Quantitative strategies to fuel the merger of discovery and hypothesis-driven shotgun proteomics

Kelli G. Kline, Greg L. Finney and Christine C. Wu

Abstract
The ultimate goal of most shotgun proteomic pipelines is the discovery of novel biomarkers to direct the development of quantitative diagnostics for the detection and treatment of disease. Differential comparisons of biological samples identify candidate peptides that can serve as proxies of candidate proteins. While these discovery approaches are robust and fairly comprehensive, they have relatively low throughput. When merged with targeted mass spectrometry, this pipeline can fuel hypothesis-driven studies and the development of novel diagnostics and therapeutics.

Keywords: quantitative shotgun proteomics; biomarker discovery; targeted mass spectrometry; human tissue

INTRODUCTION
A common goal of most shotgun proteomic studies is the discovery of novel biomarkers of disease. However, discovery alone is insufficient and proteomic pipelines that streamline the process from discovery to diagnostics are beginning to take shape. This review will focus on several robust quantitative strategies that fuel the merger of discovery and hypothesis-driven shotgun proteomics as it applies to the analysis of human samples. As illustrated in Figure 1, biological samples are first compared using various quantitative platforms [e.g. spectral count comparisons, chemical derivatizations and label-free differential mass spectrometry (MS)]. These discovery approaches serve to identify differentially expressed peptides which may represent candidate biomarkers of disease and/or lead to enhanced insight in the molecular mechanisms of disease. While these discovery approaches are robust and can monitor thousands of proteins in each analysis, they have relatively low throughput. When merged with targeted MS, capable of monitoring hundreds of proteins with high-throughput capacity, this pipeline can serve to drive hypothesis-driven studies, critical for the development of quantitative assays and novel diagnostics and therapeutics.

DISCOVERY-DRIVEN SHOTGUN PROTEOMICS
Shotgun proteomic platforms provide robust alternatives to gel-based approaches for the quantitative analysis of complex protein samples. Because the analyses are performed at the peptide level, the biochemical heterogeneity of intact proteins is dramatically reduced [1]. A popular example of a shotgun proteomic platform is MudPIT (Multi-dimensional Protein Identification Technology), which uses biphasic microcapillary columns typically packed with both strong cation exchange (SCX) and reverse phase (RP) material to resolve protein mixtures that have been digested and loaded directly onto the column [1–4]. The solid phase column packing material allows for the separation of peptides based...
on both charge and hydrophobicity. The column is then placed inline with a tandem mass spectrometer to facilitate electrospray ionization (ESI) followed by mass analysis of the ions as they are eluted off the column using alternating salt pulses and organic mobile gradients. MudPIT allows for the identification of proteins with variable pI, molecular weight and hydrophobicity, making this method largely unbiased.

**SPECTRAL COUNTING**

A simple differential measure of relative protein abundance known as ‘spectral counting’ is commonly used in conjunction with MudPIT analyses (Figure 2). The total number of spectra that identify peptides originating from a given protein shows good linear correlation with the abundance of that protein [5–7]. This observation allows for relative quantitative comparisons of protein levels between samples by summing all the MS/MS spectral observations for all peptides of a given protein, including spectra from the same peptide at multiple charge states. The major analytical caveat to using this approach is that spectral count ratios can be biased by protein size, peptide length and other peptide physiochemical properties that affect MS detection. Therefore, appropriate statistical analyses should be considered [8, 9].

Spectral counting has shown high utility for simple (tandem affinity purified complexes) and moderately complex samples (such as fractionated tissues and yeast lysates) [7, 10, 11]. However, when analyzing extremely complex samples (such as unfractionated human tissue), the low overlap of protein identifications that is observed between samples can significantly limit the depth of quantitative coverage. A study by Kline et al. [12] reported only 51% overlap in protein identifications from the analysis of total unfractionated protein from human cardiac explants. Though this study was practically limited in depth of relative comparisons between samples, it illustrated the use of spectral counts as a metric for peptides that are detectable within a particular biological matrix or sample. Indeed, a list of proteotypic peptides (defined as peptides that are most representative and most likely to be observed for a protein of interest in a particular biological matrix) was generated in this study, and 98.3% of the peptides identified with high spectral counts (>200 spectra) had calculated observation frequencies >0.5 (i.e. the peptide was identified in at least 50% of the MS runs where the corresponding protein was identified). Databases such as these provide a valuable resource for the selection of peptides which may serve as quantitative proxies, representative for proteins of interest in a targeted MS analysis (further discussed in ‘Targeted MS’ section).

