and triple-charged precursors with a minimum signal of 8,000 between 375 - 1,600 m/z were subjected to collision-induced dissociation (CID) with 35% normalized collision energy, 30 ms activation time, and activation Q at 0.25. To reduce repeat analyses, dynamic exclusions were established for 60 seconds with an isolation width of 1.6 m/z units, for low and high mass exclusion of 0.8 m/z units each per precursor.

Database Searching

Unprocessed MS and MS/MS data in a RAW file format were converted to the mzXML format using ReAdW (from TPP version 4.1, http://sourceforge.net/projects/sashimi/files/) and imported into the CPAS organization and analysis application database (version 9.10).[34] X! Tandem (version 2.007.01.01.1, http://www.thegpm.org/)[4] identified peptides and proteins from fragment ion spectra of selected precursors using the non-redundant human international protein index (version 3.53) maintained at the European Bioinformatics Institute. Parent ions were allowed an error of up to 2.5 Da above and 1 Da below the monoisotopic mass, while fragment ions were allowed a mass tolerance of 0.5 Da from the monoisotopic mass. Parent ions required less than 20 ppm mass accuracy and greater than 90% matched molecular weight against expected values based on the PeptideProphet algorithm (http://peptideprophet.sourceforge.net/).[35] Peptide identifications were statistically validated using PeptideProphet and filtered in CPAS using a PeptideProphet minimum probability cutoff that generates a false discovery rate of 1%.

Search parameters specified tryptic digestion, with cleavage at arginine or lysine, and allowed two missed cleavages per peptide. Cysteine alkylation from iodoacetamide treatment was set as a fixed modification. S-carbamoylmethylcysteine cyclization at the amino-terminus, pyroglutamic acid formation from glutamine and glutamate, oxidation of methionine, and single and double isotope label incorporation at lysine and arginine were considered variable modifications. Peptides were considered distinct if they differed in sequence or modifications. Although distinct proteins within a family may share identical peptides, ambiguous assignments were grouped by a single protein identifier based on a representative group member following the law of parsimony.

Quantitation by Ion Current Integration

The XPRESS software (version 2.1, from TPP version 3.4) was used within CPAS to reconstruct peptide elution profiles.[36] Peptide signal intensity was integrated over the number of MS scans in which an identified peptide ion was observed, thereby providing quantitative areas for unlabeled and ^{18}O labeled peptides. The mass tolerance for the selection of peptide pairs was set to 0.05 Da. XPRESS was not used to calculate protein abundance ratios from these peptide elution peak areas.

Software Setup

The software described here interacted with a MySQL database that was populated with reference data used to filter and organize results. Proteins were defined by the 40,788 available gene names from the SwissProt catalog of the Universal Protein Resource (UniProt). Gene names were associated with IPI identifiers from X! Tandem searches. Keywords were defined by the total set of 32,378 terms in 13 categories from UniProt and Gene Ontology catalogs (available at http://www.uniprot.org/docs/keywlist and

http://www.geneontology.org/GO.downloads.ontology.shtml). The complete human repository of proteins from the UniProt knowledgebase, including protein-specific accession numbers, molecular weight information, and keyword associations (available at http://www.uniprot.org/downloads), was loaded into the MySQL database.

Software Implementation

The software was written in Java to facilitate platform independence. Reference and experimental data were stored in a MySQL database. The Apache POI library was used to read and write Excel files. The Apache Commons Math library was used as a standard resource to compare implemented statistical calculations and to calculate p-values from t-statistics. The standard analytical software R was also used to compare and validate implemented statistical calculations, to perform cluster analysis using Pearson correlation coefficients[37] and to generate heat maps. Prism (version 4.0a) was used to calculate frequency distributions and to produce histograms. The software was run on standard desktop computers running Linux or Mac OS X but can also be run with Microsoft Windows.

RESULTS AND DISCUSSION

Activation of Neutrophils with Lipopolysacccharide

Figure 1 illustrates the series of steps taken to generate data for LC-MS/MS analysis. To control the standardization of experimental variables, HL60 cells were differentiated along the neutrophil lineage in culture and split into two groups prior to treatment with LPS. The control group and LPS-treated group were lysed and enriched for phospho-protein complexes on separate affinity columns. The eluate from each column was loaded with equal total protein content and separated by gel electrophoresis. Equal protein loading of each gel was important to ensure accurate relative ratios between samples for quantitative analysis. 11 slices were excised from each gel lane, and each slice was digested and labeled with ^{18}O at the peptide COOH-terminus using trypsin.

Two technical replicates were performed. In one case, the LPS-treated sample was labeled with ^{18}O, while the control remained unlabeled. In the other case, the labeled state of the treated cells was switched so that the control was labeled with ^{18}O while the LPS-treated sample remained unlabeled. This strategy was intended to provide validation for peptide quantitation, independent from any bias in labeling efficiency with different peptide sequences.[38] For each gel slice, the differentially labeled samples were combined in equal volumes and analyzed by LCMS/MS.

Initial LC-MS/MS Data Processing

The CPAS platform[34] was used to manage the variety of tasks involved in processing LCMS/MS data to generate validated peptide identities and measures of relative abundance. In CPAS, X! Tandem[4] was used to identify peptide sequences and assign peptides to proteins by gene name. Also in CPAS, XPRESS[36] was used to calculate peptide elution peak areas by integrating the intensity of each peptide parent ion over chromatographic time.

Annotator Input

Data were exported from CPAS as Microsoft Excel files and loaded into a MySQL database. Each excel file combined the 11 LC-MS/MS runs from a single experiment and contained all of the information available from CPAS analyses, including columns for peptide sequence, gene name, LC-MS/MS run/fraction name, Peptide Prophet score, protein accession number, scan number, retention time, and quantitative analysis fields. Additional descriptions of these fields are available in the CPAS documentation

(https://www.labkey.org/wiki/home/Documentation/page.view?name=pickPeptideColumns). An example of program input, generated directly by CPAS, is provided in **Supplementary Table 1**.

In addition to this experimental input, the human protein repository was downloaded from UniProt and Gene Ontology catalogs as a reference. Every protein that was identified in an experiment was queried against this reference.

The program input was filtered to ensure quality, then grouped by gene name. For each group, descriptive and normality statistics were calculated and one-sample t-tests were performed. The data was regrouped by keyword and a separate set of descriptive statistics, normality statistics and t-tests were calculated for all keyword groupings. Finally, the program compared all pairs of keywords to determine the degree of overlap, thereby highlighting shared functions. The details of these steps and their theoretical underpinnings are discussed below.

Data Filtering Strategy

Each file contained over 15,000 peptide identifications per experiment. Manual searches of the data highlighted several key aspects that were misinterpreted, thereby introducing significant errors in the final evaluation. For example, in the Excel file generated by CPAS using X! Tandem and XPRESS, identical peptide sequences appeared multiple times (**Supplementary Table 1**, rows 5 – 6, 7 – 8, 15 – 16). Whether XPRESS used an identical set of MS scans or a slightly different range of MS scans to quantify the same peptide multiple times, the effect was to bias the calculated average toward the abundance of that particular peptide. For example, suppose that the relative abundance of protein X was calculated from peptides a, b and c with relative abundances of 2, 2 and 5, respectively. It is our opinion that the abundance of protein X should be 3 because (2+2+5)/3 = 3. If peptide c is counted twice, because it was sampled twice, then the abundance of protein X is 3.5 because (2+2+5+5)/4 = 3.5. This problem is exaggerated if peptide c has a value of 10 instead of 5, or if peptide c is counted more than twice. We realized that to calculate an accurate average for protein abundance, we would have to filter the data so that each unique peptide sequence was considered only once.

The protein myosin heavy chain 9 (MYH9) provides a prime example for illustrating the utility of considering the contribution of repetitive measurements from a single peptide sequence during protein quantitation. A highly abundant intracellular protein, MYH9 was initially represented in the dataset by 778 unique measurements. Each measurement represented a unique instance where a peptide parent ion was fragmented for identification by MSMS and the ratio of the labeled parent ions in MS was calculated from integrated MS peak intensities over the time of chromatographic elution. Many of these measurements were redundant; for instance, the peptide sequence TDLLLEPYNK was quantified 8 times over its total elution period. These 8 unique measurements for TDLLLEPYNK were distinguished by independent fragmentation spectra with PeptideProphet scores ranging from 0.9150 to 0.9968 (Supplementary Table 1, rows 568, 1861, 1865, 1894, 1898, 10472, 10473 and 13176). In an effort to reduce the number of times that any one peptide was sampled by MSMS, dynamic exclusions were set during data acquisition for 60 seconds with an isolation width of 1.6 m/z units for each parent peptide ion. Nevertheless, for high-abundance peptides that elute over a duration longer than 1 min. this dynamic exclusion setting is not sufficient and the peptide is sampled repeatedly by the spectrometer. Setting the dynamic exclusion window to a longer duration risks missing separate and unique peptides that may be identified within that mass range. Therefore, to differentiate among these eight independent measurements of the same peptide we established a set of postprocessing filters that could be used to describe the quality of each measurement; these

filters will be discussed in detail in sections to follow. For the peptide TDLLLEPYNK, 4 of the 8 independent measurements were considered valid because the peptides had been identified in fractions from gel slices that corresponded to the high molecular weight of MYH9. These 4 independent measurements reported a relative abundance for TDLLLEPYNK that ranged from (0.66 to 0.71), where 0.5 represents equal relative abundance of the peptide in the unlabeled and heavy-labeled samples. The measurement for TDLLLEPYNK with the highest PeptideProphet score (0.98) generated a ratio of 0.71, which was very close to the average ratio for all 4 valid measurements (0.68). TDLLLEPYNK is just one peptide from the MYH9 protein and each peptide that was observed from the total MYH9 protein sequence was analyzed in this way to calculate the relative protein abundance for MYH9.

