# PEPPeR, a Platform for Experimental Proteomic Pattern Recognition*[S]

**Jacob D. Jaffe‡, D. R. Mani‡, Kyriacos C. Leptos§, George M. Church§, Michael A. Gillette‡, and Steven A. Carr‡¶**

**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.** ***Molecular & Cellular Proteomics 5:1927–1941, 2006.***

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 proteomics-based biomarker discovery where comparative data across multiple patient samples should be obtained (3). Biomarker discovery usually uses 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 patterns 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,[1] and iTRAQ (isobaric tags for relative and absolute quantitation) 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 semiquantitative data may be obtained directly from the MS/MS data by spectrum counting (20).

From ‡The Broad Institute of Harvard and the Massachusetts Institute of Technology, Cambridge, Massachusetts 02142 and §Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115

Received, June 13, 2006, and in revised form, July 19, 2006
Published, MCP Papers in Press, July 19, 2006, DOI 10.1074/mcp.M600222-MCP200

[1] The abbreviations used are: SILAC, stable isotope labeling with amino acids in cell culture; PEPPeR, Platform for Experimental Proteomic Pattern Recognition; RT, retention time; IQR, interquartile range; BIC, Bayesian Information Criterion.

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 usable 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). Although 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 electrospray data. In LC-MS, retention time is also used as a coordinate in uniquely locating peaks 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 utilizes 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.* 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 Ref. 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 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, ultrahigh 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 that 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

TABLE I
*Protein mixture analysis concentrations*

All values are in fmol/$\mu$l.

| | Scale Mix | | | | | | | | | Variability Mix | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | I | $\alpha$ | $\beta$ |
| Aprotinin | 1 | 2 | 3 | 10 | 20 | 30 | 100 | 200 | 300 | 100 | 5 |
| Ribonuclease A | 300 | 1 | 2 | 3 | 10 | 20 | 30 | 100 | 200 | 100 | 100 |
| Myoglobin | 200 | 300 | 1 | 2 | 3 | 10 | 20 | 30 | 100 | 100 | 100 |
| $\beta$-Lactoglobulin | 100 | 200 | 300 | 1 | 2 | 3 | 10 | 20 | 30 | 50 | 1 |
| $\alpha$-Casein | 30 | 100 | 200 | 300 | 1 | 2 | 3 | 10 | 20 | 100 | 10 |
| Carbonic anhydrase | 20 | 30 | 100 | 200 | 300 | 1 | 2 | 3 | 10 | 100 | 100 |
| Ovalbumin | 10 | 20 | 30 | 100 | 200 | 300 | 1 | 2 | 3 | 5 | 10 |
| Fibrinogen | 3 | 10 | 20 | 30 | 100 | 200 | 300 | 1 | 2 | 25 | 25 |
| BSA | 2 | 3 | 10 | 20 | 30 | 100 | 200 | 300 | 1 | 200 | 200 |
| Transferrin | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 10 | 5 |
| Plasminogen | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 2.5 | 25 |
| $\beta$-Galactosidase | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 1 | 10 |

