Title:

PEPPeR: A Platform for Experimental Proteomic Pattern Recognition*

Authors:

Jacob. D. Jaffe1
D. R. Mani1
Kyriacos C. Leptos2
George M. Church2
Michael A. Gillette1
Steven A. Carr1†

Affiliations:

1. The Broad Institute of Harvard and MIT, Cambridge, MA 02142
2. Harvard Medical School, Department of Genetics, Boston, MA 02115
† corresponding author

Running title:

PEPPeR: Proteomic Pattern Recognition and Relative Quantitation
Summary

Quantitative proteomics holds considerable promise for elucidation of basic biology and for clinical biomarker discovery. However, it has been difficult to fulfill this promise due to over-reliance on identification-based quantitative methods and problems associated with chromatographic separation reproducibility. Here we describe new algorithms termed ‘Landmark Matching’ and ‘Peak Matching’ that greatly reduce these problems. Landmark Matching performs time base-independent propagation of peptide identities onto accurate mass LC-MS features in a way that leverages historical data derived from disparate data acquisition strategies. Peak Matching builds upon Landmark Matching by recognizing identical molecular species across multiple LC-MS experiments in an identity-independent fashion by clustering. We have bundled these algorithms together with other algorithms, data acquisition strategies, and experimental designs to create a Platform for Experimental Proteomic Pattern Recognition (PEPPeR). These developments enable use of established statistical tools previously limited to microarray analysis for treatment of proteomics data. We demonstrate that the proposed platform can be calibrated across 2.5 orders of magnitude and can perform robust quantification of ratios in both simple and complex mixtures with good precision and error characteristics across multiple sample preparations. We also demonstrate de novo marker discovery based on statistical significance of unidentified accurate mass components that changed between two mixtures. These markers were subsequently identified by accurate mass driven MS/MS acquisition and demonstrated to be contaminant proteins associated with known proteins whose concentrations were
designed to change between the two mixtures. These results have provided a real-world validation of the platform for marker discovery.
Introduction

There is tremendous interest in the use of mass spectrometry as a quantitative technology to measure peptide and protein abundances for comprehensive, system-wide biological research (1, 2). Quantitative proteomics may be used to systematically identify and quantify proteins and their modifications as a function of cell cycle, differentiation, or chemical treatment to obtain novel insights into basic cellular biology. Proteomics also holds promise for discovery of proteins in readily accessible biofluids that are diagnostic or prognostic of a disease condition. Such proteins are termed ‘biomarkers’.

The need for robust methods to obtain relative quantification is particularly acute in proteomic-based biomarker discovery where comparative data across multiple patient samples should be obtained (3). Biomarker discovery usually employs biofluids that greatly increase the magnitude of the challenge for quantitative proteomics due to the very high dynamic range of protein abundance ($\approx 10^{12}$ for blood) and the enormous diversity of proteins present in such samples (4).

Currently available MS platforms for quantitative proteomics fall roughly into three categories: (1) identity-based methods that rely on proteolytic digestion of proteins to peptides with analysis by LC-MS/MS (5-8); (2) pattern-only methods that focus on production of MS-derived protein pattern that are more useful for sample classification than protein quantification (9-12); and (3) hybrid identity/pattern-based methods using
peptide-derived LC-MS data from FTMS or Orbitrap mass spectrometers with very high resolution and mass accuracy (13, 14).

Identity-based approaches rely on sequencing peptides by data-dependent LC-MS/MS and identifying the proteins by database searching (15-19). A variety of chemical tagging and metabolic labeling methods for differentially labeling peptides with stable isotopic labels such as ICAT, SILAC, and iTRAQ have been developed to facilitate obtaining relative quantitative data from a limited number of samples in LC-MS/MS experiments (1, 2). These approaches use extracted ion currents for identified peptides to compute chromatographic abundances at the MS or MS/MS level. Alternatively, semi-quantitative data may be obtained directly from the MS/MS data by spectrum counting (20).

However, a significant problem for these identity-based methods is the limited and stochastic nature of MS/MS sampling for peptide sequencing on the chromatographic time scale. This is exacerbated in complex samples where the MS/MS sampling process favors acquisition of spectra for the peptides of highest abundance that represent only a fraction of the detectable m/z peaks in the mass spectrum. This means that a considerable amount of useable data is discarded by data-dependent experiments and only the most abundant peptides (and therefore proteins) in these experiments will be reliably quantified. Lower abundance proteins will be sampled for MS/MS at lower frequency, resulting in poor reproducibility across samples (21). Researchers counter this effect by either analyzing the same sample multiple times (21), or by fractionating highly complex samples at the protein and peptide level prior to the final LC-MS/MS analysis (22). While these approaches often
improve detection of lower abundance components, they also greatly decrease throughput.

Pattern-only approaches directly use raw m/z values or centroided peaks together with intensity information to define a mass spectrometric pattern from MALDI or ES data. In LC-MS, retention time is also used as a coordinate in uniquely locating peaks so as to disambiguate peptides with similar mass eluting at different times, although this dimension introduces a strict requirement for highly reproducible chromatography or complex methods to account for retention time variation. The extent of sample fractionation used in pattern-based studies has generally been very limited (12, 13). As a result, pattern-based methods can achieve higher sample throughput than identity-based approaches, thereby enabling the analysis of larger numbers of patient samples for a given study. Furthermore, pattern-based biomarker discovery usually employs powerful multivariate pattern recognition methods (23, 24). However, identifying the peptides and proteins that constitute the pattern remains essential but is often difficult or impossible using these methods. Knowledge of the identity of the peptides and proteins constituting the pattern increases confidence in the robustness of the assay, provides biological insight into disease pathogenesis, suggests therapeutic targets, and creates the opportunity for transfer of the assay to an alternative technology platform (e.g., enzyme-linked immunosorbent assays (ELISAs)). The latter is especially important, as the ability to utilize an MS platform in a clinical chemistry setting for MS-pattern based diagnostics remains untested (3).

The hybrid identity/pattern approach was introduced by Richard Smith’s group (for a recent review, see (13)). This group demonstrated the power of using high mass
accuracy data from FTICR-MS for high throughput pattern-based analysis (25, 26). The accurate mass and time tag (AMT) strategy developed by this group exploits the fact that multiple peptide species are highly unlikely to have both the same mass (to within a few ppm) and LC retention time (especially in relatively simple genomes). Using their method, databases of peptide sequence, mass, and retention time can be built from multiple experiments and searched for ex post facto assignment of identity to a chromatographic LC-MS peak based on these parameters. Abundance of the peaks enables large-scale relative quantification of peptides among groups of samples. This in turn enables statistical analysis for biological characterization or biomarker discovery. Other approaches that exploit high-resolution pattern have recently been described as well (27-29), including a hybrid MS, MS/MS deconvolution method developed by Silva et al. (30, 31).