The main point to note is that two strategies are possible in calculating a protein average: all 4 independent measurements for the single peptide TDLLLEPYNK can be used in a weighted average for the protein MYH9 or it is possible to choose only 1 of the measurements for TDLLLEPYNK in the calculation of a simple protein average. We chose the latter strategy, using the PeptideProphet score to select one relative abundance ratio for the peptide TDLLLEPYNK. Our concern was that a weighted average, representing multiple independent samplings of TDLLLEPYNK, could skew the calculated protein abundance toward this one peptide. This peptide sequence appears only once in the protein MYH9, so it seemed sensible to let each unique peptide sequence within the protein make an equal contribution toward the calculated average. Indeed, in the case of MYH9, although 778 unique measurements were recorded for MYH9 we found that only 140 unique peptide sequences from MYH9 were observed (Supplementary Table 2). Therefore, to calculate the relative abundance of MYH9 we used the average of 140 unique peptide sequences, rather than the average of all 778 redundant measurements.

In the exported data, we also noticed that often only one peptide from a labeled pair was fragmented for identification. Pairs were defined by a difference in molecular weight of either 2.004 or 4.008 Daltons for singly- or doubly-labeled COOH-termini. Search parameters in X! Tandem were not able to differentiate between COOH-terminal residues that were modified by enzymatic transfer of ^{18}O and internal residues resulting from missed tryptic cleavages that could not have been modified. As a result, pairs were assigned with differences of 4.008, 8.016, and 12.024 Da, although a difference of only 2.004 or 4.008 Da between heavy and light pairs was feasible with this labeling scheme. As a third example of a common problem in data handling, several peptides were quantified as having a relative abundance ratio exceeding 1:100. Manual inspection of the RAW data confirmed that these values were calculated based on the inappropriate assignment of peptide pairs.

Several techniques have been developed that make use of the fragmentation properties of ^{18}O labeled peptides to validate peptide identification.[33, 39-42] Instead, our strategy was to use a set of post-processing filters to remove peptides that were not members of a well-defined pair. During the subsequent development of Annotator we included additional filters to improve the quality of data used for protein quantitation.

Six optional filters with adjustable parameters were implemented (**Table 1**). Together these filters were used to ensure that a protein was defined by peptides that were properly paired and adequately sampled. Proteins and peptides that passed all six filters were used for quantitative analysis. Every unique peptide sequence was analyzed once; repetitive quantitation of peptides was removed by including only the measurement with the highest PeptideProphet score.[35] Whether data passed or failed a filter, all data were clearly listed in the final reports. This transparency allowed comparisons between reports using different

threshold values. Thresholds for each filter were manipulated during the analysis of the datasets, and the results are summarized in the proceeding sections.

Four of these optional filters were designed to accommodate stable isotope labeling and gel excision and may not be suitable for all experimental approaches. We encourage all users to consider the conditions used in sample preparation to guide the selection of appropriate data filters.

Filters 1 and 2: Chromatographic Elution Profile

The chromatographic elution profile of peptides provides two characteristics that can be used to enforce accurate quantitative analysis: the duration of elution and co-elution of peptide pairs.

The Scan Cutoff filter limits the minimum number of scans over which a peptide must be observed to ensure a reliable elution profile. Exploratory studies have determined that higher smoothness of a peptide elution profile increases the accuracy of measurements of peptide abundance.[43] By requiring a minimum duration for which a peptide is observed in MS, ions with very short and sporadic appearances can be filtered out. Maximizing the duration of peptide elution was used as a proxy for continuous peptide elution, an important characteristic of peptide chromatography and one that is used by many proteomic tools, including the Trans-Proteomic Pipeline (TPP).[5] Even very low limiting thresholds for the duration of peptide elution were successful at removing input from sporadic ions (**Table 2**, columns 1 and 2, rows 2 - 5). This had the effect of reducing the number of ion peaks that were inappropriately paired and improving the quality of data that were included in the final analysis.

The Light-Heavy Scan Cutoff filter limits the number of scans in which either the light or heavy isotope-labeled peptide is absent, thereby requiring co-elution. The co-elution of isotope-labeled heavy and light peptides by liquid chromatography is one confirmation that they share identical peptide sequences. Co-elution and subsequent analysis in a shared set of MS scans is also a requirement for the accurate comparison of relative ion abundance. To act as a true internal reference that minimizes the influence of variability in ion intensity between MS scans, peptide pairs should be present in the same MS scan. This user-defined threshold limits the number of MS scans that are not shared between peptide pairs, thereby maximizing the duration of co-elution. By this method, peptides without a co-eluting mate are not considered for quantitative analysis; they require special treatment since no relative ratio exists. Therefore, our strategy fails to identify cases of “present/absent” that may be very informative from a biological perspective. In these cases, a focused manual investigation of the data would be warranted.

This approach to filtering data was not computationally intensive and allowed end-user control over the filter parameters; however, there were disadvantages to using filters at the level of post-processing data analysis. For example, the number of scans reported by XPRESS for heavy and light peptides in a pair were always identical. Therefore, the Light-Heavy Scan Cutoff filter that we designed to confirm co-elution was not useful because data generated upstream by XPRESS did not differentiate between unique start and end scans for light versus heavy peptides in a pair. Inspection of the unprocessed RAW files clearly showed different start and end scans for each peptide in every pair. This highlights the importance of transparency in processing software and presents a case for permissive and information-rich analyses during early processing steps followed by more stringent analyses based on user-defined parameters in later steps.

Filters 3 and 4: Relative Quantitative Analysis From Labeled Pairs

Peptide pairs were defined by the difference in mass between unlabeled and ^{18}O-labeled samples. The Delta Mass Cutoff filter was used to limit the difference in mass between heavy- and light-labeled peptide pairs to either 2 or 4 Da for incomplete or complete labeling. Incomplete labeling with ^{18}O can lead to several challenges for accurate peptide quantitation and requires the use of a specialized application for data processing.[44] To test for incomplete labeling in these experiments, we imposed a maximum threshold value of 2 Da. This excluded all paired peptide sequences from the analysis and confirmed the inclusion of only completely labeled peptides from both experiments (**Table 2**, columns 3 and 4, row 2). A threshold value of 4 Da resulted in the exclusion of several hundred peptides (**Table 2**, columns 3 and 4, rows 3 - 7). This confirmed that internal lysine and arginine residues from missed proteolytic cleavages were allowed labeling modifications under the search parameters used in upstream processing software. This Delta Mass Cutoff filter was useful for defining peptide pairs by the difference in mass between heavy- and light-labeled peptides. It can also be considered a second independent validation of chromatographic co-elution. It confirms that both peptides are present in the same set of spectra, which is a result of chromatographic co-elution and shared peptide sequence identity.

To accommodate use with any labeling scheme, the user inputs any list of possible values to define peptide pairs. Although the data used in these experiments were generated with high mass accuracy so that an input threshold of 4.008 Da would be appropriate, the software was also designed for use with data from instruments that provide less confidence. Therefore, any difference in mass that was within 0.1 Da of the input value was retained. Because much of the work toward the identification and quantitation of peptide pairs was performed by upstream software, a strict threshold did not provide any additional benefit in this analysis.

A fourth filter, Ratio Cutoff, was imposed to limit inaccuracy in the relative quantitation of peptide pairs. During initial analysis, the relative areas of heavy and light peptide ions occasionally reached values nearing 1:100 and 1:1000. These outliers significantly broadened the standard deviation of relative peptide ratios for each protein, reducing confidence in the quantitative analysis. To exclude these values from the analysis, the user provides a minimum relative ratio between the treated and control peptide elution peak areas. Algebraically this implies that all ratios must be less than the reciprocal of the threshold value, thereby also providing a limit on the maximum value for peptide fold-change ratios.

This Ratio Cutoff filter removed several hundred peptide sequences in our data sets that demonstrated a greater than 20-fold difference in relative elution peak areas (**Table 2**, columns 5 and 6, row 4). Interestingly, around half of the total peptide sequences demonstrated a greater than 2-fold difference in relative elution peak areas. This filter was valuable for investigating the distribution of relative differences in peptide abundance across the entire experiment. The removal of outliers with extreme values increased the precision and accuracy of relative peptide ion quantitation and subsequent protein quantitation;[45] however, caution should be exercised to select against only the most extreme outliers, which are generally caused by inaccurate peak selection.

Filters 5 and 6: Protein Assignments

Peptide sequences were organized by the gene name of proteins to which they were assigned by X! Tandem. Organization by gene name provided the basis for two additional filters limiting the inclusion of peptides in quantitative analyses. The first of these filters, Molecular Weight Cutoff, took advantage of the range of molecular weights defined by the

gel slice from which a protein was excised. This filter was intended to limit the analysis to peptides that were digested by trypsin and were not the result of protein degradation. It was also intended to prevent oversampling of contaminating proteins that were present in every gel slice.

Identified proteins were referenced against the UniProt database, and molecular weight information was matched against fraction definitions provided by the user. The Molecular Weight Cutoff filter established a percentage of error that would be tolerated for the protein molecular weight, as determined by UniProt. Peptides from proteins that were identified in appropriate fractions, with added or subtracted error, were retained. The limitation of this filter was that the molecular weight noted by UniProt pertains to the protein precursor and not to the active form of expressed and modified proteins. Despite this limitation, a broad error allowing two times, or 100% difference above and below, the expected protein molecular weight removed over six hundred peptides from the total analysis (**Table 2**, columns 7 and 8, row 4). Indeed, intracellular protein processing and frequent post-translational modifications were not expected to affect expected molecular weights by more than 100%. Therefore, for the quantitative analysis of biological significance in this study, a filter threshold of 50% error above and below the reported molecular weight was used.

The sixth and final filter, Peptide Sequence Count Cutoff, removed proteins whose total number of peptide sequences was less than or equal to the cutoff value. This filter ensured that each protein was characterized by a minimum number of peptides. For example, the quantitative analysis of a protein from the relative ratio of one peptide between samples cannot be counted with confidence, and that protein should be excluded. The advantage of this filter was that it could be used to limit the analysis to proteins that were sampled frequently and therefore identified and analyzed with high confidence.

Statistics And Heuristics To Guide The Selection Of Biologically Significant Proteins

Peptides were included for quantitative analysis using the following filter parameters:

1. A peptide ion was observed for a minimum of 20 scans (Scan Cutoff),
2. Heavy and light pairs were defined by a 4 Da difference in molecular weight (Delta Mass),
3. Heavy and light pairs were present at a maximum direct ratio of their areas of 100:1 (Ratio Cutoff),
4. The protein from which a peptide was derived was identified in a gel slice within 50% error from the reported molecular weight (Molecular Weight Cutoff), and
5. More than one peptide was observed per protein (Peptide Sequence Count).