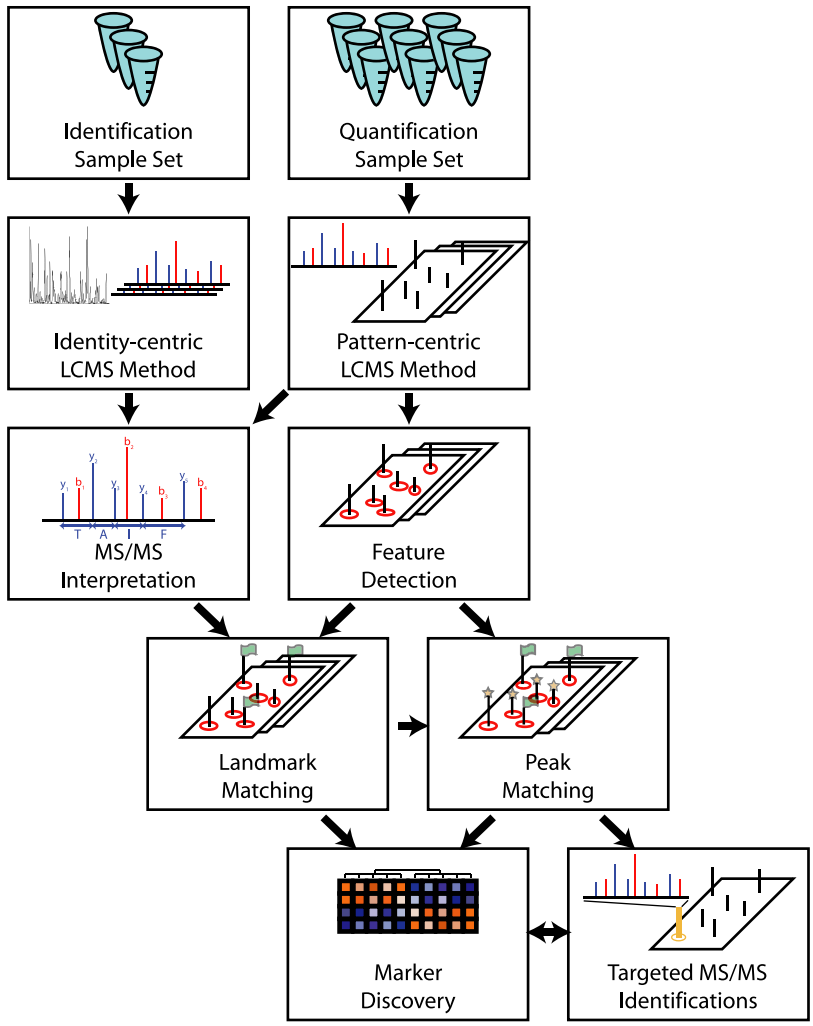



FIG. 1. **Schematic of the PEPPeR process.** The platform allows for independent deep discovery experiments (Identification Experiments) that can increase the depth of coverage for protein level biomarker discovery and high throughput Quantification Experiments. However, dedicated Identification Experiments are entirely optional. Peptide sequence information may be incorporated from a search engine of choice, whereas 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.

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. 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/$\mu$l stocks in water (Table I) to form new substocks. Each substock was diluted 1:1 with 6 M 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 10-fold with water. Trypsin (Roche Applied Science proteomics grade) was added at a mass ratio of 1:50 trypsin to total protein. Digestion proceeded at 37 °C for 18 h with shaking at 600 rpm. 10 $\mu$l 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 Mixes $\alpha$ and $\beta$ were distributed to five 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 $tC_{18}$ solid phase extraction cartridge (Waters) using the following procedure: wet with 2 × 1 ml of ACN, equilibrate 2 × 1 ml of 0.1% formic acid, load 1 ml of substock, wash 3 × 1 ml of 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 I.

### *Preparation of Mitochondrial Protein Extracts*

Preparation of mitochondrial extracts from C57BL6/J mice aged either 2 or 6 weeks was as described previously (34). The 2-week extract was spiked with Variability Mix $\alpha$ and the 6-week extract was spiked with Variability Mix $\beta$ prior to digestion such that the final concentrations in the analyzed samples would be the same as those shown in Table I. 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 Fig. 1. The components of the work flow are discussed in detail below. Capitalized terms are defined in 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

BOX 1. PEPPeR terminology.

#### *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 I).

LC parameters were common to both strategies. Chromatography was performed using an Agilent 1100 nanoflow 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 $\mu$m inner diameter, 15 nm tip opening; New Objective; Woburn, MA) was packed with 12.5 cm of ReproSil-Pur $C_{18}$-AQ 3-$\mu$m resin (Dr. Maisch GmbH) and directly interfaced to an LTQ-FT mass spectrometer fitted with a nanoelectrospray ionization source (ThermoElectron, Waltham, MA). 1 $\mu$l 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, and 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. One 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 (ion trap MS). 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 $5 \times 10^5$ ions to minimize

harmonic noise (target value raised to $1 \times 10^6$ for mitochondrial samples). Parameters for the Quantification Experiment acquisition strategy were identical except that the top three 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 five technical replicates.

*Peptide Spectrum Interpretation*

MS/MS spectra were extracted from the raw data and interpreted using SpectrumMill Data Extractor and MS/MS Search Revision B.03.02.059 (Agilent). 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 National Center for Biotechnology Information (NCBI) non-redundant protein database dated August 1, 2005 and containing 2,724,841 entries. Data from the mitochondrial preparations were searched against the International Protein Index (IPI) mouse database version 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 1 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 as a module of GenePattern at www.broad.mit.edu/tools/software.html.

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 retention time (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 supplemental information, and an example parser for SpectrumMill results is provided. An XML (extensible markup language) format will be available in the future.