Several key problems face practitioners of quantitative proteomics. Label-free approaches such as Smith's do not necessarily rely on direct MS/MS sequencing for quantification, but then the issues of chromatographic reproducibility and alignment become manifest. Many approaches such as dynamic time warping have been suggested as possible remedies to these problems but none has gained wide acceptance (12, 32). Finally, in approaches that do not rely on sequencing for quantification, feature definition and consistent recognition across multiple experiments becomes challenging. Features are usually defined based on parameters such as m/z and retention time, but being able to match peaks with the same identity in large data sets of many LC-MS runs is difficult even when chromatography is highly reproducible.
Here we describe a novel pattern-based biomarker discovery system (PEPPeR) where peptide identity is used to guide the alignment and analysis of MS pattern. The method leverages high mass accuracy, high resolution LC-MS (and MS/MS) data produced by hybrid, ultra-high performance FTICR or Orbitrap mass spectrometers together with intelligent data acquisition strategies and newly developed algorithms to address the problem of quantifying a high percentage of the molecular species present in a sample without strict regard to their identities. We utilize limited MS/MS sequencing during quantification experiments to establish chromatographic landmarks that can subsequently be used for time-base independent peptide identity alignment. This approach allows (but does not require) large historical databases of peptide identifications done during discovery phase experiments to be mapped onto higher throughput quantification experiments. These landmarks also serve to calibrate peak matching via a Gaussian mixture model which allows consistent recognition of both previously identified and novel, unidentified features across multiple LC-MS experiments. We show that label-free quantification is robust across 2.5 orders of magnitude in both simple and complex mixtures using our platform for both calibrating measurements and calculating ratios. Robust ratio calculation is subsequently used to recognize novel peptides that are indicative of a change of biological state; that is to say, biomarkers. Importantly, the methods we have developed do not rely on highly reproducible or predictable LC retention time for peptides. PEPPeR enables recovery of a significant amount of currently unexploited quantitative information in the MS data obtained from these instruments, while greatly increasing the accuracy and efficiency of obtaining the sequence identities of even minor m/z peaks found to change across
samples. Finally, it will allow for use of established statistical tools on proteomics data for biomarker discovery and provides a clear reverse path for identification of unknown molecular species via accurate mass-driven, targeted MS/MS experiments (33).
Experimental Procedures

Reagents and Chemicals

Proteins were obtained from Sigma-Aldrich (St. Louis, MO). Mitochondrial extracts were kindly provided by Dr. Vamsi Mootha of the Broad Institute, Cambridge, MA. All other reagents (including water) were of HPLC or Proteomics grade.

Preparation of Protein Mixtures and Digests

Two sets of protein mixtures (‘Scale Mixes’ and ‘Variability Mixes’) were prepared by serial dilution with water from 1 nmol/μl stocks in water (Table 1) to form new substocks. Each substock was diluted 1:1 with 6M urea/50 mM Tris pH 8.0 and reduced with 10 mM DTT for 30 min at 37 °C. Cysteines were subsequently alkylated with 50 mM iodoacetamide for 30 min at 37 °C in the dark. Each substock was further diluted to 1/10th its original volume with water. Trypsin (Roche Proteomics Grade) was added at a mass ratio of 1:50 trypsin to total protein. Digestion proceeded at 37 °C for 18 hours with shaking at 600 RPM. 10 μL of formic acid was added to terminate digestion and substocks were subsequently frozen at -80 °C. In the case of the Variability Mixes, aliquots of mix α and β were distributed to 5 different members of our
laboratory for independent digest preparation. This was done to provide a measure of digest variability.

Substocks were desalted with a SepPak 100 mg tC18 solid phase extraction cartridge (Waters) using the following procedure: wet with 2 x 1 mL ACN, equilibrate 2 x 1 mL of 0.1% formic acid, load 1mL substock, wash 3 x 1 mL 0.1% formic acid, elute 1.5 mL of 70% ACN / 0.1% formic acid. Eluates were frozen at -80 °C, lyophilized to dryness with a vacuum concentrator, and subsequently resuspended at concentrations 100-fold higher than those stated in Table 1.

**Preparation of Mitochondrial Protein Extracts**

Preparation of mitochondrial extracts from C57BL6/J mice aged either 2 or 6 weeks was as previously described (34). The 2 week extract was spiked with Variability Mix α and the 6 week extract was spiked with Variability Mix β prior to digestion such that the final concentrations in the analyzed samples would be the same as those shown in Table 1. Digestion proceeded as above, although samples were desalted using an OASIS HLB 10 mg cartridge in this case.
The PEPPeR Pipeline

The complete PEPPeR pipeline consists of experimental design considerations, data acquisition strategies, and computational analysis. A flow chart illustrating the pipeline is shown in Figure 1. The components of the work flow are discussed in detail below. Capitalized terms are defined in Box 1.

LC-MS Procedures

Samples were subjected to two types of LC-MS analysis that differed in purpose. One strategy was to comprehensively identify as many unique peptides as possible by using a relatively higher number of data-dependent MS/MS scans for every precursor MS scan (‘Identification Experiment’). For this strategy, each substock was diluted 10-fold prior to analysis. The other strategy was to optimize quantitative information by increasing the frequency of precursor MS scans and performing more replicated analyses (‘Quantification Experiment’). Substocks were diluted 100-fold for this purpose (final concentrations shown in Table 1).

LC parameters were common to both strategies. Chromatography was performed using an Agilent 1100 nano-flow chromatograph (Agilent; Palo Alto, CA) with Buffer A 0.1% formic acid and Buffer B 90% ACN / 0.1% formic acid. A PicoFrit column (75 µm i.d., 15 nm tip opening; New Objective; Woburn, MA) was packed with 12.5 cm of ReproSil-Pur C18-AQ 3-µm resin (Dr. Maisch GmbH) and directly interfaced to an LTQ-FT mass spectrometer fitted with a nano-electrospray ionization source (ThermoElectron; Waltham, MA). 1 µl of sample was injected for each analysis and the
following gradient was used: 0-20 min 3% B at 600 nl/min, 20-30 min 3-15% B at 200 nl/min, 30-80 min 15-45% B at 200 nl/min, 80-85 min 45-90% B at 200 nl/min, followed by normal regeneration and re-equilibration procedures.

MS analysis parameters for the Identification Experiments were as follows. 1 precursor MS scan (FTMS, resolution 100,000) was followed by data-dependent MS/MS scans of the top 10 most abundant ions executed in reverse order (ITMS). Dynamic exclusion was enabled with a repeat count of 1 and an exclusion duration of 40 s with the maximum possible list size and a +/- 25 ppm rejection window. Charge state screening and monoisotopic precursor selection were both enabled. Nano-ESI voltage was 2.1 kV and the precursor MS scan target value was set to 5 x 10⁵ ions to minimize harmonic noise (target value raised to 1 x 10⁶ for mitochondrial samples). Parameters for the Quantification Experiment acquisition strategy were identical except that the top 3 most abundant ions were selected rather than 10. For Identification Experiments, each sample was subjected to LC-MS analysis once. For Quantification Experiments, each sample was analyzed with 5 technical replicates.

**Peptide Spectrum Interpretation**

MS/MS spectra were extracted from the raw data and interpreted using SpectrumMill Data Extractor and MS/MS Search Rev B.03.02.059 (Agilent; Palo Alto, CA). SpectrumMill parameters for extraction, searching, and autovalidation can be seen in the accompanying supplemental information. Data from the Scale Mixes and Variability Mixes were searched against a small protein database consisting of only those proteins that composed the mixtures and common contaminants (52 proteins
total). Subsequently, novel feature spectra acquired by targeted means were also searched against the NCBI non-redundant protein database dated August 1, 2005 and containing 2724841 entries. Data from the mitochondrial preparations were searched against the IPI mouse database v. 3.01 (35) and the small database mentioned above.

**Peptide LC-MS Feature Detection**

Peptide features were extracted and deisotoped from the raw data using MapQuant, a program that uses image processing techniques to identify and quantify organic species present in LC-MS runs (36). The MapQuant processing script used is provided as supplemental information. MapQuant processing and the subsequent algorithmic steps described below were carried out on a 464 processor Beowulf Linux cluster to take advantage of parallel processing opportunities. Typical combined run time for all algorithms using our hardware configuration is less than one day.