The ability to accurately quantify proteins in an extremely complex mixture by spectral counting largely depends on the number of spectra obtained and the coverage of sampling. This presents a large obstacle for those proteins present at low abundance, such as integral membrane proteins, which may never be selected for data-dependent acquisition due to limitations associated with depth of sampling of complex mixtures. A common solution to these limitations is sample fractionation to reduce complexity and increase the depth of sampling in
MudPIT experiments. However, fractionation increases the overall number of samples to be analyzed and compared, drastically increasing the amount of MS time required. Furthermore, comprehensive coverage by MudPIT analyses of unfractinated complex samples typically requires multiple technical replicates from each sample to achieve saturation of unique protein identifications and obtain good statistical analyses for spectral count comparisons [12]. This requires significant commitments of instrument and data processing time. Furthermore, sample fractionation requires significant amounts of starting material, which is often incompatible with human samples and thereby limits the practical utility of spectral counting for human studies.

**STABLE ISOTOPE LABELING STRATEGIES**

A common strategy to increase the quantitative precision of a proteomic analysis is the incorporation of stable isotope-labeled internal standards. There are two general approaches to incorporate stable isotopes into proteins, metabolic labeling and chemical derivatization. The method of stable isotope metabolic labeling has successfully been used in both cell culture and mammals [13–15]. However, given the practical limitations of this approach in the analysis of human samples, it will not be discussed further in this review. Chemical derivatization platforms utilize various isotopically labeled tags or reagents (Review available [16]) [17–20]. Two popular tagging methods are discussed below.

In the ICAT (isotope-coded affinity tag) method, proteins from two different biological samples are labeled with either an isotopically ‘heavy’ (typically deuterium or $^{13}$C) or isotopically ‘light’ (native) ethylene glycol linker with a biotin affinity tag and a thiol-specific reactive group that selectively couples to the side chain of a reduced cysteine residue (Figure 3) [19]. Following covalent modification of all the cysteine residues in the samples with either a ‘heavy’ or ‘light’ isotope tag, the samples are mixed together, digested with protease and incubated on an avidin column to allow for enrichment of labeled peptides by the biotin moiety located on the isotope tag [21]. Following MS analysis, the relative abundance of peptides is determined by the ratio of the signal intensities from the ‘heavy’ and ‘light’ forms of each peptide. Individual peptide ratios from the same protein are then combined to produce abundance ratios of identified proteins in the sample. A major advantage of this tagging system is that it facilitates the enrichment of the modified peptides via affinity purification of the biotin moiety, thereby enhancing the detection of low-abundance proteins. However, because ICAT reagents selectively label proteins/peptides containing cysteine residues, those proteins which do not contain a cysteine residue will not be quantified using this method.

![Diagram](image-url)
More recently, the iTRAQ method was developed, which solved some of the limitations of ICAT. The iTRAQ method uses amine-specific isobaric reagents to label the primary amines of peptides for the concurrent quantitative comparison of up to eight different biological samples (Figure 4) [22, 23]. Labeled peptides from the different samples are mixed together to collectively undergo mass analysis. The MS spectrum of each peptide in the sample does not increase in complexity due to the isobaric nature of the tagging reagents, as compared to the ICAT method which results in multiple MS peaks per peptide. During peptide fragmentation, the isobaric amine groups are also fragmented to release reporter ions with distinct $m/z$'s (e.g. 114.1, 115.1, 116.1 and 117.1 in the 4-plex reaction shown in Figure 4). Relative peptide abundance measurements between the four samples are then made based on the relative intensity of each reporter ion. Similar to the ICAT method, multiple peptide measurements can then be combined to result in relative abundance measurements at the protein level. The iTRAQ method has recently been used to monitor small molecule binding to proteins and map drug-induced changes in protein phosphorylation states in human cells, illustrating this method’s potential use in drug discovery experiments [24]. However, although iTRAQ labeling allows for incorporation of labels into all peptides in a mixture (in contrast to ICAT), this inclusive feature inherently requires that the labeling be done after protease digestion, thereby limiting their use as internal standards because experimental variations at prior sample preparation steps are not accounted for.