Of the 15,000 peptides originally identified for quantitation from one experiment, 50% were thrown out for redundancy and 17% were excluded using these filtration parameters. This left an average of 5000 unique peptides per experiment for quantitative analysis.

The control flow for quantitative analysis is summarized in **Figure 2**. Populations of peptides were analyzed in three major groupings: at the Experiment Level, the Keyword Level, and the Protein Level. Each group was examined to get a global view of abundance distributions. The average abundance for peptide heavy and light ratios was close to equal at the Experiment Level, confirming equal total protein abundance between the two differentially labeled samples. Nevertheless, each peptide ratio was normalized by the median of all labeled peptide ratios to ensure that abundance comparisons were based on a stable baseline.

The total population of labeled peptide ratios at the Experiment Level was used to describe patterns and trends in the data at the Protein and Keyword Levels. For example, the standard deviation and standard error of labeled peptide ratios at the Experiment Level was used to highlight individual proteins that showed a change in abundance relative to the whole population. Therefore, proteins that changed in abundance following LPS treatment were selected in both experiments using a standardized threshold established by the distribution of the observed population.

A one-sample t-test was also incorporated to compare the means of peptide ratios at the Protein and Keyword Levels against the mean at the Experiment Level. The number of peptides in each group determined the degrees of freedom for the t-test. In this way, we selected for proteins and keywords that were most affected by LPS treatment by identifying those whose means were significantly different from the mean at the Experiment Level.

The complementary analytical approach of using heuristic thresholds and statistical t-tests to select for relevant data is summarized in **Figure 3**. This statistic- and heuristic-based approach for selecting proteins that are functionally significant from proteomic data is unique. Our goal was to implement a set of rules that could be applied uniformly across data sets to determine significance for changes in abundance using measurements from the total population in an experiment.

Software solutions have been developed to determine which identified proteins merit further investigation. For example, ASAPRatio uses a log-transformed fitted normal, justified by the central limit theorem for large sample sizes, and an error function to generate p-values.[5] Meanwhile, GOMiner uses Fisher's exact test and q-values to manage small sample sizes.[46] In general, existing software encourages the heuristic use of statistical tests because small sample sizes are the most common condition for proteomic data that is organized by protein identity. Therefore, additional conditions, such as the assurance of normality, limit the applicability of rigorous statistical tests of significance.

This application uses the basic t-test for measurement confidence because it is easy to understand and therefore has more practical value and transparency for use as a common metric. It can be used heuristically but the inclusion of sample size and normality data also allows for a more rigorous test of significance. T-tests are performed for every group of peptides at the Protein and Keyword Levels. This multiple testing may result in an inflated number of false positives at the Experiment Level. To compensate for this, the format of the software output easily allows for the calculation of a Bonferroni correction using a significance threshold calculated from the total number of proteins or keywords identified in each experiment.[47] Alternatively, the user can extract a column of p-values for every protein or keyword, and calculate corresponding q-values using an external program. The current version of the software was used in an exploratory manner; however, future versions of the software will incorporate more explicit features for family-wise error-rates and false-discovery rates.

Data Management to Support Statistics and Heuristics

The t-test is widely used to show significance within data sets but relies on normality assumptions. Previous use of the t-test for proteomic data obtained by mass spectrometry is associated with several challenges, in particular low power due to small sample sizes[47] and the need to satisfy normality assumptions.[48]

To enable a more rigorous use of the t-test for identifying proteins and keywords that are significantly different from the majority of the population, we implemented a set of statistics related to descriptions of normality. Normality was described heuristically by values

calculated for skewness and kurtosis. As a more formal test for normality, we implemented D'Agostino-Pearson Omnibus K2 scores and p-values.[49]

Relative peptide abundance was calculated from the direct ratio of integrated heavy- to light-labeled peptide elution peak areas: $\frac{Heavy_{area}}{Light_{area}}$ or $\frac{Light_{area}}{Heavy_{area}}$ The frequency distribution of these fold-change ratios at the Experiment Level is truncated at 0 with an arbitrarily long right tail, which introduces skewness. Fold-change ratios have been log-transformed to correct for this skewness; however, the application of these normality statistics showed that applying a log-transformation to fold-change ratios does not imply normality (**Figure 4**).

Log-transformations facilitate an intuitive understanding of fold-change ratios by centering them at 0 and providing the same absolute value for an increase or decrease in abundance. Nevertheless, several complications are related to the log-transformation of proteomic data. After performing a log-transformation of the original data points, it becomes difficult to make statistical statements in terms of the original data. For example, the arithmetic mean of log-transformed data becomes the geometric mean under the reverse transformation, and more complex statistics become even less straightforward under reverse transformations.

An alternative formulation to fold-change was adopted, resulting in better normality statistics and a clearer understanding of system-wide calculations. This metric calculates the ratio of heavy- or light-labeled peptides in relation to the total sum of both peptide elution peak areas: $\frac{Heavy_{area}}{(Heavy_{area}+Light_{area})}$ or $\frac{Light_{area}}{(Heavy_{area}+Light_{area})}$ By this formulation, the relative ratio is always between 0 and 1 for every peptide pair, and a value of 0.5 represents equal abundance. This simplifies computation and facilitates comparisons within and between peptide pairs and among experiments. It is a correction often used to avoid sloping baselines, which prevent accurate comparisons and severely affect the precision of every measurement. This correction establishes an internally normalized scale and is particularly suited to mass spectrometry where comparative measurements between peaks is less precise for ratios that are far from 1:1.[45]

This method for calculating relative peptide abundance is convenient for comparing treated and control samples because abundance ratios for each share the same denominator and are defined by their inter-dependence (**Figure 4A**). Whether an experiment emphasizes the light- or the heavy-labeled set of peptides, the relative ratio of the control versus the treated sample is consistent. On the other hand, the mean of ratios calculated by a direct fold-change comparison does not have an obvious relationship to the mean of the reciprocal ratios (**Figure 4B**). Although a direct ratio of heavy- and light-labeled peptides is consistent with the reciprocal at the Peptide Level, the reciprocal ratios are not interchangeable at the Protein Level.

Using fold-change ratios that define a consistent relationship between a treated sample and its control resulted in data with a higher tendency toward normality (**Supplementary Figure 1**) and a lower degree of estimated error (**Figures 4C, 4D, 4E, 4F**). Using this method to calculate relative peptide ratios, the frequency distribution at the Experiment Level was constrained at both the right and left tail, thereby minimizing the standard deviation of the total population. This constraint also had the effect of producing low kurtosis scores that increased the power of one-sample t-tests.[50]

Using relative peptide ratios calculated from the total area of the peptide pair, we imposed requirements for normality and for statistic and heuristic significance (**Table 3**). Per experiment, roughly 5,000 labeled peptide pairs were used to calculate the relative

abundance of just over 700 proteins. Of these, 85% of the proteins failed normality, and 65% of those remaining did not significantly change in abundance when compared to the total population of proteins sampled ($p \leq 0.01$). This left less than 5% of the proteins as statistically significant indicators of LPS activation in differentiated HL60 cells. Using a less stringent threshold, 15% of the proteins originally identified were at least 1 standard deviation from the mean at the Experiment Level. Quantitative validation required that proteins were selected as significant in both technical replicates. Only 7 proteins were statistically significant in both experiments and 26 proteins were heuristically significant in both. The proteins that passed these tests for having significantly changed in abundance in response to LPS are presented in **Figure 5A**.

Quantitative Analysis of Keywords Selects for Unique Sets of Proteins

Traditionally, peptides are grouped under the gene names of the proteins from which they were derived. Gene names provide a natural classification system for proteomic data; however, they also create challenges for statistical validation and biological inference. For example, in one experiment 55.3% of the proteins used for quantitative analysis were identified by less than 10 unique peptides and 88.3% were identified by less than 20 peptides (**Supplementary Figure 2**). Sample sizes this small make it difficult to reach valid statistical conclusions. Whether the final list contains 10 or 100 proteins, it can be time consuming and difficult to derive a central trend that describes changes in cellular function. Keyword categories, which encompass multiple gene names, can be used to generate sample sizes that are more conductive to hypothesis testing.

Peptides were grouped by keyword term and the mean of their relative abundance ratios was used to determine which biologically functional categories were influenced by LPS to the greatest extent. The richness of the UniProt and Gene Ontology keywords, which included categories such as molecular function, cellular compartment, post-translational modification and associated ligand, provided a strong set of categories and terms to work with “out of the box.” All 13 keyword categories were used, containing a total of 32,378 terms that describe various properties and functional characteristics of identified proteins.

Our original hypothesis was that groupings by keyword would allow for loose comparisons between experiments. For example, instead of requiring the same protein to be sampled in each experiment, proteins within the same keyword term could be observed and grouped for a summary effect. Contrary to our expectations, significant keywords emphasized proteins that, when analyzed independently, were not selected as significant. Significance at the Protein Level did not determine quantitative significance at the Keyword Level.

Out of 32,000 possible keyword terms, only 1% were represented by peptides in this analysis (**Table 3**). Of those represented in these experiments, 70% failed normality and only 20% significantly changed in abundance ($p \leq 0.01$). Using a more permissive threshold for heuristic significance, an average of 15% of the sampled keywords changed in abundance in response to LPS. For validation, we required significant keywords to be selected in both technical replicates. Only 14 keywords were statistically significant in both experiments and 12 keywords were heuristically significant in both. These are presented in **Figure 5B**.

Quantitative Analysis of Keyword Overlap Demonstrates Functional Signatures

To overcome the fact that proteins rarely perform a discrete function, we also measured the degree of observed overlap between functional categories that were identified in the experiment. The degree of overlap between keywords was calculated by the number of

proteins that are shared between terms: $\frac{P_{ab}}{P_a+P_b+P_{ab}}$ where P_{ab} is the number of proteins shared between keywords a and b, P_a is the number of proteins associated with keyword a, and P_b is the number of proteins associated with keyword b. The formula describes the number of proteins that are common between any two terms divided by the total number of distinct proteins in both terms. Described in the language of set theory, the percentage of overlap between keyword terms is defined as the cardinality of the intersection between terms divided by the cardinality of the union. The resulting score is a number between 0 and 1, where 1 represents complete overlap.