**Propagation of Peptide Identities Across Multiple LC-MS Experiments**

*(Landmark Matching)*

Landmark matching attempts to propagate peptide identifications across multiple LC-MS runs using a combination of accurate mass measurement and relative retention time information. Terminology used in this section is defined in Box 1 and capitalized for easy reference. The computer programs that implement landmark matching are written in Perl and are available at http://www.broad.mit.edu/software.
Peptides sequenced from related LC-MS experiments become part of the BASIS SET. Information about confidently identified peptides is retained in the BASIS SET, namely: peptide sequence, charge in which it was observed, experiment in which it was observed, and scan boundaries composing the MS/MS scans grouped together for sequence identification. These scan boundaries become the basis for absolute or relative retention time comparisons in later landmark matching. This information is easily culled from the spectrum filename for spectrum extractor outputs that follow the same file naming convention as extract_msn (ThermoElectron). In the case of our experiments, an example filename VARMIX_A_01.3645.3675.2.pkl would translate to experiment VARMIX_A_01, scan boundaries 3645-3675, charge state z=2. If a peptide was confidently identified from this spectrum, its sequence and spectrum name would become part of the BASIS SET. In cases where absolute retention time is required, scan numbers can be translated into time units using software libraries provided by the instrument manufacturer or ‘looked-up’ in the RT ruler generated by MapQuant during xr2or data extraction (36). These data are stored in a simple text file that can be parsed by subsequent programs in the pipeline. This format (‘goodouts’) is described in the Supplementary Information and an example parser for SpectrumMill results is provided. An XML format will be available in the future.

Landmark matching is an sequential process. Limited MS/MS data acquisition during LC-MS experiments is used to confidently identify peptides which can then be used as registration marks with other experiments. First, peptides sequenced during the CURRENT EXPERIMENT are mapped onto features identified by MapQuant in that experiment using a loose m/z matching TOLERANCE (+/- 25 ppm) and an absolute
retention time RADIUS (typically 0.3 min). Second, an m/z recalibration is calculated using a least-squares quadratic fit based on these preliminary matches and a more stringent TOLERANCE is computed based on the distribution of residual m/z errors after recalibration (+/- 3σ, typically < 5 ppm). Third, peptides are re-mapped onto the features from the experiment using the new m/z calibration and stringent m/z TOLERANCE. These become LANDMARKS for the single experiment. Finally, peptides observed in any related experiment are mapped onto features in the experiment under consideration using a relative retention time heuristic, the LANDMARK SCORE.

Algorithms for each type of matching exercise are shown in Box 2.

A walk-through of LANDMARK LIST selection and scoring is shown in Fig. 2. A LANDMARK LIST is selected via a common overlap between peptides observed in the CURRENT EXPERIMENT (the LANDMARKS) and some other experiment in the BASIS SET (the COMPARISON EXPERIMENT) with the requirement that the sequence of the PUTATIVE ASSIGNMENT made to the feature being examined was confidently identified in the COMPARISON EXPERIMENT. In this way, each PUTATIVE ASSIGNMENT has its own unique LANDMARK LIST and multiple LANDMARK LISTS may be possible, but the 'best' LANDMARK LIST is derived from a single COMPARISON EXPERIMENT. Currently, the COMPARISON EXPERIMENT is selected where: A) the standard deviation of the centroided scans leading to the confident identification of the PUTATIVE ASSIGNMENT in the COMPARISON EXPERIMENT is < a constant κ (typically, κ = 200 scans indicates a sharply eluting peak for our acquisitions methods), B) there is at least one confidently identified peptide in common between the two experiments both before and after the observation of the PUTATIVE ASSIGNMENT, C) the standard deviation of the centroided scans leading to the
confident identification of the peptides before and after the PUTATIVE ASSIGNMENT in the COMPARISON EXPERIMENT is < a constant $\omega$ (typically, $\omega = 500$ scans leads to acceptable performance), and D) the two experiments share the greatest common overlap of confidently identified peptides. Using these criteria, COMPARISON EXPERIMENT 2 would be selected in the example shown in the left panel of Fig. 2.

After the best COMPARISON EXPERIMENT is selected, the LANDMARK SCORE is computed according to the following heuristic:

Let:
- $\Lambda$ be a list of peptides observed in the COMPARISON EXPERIMENT ordered by elution time. Here, elution time is defined by the centroid of all MS/MS scans leading to the identification of the peptide. $\Lambda_0$ is defined as the position of the putative assignment in $\Lambda$
- $\mu(x)$ be the centroid of elution time of peptide $x$ in the comparison experiment (in scans)
- $\sigma(x)$ be the standard deviation of elution time of peptide $x$ in the comparison experiment (in scans)
- $\tau(x)$ be the centroid of elution time of peptide $x$ in the current experiment (in seconds)
- $\delta$ be the average retention time peak width, such that peptides eluting within $\delta$ sec are considered to be co-eluting (typically $\delta = 30$ s)
- $w$ the number of peptides to consider before and after the putative assignment on the landmark list (typically $w = 3$)

Then, the landmark score $S$ (range $-w..w$) is defined as:

$$S = \sum_{i=1}^{w} \left[ \xi(\Lambda_{-i}, \Lambda_0) + \xi(\Lambda_0, \Lambda_i) \right]$$

$$\xi(m,n) = \begin{cases} 1 & \text{if } \tau(m) < \tau(n) \\ 0.5 & \text{if } \tau(n) - \tau(m) < \delta \text{ and } \mu(m) + \sigma(m) > \mu(n) - \sigma(n) \\ -1 & \text{else} \\ 0 & \text{else} \end{cases}$$
A passing score (\text{THRESHOLD}) may be set empirically by evaluation of matching probabilities (see below). Our current \text{THRESHOLD} is set to $s = 2$. Use of this threshold is also supported by an alternative regression-based landmark matching system that we have also implemented. A example of the scoring system is shown in the right panel of Fig. 2.

All \text{PUTATIVE ASSIGNMENTS} that pass the landmark test \text{THRESHOLD} are then reported. A match is reported as a peptide sequence with corresponding \text{m/z}, retention time, and abundance information as derived from the corresponding \text{MapQuant features}. Processing an entire experiment results in a list of such matches. Any directly sequenced features for that experiment (\text{LANDMARKS}) that were not included in the final match list are merged back into the data set. This final process also provides an estimate of the false negative rate for landmark matching.

\textit{Probabilistic Evaluation of Landmark Matches by Bootstrapping}

The probability that a given match is by chance can be evaluated using the combined probabilities of that the \text{m/z} assignment is by chance and that passing the landmark threshold filter is by chance. Both probabilities are dependent on the assumptions that identifications in the BASIS SET are correct and that the BASIS SET is complete. A p-value can be computed as follows:

$$P_{\text{overall}} = P_{\text{m/z}}P_{(\text{landmark}\mid\text{m/z})}$$
\( P_{m/z} \) can be computed analytically by counting the total number of BASIS SET peptides that fall within the tolerance of the matched m/z at the specified z and dividing by the total number of BASIS SET peptides. \( P_{(\text{landmark} \mid m/z)} \) can be computed according to Bayes' Rule:

\[
P(\text{landmark} \mid m/z) = \frac{P(m/z \mid \text{landmark}) P(\text{landmark})}{P(m/z \mid \text{landmark}) P(\text{landmark}) + (P(m/z \mid \sim \text{landmark})(1 - P(\text{landmark}))}
\]

(Eq. 3)

Because these probabilities can not all be calculated empirically they are assigned by BOOTSTRAPPING (100 BOOTSTRAPS). \( P(\text{landmark}) \) is estimated by selecting 100 FEATURES at random, assigning their retention time to the PUTATIVE ASSIGNMENT, and determining what fraction pass the landmark test using the same LANDMARK LIST that was selected for the PUTATIVE ASSIGNMENT. \( P(m/z \mid \text{landmark}) \) and \( P(m/z \mid \sim \text{landmark}) \) are calculated by determining what fraction of these random FEATURES fall within the m/z TOLERANCE of the PUTATIVE ASSIGNMENT. The overall probability is corrected for multiple hypothesis testing by multiplying by the number of peptides in the BASIS SET that fall within m/z TOLERANCE of the putative assignment.