**DIFFERENTIAL MS QUANTITATION**

Recently, label-free approaches (collectively referred to as differential MS or dMS) have been developed, which utilize computational methods to find differences in the intensity of peptides in the MS signals, rather than the total number of MS/MS spectra that were stochastically sampled. These approaches rely on the comparison of distinct regions of $m/z$ and retention time which represent the same analyte across multiple analyses, and are based on the concept of comparing samples via MS ‘images’ to detect feature differences after computational processing. In general, these computational methods can be divided into two groups. In the first group, spectral features (typically isotopic clusters indicating a peptide) are detected first and then statistically compared [25–29]. Alternatively, the second group applies statistics first so that only differential features are detected from the extracted ion chromatograms [30–33]. Both computational methods rely on the ability to chromatographically align multiple LC-MS runs from different samples, identify differences between the samples and determine the magnitude of the change (Figure 5). The absence of internal standards requires the careful use of
experimental and computational methods to normalize and correct experimental variances which affect retention time and MS intensity.

Regardless of the approach used, chromatographic alignment is an essential component to dMS analyses and also minimizes the chromatographic error so that signals measured at a given time and m/z in the mass spectrometer from different runs can be assumed to originate from the same molecular species. This can be addressed computationally by two classes of algorithms. The first class uses full scan MS data to ‘align’ the retention times of one run to another. This alignment is performed by algorithms such as dynamic time warping or correlation optimized warping which find an optimal mapping of retention times between runs that maximizes their similarity [34–36]. The second class matches detected features (such as peptide isotope...
distributions or MS/MS spectra) between runs, and applies algorithms such as regression to fit a time correction function to the matched markers [29, 37, 38]. Additionally, variations in signal intensity can be addressed by normalization. Callister et al. [39] assessed multiple normalization techniques for the label-free analysis of LC-FTMS data, and Wang et al. [40] use global normalization ratios while adjusting for low-intensity signals missing in some runs. Methods for the statistical discovery of differences in label-free MS data have been reviewed recently [41, 42]. The identification of peptide isotope distributions, or regions of $m/z$ and RT, which significantly differ between samples have been implemented in numerous software packages, and summarized in multiple reviews [42–44]. Figure 6A shows an example of base peak chromatograms from two human heart biopsy samples following chromatographic alignment and normalization. Differences are then detected by comparing peptide maps of two mean, aligned human samples based on their $m/z$ intensities and chromatographic retention time (obvious visual differences highlighted by black boxes in Figure 6B).

Importantly, label-free MS quantitation requires reproducible sample preparation and chromatographic separations, which can be difficult when working with extremely complex samples such as human tissues. Also, because differences between samples are identified from the MS scans, the mapping of differentially expressed features to peptide identifications continues to be limited by the scan speed of the mass analyzer, causing only a fraction of identified features to be mapped back to peptide identifications by MS/MS. To by-pass this obstacle, ‘inclusion’ lists can be generated to modify

---

**Figure 5:** Relative quantitation using dMS. Samples are digested and chromatographically separated. Samples are then sequentially analyzed using LC-MS/MS. The collected data are then chromatographically aligned to allow for peak or feature identification that differ between the samples. Differential features are then mapped back to MS/MS spectra to obtain peptide identifications.
subsequent data-dependant MS/MS analyses to focus instrument time on the identification of previously unidentified ions that gave rise to significant differences [45].

Universal to all the quantitative platforms described above is the preferential identification of high-abundance proteins in a complex sample, with lower sampling of low-abundance protein species. In the case of the isotope tagging strategies, quantitative information is further restricted to peptides that contain the chemical derivatization. While many differentially expressed features or regions may be identified using dMS, only a subset of those may be linked to peptide identifications by MS/MS. For the analysis of highly complex samples, such as unfractionated human tissue, these techniques are invaluable for discovering differentially expressed peptides (protein candidates), but remain limited in utility for the targeted quantitation of peptides (especially those derived from proteins of low abundance) in a high-throughput manner.

**TARGETED MS**

The necessity for highly sensitive and selective detection strategies for analytes in complex human samples was recognized early by researchers interested in the detection of small molecule analytes, such as drug metabolites and hormones. In fact, the development of selective reaction monitoring (SRM) was largely pioneered by those interested in detecting these types of compounds in highly complex and typically unfractionated samples [46–51]. The use of SRM for the robust detection of a multitude of compounds has been reported in human plasma [49, 52–54] and other complex human samples [55, 56]. Indeed, the use of isotope-labeled internal standards using SRM is well established for the robust and high-throughput analysis of metabolites in complex human samples. Only more recently has the proteomics field begun using the SRM technique as a selective way to detect and quantitatively monitor unique peptides from proteins of interest [57–60]. This shift in proteomics to targeted MS...
methods was largely based upon the difficulty in the detection of low-abundance proteins within the context of unfractionated complex samples such as human tissue and plasma. In fact, the accurate quantitation of peptides and proteins in human plasma is incredibly challenging due to the extreme dynamic range of $\sim 10^9$ [61]. Furthermore, monitoring low-abundance proteins in human tissue biopsies, such as the heart, which combines the complexity of a heterogeneous tissue with blood contamination, can be a daunting task.