Landmark matching is a sequential process. Limited MS/MS data acquisition during LC-MS experiments is used to confidently identify peptides that 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 ($\pm$25 ppm) and an absolute retention time RADIUS (typically 0.3 min). Second, an *m/z* recalibration is calculated using a least squares quadratic fit based on these preliminary matches, and a more stringent TOLERANCE is computed based on the distribution of residual *m/z* errors after recalibration ($\pm 3\sigma$, typically $<$5 ppm). Third, peptides are remapped 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

| ***Absolute Time Matching*** | ***Landmark Matching*** |
|---|---|
| - for each **FEATURE** in experiment {<br>- if peptide in **BASIS SET** with same m/z within **TOLERANCE**, z as **FEATURE** {<br>- Make **PUTATIVE ASSIGNMENT** to **FEATURE**<br>- If RT of **FEATURE** within **RADIUS** of sequencing event {<br>- Report match<br>- Report confidence<br>}<br>}<br>} | - for each **FEATURE** in experiment {<br>- if peptide in **BASIS SET** with same m/z within **TOLERANCE**, z as **FEATURE** {<br>- Make **PUTATIVE ASSIGNMENT** to **FEATURE**<br>- if **LANDMARK LIST** can be found {<br>- if **LANDMARK SCORE** > **THRESHOLD** {<br>- Report match<br>- Do **BOOTSTRAP**<br>- Report confidence<br>}<br>}<br>}<br>} |

BOX 2. Matching algorithms.

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 less than a constant $\kappa$ (typically $\kappa = 200$ scans indicates a sharply eluting peak for our acquisition 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 less than 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 Fig. 2, *left panel*.

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 $\delta$ seconds 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 follows.

$$S = \sum_{i=1}^{w} (\xi(\Lambda_{-i}, \Lambda_0) + \xi(\Lambda_0, \Lambda_i)) \quad \text{(Eq. 1)}$$

$$\xi(m, n) = \begin{cases} 1 \text{ if } \tau(m) < \tau(n) \\ \text{if } \tau(m) > \tau(n) \begin{cases} 0.5 \text{ if } \tau(n) - \tau(m) < \delta \text{ and } \mu(m) + \sigma(m) > \mu(n) - \sigma(n) \\ -1 \text{ if else} \end{cases} \\ 0 \text{ if else} \end{cases}$$

A passing score (THRESHOLD) may be set empirically by evaluation of matching probabilities (see below). Our current 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. An example of the scoring system is shown in Fig. 2, *right panel*.

All PUTATIVE ASSIGNMENTS that pass the landmark test THRESHOLD are then reported. A match is reported as a peptide sequence with corresponding *m/z*, retention time, and abundance information as derived from the corresponding MapQuant features. Processing an entire experiment results in a list of such matches. Any directly sequenced features for that experiment (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.

*Probabilistic Evaluation of Landmark Matches by Bootstrapping*

The probability that a given match is by chance can be evaluated using the combined probabilities that the *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.

$$p_{\text{overall}} = p_{m/z} p_{(\text{landmark}|m/z)} \quad \text{(Eq. 2)}$$

$p_{m/z}$ 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 charge (*z*) and dividing by the total number of BASIS SET peptides. $p_{(\text{landmark}|m/z)}$ can be computed according to Bayes' Rule.

$$P(\text{landmark}|m/z) = \frac{P(m/z|\text{landmark})\,P(\text{landmark})}{P(m/z|\text{landmark})\,P(\text{landmark}) + (P(m/z| \sim \text{landmark})(1 - P(\text{landmark})} \quad \text{(Eq. 3)}$$

Because these probabilities cannot all be calculated empirically they are assigned by BOOTSTRAPPING (100 BOOTSTRAPS). *P*(*landmark*) is estimated by selecting 100 FEATURES at random, assigning their retention time to the PUTATIVE ASSIGNMENT, and determining what fraction passes the landmark test using the same LANDMARK LIST that was selected for the PUTATIVE ASSIGNMENT. *P*(*m/z*|*landmark*) and *P*(*m/z*|~*landmark*) are calculated by determining what fraction of these random FEATURES falls 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 the *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.