Keyword overlap provided an intuitive means for detecting proteins and protein associations that serve multiple functions. For example, there was 100% overlap between the keyword terms LDL, Chylomicron and Atherosclorosis because the protein APOB48R was associated with all three keywords in both experiments. Similarly, CLIC1 was associated with the terms Chloride, Chloride Channel, and Ionic Channel, resulting in 100% overlap between those terms. Meanwhile, the term Prenylation shared 6% overlap with Cardiomyopathy, 4% overlap with Cell Adhesion, and 5% overlap with Chaperone (**Figure 6**). Similarly, Integrin shared 30% overlap with Cell Adhesion and 33% overlap with Epidermolysis bullosa. In this way, by taking into consideration the overlap between keyword terms, the inherent hierarchical grouping of gene ontology keywords does not negatively affect quanitation. During quantitative analysis, these hierarchical terms flatten out if every member of a specific term is contained within the parent term. On the other hand, if the parent term contains a more inclusive grouping of proteins than a specific term, there may be significant differences in the quantitative signature.

It is particularly important to note, that several of the proteins shared between functional terms were not selected as statistically or heuristically significant at the Protein Level. Instead, these statistically significant keywords were selected based on the observed distribution of all keywords in the total population of peptides. By selecting a few keyword terms of personal interest from the total population of keywords found in this experiment, we were able to generate a signature of protein activity (**Figure 7**). This limited signature demonstrates that functional categories can be used to quantitatively monitor changes in cell behavior, thereby providing a complete description of cell responses and how they change with specific stimuli.

Comparative Evaluation Analysis

To evaluate the effectiveness of our approach in selecting proteins of significance, we also analyzed our data using the conventional method of analysis for proteomic data sets. In the following section we investigate the consequences of this conventional approach in terms of:

1. the proteins that were selected by quantitative analysis, and
2. the holistic biological context that could be derived from functional groups to which these proteins belong.

A conventional quantitative analysis consists of calculating the average fold-change in protein abundance from a log-transformed ratio of heavy and light peptides. To select for proteins that increased or decreased in abundance relative to an experimental control, an arbitrary threshold value for the average fold-change is selected. Therefore, for this comparative evaluation, each peptide was quantified by the relative fold-change between the LPS-treated sample and the un-treated control: $Log_2\left(\frac{Area_{LPS}}{Area_{Control}n}\right)$ where $Area_{LPS}$ is the integrated peptide elution peak area of the LPS-treated sample, *AreaControl* is the integrated

peptide elution peak area of the untreated control sample and *n* is the median of all $\frac{Area_{LPS}}{Area_{Control}}$ ratios in an experiment, included as a normalization factor.

Then, the relative abundance of each protein was calculated from the average fold-change of its constituent peptides (**Supplementary Table 5**). A list was generated of all the proteins that passed the arbitrary threshold of an average fold-change greater than or equal to 1.0 or less than or equal to -1.0. To ensure correct data management using this conventional method of anlaysis, we used two basic filters prior to the quantitative analysis of peptides and proteins:

1. We removed duplicate peptide sequences associated with the same protein, keeping unique peptide sequences with the highest PeptideProphet score.
2. After grouping peptide sequences by gene name, we removed proteins that were identified by only one peptide sequence.

In Table 4 we compare the results from this conventional analysis to results obtained using the filters and data management methods developed in Annotator and described in previous sections. We found that the more robust results of the conventional analysis were corroborated by results obtained using Annotator (**Table 4**, column 2: Validated by the Conventional Method). Annotator also selected six additional proteins of quantitative significance that were missed in the conventional analysis (**Table 4**, column 1: False negatives), and screened out thirteen proteins that did not significantly change in abundance (**Table 4**, columns 3 and 4: False positives).

In this comparison, false positives fell into one of two categories (**Table 4**, columns 3 and 4). Proteins in the first category (**Table 4**, column 3: Quantitative significance) were selected by the conventional method because their relative abundance was greater than the arbitrarily set fold-change cutoff (greater than 1.0 or less than -1.0). Annotator rejected these proteins because, when viewed from the perspective of the total population of observed proteins in the experiment, their relative abundance was less than one standard deviation from the population mean. That is to say that although the fold-change of these proteins was greater than the cutoff, it did not accurately describe the quantitative significance of the proteins within the observed population. In general, the use of standard deviations from the mean instead of fold-change thresholds offers a subtle but distinct advantage for making comparisons between experiments. If data points within an experiment are modeled as a normal curve, the standard deviations will always fall on consistent points of that curve. This is not true for arbitrary fold-change thresholds, where a given threshold may capture more or fewer data points depending on the overall spread of the data.

False positives that fell in the second category (**Table 4**, column 4) were rejected by Annotator using the Delta Mass Cutoff Filter (Filter 3, refer to **Table 1**) and the Molecular Weight Cutoff Filter (Filter 5, refer to **Table 1**). These filters confirmed the accurate selection of peptides used for protein quantitation; the Delta Mass Cutoff filter required that the difference in mass between ^{18}O-labeled and unlabeled peptides was equal to 4 Da., while the Molecular Weight Cutoff Filter excluded peptides that were identified in gel slices greater or less than 50% of the reported protein molecular weight.

EIF5A was identified as a false positive in the conventional analysis (**Table 4**, column 3) and provides an interesting example of the how peptide selection can influence quantitative significance. Annotator did not select EIF5A as quantitatively significant because its average abundance was less than one standard deviation from the mean. It was borderline, however, with an ; however, it was very close with an average abundance that was 0.97

standard deviations from the mean. It is interesting to note that EIF5A was selected as significant in the conventional analysis based, in part, on the inclusion of one peptide with an extreme direct ratio between labeled pairs of -4.7 but with a difference in mass of 12.1 Da. between peptides in the pair. According to the stable isotope-labeling scheme used in this study, a peptide pair could not be defined by a difference in mass greater than 4.0 and this peptide was filtered out of the analysis using Annotator, thereby reducing the average abundance of the protein below the cutoff. This demonstrates the effectiveness of filters used to remove errors from upstream handling techniques and consequences for the selection of proteins in the final analysis.

We used DAVID (http://david.abcc.ncifcrf.gov) to determine which functional trends were overrepresented in the list of proteins that were selected as quantitatively significant by conventional fold-change analysis. From a list of 16 proteins that increased in abundance, DAVID identified 50 keyword terms that were significantly overrepresented ($p \leq 0.01$). Many of the terms were redundant, such as RNA splicing via transesterification reactions, RNA splicing via transesterification reaction with bulged adenosine as nucleophile, nuclear mRNA splicing via spliceosome, mRNA processing and RNA processing. On the whole, the trend seemed to point toward spliceosome activity and the general process of translation but an easy summary was elusive and the only option for presentation of results was a list. The responsibility of identifying trends from this list rests with the user.

Given a smaller list of only 7 proteins that decreased in abundance, DAVID identified 8 keyword terms that were significantly overrepresented ($p \leq 0.01$). There was no identifiable trend among the enriched groups, which ranged from the generation of precursor metabolites and energy, to non-membrane bounded organelle, cytosol, phosphoprotein, and acetylation.

One clear disadvantage of this approach is that a shorter input list decreased sensitivity and skewed the results toward very general terms. This comparative analysis demonstrates that more often than not inaccurate quantitation obscures biologically interesting proteins from being selected. DAVID uses a variant of Fisher's exact test to determine if proteins of a particular keyword category are disproportionately represented in a given list.[51] DAVID does not take into account the calculated relative abundance of each protein. Instead, each protein in a list is counted as a member of a keyword group and group membership is scored based on expected frequency. From lists of only 16 or 7 proteins, the statistical power of enrichment analysis is very weak. In contrast to DAVID's enrichment analysis approach, Annotator calculates the average relative abundance of all peptides grouped by a common keyword term. Therefore, keyword abundance is based on observed measurements.

There was little meaningful overlap between the keyword analysis performed by Annotator and the enrichment analysis performed by DAVID. A few of the keywords that Annotator selected as having changed most in abundance (Chloride Channel, Inflammatory Response, Plasminogen Activation, Hypusine modification) were not identified by DAVID. By comparing DAVID's enrichment analysis to keyword quantitation performed by Annotator, we were able to highlight Annotator's ability to identify meaningful keywords from the entire proteomic population, rather than by comparisons using a small group of significant proteins.

CONCLUSIONS

With an average run time of less than 1 minute, Annotator allows users to efficiently analyze large sets of LC-MS/MS data for quantitative significance in a biological context. The defining feature of this analysis is that relative peptide abundance was used to calculate the observed relative abundance and degree of similarity between functional categories. The

approach is fundamentally different from enrichment analyses or network overlays because relative abundance was calculated from direct measurements of peptide abundance, which is the basic observable unit in LC-MS/MS studies.[52] Similarly, statistical and biological significance were based only on the observed population in a given experiment. Based on our direct comparison using a traditional enrichment analysis, we show that results using Annotator differ dramatically from having a statistically significant quantitative proteomics dataset, obtained using Census[53] or MaxQuant[10], that is then subjected to GO analysis. We show that an enrichment analysis using a list of less than 100 proteins is no more than a qualitative listing of functions; the results from Annotator are not a mere list of enriched GO terms with measures of confidence. Our major aim was to avoid the selective reporting of changes to only a few proteins of interest, a process that can introduce personal bias and overlook previously unreported results. In developing a method to carry out this aim, we revealed major weaknesses in the current determination of cellular function from short lists of gene names. This reiterates the particularity of proteomic analyses using LC-MSMS as providing fundamentally different qualities and quantities of data compared to genome-wide sequencing efforts.