Targeted MS (LC-MS/MS in the SRM mode) approaches require prior knowledge of the analyte to be detected. However, when coupled with the discovery methods described above, targeted MS provides the sensitivity, selectivity and throughput required for the analysis of human clinical samples. Most targeted MS techniques utilize the discriminating power of quadrupole mass analyzers to select specific peptides in a complex sample mixture for detection. Peptide SRM uses ESI followed by two phases of mass selection. The first phase (in Q1) selects for the $m/z$ of the precursor ion (charged unique peptide). Following fragmentation of the selected peptide by collision induced dissociation (CID in q2), the second phase (in Q3) selects for the $m/z$ of fragment ion derived from the precursor ion (Figure 7).

The use of two mass filters in the triple quadrupole allows for the peptide analyte to be selectively detected, even if it is present in very low abundance in an unfractionated complex mixture. This level of mass selection increases sensitivity by only allowing the transmission of a small population of ions, thus minimizing chemical background noise [62]. Additionally, SRM measurements are easily multiplexed [63], due to the rapid duty cycle of current instrumentation (10–100 ms), facilitating hundreds of peptides to be monitored on a chromatographic time scale in a single MS run. Because SRM assays are typically optimized for peptides separated using single phase chromatography, analyses are high-throughput and rapid with minimal sample consumption (0.5–5 µg peptides/analysis) and therefore, well suited to accommodate the large sample numbers of clinical studies.

SRM data analysis involves the comparison of the area under the curve (AUC) from extracted ion chromatograms for each transition monitored. Relative changes between multiple samples can be made without the inclusion of internal standards. However, the incorporation of stable isotope-labeled internal standards allows for high precision quantitation of each unique peptide [57, 59, 60, 64]. Because the amount of internal standard peptide is known and calibrated for the linear quantitative range, the ratio between the areas under the extracted ion chromatograms of endogenous peptide and isotope-labeled peptide allows for the calculation of the absolute amount of the endogenous peptide [57]. Quantitation of proteins using this method is based on the detection of one or more peptides from that protein [65]. In contrast to global MS quantitation methods, targeted MS requires prior knowledge of the peptides (proteins) selected for quantitation. This distinction, along with the throughput and sensitivity, makes it ideal for hypothesis-driven queries, and for the verification and validation of those protein candidates identified using global MS analyses.

However, obstacles may arise due to the fact that specific unique peptides are used as quantitative proxies for the protein from which they are cleaved to yield a measure of the protein abundance. While this can be advantageous in many ways because it allows for separate detection and quantitation of multiple peptides from a single protein to provide information concerning protein abundance, or even potentially varied modification states [57, 64], it can also be limiting if the detection of multiple peptides is difficult and/or modified peptide isoforms are below the limits of detection or quantitation of each unique peptide [57, 59, 60, 64]. Because internal standards allows for high precision quantitation of proteins using this method is based on the detection of one or more peptides from that protein [65]. In contrast to global MS quantitation methods, targeted MS requires prior knowledge of the peptides (proteins) selected for quantitation. This distinction, along with the throughput and sensitivity, makes it ideal for hypothesis-driven queries, and for the verification and validation of those protein candidates identified using global MS analyses.
CONCLUSION

The proteomics field has recognized the need for quantitative methods that are compatible with the analysis of complex biological samples, particularly human samples such as tissue and plasma. However, proteomic analysis of unfractionated complex human samples is difficult due to the large dynamic range of protein expression. Currently there are a number of robust discovery platforms that are used in the field to identify differentially expressed proteins in human samples, either utilizing isotope-tags or label-free methods. When merged with targeted MS platforms, capable of monitoring hundreds of proteins with high-throughput capacity,
this pipeline (illustrated in Figure 8) can serve to drive hypothesis-driven studies, critical for the development of novel diagnostics and therapeutics.

### Key Points

- Major quantitative shotgun proteomic strategies for discovery are summarized.
- Targeted proteomics using LC-MS/MS in the SRM mode is discussed.
- An experimental pipeline merging discovery and targeted analyses is proposed for high-throughput clinical studies.

### References


47. Barbarin N, Mawhinney DB, Black R, Henion J. High-throughput selected reaction monitoring liquid
chromatography-mass spectrometry determination of methylphenidate and its major metabolite, ritalinic acid, in rat plasma employing monolithic columns. J Chromatogr B