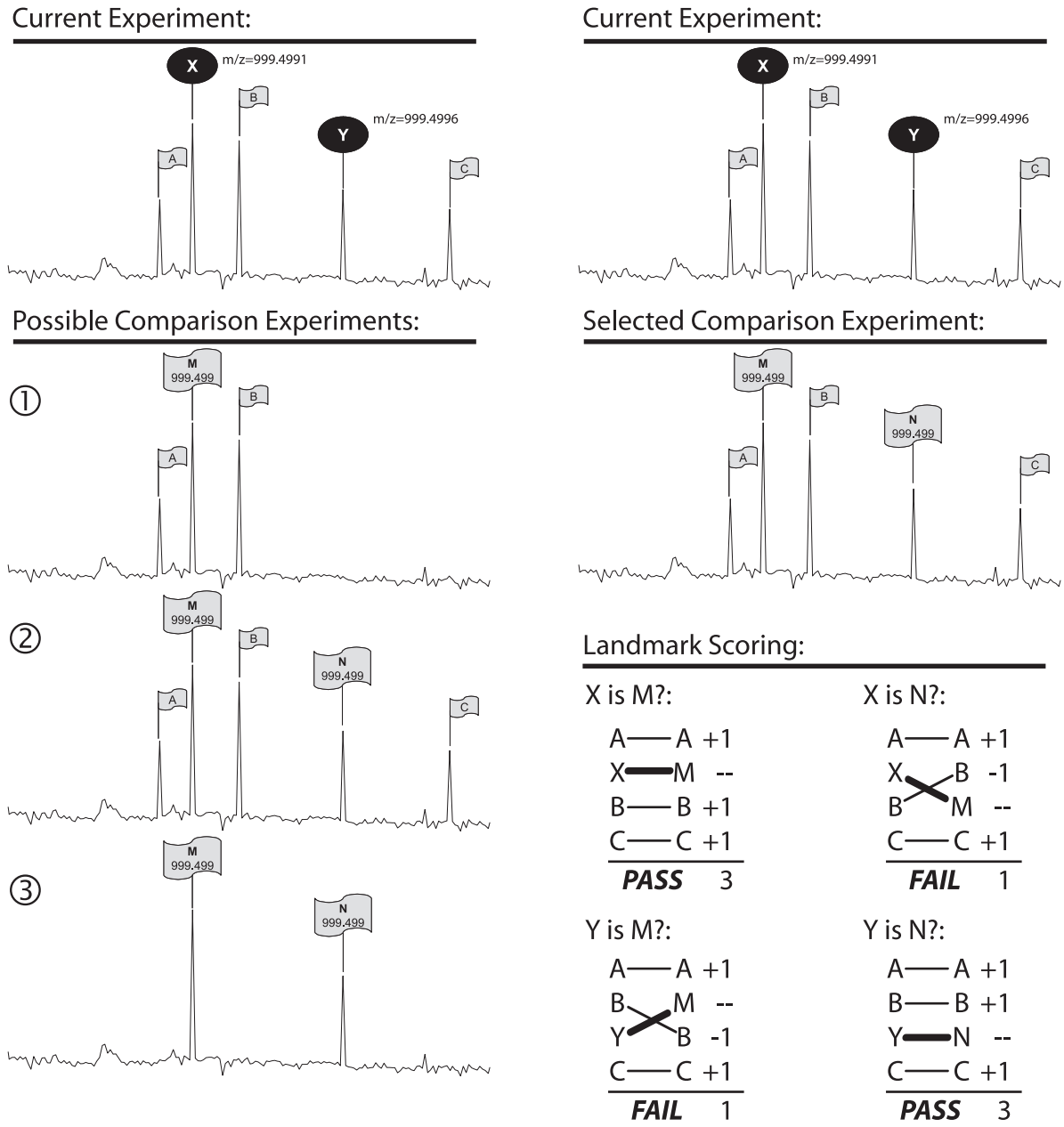



FIG. 2. **Example of LANDMARK LIST selection and scoring.** *Flags*, peaks identified by direct MS/MS sequencing. *Ovals*, accurate mass features of unknown sequence identity. *Left panel*, LANDMARK LIST selection. Consider selection of a LANDMARK LIST to interrogate peak X. Experiments ① and ② share 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 cannot 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 that are then scored with the heuristic described in Equation 1.

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 $RT_{corrected} = a_0 + a_1 \times RT + a_2 \times RT^2$ 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 corrections, 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). Because 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 $Q_3 + 1.5 \times \text{IQR}$ or less than $Q_1 - 1.5 \times \text{IQR}$ where $Q_1$ is the lower quartile, $Q_3$ is the upper quartile, and the interquartile range $\text{IQR} = Q_3 - Q_1$.