Global Peak Matching of LC-MS Features Across Multiple LC-MS Experiments
Coarse m/z and RT Corrections

Once landmark matching has been performed on all the runs in an experiment set, we use these resulting landmarked peptides to determine coarse mass and retention time corrections. In preparation for peak matching and pattern recognition, mass and retention time corrections are applied to all charge identified peaks in the MapQuant output irrespective of whether the peak is a landmark or not. Masses are corrected using a least-squares quadratic fit, as above.

Because absolute retention time is difficult to reproduce, we apply a coarse retention time correction to all runs prior to clustering. The aim is not to achieve perfect chromatographic alignment, but merely to improve the efficiency of clustering and derive appropriate RT tolerances. We start by selecting an arbitrary run as the reference. For each of the remaining runs, we correct retention time as follows: Identify the common set of landmarks in the reference run and the run under consideration. Using these landmarks, the corrected retention time for all (landmark or unidentified) peptides is calculated as \[ R_{\text{corrected}} = a_0 + a_1 \times RT + a_2 \times RT^2 \], where the constants are estimated by comparing the retention time of common landmarks in the run to their retention times in the reference.

In both mass and retention time correction, there may be landmarks that have unusually large variation in mass or retention time respectively. Mass outliers can result from potentially incorrect peptide identification, artifacts arising from MapQuant peak detection, erroneous landmark matching, or stochastic MS variation. Retention time outliers can be caused by chromatographic variation, occasional elution of a peptide
over multiple time points or the gradual elution of a peptide over a long time period (i.e., a very broad chromatographic base peak). Since such variation is the exception rather than the rule, we have designed our mass and retention time correction algorithms to be robust to such variation by excluding outlier landmarks. Outliers are defined as those landmarks whose mass or retention times (for mass or retention time correction, respectively) are greater then $Q_3 + 1.5 \times IQR$ or less than $Q_1 - 1.5 \times IQR$, where $Q_1$ is the lower quartile, $Q_3$ is the upper quartile, and the interquartile range $IQR = Q_3 - Q_1$.

**Clustering of Features with a Gaussian Mixture Model**

With mass and retention time corrected data that includes all charge-identified peaks in all the runs detected by MapQuant, we now address the peak matching and alignment problem. The challenge here is to match identical LC-MS features across multiple sample runs, taking into account m/z and retention time variation. Given the high performance MS instrumentation we are using, m/z variation is minimal, but still needs to addressed due to the fact that complex protein mixtures can give rise to peptides with very similar m/z values. Retention time, however, can vary significantly from run to run for a given peptide, though retention time correction offsets this variation to a degree.

The peak matching process starts with the union of all charge-identified peaks from all the LC-MS runs under consideration. Each peak is defined by its m/z, retention time (RT) and charge (z). Each peak also has an observed intensity in its respective run, and may be a sequence-identified landmark.
The first step in the peak matching process is to use the sequence-identified landmark peptides \( \{p_1, p_2, \ldots, p_L\} \) to determine the m/z and RT tolerances. For each peptide \( p_i \), we calculate the m/z and RT variation as the difference of the minimum and maximum observed m/z or RT value, respectively. Let the m/z variation for the \( L \) peptides be \( M = \{m_1, m_2, \ldots, m_L\} \), and the corresponding RT variation be \( R = \{r_1, r_2, \ldots, r_L\} \). Let \( M_1 \) and \( M_3 \) be the lower and upper quartile of M. Similarly let, \( R_1 \) and \( R_3 \) be the lower and upper quartiles of R. Defining outliers as those points that lie beyond 1.5 times the interquartile range, we define the m/z and RT tolerances as the m/z and RT variation of landmarks that are not outliers:

\[
m/z_{\text{tolerance}} = \frac{1}{2} (M_3 + 1.5 \times (M_3-M_1)) - \frac{1}{2} (M_1 - 1.5 \times (M_3-M_1)) = 4 \times (M_3-M_1) \tag{Eq. 4}
\]

\[
RT_{\text{tolerance}} = \frac{1}{2} (R_3 + 1.5 \times (R_3-R_1)) - \frac{1}{2} (R_1 - 1.5 \times (R_3-R_1)) = 4 \times (R_3-R_1) \tag{Eq. 5}
\]

The second step is to sort the peaks by m/z and partition them into m/z-strips such that adjacent m/zs within a strip are separated by less than m/z_{tolerance}, while m/zs in different strips are separated by more than m/z_{tolerance}. Thus, m/z strips are adaptive, and have boundaries when adjacent m/zs differ by more than the m/z tolerance, and hence, by definition, belong to different peptides. Creation of m/z strips divides the peak alignment process into smaller, independent tasks making the matching process more tractable and parallelizable, in addition providing more reliable matching.

Each m/z strip represents one or more peptides, spanning the entire RT range of the LC-MS run, and can contain anywhere from a few peaks to several hundred peaks.
Peak matching and alignment is performed within each m/z strip, and entails: (i) determining the number of peaks in the strip, and (ii) computing the location of these peaks in (m/z, RT). Charge states are kept separate, and the process is repeated for each observed charge state.

Peak matching is performed using model-based clustering (37, 38), using the Expectation Maximization (EM) algorithm (39) for Gaussian mixture model parameter estimation. Peak matching takes only m/z and RT of the peak into account, and does not use intensity during the peak clustering operation. For each matched peak, we assume that its true position in (m/z, RT) space is represented by $\mu$, and the observed coordinates of this peak vary around $\mu$ in a Gaussian manner with mean $\mu$ and covariance $\Sigma$, so that it is represented by

$$f_k(x | \mu_k, \Sigma_k) = \frac{\exp \left\{ \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) \right\}}{(2\pi)^{d/2} |\Sigma_k|^{1/2}}$$

(Eq. 6)

where $x$ represents the data (constituent peaks subsumed by this matched peak), $k$ is an integer subscript specifying the matched peak, and $d$ is the number of dimensions ($=2$ in our case). Assuming that the m/z strip has $K$ matched peaks, i.e., $k \in \{1, 2, \ldots, K\}$, the strip is represented by a mixture of $K$ Gaussians. The likelihood function (37) for this Gaussian mixture model is

$$L(\mu_1, K, \mu_K; \Sigma_1, \Sigma_K; \tau_1, K, \tau_K | x) = \prod_{i=1}^{n} \sum_{k=1}^{K} \tau_k f_k(x_i | \mu_k, \Sigma_k).$$
(Eq. 7)

Here $\tau$ represents the mixing proportion, i.e., the probability that a given observed peak comes from the $k$-th matched peak, and $n$ is the number of observations in the data, i.e., the number of peaks in the m/z strip under consideration. The location of the $K$ peaks in the m/z strip is determined by maximizing this likelihood function (or equivalently maximizing the logarithm of the likelihood, or log likelihood) in an iterative manner using Expectation Maximization.

In the above clustering, we have assumed that the number of matched peaks $K$ in an m/z strip is known in order to estimate the location of these peaks based on the observed data. The optimal number of matched peaks $K_{opt}$ for a strip is determined using the Bayesian Information Criterion (BIC). The BIC is the value of the maximized log likelihood with a penalty for the number of parameters in the model: $\text{BIC} = 2l - r \log(n)$, where $l$ is the maximized log likelihood for the model, $r$ is the number of parameters and $n$ is the number of data points (40).

Furthermore, in model-based clustering the covariance matrix $\Sigma_k$ is parameterized by eigenvalue decomposition $\Sigma_k = \lambda_k D_k A_k D_k^T$ (38). In this representation, the orientation of the principal components of $\Sigma_k$ is determined by $D_k$, while $A_k$ determines the shape of the density contours, and $\lambda_k$ specifies the volume of the corresponding ellipsoid. These orientation, volume and shape characteristics are determined from the data and the fit for various constrained parameterizations—e.g., equal-volume spherical variance, constant variance, unconstrained variance, etc. (see (38) for details)—evaluated to determine the best model for the observed peaks in the m/z strip.
Once the optimal number of clusters and the location of these clusters have been estimated for an m/z strip, we coalesce clusters whose centroids fall within the m/z and RT tolerance limits. This is to avoid unnecessary splitting of a single peptide into multiple matched peaks, which tends to happen occasionally with the clustering methodology described above.