Using Annotator, quantitation by keyword terms provided access to biologically relevant signatures that could be statistically validated.[54, 55] These functional categories may be thought of as annotated subsystems of proteins that share a common biological role.[56] Annotator presents these subsystems without a direct connection to network models; however, by exploring observed relationships among keywords Annotator may provide evidence that supports current network models.[57, 58]

In comparative genome analysis, the function-based subsystem approach is very efficient for highlighting promising drug targets and is especially robust in cases where therapies are directed at the whole organism, such as during infection.[59] Our results show that by using function-based signatures in large-scale proteomic studies we may be able to infer essentiality, vulnerability, and conservation. Our analysis shows that the key to applying the approach successfully is to base functional signatures on quantitative measurements and use statistics to standardize the selection of significant categories.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded in part by NIH grants GM60443 and GM074691 and a Spark award from the Chicago Biomedical Consortium with support from The Searle Funds at The Chicago Community Trust. J.E.S. was supported by an American Heart Association Midwest Affiliate predoctoral fellowship. Donald Wolfgeher is acknowledged for tryptic digestions and ^{18}O labeling. We thank Kenneth Johnson and Robert Bergen at the Mayo Proteomics Research Center for Orbitrap mass spectrometry.

ABBREVIATIONS

FBS	Fetal bovine serum
TPP	Trans-Proteomic Pipeline
LPS	Lipopolysaccharide
PS	peptide sequence
UniProt	Universal Protein Resource

IPI	International Protein Index
MALDI-TOF MS	Matrix-assisted laser-desorption ionization time-of-flight mass spectrometry
LC	liquid chromatography
MS	mass spectrometry
MS/MS	tandem mass spectrometry
LC-MS/MS	liquid chromatography tandem mass spectrometry
ppm	parts per million

REFERENCES

1. Jacob RJ. Bioinformatics for LC-MS/MS-based proteomics. Methods Mol Biol. 2010; 658:61–91. [PubMed: 20839098]
2. Eng JK, McCormack AL, Yates JR III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of The American Society for Mass Spectrometry. 1994; 5(11):976–989.
3. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999; 20(18):3551–3567. [PubMed: 10612281]
4. Craig R, Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun Mass Spectrom. 2003; 17(20):2310–6. [PubMed: 14558131]
5. Li XJ, Zhang H, Ranish JA, Aebersold R. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. Anal Chem. 2003; 75(23):6648–57. [PubMed: 14640741]
6. Keller A, Eng J, Zhang N, Li XJ, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol. 2005; 1:0017. 2005. [PubMed: 16729052]
7. Halligan BD, Slyper RY, Twigger SN, Hicks W, Olivier M, Greene AS. ZoomQuant: an application for the quantitation of stable isotope labeled peptides. J Am Soc Mass Spectrom. 2005; 16(3):302–6. [PubMed: 15734322]
8. Kohlbacher O, Reinert K, Gropl C, Lange E, Pfeifer N, Schulz-Trieglaff O, Sturm M. TOPP--the OpenMS proteomics pipeline. Bioinformatics. 2007; 23(2):e191–7. [PubMed: 17237091]
9. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem. 2007; 389:1017–1031. [PubMed: 17668192]
10. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008; 26(12):1367–72. [PubMed: 19029910]
11. Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, Sun Z, Nilsson E, Pratt B, Prazen B, Eng JK, Martin DB, Nesvizhskii AI, Aebersold R. A guided tour of the Trans-Proteomic Pipeline. Proteomics. 2010; 10(6):1150–9. [PubMed: 20101611]
12. Malik R, Dulla K, Nigg EA, Korner R. From proteome lists to biological impact--tools and strategies for the analysis of large MS data sets. Proteomics. 10(6):1270–83. [PubMed: 20077408]
13. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003; 422(6928):198–207. [PubMed: 12634793]
14. Breitkreutz A, Choi H, Sharom JR, Boucher L, Neduva V, Larsen B, Lin ZY, Breitkreutz BJ, Stark C, Liu G, Ahn J, Dewar-Darch D, Reguly T, Tang X, Almeida R, Qin ZS, Pawson T, Gingras AC, Nesvizhskii AI, Tyers M. A global protein kinase and phosphatase interaction network in yeast. Science. 2010; 328(5981):1043–6. [PubMed: 20489023]
15. Oda K, Kitano H. A comprehensive map of the toll-like receptor signaling network. Mol Syst Biol. 2006; 2:0015. 2006. [PubMed: 16738560]

16. Ekins S, Nikolsky Y, Bugrim A, Kirillov E, Nikolskaya T. Pathway mapping tools for analysis of high content data. Methods Mol Biol. 2007; 356:319–50. [PubMed: 16988414]
17. Levy ED, Landry CR, Michnick SW. How perfect can protein interactomes be? Sci Signal. 2009; 2(60):pe11. [PubMed: 19261595]
18. Levy ED, Landry CR, Michnick SW. Cell signaling. Signaling through cooperation. Science. 2010; 328(5981):983–4. [PubMed: 20489011]
19. Lin CC, Juan HF, Hsiang JT, Hwang YC, Mori H, Huang HC. Essential core of protein-protein interaction network in Escherichia coli. J Proteome Res. 2009; 8(4):1925–31. [PubMed: 19231892]
20. Janga SC, Babu MM. Network-based approaches for linking metabolism with environment. Genome Biol. 2008; 9(11):239. [PubMed: 19040774]
21. Huang D, Sherman B, Tan Q, Collins J, Alvord WG, Roayaei J, Stephens R, Baseler M, Lane HC, Lempicki R. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biology. 2007:8.
22. Hosack DA, Dennis G Jr. Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biol. 2003; 4(10):R70. [PubMed: 14519205]
23. Searle BC. Scaffold: A bioinformatic tool for validating MS/MS-based proteomic studies. Proteomics. 2010; 10(6):1265–1269. [PubMed: 20077414]
24. Bhatia VN, Perlman DH, Costello CE, McComb ME. Software tool for researching annotations of proteins: open-source protein annotation software with data visualization. Anal Chem. 2009; 81(23):9819–23. [PubMed: 19839595]
25. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research. 2003; 13:2498–2504. [PubMed: 14597658]
26. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang P-L, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD. Integration of biological networks and gene expression data using Cytoscape. Nature Protocols. 2007; 2:2366–2382.
27. Hammond DE, Hyde R, Kratchmarova I, Beynon RJ, Blagoev B, Clague MJ. Quantitative analysis of HGF and EGF-dependent phosphotyrosine signaling networks. J Proteome Res. 2010; 9(5): 2734–42. [PubMed: 20222723]
28. Domon B, Aebersold R. Mass spectrometry and protein analysis. Science. 2006; 312(5771):212–7. [PubMed: 16614208]
29. Cravatt BF, Wright AT, Kozarich JW. Activity-based protein profiling: from enzyme chemistry to proteomic chemistry. Annu Rev Biochem. 2008; 77:383–414. [PubMed: 18366325]
30. Ear PH, Michnick SW. A general life-death selection strategy for dissecting protein functions. Nat Methods. 2009; 6(11):813–6. [PubMed: 19820714]
31. Breitman TR, Selonick SE, Collins SJ. Induction of differentiation of the human promyelocytic leukemia cell line (HL-60) by retinoic acid. Proc Natl Acad Sci U S A. 1980; 77(5):2936–40. [PubMed: 6930676]
32. Kristjansdottir K, Wolfgeher D, Lucius N, Angulo DS, Kron SJ. Phosphoprotein profiling by PA-GeLC-MS/MS. J Proteome Res. 2008; 7(7):2812–24. [PubMed: 18510356]
33. Heller M, Mattou H, Menzel C, Yao X. Trypsin catalyzed 16O-to-18O exchange for comparative proteomics: tandem mass spectrometry comparison using MALDI-TOF, ESI-QTOF, and ESI-ion trap mass spectrometers. J Am Soc Mass Spectrom. 2003; 14(7):704–18. [PubMed: 12837592]
34. Rauch A, Bellew M, Eng J, Fitzgibbon M, Holzman T, Hussey P, Igra M, Maclean B, Lin CW, Detter A, Fang R, Faca V, Gafken P, Zhang H, Whiteaker J, States D, Hanash S, Paulovich A, McIntosh MW. Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. J Proteome Res. 2006; 5(1):112–21. [PubMed: 16396501]

35. Keller A, Nesvizhskii AI, 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(20):5383–92. [PubMed: 12403597]
36. Han DK, Eng J, Zhou H, Aebersold R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. Nat Biotechnol. 2001; 19(10):946–51. [PubMed: 11581660]
37. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998; 95(25):14863–8. [PubMed: 9843981]
38. Ramos-Fernandez A, Lopez-Ferrer D, Vazquez J. Improved method for differential expression proteomics using trypsin-catalyzed 18O labeling with a correction for labeling efficiency. Mol Cell Proteomics. 2007; 6(7):1274–86. [PubMed: 17322307]
39. Takao T, Hori H, Okamoto K, Harada A, Kamachi M, Shimonishi Y. Facile assignment of sequence ions of a peptide labelled with 18O at the carboxyl terminus. Rapid Commun Mass Spectrom. 1991; 5(7):312–5. [PubMed: 1841649]
40. Schnolzer M, Jedrzejewski P, Lehmann WD. Protease-catalyzed incorporation of 18O into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry. Electrophoresis. 1996; 17(5):945–53. [PubMed: 8783021]
41. Zhang N, Aebersold R, Schwikowski B. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. Proteomics. 2002; 2(10): 1406–12. [PubMed: 12422357]
42. Volchenboum SL, Kristjansdottir K, Wolfgeher D, Kron SJ. Rapid validation of Mascot search results via stable isotope labeling, pair picking, and deconvolution of fragmentation patterns. Mol Cell Proteomics. 2009; 8(8):2011–22. [PubMed: 19435713]
43. Yang C, Yang C, Yu W. A regularized regression method for peptide quantification. Journal of Proteome Research. 2010 In press.
44. Mason CJ, Therneau TM, Eckel-Passow JE, Johnson KL, Oberg AL, Olson JE, Nair KS, Muddiman DC, Bergen HR 3rd. A method for automatically interpreting mass spectra of 18O-labeled isotopic clusters. Mol Cell Proteomics. 2007; 6(2):305–18. [PubMed: 17068186]
45. MacCoss, MJ.; Wu, CC. Computational Analysis of Quantitative Proteomics Data Using Stable Isotope Labeling.. In: Sechi, S., editor. Quantitative Proteomics by Mass Spectrometry. Vol. 359. Humana Press; 2007. p. 177-189.
46. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. 2003; 4(4):R28. [PubMed: 12702209]
47. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 2003; 4(4):210. [PubMed: 12702200]
48. Karp NA, Lilley KS. Design and analysis issues in quantitative proteomics studies. Proteomics. 2007; 7(Suppl 1):42–50. [PubMed: 17893850]
49. D'Agostino RB, Belanger A, D'Agostino RBJ. A Suggestion for Using Powerful and Informative Tests of Normality. The American Statistician. 1990; 44(4):316–321.
50. Reineke DM, Baggett J, Elfessi A. A Note on the Effect of Skewness, Kurtosis, and Shifting on One-Sample t and Sign Tests. Journal of Statistics Education. 2003; 11(3):1–13.
51. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4(1):44–57. [PubMed: 19131956]
52. Kuster B, Schirle M, Mallick P, Aebersold R. Scoring proteomes with proteotypic peptide probes. Nat Rev Mol Cell Biol. 2005; 6(7):577–83. [PubMed: 15957003]
53. Park SK, Venable JD, Xu T, Yates JR 3rd. A quantitative analysis software tool for mass spectrometry-based proteomics. Nat Methods. 2008; 5(4):319–22. [PubMed: 18345006]
54. Yang D, Li Y, Xiao H, Liu Q, Zhang M, Zhu J, Ma W, Yao C, Wang J, Wang D, Guo Z, Yang B. Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories. Bioinformatics. 2008; 24(2):265–71. [PubMed: 18006543]