*Clustering of Features with a Gaussian Mixture Model*—With mass and retention time corrected data that include 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*, RT, and *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.

$$m/z_{\text{tolerance}} = (M_3 + 1.5 \times (M_3 - M_1)) - (M_1 - 1.5 \times (M_3 - M_1)) = 4 \times (M_3 - M_1) \quad \text{(Eq. 4)}$$

$$\text{RT}_{\text{tolerance}} = (R_3 + 1.5 \times (R_3 - R_1)) - (R_1 - 1.5 \times (R_3 - R_1)) = 4 \times (R_3 - R_1) \quad \text{(Eq. 5)}$$

The second step is to sort the peaks by *m/z* and partition them into *m/z strips* such that adjacent *m/z* features within a strip are separated by less than $m/z_{\text{tolerance}}$, whereas *m/z* assignments in different strips are separated by more than $m/z_{\text{tolerance}}$. Thus, *m/z* strips are adaptive and have boundaries when adjacent *m/z* assignments 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 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

$$f_k(x|\mu_k, \Sigma_k) = \frac{\exp\left\{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right\}}{(2\pi)^{d/2}|\Sigma|^{1/2}} \quad \text{(Eq. 6)}$$

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* Gaussian peaks. The likelihood function (37) for this Gaussian mixture model is as follows.

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

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 to estimate the location of these peaks based on the observed data. The optimal number of matched peaks $K_{\text{opt}}$ 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 $D_k$, $A_k$ 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 Ref. 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 $\log_2$-transformed. $\text{Log}_2$ 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 using 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 $\pm$ 27 peaks were directly sequence-identified in these experiments. After landmark matching, an average of 281 $\pm$ 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 Equations 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 "Experimental Procedures"). We can estimate the utility of such external data sets by tracking the use of Comparison Experiments during landmark matching. In general, over one-third 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

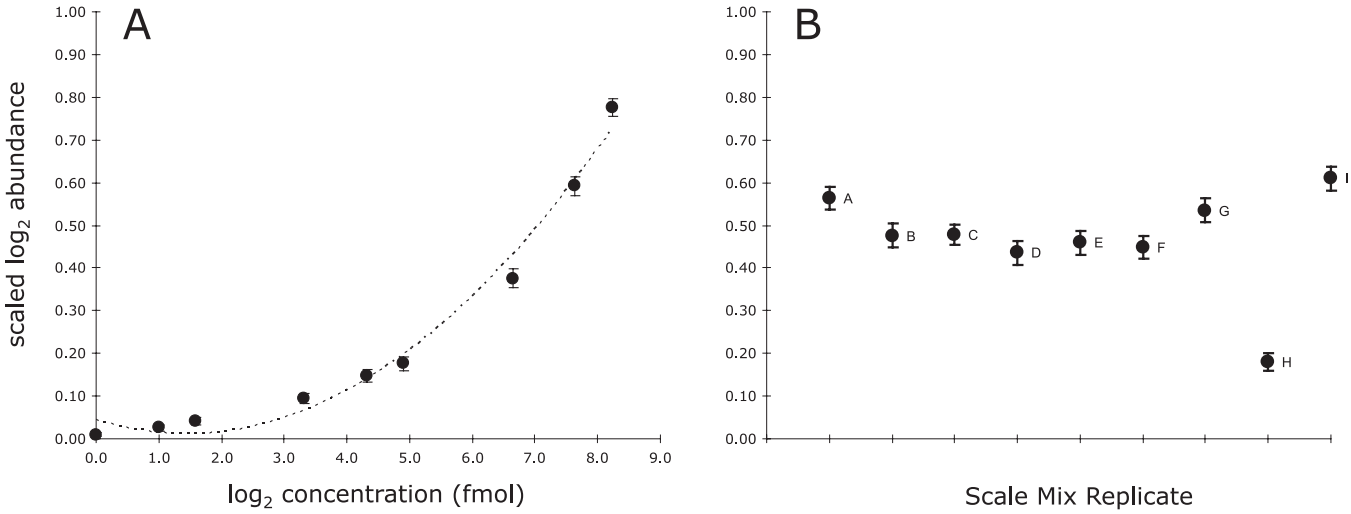



FIG. 3. **Calibration of peptide data in Scale Mixes.** *A*, composite averages for scaled peptide abundances across all variable proteins in the Scale Mixes. Scaling and averaging is described under "Experimental Procedures." A quadratic fit is shown with $R^2 = 0.98$. *B*, composite averages for scaled peptide abundances across all constant proteins in the Scale Mixes. The *Mix letter* (see Table I) is shown next to the data point. The mean of all nine points is 0.50 with a coefficient of variation of 12%. *Error bars* show S.E. in *A* and *B*.