Applying the model-based clustering methodology to all the m/z strips results in a final set of matched peaks that are deemed the common feature space across all the LC-MS runs. Each matched peak is represented by an (m/z, RT, z) triplet, and is constituted by peaks from one or more runs that cluster together. Once the common peak set is established, the intensities of these peaks in the various samples are determined based on the intensities of the observed peaks constituting each matched peak. There may be multiple peaks from a single sample run that fall into a single matched peak—in such cases, the final intensity of the matched peak for that sample is the sum of the multiple peak intensities. If a single peak from a run falls into a matched peak, the corresponding intensity is carried over; matched peaks missing from a specific run are marked missing. Furthermore, any peptide identities from landmarking are carried over, and the matched peak is marked with the identity of the landmark.

The final result of the peak matching and alignment process is an intensity table whose rows represent features that are the matched peaks, and whose columns represent sample (or runs).

The peak matching algorithms are implemented primarily as an R language library (41), with supporting shell and Perl scripts. The algorithms have been parallelized to efficiently utilize cluster computing environments.
Quantitative Data Analysis

Peptide sequences from matched features were assigned to their parent proteins by string matching. The matches were then sorted and grouped by parent protein. All abundance values were normalized to the summed abundance of all features identified by MapQuant for the experiment and log₂ transformed. Log₂ transformation effectively redistributes the abundance values observed into a normal distribution. This transformation is supported by Box Cox analysis (data not shown) (42). Average peptide abundance was computed by averaging the technical replicates for a sample.

Calibration curves were constructed by scaling each abundance value for a given peptide by its maximum observed intensity value across all concentrations. Peptides not observed at a given concentration were assigned the abundance value of 0. All peptides for a given protein or set of proteins were then averaged at each measured concentration.

Ratios were computed in log-space by subtraction using average peptide abundances. Subsequently, protein ratios were computed by averaging all of the constituent peptide ratios and taking 2 to the power of this average.

Significantly changing matched peaks were detected by taking the median of all technical replicates (LC-MS runs) for a given sample and then employing a signal-to-noise marker selection algorithm with a multiple-hypothesis corrected false discovery rate threshold of 1% across known sample groups (43, 44).
Results

Propagation of Identities by Landmark Matching

Landmark matching provides an alternative to traditional chromatographic alignment by using relative chromatographic elution order information and sequence-identified ‘landmarks’ to assign peptide identities to LC-MS peaks and propagate them across multiple LC-MS experiments. The process is illustrated in Fig. 2. This is achieved by performing a limited number of data-dependent MS/MS scan acquisitions in an LC-MS experiment that is primarily designed for chromatographic resolution and quantification. In general, the number of assignments propagated by landmark matching is dependent on the complexity of the sample subjected to LC-MS. For simple samples, many peptides will be sampled by MS/MS scans and relatively few assignments will be made by landmark matching. However, for complex samples such as cell lysates, landmark matching may be responsible for the assignment of peptide sequences to a significant number of LC-MS peaks, especially those of low abundance that were not selected for MS/MS sequencing.

Landmark matching adds value even for the simple protein mixtures used to calibrate the system (Scale Mixes) and determine the robustness of ratios (Variability Mixes). An average of 165 +/- 27 peaks were directly sequence identified in these experiments. After landmark matching, an average of 281 +/- 44 peaks had sequence assignments, an increase of 70%. The false positive rate of these matches can be evaluated according to the formulas described in Eqs. 2 and 3. In a randomly selected LC-MS experiment, 93% of matched features have a multiple hypothesis-corrected p-value of < 0.005, but the p-values of each individual match are preserved and could be
used for further filtering. The false negative rate can be evaluated by counting the number of peaks that were directly sequence identified but not assigned by landmark matching. We calculate this rate at < 2%. In any case, we can merge these directly sequence identified peaks back into the data set if they were missed during landmark matching. Summary landmark matching statistics can be seen in the Supplemental Material.

Landmark matching proved extremely useful in propagating peptide identities across LC-MS experiments in the more complex mitochondrial samples. Here, an average of 1083 peaks/experiment were directly sequence identified. Landmark matching adds an average of an additional 685 peaks/experiment. Again the probability of a match by chance was extremely low. In a randomly selected experiment, 98% of assignments had $p < 0.005$ and all assignments had $p < 0.04$. The false negative rate was slightly higher (13%) in this case, perhaps suggesting that the landmark matching score threshold is set too high, but again, the directly sequenced peptides are easily merged back into the data set.

The use of a relative (rather than absolute) chromatographic retention time criterion allows propagation of peptide identities from related LC-MS experiments with different designs. In practice, one may utilize longer separations with more MS/MS spectra to achieve ‘coverage’ of a sample and increase the BASIS SET. Alternatively, one may use data from lower performance instrumentation to enhance the BASIS SET. Indeed, we executed LC-MS experiments with higher concentrations of analytes and an MS to MS/MS scan ratio of 1:10 rather than 1:3 to populate our BASIS SET in certain cases (Identification Experiments, see methods). We can estimate the utility of such
external datasets by tracking the use of COMPARISON EXPERIMENTS during landmark matching. In general, over 1/3 of peptide identity assignments were made through landmark matching to an external COMPARISON EXPERIMENT. Certain peptides may only have been sequenced in Identification Experiments but were assignable to features in Quantification Experiments through landmark matching. Landmark matching enables robust quantification of peptides and proteins across multiple samples, conditions, etc, in a label free manner as presented below.

Another landmark matching technique was used to increase coverage of the Variability Mix peptides present in the mitochondrial extracts. Due to the presence of thousands of mitochondrial peptides, not as many Variability Mix peptides were directly sequenced in the Identification or Quantification LC-MS experiments for these samples. Therefore, the BASIS SET used for the mitochondrial experiments had a paucity of these peptides. However, by using the data from the Variability Mix and Scale Mix LC-MS experiments as the BASIS SET, we were able to increase the number of Variability Mix proteins detected from 7 to 12 (the total number of proteins in the mixture) and the number of Variability Mix peptides mapped on features from 101 to 155. Ultimately, this allowed quantification of ratios of all 12 Variability Mix proteins in a complex mixture.

**Label-free Quantification across 2.5 Orders of Magnitude**

We constructed simple mixtures of proteins across 3 decades of molar abundance to measure system performance in a typical experimental dynamic range. Each mixture was independently prepared rather than made by serial dilution of a
master mixture to introduce real-world sample preparation noise. Shown in Fig. 3A is the composite calibration for all peptides belonging to proteins whose abundances change across the Scale Mixes. A quadratic fit is shown with $R^2 = 0.98$. Results for individual proteins in the Scale Mixes can be seen in Table 2. Fig. 3B shows the reproducibility of the peptides belonging to the proteins whose abundance is held constant in all of the Scale Mixes. As is immediately evident, Mix H is an outlier compared to the other mixes and despite repeated attempts at preparing this mixture we could not get it to agree with the other mixtures. However, it is the robust performance of the procedure in general (as seen with the other mixtures) that makes easy outlier identification possible.

These results demonstrate the ability of the platform to perform quantification at the peptide level over at least 2.5 orders of magnitude of concentration. The integrated MS signal peak volumes (abundances) of the features used in quantification ranged over 4 orders of magnitude demonstrating a theoretical dynamic range of > 10,000 for the platform.