55. Minguez P, Al-Shahrour F, Dopazo J. A function-centric approach to the biological interpretation of microarray time-series. Genome Inform. 2006; 17(2):57–66. [PubMed: 17503379]

56. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005; 33(17):5691–702. [PubMed: 16214803]

57. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol. 2007; 3:88. [PubMed: 17353930]

58. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004; 5(2):101–13. [PubMed: 14735121]

59. Osterman AL, Begley TP. A subsystems-based approach to the identification of drug targets in bacterial pathogens. Prog Drug Res. 2007; 64:131, 133–70. [PubMed: 17195474]

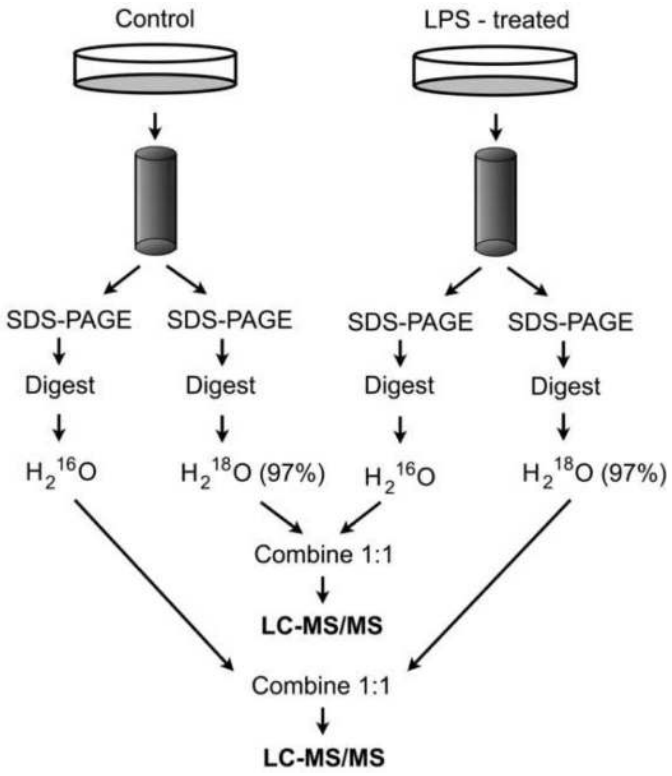

Figure 1. Technical replicates, in which the state of ^{18}O labeling for the sample and the control are switched

Differentiated HL60 cells were treated in culture with LPS or water for 30 minutes. Cell lysates from each population were enriched for phosphorylated protein complexes, which were separated by gel electrophoresis and divided into 11 fractions. Proteins in each fraction were digested with trypsin in-gel. Peptides were extracted and labeled with ^{18}O at the COOH-terminus, using trypsin immobilized on beads, or left unlabeled in water. Labeled samples were combined with the equivalent fractions from the unlabeled sample for equal relative abundance. To confirm complete ^{18}O labeling and validate relative peptide abundance, a technical replicate was performed with duplicate gel electrophoresis and digestion but a switched ^{18}O labeling state.

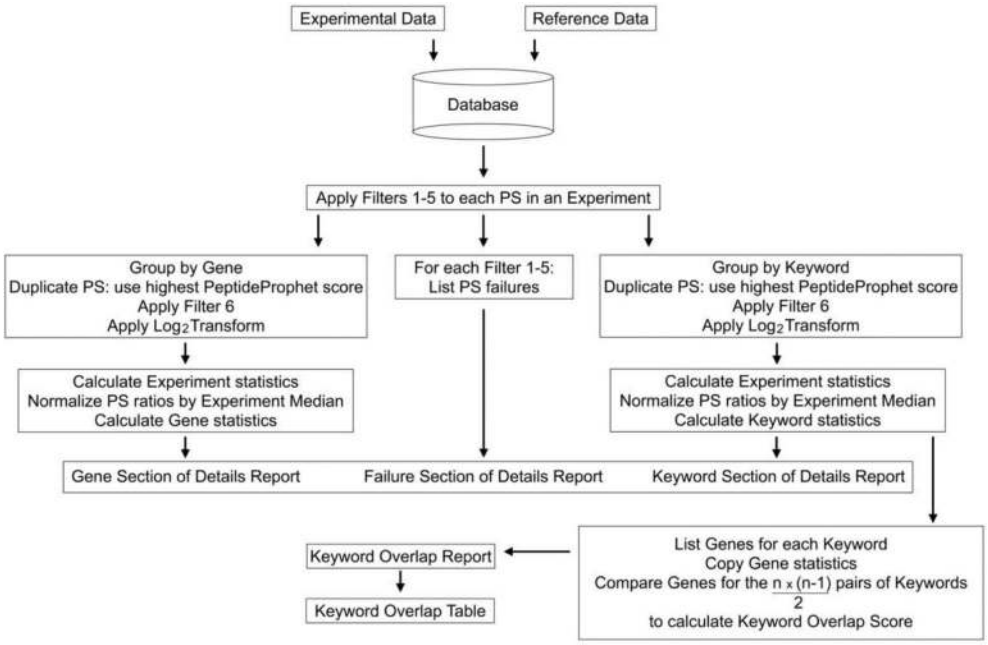

Figure 2. Computation control flow

Experimental data were loaded into a MySQL database and compared against proteins and keywords referenced in the UniProt and Gene Ontology catalogues. Five filters based on observations relating to the chromatographic elution profile, the relative quantitation of labeled peptide pairs, and the expected protein molecular weight excluded peptide sequences from the analysis. Only unique peptide sequences with the highest PeptideProphet score were retained. Peptides were grouped by gene name or keyword term, and only those with more than one peptide were used for quantitative analysis. Initial statistics were calculated and the median of all peptide ratios was used to normalize for equal relative abundance at the Experiment Level. Statistics were re-calculated for all relative peptide ratios and for proteins and keywords. These statistics made up the Details Report, which listed peptide sequences under the appropriate gene name or keyword term. Proteins were also evaluated for multiple functional roles by calculating the percentage of overlap between all keyword terms. Proteins that were shared and excluded by each keyword pair were listed in the Keyword Overlap Report or presented in a tubular format in the Keyword Overlap Table.

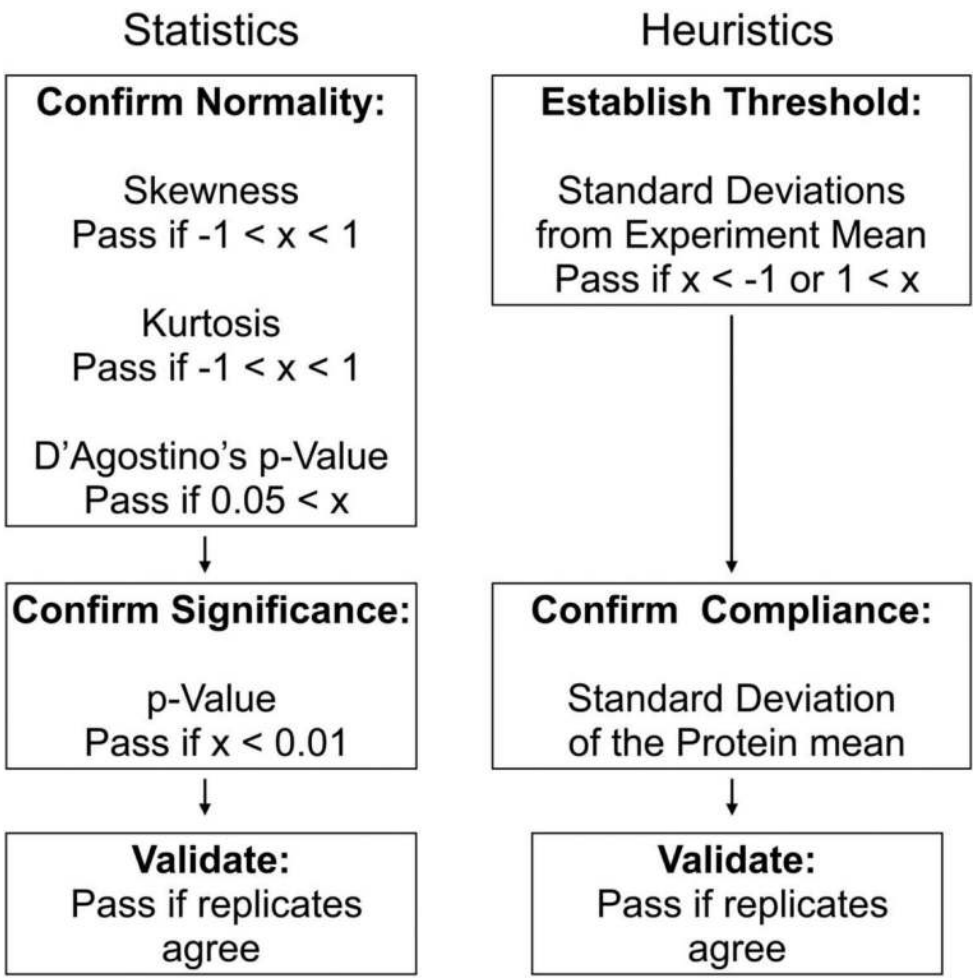

Figure 3. Schematic for the systematic use of statistics and heuristics to guide the selection of proteins and keywords that significantly change in abundance