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 seven 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. 3*A* 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 Tables II and III. Fig. 3*B* 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 with the other mixtures, 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

TABLE II
*Proteins with variable concentrations*

$R^2$ is shown for a quadratic fit of abundance *versus* concentration. LOQ is the limit of quantification in fmol. No. of peptides refers to how many peptides were used in the calibration fit.

| Protein name | $R^2$ | LOQ | No. of peptides |
|---|---|---|---|
| Aprotinin | 0.87 | 30 | 4 |
| Carbonic anhydrase | 0.90 | 10 | 10 |
| Ovalbumin (chicken) | 0.95 | 10 | 13 |
| Albumin | 0.92 | 2 | 155 |
| Ribonuclease A | 0.83 | 100 | 6 |
| Casein ($\alpha$-S1 and -S2) | 0.99 | 3 | 45 |
| $\beta$-Lactoglobulin | 0.88 | 3 | 27 |
| Myoglobin (horse) | 0.89 | 1 | 45 |
| Fibrinogen $\beta$ chain | 0.96 | 3 | 26 |
| Composite | 0.98 | 1 | 331 |

TABLE III
*Proteins held constant in all mixtures*

"Average" refers to average scaled abundance for the peptides. All proteins are of bovine origin unless noted. CV, coefficient of variation; N/A, not applicable.

| Protein name | Abundance | Average | CV | No. of peptides |
|---|---|---|---|---|
| | *fmol* | | % | |
| Transferrin (human) | 10 | 0.53 | 26 | 181 |
| Plasminogen (human) | 3 | 0.32 | 56 | 39 |
| $\beta$-Galactosidase (*E. coli*) | 1 | 0.28 | 94 | 28 |
| Composite | N/A | 0.50 | 12 | 248 |

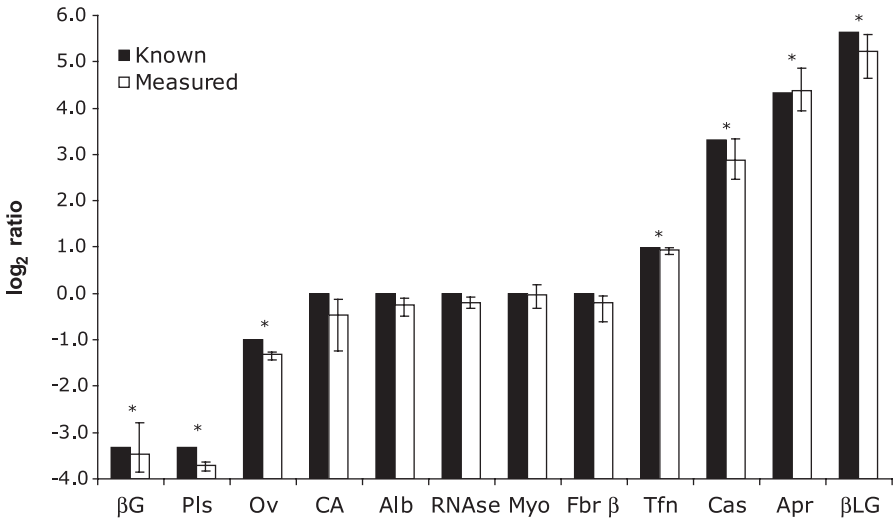



FIG. 4. **Measurement of ratios in Variability Mixes.** Known ratios are shown in *black* next to the measured ratio shown in *white*. * indicates statistical significance $p < 0.01$ that the ratio is different then 1. *Error bars* represent high and low ranges observed (among independent preparations). *βG*, β-galactosidase; *Pls*, plasminogen; *Ov*, ovalbumin; *CA*, carbonic anhydrase; *Alb*, serum albumin; *RNAse*, ribonuclease A; *Myo*, myoglobin; *Fbr β*, fibrinogen β; *Tfn*, transferrin; *Cas*, α-casein; *Apr*, aprotinin; *βLG*, β-lactoglobulin.

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 α and β). This was done to mimic real world sample variability. These samples were then analyzed with five replicates each for a total of 50 analyses. The ratios computed from the proteomics data were in good agreement with the known ratios (Fig. 4) with average absolute deviation across all proteins and laboratory 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 α and β, 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 from 1.