**Robust Label-free Quantification of Ratios**

We asked five different members of our laboratory to independently prepare tryptic digests of two mixtures of proteins with set ratios (Variability Mixes $\alpha$ and $\beta$). This was done to mimic real world sample variability. These samples were then analyzed with 5 replicates each for a total of 50 analyses. The ratios computed from the proteomic data were in good agreement with the known ratios (Fig. 4), with average absolute deviation across all proteins and lab members < 20%. We measured peptide
ratios from 0.1 to 50 (a 500-fold range) in a single experiment, once again illustrating the high dynamic quantitation range of the platform. For proteins that changed between mixes $\alpha$ and $\beta$, the correct direction of change was always computed and the ratios were significantly different from 1 in all cases ($p < 0.01$). In contrast, none of the ratios for the proteins held constant were significantly different than 1.

Importantly, the platform was able to accurately quantify ratios across a wide range of absolute molar concentrations. The ratio between 1 and 10 fmol of $\beta$-galactosidase was just as easily seen as the ratio between 10 and 100 fmol of $\alpha$-casein. Two-fold ratios were readily detected in the 5-10 fmol range (ovalbumin and transferrin, $p = 2.3 \times 10^{-11}$ and $9.3 \times 10^{-29}$ respectively). Moreover, the range of concentrations spanned in a single LC-MS experiment ranged from 1 to 200 fmol, illustrating the independence of ratio calculation at various analyte levels present in the samples themselves.

The robustness of the calculations is due in part to the presence of measured values for peptides across a large fraction of biological and technical replicates. ‘Sparser’ data would lead to higher error estimates and decreased statistical significance. Peptide assignment via landmark matching is the principal technology that enables ‘filling out’ of the data matrix for features without supporting MS/MS data. Robustness is important for the purpose of recognizing significantly changing proteins composed of multiple peptide measurements (i.e., ratio clustering). Ratio clustering in combination with peak matching via a Gaussian mixture model (see below) can lead to the discovery of novel, unidentified features that are indicative of a change of state (i.e., a biomarker).
Peak Matching via a Gaussian Mixture Model and Pattern Recognition

We analyzed the 50 Variability Mix LC-MS runs with the peak matching algorithms described above. Through this process, 157,774 LC-MS features were coalesced into 14,109 matched peaks that form the common feature set across all LC-MS runs, and were present in 1 or more runs (see Supplemental Information for a distribution of features by number of runs matched). The tolerance in the m/z dimension for these clusters was 2.9 ppm and in RT dimension was 2.71 minutes. 431 peaks were previously identified at the sequence level by landmark matching. A small number (17) of landmark peptides were divided into 2 or more peaks, indicating that occasionally clusters of LC-MS features did not fully coalesce. However, we estimate this occurs with a frequency of < 9%, and can be remedied by employing slightly looser tolerances in the m/z and RT dimensions. The matched peaks were identified by the triple of m/z-RT-z (with associated sequence, when known). The intensities of the matched peaks in each run were used for subsequent pattern recognition.

The results of marker selection for peaks distinguishing Mix $\alpha$ from Mix $\beta$ are shown in Fig. 5. It is clear that the peptides selected as markers are consistently differentially abundant between samples classes. While some of the differential matched peaks have already been sequence identified by landmark matching, the vast majority are of unknown provenance and represent an opportunity for novel marker discovery (see below). The contrasting abundances of the marker-selected peaks easily visually classify Variability Mixes $\alpha$ and $\beta$. 
Only 1 of 119 sequence identified peaks (SLHTLFDELcK) was found to be statistically different between the two mixes but actually came from a protein (BSA) whose composition was the same in both mixes. In fact, its ratio is measured at 1:1.1 between the two mixtures. This happens to be consistent enough to pass our p-value threshold but most likely would not be considered a real-world marker.

**Discovery of ‘Novel’ Marker Candidates via Peak Matching**

We used the PEPPeR pipeline to discover peaks that were found to be significantly changing between the $\alpha$ and $\beta$ compositions of the Variability Mixes with a minimum of 5-fold change (the 5-fold threshold was chosen for simplifying subsequent targeted MS/MS identification experiments). 232 peaks with a false discovery rate of $<$ 1% were identified (114 peaks higher in the $\alpha$ mix and 118 peaks higher in the $\beta$ mix – see supplemental material). Of these, 65 peaks had been identified via landmark marching and/or direct sequencing. We designed LC-MS methods with accurate mass-targeted inclusion lists (+/- 10 ppm) to attempt to acquire MS/MS spectra for each of the 232 peaks and tested them on representative mixtures of the $\alpha$ and $\beta$ compositions. Ultimately, MS/MS spectra were acquired for 171 of the targets with 119 yielding confident identifications via SpectrumMill searching. Of the 65 previously identified peaks, new MS/MS spectra were obtained for 63 of them. 59 yielded confident identification which agreed in every case with the prior identification (100% specificity, 91% sensitivity); the other 4 spectra did not yield confident identifications. Of the remaining 60 confident peak identifications, 25 belong to proteins included in the mixtures by design. All but one of these were derived from proteins whose proportions
change from mix $\alpha$ to $\beta$. This represents a 38% increase in the number of marker peptides identified.

The 35 additional peaks were revealed to be derived from contaminants in our protein stocks. For example, multiple peptides from *E. coli* RNA polymerase complex members and chaperones were detected as significantly changing (see Supplemental Information for proteins identified). These must be due to contamination of our $\beta$-galactosidase stock (the only protein in the mix derived from *E. coli*), which itself changes by 10-fold between the two mixes (we subsequently verified that the powdered stock was contaminated, see Supplemental data). Not a single peptide from the common laboratory contaminant keratin was identified as significantly changing between the two mixtures. These results underscore the importance of quality assessment when purchasing and using protein standards. More significantly, it demonstrates discovery of novel proteins that could be considered ‘biomarkers’ by ex post facto identification of LC-MS features shown to be changing between two sample groups without prior knowledge of their identities or even presence in the mixture. The unbiased discovery of these contaminant proteins is akin to real-world validation of the system as a whole.

**Measurement of Ratios in a Complex Mixture**

We were able to quantify ratios for all 12 proteins from the Variability Mixes that were spiked into protein extracts prepared from mouse liver mitochondria. Coverage and estimated errors were not as good as in simple mixtures, due in part to the fact that
this was a single total experimental replicate versus the 5 that were done for the
Variability Mixes alone. Nevertheless, we were still able to accurately measure the
known ratios of all of the Variability Mix proteins (Fig. 6). For proteins whose ratio
varied between the two mixtures, the correct direction of change was always detected.

3 of the 7 variable proteins demonstrated strong statistical significance (3/7 with \( p < 0.05 \), 5/7 with \( p < 0.1 \)). Proteins whose ratios were 1 between the two mixtures were
never found to have ratios significantly different from unity. These results were obtained
despite the background of thousands of mitochondrial peptides present in the samples.

We simultaneously quantified ratios for ~500 mitochondrial proteins from ~3000
peptides. Analysis of these results will be the subject of another manuscript (in
preparation), but ~50 proteins were found to vary with statistical significance from 2 to 6
weeks of age. We plan to complete our pipeline by performing Peak Matching and then
directed sequencing to identify or reinforce our results.
Discussion

In recent years, DNA microarray technology for the study of gene expression has revolutionized the discipline of biology. Numerous tools have been developed to analyze microarray data and countless discoveries have been made with these new techniques. Proteomics now stands poised to make similar contributions, and it would be highly desirable to leverage all of the previous efforts at building data mining and analysis tools. The key difference is that proteomics experiments are not ‘addressable’ in that we can not pre-design the m/z values, elution times and identities of peptides in an LC-MS experiment in the way that a microarray designer might synthesize a particular piece of DNA in a specific location on a microarray for hybridization.