The t-test forstatistical significance at the Protein and Keyword Levels requires that those groups of peptides display normal distributions. Three tests were used to confirm normality before the t-test was applied. Only populations with skewness and kurtosis scores between -1 and 1, and with a D'Agostino's p-value greater than 0.05, were analyzed for statistical significance ($p \leq 0.01$) against the mean of relative peptide ratios at the Experiment Level. A more permissive scale was also used to compare proteins and keywords to the total sampled population. Proteins and keywords whose mean was greater than one standard deviation above or below the mean of relative peptide ratios at the Experiment Level were considered heuristically significant. In this evaluation, the only requirement was that the standard deviation of proteins and keywords did not exceed this threshold cutoff. For both measures of significance, only proteins and keywords that were selected as significant in both technical replicates were considered valid.

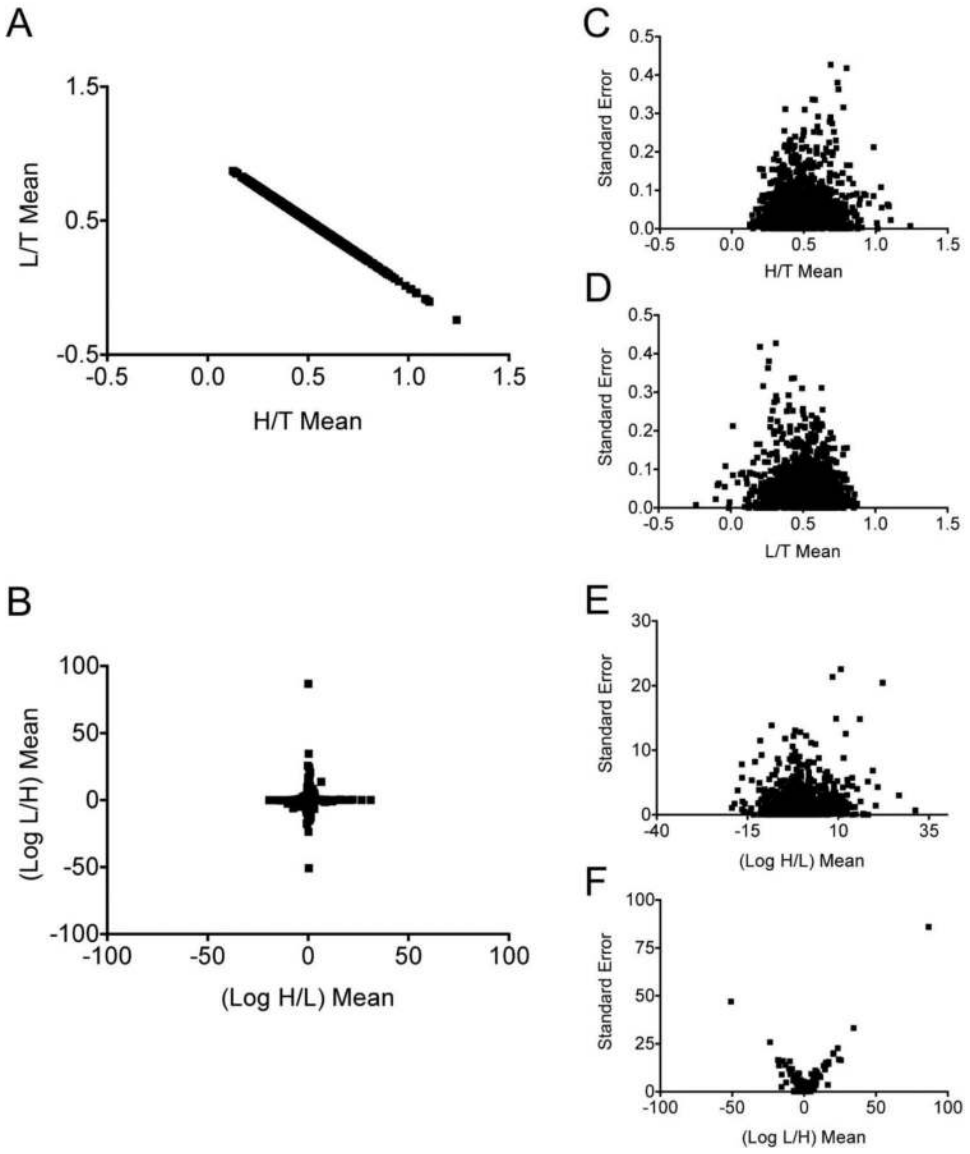

Figure 4. The estimation of error is influenced by the calculation of relative protein abundance

A consistent relationship defines heavy and light peptide pairs analyzed in the context of the total sum of both peptide elution peak areas (**A**). Although the direct ratio of heavy and light peptide elution peak areas is clearly related to the reciprocal ratio at the peptide level, the mean of the direct ratio and the mean of the reciprocal ratio are not related in an obvious way (**B**). Therefore, it can be difficult to compare relative protein abundance between experiments using the direct ratio of heavy to light peptide elution peak areas. Ratios calculated from the relationship between heavy or light peptide elution peak areas to the total area of the pair are constrained between 0 and 1. The standard error of the mean of these ratios is small (**C**, **D**). By comparison, the direct ratio of heavy and light peptide elution peak areas is not bounded. The error associated with the mean of heavy to light ratios and the mean of the reciprocal ratios are relatively large and vary considerably according to the ratio calculated (**E**, **F**).

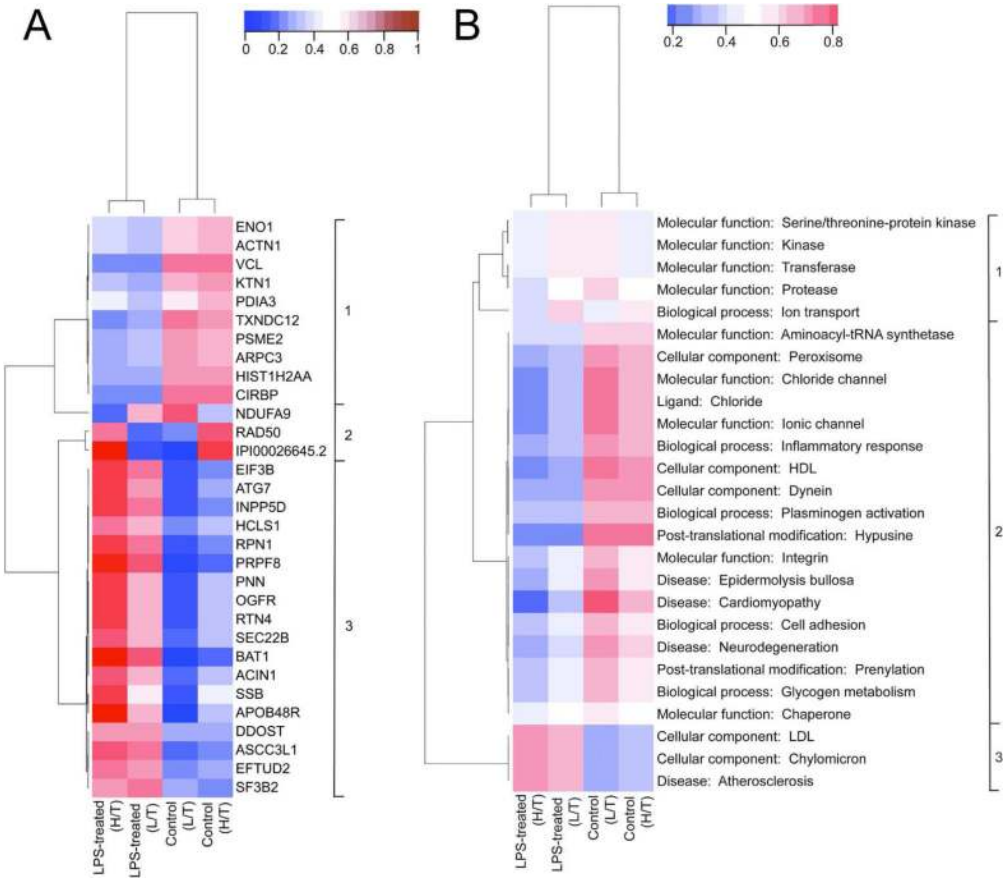

Figure 5. Proteins and Keywords that significantly changed in abundance in response to LPS

Cluster analysis and heat maps provide a rapid and intuitive method for defining relationships between observed proteins, functional categories, and samples. Columns highlight consistency between technical replicates. Rows emphasize similarities in response patterns between groups of proteins and keywords. Only proteins and keywords that were statistically ($p \leq 0.01$) or heuristically (greater than 1 standard deviation from the Experiment Level mean) significant in both technical replicates are shown here. A mean ratio of 0.5 (**white**) represents equal relative abundance between the LPS-treated sample and the control. An increase in abundance in response to LPS is highlighted in **red**, while a decrease in abundance is highlighted in **blue**. Proteins and keywords are hierarchically clustered by Pearson correlation values to reveal similar response patterns. **A**) The proteins in **Group 1** decreased in abundance, while the proteins in **Group 3** increased in abundance in response to LPS. The proteins in **Group 2** are characterized by values that did not agree between technical replicates and these proteins are considered contaminants from sample processing. **B**) The keywords in **Group 2** decreased in abundance in response to LPS, while the keywords in **Group 3** increased in abundance. The keywords in **Group 1** are shown to be poor groupings for quantitative analyses in these data sets because differences in protein membership between technical replicates may have diverse functional effects. These groupings provide clues regarding the interdependence of protein and cellular functions, suggesting possible avenues of further research.