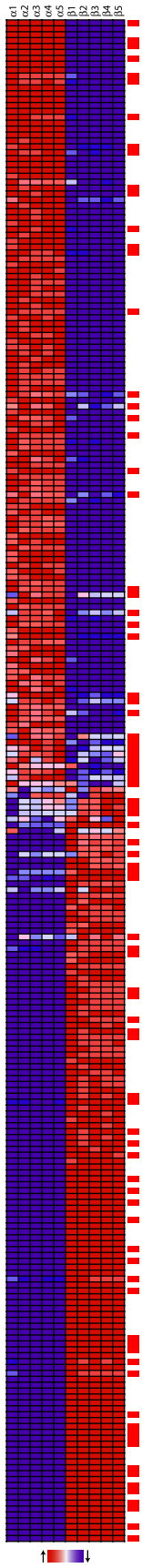
Importantly the platform was able to accurately quantify ratios across a wide range of absolute molar concentrations. The ratio between 1 and 10 fmol of β-galactosidase was just as easily seen as the ratio between 10 and 100 fmol of α-casein. 2-Fold ratios were readily detected in the 5–10 fmol range (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 one 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 the RT dimension was 2.71 min. 431 peaks were previously identified at the sequence level by landmark matching. A small number (17) of landmark peptides were divided into two 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 using slightly looser tolerances in the *m/z* and RT dimensions. The matched peaks were identified by the (*m/z*, RT, *z*) triplet (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 α from Mix β are shown in Fig. 5. It is clear that the peptides selected as markers are consistently differentially abundant between samples classes. Although 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 α and β.

Only one of 119 sequence-identified peaks (SLHTLFGDELCK) was found to be statistically different between the two mixtures but actually came from a protein (BSA) whose composition was the same in both mixtures. 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 Mix $\alpha$ and 118 peaks higher in Mix $\beta$; see supplemental material). Of these, 65 peaks had been identified via landmark matching and/or direct sequencing. We designed LC-MS methods with accurate mass-targeted inclusion lists ($\pm$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 four 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 *Escherichia 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 mixture derived from *E. coli*), which itself changes by 10-fold between the two mixtures (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 their presence in the mixture. The unbiased discovery of

FIG. 5. **Marker selection in variability.** Shown is a heat map showing significantly changing peaks (with identities where known) between Variability Mixes $\alpha$ and $\beta$. Markers selected have $p < 0.01$. *Rows* are individual matched peaks. The five *columns* for each mixture represent the five individual preparations. A *red block* to the *right* of each *row* indicates a landmark-matched peak. The *scale* is shown at *bottom*.

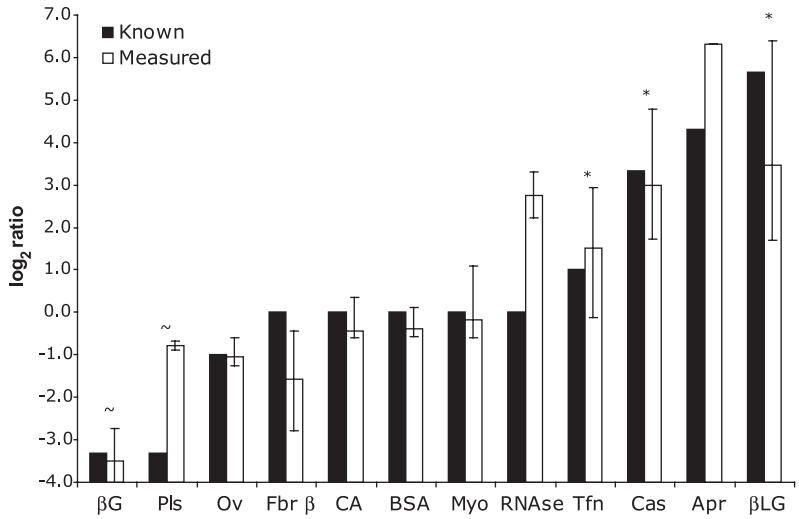



FIG. 6. **Measurement of ratios in a complex mixture.** Known ratios are shown in *black* next to the median measured ratio shown in *white*. * indicates statistical significance $p < 0.05$ that the ratio is different then 1; ~ indicates $p < 0.1$. *Error bars* represent the interquartile range observed among the peptides measured for that protein (there was only one measurement for aprotinin (*Apr*), so no error bar is shown). Abbreviations are as in Fig. 4.

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 five 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. Three of the seven variable proteins demonstrated strong statistical significance (three of seven with $p < 0.05$; five of seven 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 study,[2] 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 predesign 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, the goal of PEPPeR 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.

[2] B. Chang and S. A. Carr, manuscript in preparation.

At the same time, the utilization of high performance data and blind discovery of potential markers by PEPPeR 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*, and its ratio did in fact change between the two Variability Mixes. 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 1 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 the concentration of a protein 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.