PEPPeR seeks to address some of these difficulties through two new strategies: landmark matching and peak matching. Rather than attempting to perform complex chromatographic alignment, PEPPeR’s goal is to recognize LC-MS features that are the same across multiple experiments with healthy tolerances for experimental variation. Landmark matching relies on the observation that peptides in a mixture tend to elute in the same order from chromatography run to run regardless of deviations in retention time or even length of gradient. In this way, landmark matching uses relative retention properties of one peptide to another to propagate identities across LC-MS runs. It is unique in that it is time base-independent and utilizes confidently identified peptides within experiments as registration marks. It is assisted by but not dependent upon MS/MS acquisitions for quantification. At the same time, it can utilize data streams of previously identified peptides from LC-MS experiments with different chromatographic gradients or data acquisition strategies and thus is highly integrative. We have
demonstrated its adaptability and robustness in both simple and complex mixtures. Peak matching builds upon landmark matching by using propagated identities to perform coarse alignments and derive tolerances in preparation for clustering of LC-MS features across runs. The use of clustering rather than fine time-based alignment is what fundamentally allows PEPPeR to recognize even unidentified peptides in a set of related experiments. We expect this technology to scale to more complex samples with attendant fractionation schemes by collapsing multiple LC-MS experiments in silico. By carefully selecting appropriate BASIS SETS for landmark matching, performance should approach that of unfractionated samples. At the same time, the PEPPeR’s utilization of high performance data and blind discovery of potential markers may lessen the need for sample fractionation.

We demonstrated an unintentional but remarkable validation of the PEPPeR system by detecting significantly changing peaks that turned out to correspond to peptides from *E. coli* proteins. There was only a single protein in our mixtures sourced from *E. coli* whose ratio did in fact change between the two Variability Mixes, and these peaks were derived from contaminants present in the powdered stock. A secondary benefit to using the high performance instrumentation as used in this study is that we had an accurate measurement of the m/z values of the novel candidate markers. This allowed us to rapidly close the loop and identify the candidates with the use of accurate mass driven precursor dependent MS/MS acquisition. We obtained the identities of the novel markers literally within one day of computational marker selection. The unbiased marker selection also resulted in greater peptide coverage (and therefore more quantitative measurements) for those proteins known to be present in different
concentrations in the mixture. This is critical for accumulating statistical evidence that a protein's concentration is really changing in different biological states.

Importantly, PEPPeR is based on freely available tools with transparent algorithms open to examination and/or modification. Although it is best used with high performance instrumentation, it may be adaptable to a variety of instrument types. It should also be adaptable to other quantification strategies, such as SILAC. This is in contrast to commercially available tools that are comparative black boxes. PEPPeR does require a significant computational infrastructure to run efficiently, but highly parallelized multiprocessor systems are now commonplace, and PEPPeR is intended to work in highly distributed cluster computer environments. The amount of computational power required is on par with the magnitude of the extremely difficult task at hand.

Using PEPPeR as a basis, we can now begin to explore other challenges in quantitative proteomics such as computational techniques for normalizing protein expression, and more applied endeavors such as biomarker discovery. We fully expect that the platform will enable recovery of a significant amount of currently unexploited information in the MS data obtained from these instruments, while greatly increasing the accuracy and efficiency of obtaining the sequence identities of even minor m/z peaks found to change across samples. We expect these features to result in a large increase in the number of lower abundance proteins identified as differentially regulated in disease versus health. Recovery of this information along with increased depth of coverage may lead to an increased rate of discovery of useful candidate biomarkers.
**Acknowledgements**

We would like to thank Dr. Vamsi Mootha and Ms. Betty Chang for their kind gift of mitochondrial protein extract. As well, thanks are extended to members of the Carr laboratory for performing protein digests for this the study, and especially Eric Kuhn for his data on β-galactosidase contamination.

* This work was supported by grants from The Gates Foundation and the Entertainment Industry Foundation to S. A. Carr and the Department of Energy (DOE-GTL) to G.M. Church.
References


Figure Legends

Figure 1. Schematic of the PEPPeR Process. The platform allows for independent deep discovery experiments (Identification Experiments) that can increase the depth of coverage for protein level biomarker discovery, and high-throughput Quantification Experiments. However, dedicated Identification Experiments are entirely optional. Peptide sequence information may be incorporated from a search engine of choice, while feature detection is accomplished with MapQuant (36). Incorporation of parallel information streams is accomplished by Landmark Matching. Peak matching recognizes identical molecular species across multiple experiments. Parameterized peaks with associated abundances allow for statistical methods of marker discovery, and accurate mass-driven, targeted follow-up experimentation enables rapid loop closure for marker identification. Red circles=LC-MS features detected by MapQuant; Green flags=landmark matched peptides; Orange stars=matched peaks.

Figure 2. Example of Landmark List Selection and Scoring. Flags: peaks identified by direct MS/MS sequencing. Ovals: accurate mass features of unknown sequence identity. Left panel: Landmark List Selection. Consider selection of a LANDMARK LIST to interrogate peak X. Experiments ① and ② share identified peaks with the CURRENT EXPERIMENT and contain a peptide that satisfies the m/z requirement of peak X; experiment ③ contains a peptide that satisfies the m/z requirement of peak X but has no other peaks in common. Therefore, ③ can not be a valid COMPARISON EXPERIMENT. Experiment ② shares more sequenced peptides with the CURRENT EXPERIMENT than ①, so ② is selected as the COMPARISON EXPERIMENT. Right panel: Landmark Scoring.
Bonus and penalty points are assessed based on substitution of the PUTATIVE ASSIGNMENT for the unknown peak and comparison of relative elution order. Thick lines indicate substitution for the PUTATIVE ASSIGNMENT. Thin lines indicate landmark alignments which are then scored with the heuristic described in Eq. 1.

Figure 3. Calibration of Peptide Data in Scale Mixes. A) Composite averages for scaled peptide abundances across all variable proteins in the Scale Mixes. Scaling and averaging is described in the methods. A quadratic fit is shown with $R^2 = 0.98$. B) Composite averages for scaled peptide abundances across all constant proteins in the Scale Mixes. The mix letter (c.f. Table 1) is shown next to the data point. The mean of all 9 points is 0.50 with a CV of 12%. Error bars show standard error of the mean in A and B.

Figure 4. Measurement of Ratios in Variability Mixes. Known ratios shown in black next to the measured ratio shown in white. * indicates statistical significance $p < 0.01$ that the ratio is different then 1. Error bars represent high and low ranges observed (among independent preparations). Abbreviations: $\beta$G = $\beta$-galactosidase; Pls=plasminogen; Ov=ovalbumin; CA=carbonic anhydrase; Alb=serum albumin; RNAse=ribonuclease A; Myo=myoglobin; fibrinogen $\beta$; Tfn=transferring; Cas=\(\alpha\)-casein; Apr=aprotinin; $\beta$LG=\(\beta\)-lactoglobulin
Figure 5. Marker Selection in Variability. Heat map showing significantly changing peaks (with identities, where known) between Variability Mixes α and β. Markers selected have p < 0.01. Rows are individual matched peaks. The 5 columns for each mix represent the 5 individual preparations. A red block to the right of each row indicates a landmark matched peak. Scale at bottom.