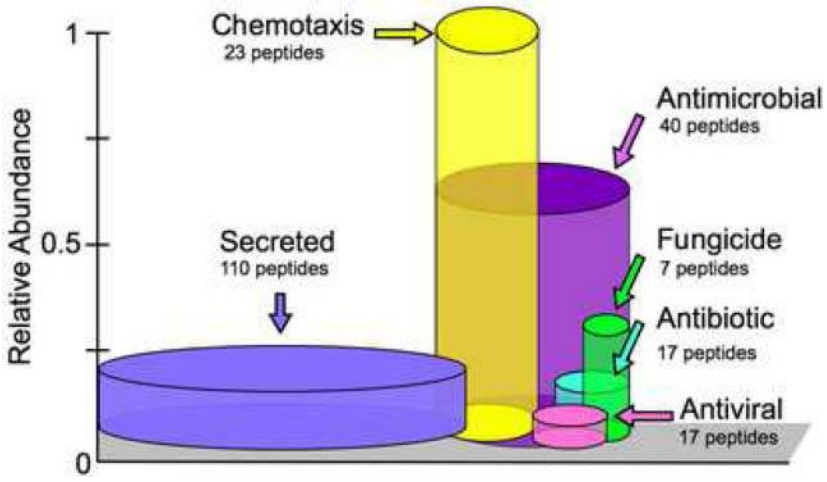

Figure 6. A Quantitative Signature of Protein Function in the Cell

A selective subset of functional categories was used to generate a quantitative protein activity signature for a human neutrophil stimulated with lipopolysaccharide from E.coli. Peptides were grouped into functional categories and relative abundance was calculated directly from quantitative LCMS/MS measurements. This signatures demonstrates that of the 40 antimicrobial peptides observed, 7 peptides had a dual-function as fungicidal peptides, and 17 were also antiviral. Peptides with multiple functionalities may be particularly interesting drug targets for infection. The quantitative signature, however, demonstrates specificity toward an antimicrobial response because the antimicrobial category increased in abundance to a greater extent than the fungicidal and antiviral categories. The signature also reveals that one of the predominant responses to bacterial infection is a significant increase in chemotaxis.

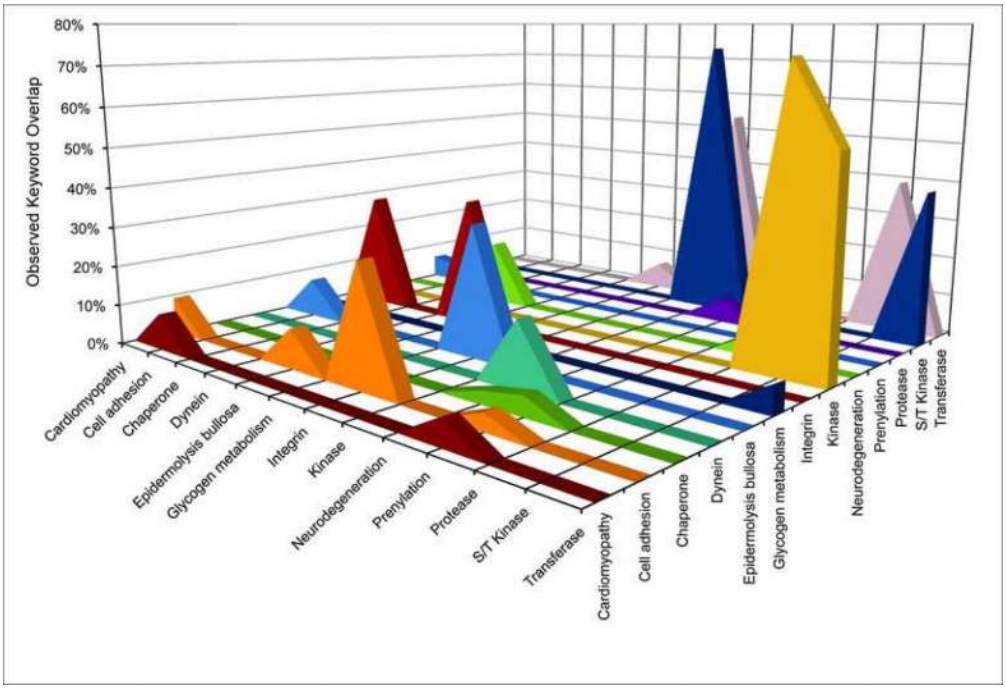

Figure 7. Percentage of overlap between significant keyword terms

Specific keyword categories are selected to determine co-relationships, as an indication of shared function. The matrix lists the same keywords on the X- and Y- axes, and the number of proteins that overlap between categories is shown as a percentage on the vertical Z-axis. The terms shown here were selected as significant, according to tests for normality and statistical significance at $p \leq 0.01$, or according to a mean peptide ratio greater than one standard deviation from the mean at the Experiment Level. This figure shows that Cell Adhesion and Integrin share 30% functional overlap. Also, 70% of the Kinases identified were S/T Kinases, and 50% of the Transferases were classified as Kinases.

Table 1

Six optional post-processing filters control the quality of data used for quantitative analysis.

Filter	Name	Description
1	Scan Cutoff	Minimum number of scans per peptide
2	Light/Heavy Scan Cutoff	Maximum number of scans in which the light and heavy peptides from a pair are not both present
3	Delta Mass Cutoff	Difference in Daltons between heavy and light peptide molecular weights
4	Ratio Cutoff	Minimum ratio of integrated ion current areas between light and heavy peptide pairs
5	Molecular Weight Cutoff	Maximum percentage of error outside of the defined molecular weight range
6	Peptide Sequence Count Cutoff	Minimum number of unique peptides per gene

Table 2

The effect of multiple filter thresholds on peptide and protein counts.

Scan Cutoff		Delta Mass		Ratio		Mol. Weight		PS Count		# Pass	
Filter	Fail	Filter	Fail	Filter	Fail	Filter	Fail	Filter	Fail	PS	Proteins
0	0	None	0	0	0	Infinity	0	1	0	6570	1234
5	59	2	6570	0.01	58	1.5	362	1	0	0	0
5	59	4	619	0.01	58	1.5	362	1	343	5483	806
10	59	4	619	0.05	349	1	695	1	336	5354	795
25	299	4	619	0.1	586	0.5	1578	1	302	4950	742
50	1769	4	619	0.2	1011	0.2	2892	2	607	3246	378
100	5882	4	619	0.5	3125	0	5004	4	214	174	17

This table demonstrates the effects of various filter settings on a representative data set. Each user is encouraged to explore various threshold values to find the appropriate parameters for their data. 15,825 peptides were identified and quantified by X! Tandem and XPRESS in CPAS. Redundant measurements were excluded by requiring only unique peptide sequences (PS) with the highest PeptideProphet score, leaving 6570 peptides for quantitative analysis. The Filter column reports the threshold value and the Fail column reports the number of peptides that were excluded from the analysis using that threshold. The Scan Cutoff, Delta Mass, Ratio and Molecular Weight filters were applied independently and may contain identical peptide sequences that failed more than one filter. The PS (peptide sequence) Count filter was applied last and results are dependent on the first 4 filters. The number of unique peptide sequences (PS) and proteins that passed all five thresholds are shown in the final column.

Table 3

The effect of significance filters on protein and keyword counts.

	Experiment 1	Experiment 2
Total Peptides Used	4976	5046
Total Proteins	720	750
Fail skewness	444	476
Fail kurtosis	150	179
Fail D'Agostino's p-value	37	31
Fail p-value	59	38
Statistically significant proteins	**29**	**26**
Heuristically significant proteins	**114**	**130**
Total Keywords	340	349
Fail skewness	104	108
Fail kurtosis	84	82
Fail D'Agostino's p-value	49	33
Fail p-value	21	73
Statistically significant keywords	**83**	**53**
Heuristically significant keywords	**72**	**28**

Peptides were grouped by gene name and keyword term. This table provides the number of proteins and keywords that failed a series of tests for normality, where it was required that skewness and kurtosis scores were between -1 and 1 and D'Agostino's p-value was above 0.05. Proteins and keywords whose peptide populations passed these normality tests were evaluated for statistical significance based on a one-sample t-test ($p \leq 0.01$). Using a more permissive threshold for heuristic significance, proteins and keywords were selected based on whether their abundance had changed by more than one standard deviation from the mean of the total population of peptides in an experiment.

Table 4

A comparison of results obtained by Annotator versus a conventional method of quantitative analysis.

	False Negatives from the Conventional Method	Validated in the Conventional Method	False Positives from the Conventional Method	
	Selected by Annotator with Statistical significance at $p \leq 0.01$	Selected by both methods	No heuristic significance $\sigma_{Mean} > 1$	Inaccurate selection of peptides
Increased in Abundance	EFTUD2 SSB	ASCC3L1 BAT1 DDOST INPP5D OGFR PNN PRPF	ACBD3 BCAP31 CDC37 DAP3 EIF3G RAN TUBA1A	RPL3 RTN3
Decreased in Abundance	ACTN1 ENO1 KTN1 PDIA3	CIRBP TXNDC12 VCL	EIF5A EZR ITGB1	ENO2

A conventional quantitative analysis calculates average protein abundance from a log-transformation of the direct ratio of heavy and light peptide elution peak areas. Then, an arbitrary cutoff of fold-change ≥ 1 was used to select for proteins that significantly increased or decreased in abundance. In contrast, the analysis using Annotator was performed using filters to exclude peptides that were introduced in error by upstream software. Then, the average protein abundance was calculated from the ratio of heavy- or light-labeled peptides in relation to the sum of both peptide elution peak areas in a labeled pair. In Annotator, proteins were selected as having significantly changed in abundance if they were found to be statistically significant ($p \leq 0.01$) or the average abundance was greater than one standard deviation from the mean of all peptides in an Experiment (heuristic significance, $\sigma_{Mean} > 1$). This comparison between the two methods demonstrates the effectiveness of Annotator in removing false positives and recognizing the statistical significance of false negatives in the conventional method.