Figure 6. Measurement of Ratios in a Complex Mixture. Known ratios shown in black next to the median measured ratio shown in white. * indicates statistical significance p < 0.05 that the ratio is different then 1, ~ indicates p < 0.1. Error bars represent the interquartile range observed among the peptides measured for that protein (only one measurement for Apr, so no error bar is shown). Abbreviations as in Fig. 3.
## Tables

**Table 1. Protein Mixture Analysis Concentrations**

<table>
<thead>
<tr>
<th>Protein</th>
<th>Scale Mix</th>
<th>Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A B C D E F G H I</td>
<td>α β</td>
</tr>
<tr>
<td>Aprotinin</td>
<td>1 2 3 10 20 30 100 200 300</td>
<td>100 5</td>
</tr>
<tr>
<td>Ribonuclease A</td>
<td>300 1 2 3 10 20 30 100 200</td>
<td>100 100</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>200 300 1 2 3 10 20 30 100</td>
<td>100 100</td>
</tr>
<tr>
<td>beta-Lactoglobulin</td>
<td>100 200 300 1 2 3 10 20 30</td>
<td>50 1</td>
</tr>
<tr>
<td>alpha Casein</td>
<td>30 100 200 300 1 2 3 10 20</td>
<td>100 10</td>
</tr>
<tr>
<td>Carbonic anhydrase</td>
<td>20 30 100 200 300 1 2 3 10</td>
<td>100 100</td>
</tr>
<tr>
<td>Ovalbumin</td>
<td>10 20 30 100 200 300 1 2 3</td>
<td>5 10</td>
</tr>
<tr>
<td>Fibrinogen</td>
<td>3 10 20 30 100 200 300 1 2</td>
<td>25 25</td>
</tr>
<tr>
<td>BSA</td>
<td>2 3 10 20 30 100 200 300 1</td>
<td>200 200</td>
</tr>
<tr>
<td>Transferrin</td>
<td>100 100 100 100 100 100 100 100 100</td>
<td>10 5</td>
</tr>
<tr>
<td>Plasminogen</td>
<td>30 30 30 30 30 30 30 30 30</td>
<td>2.5 25</td>
</tr>
<tr>
<td>beta-Galactosidase</td>
<td>10 10 10 10 10 10 10 10 10</td>
<td>1 10</td>
</tr>
</tbody>
</table>

All values in fmol/μl
Table 2a

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>$R^2$</th>
<th>LOQ</th>
<th># of peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aprotinin</td>
<td>0.87</td>
<td>30</td>
<td>4</td>
</tr>
<tr>
<td>Carbonic Anhydrase</td>
<td>0.90</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Ovalbumin (chicken)</td>
<td>0.95</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>Albumin</td>
<td>0.92</td>
<td>2</td>
<td>155</td>
</tr>
<tr>
<td>Ribonuclease A</td>
<td>0.83</td>
<td>100</td>
<td>6</td>
</tr>
<tr>
<td>Casein ($\alpha$-S1 and S2)</td>
<td>0.99</td>
<td>3</td>
<td>45</td>
</tr>
<tr>
<td>$\beta$-Lactoglobulin</td>
<td>0.88</td>
<td>3</td>
<td>27</td>
</tr>
<tr>
<td>Myoglobin (horse)</td>
<td>0.89</td>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td>Fibrinogen $\beta$ chain</td>
<td>0.96</td>
<td>3</td>
<td>26</td>
</tr>
</tbody>
</table>

**Composite** 0.98 1 331

$R^2$ is shown for a quadratic fit of abundance vs. concentration. LOQ is the limit of quantification in fmol. # of peptides refers to how many peptides were used in the calibration fit.

Table 2b

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>Abundance (fmol)</th>
<th>Average</th>
<th>CV</th>
<th># of peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>transferrin (human)</td>
<td>10</td>
<td>0.53</td>
<td>26%</td>
<td>181</td>
</tr>
<tr>
<td>plasminogen (human)</td>
<td>3</td>
<td>0.32</td>
<td>56%</td>
<td>39</td>
</tr>
<tr>
<td>$\beta$-galactosidase (E. coli)</td>
<td>1</td>
<td>0.28</td>
<td>94%</td>
<td>28</td>
</tr>
</tbody>
</table>

**Composite** n/a 0.50 12% 248

“Average” refers to average scaled abundance for the peptides.

All proteins of bovine origin unless noted.
Box 1.

**Basis Set**: set of all peptides sequenced in all experiments with information retained about the specific experiment in which they were sequenced

**Landmark**: a peptide sequenced in a quantification experiment and matched to a **Feature**

**Feature**: a fitted m/z-RT peak derived from a quantification experiment, no ID assigned, derived from MapQuant

**Putative Assignment**: assignment of a sequence from the basis set to a **Feature** on the basis of m/z alone

**Radius**: Time window of RT allowed for absolute time matching

**Tolerance**: ppm window of m/z tolerance for making a putative assignment

**Current Experiment**: an LC-MS experiment under consideration during landmark matching into which peptide identities are being propagated

**Comparison Experiment**: an LC-MS experiment in a group of experiments related by sample source or study (or, an experiment represented in the **Basis Set**) in which the **Putative Assignment** peptide has been confidently identified via MS/MS sequencing and which also shares identifications of some landmark peptides from the **Current Experiment**; an experiment from which peptide identities may be propagated (see Fig. 2)

**Landmark List**: a list of sequenced peptides observed in the **Current Experiment** also observed in a **Comparison Experiment** where the **Putative Assignment** was identified, sorted in elution order with retention times

**Landmark Score**: a computed score describing how well elution order in the **Comparison Experiment** matched the elution order in the **Current Experiment**

**Threshold**: **Landmark Score** needed to be reported as a match

**Bootstrap**: repeated randomization done to assess confidence in result
<table>
<thead>
<tr>
<th><strong>Absolute Time Matching</strong></th>
<th><strong>Landmark Matching</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>for each FEATURE in experiment {</td>
<td>for each FEATURE in experiment {</td>
</tr>
<tr>
<td>if peptide in BASIS SET with same</td>
<td>if peptide in BASIS SET with same</td>
</tr>
<tr>
<td>m/z within TOLERANCE, z as FEATURE</td>
<td>m/z within TOLERANCE, z as FEATURE</td>
</tr>
<tr>
<td>Make PUTATIVE ASSIGNMENT to FEATURE</td>
<td>Make PUTATIVE ASSIGNMENT to FEATURE</td>
</tr>
<tr>
<td>If RT of FEATURE within RADIUS of sequencing event {</td>
<td>if LANDMARK LIST can be found {</td>
</tr>
<tr>
<td>Report match</td>
<td>if LANDMARK SCORE &gt; THRESHOLD {</td>
</tr>
<tr>
<td>Report confidence</td>
<td>Report match</td>
</tr>
<tr>
<td>}</td>
<td>Do BOOTSTRAP</td>
</tr>
<tr>
<td>}</td>
<td>Report confidence</td>
</tr>
<tr>
<td>}</td>
<td>}</td>
</tr>
<tr>
<td>}</td>
<td>}</td>
</tr>
</tbody>
</table>
Identification Sample Set

Quantification Sample Set

Identity-centric LCMS Method

Pattern-centric LCMS Method

Feature Detection

MS/MS Interpretation

Landmark Matching

Peak Matching

Marker Discovery

Targeted MS/MS Identifications

Figure 1
Current Experiment:

Possible Comparison Experiments:

Selected Comparison Experiment:

Landmark Scoring:

X is M?:

A — A +1
M — —
B — B +1
C — C +1

PASS 3

X is N?:

A — A +1
X — B -1
B — M --
C — C +1

FAIL 1

Y is M?:

A — A +1
B — M --
Y — B -1
C — C +1

FAIL 1

Y is N?:

A — A +1
B — B +1
Y — N --
C — C +1

PASS 3
Figure 3

A

scaled log2 abundance

log2 concentration (fmol)

B

Scale Mix Replicate

0.00

0.10

0.20

0.30

0.40

0.50

0.60

0.70

0.80

0.90

1.00

1.00

0.90

0.80

0.70

0.60

0.50

0.40

0.30

0.20

0.10

0.00
Figure 4
Figure 5
Figure 6

[Graph showing log2 ratio for various proteins: βG, Pls, Ov, Fbr β, CA, BSA, Myo, RNAse, Tfn, Cas, Apr, βLG. The graph compares known and measured values, with error bars indicating variability.]