*Acknowledgments*—We thank Dr. Vamsi Mootha and Betty Chang for the kind gift of mitochondrial protein extract. We also thank members of the Carr laboratory for performing protein digests for this study and especially Eric Kuhn for data on β-galactosidase contamination.

* This work was supported by grants from The Gates Foundation and the Entertainment Industry Foundation (to S. A. C.) and the United States Department of Energy (DOE-GTL) (to G. M. C.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

¶ To whom correspondence should be addressed. Tel.: 617-324-9762; Fax: 617-252-1902; E-mail: scarr@broad.mit.edu.

REFERENCES

1. MacCoss, M. J., and Matthews, D. E. (2005) Quantitative MS for proteomics: teaching a new dog old tricks. *Anal. Chem.* **77,** 294A–302A
2. Ong, S. E., and Mann, M. (2005) Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **1,** 252–262
3. Gillette, M. A., Mani, D. R., and Carr, S. A. (2005) Place of pattern in proteomic biomarker discovery. *J. Proteome Res.* **4,** 1143–1154
4. Anderson, N. L., and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1,** 845–867
5. Shen, Y., Moore, R. J., Zhao, R., Blonder, J., Auberry, D. L., Masselon, C., Pasa-Tolic, L., Hixson, K. K., Auberry, K. J., and Smith, R. D. (2003) High-efficiency on-line solid-phase extraction coupling to 15–150-μm-i.d. column liquid chromatography for proteomic analysis. *Anal. Chem.* **75,** 3596–3605
6. Adkins, J. N., Varnum, S. M., Auberry, K. J., Moore, R. J., Angell, N. H., Smith, R. D., Springer, D. L., and Pounds, J. G. (2002) Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. *Mol. Cell. Proteomics* **1,** 947–955
7. Tirumalai, R. S., Chan, K. C., Prieto, D. A., Issaq, H. J., Conrads, T. P., and Veenstra, T. D. (2003) Characterization of the low molecular weight human serum proteome. *Mol. Cell. Proteomics* **2,** 1096–1103
8. Shen, Y., Jacobs, J. M., Camp, D. G., II, Fang, R., Moore, R. J., Smith, R. D., Xiao, W., Davis, R. W., and Tompkins, R. G. (2004) Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome. *Anal. Chem.* **76,** 1134–1144
9. Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359,** 572–577
10. Petricoin, E. F., Zoon, K. C., Kohn, E. C., Barrett, J. C., and Liotta, L. A. (2002) Clinical proteomics: translating benchside promise into bedside reality. *Nat. Rev. Drug Discov.* **1,** 683–695
11. Villanueva, J., Philip, J., Entenberg, D., Chaparro, C. A., Tanwar, M. K., Holland, E. C., and Tempst, P. (2004) Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry. *Anal. Chem.* **76,** 1560–1570
12. Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T. A., Hill, L. R., Norton, S., Kumar, P., Anderle, M., and Becker, C. H. (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* **75,** 4818–4826
13. Zimmer, J. S., Monroe, M. E., Qian, W. J., and Smith, R. D. (2006) Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrom. Rev.* **25,** 450–482
14. Page, J. S., Masselon, C. D., and Smith, R. D. (2004) FTICR mass spectrometry for qualitative and quantitative bioanalyses. *Curr. Opin. Biotechnol.* **15,** 3–11
15. Eng, J. K., McCormack, A. L., and Yates, J. R., III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989
16. Schutz, F., Kapp, E. A., Simpson, R. J., and Speed, T. P. (2003) Deriving statistical models for predicting peptide tandem MS product ion intensities. *Biochem. Soc. Trans.* **31,** 1479–1483

17. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392
18. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75,** 4646–4658
19. Ulintz, P. J., Zhu, J., Qin, Z. S., and Andrews, P. C. (2006) Improved classification of mass spectrometry database search results using newer machine learning approaches. *Mol. Cell. Proteomics* **5,** 497–509
20. Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A., and Yates, J. R. (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **1,** 39–45
21. Washburn, M. P., Ulaszek, R. R., and Yates, J. R., III (2003) Reproducibility of quantitative proteomic analyses of complex biological mixtures by multidimensional protein identification technology. *Anal. Chem.* **75,** 5054–5061
22. Tang, H. Y., Ali-Khan, N., Echan, L. A., Levenkova, N., Rux, J. J., and Speicher, D. W. (2005) A novel four-dimensional strategy combining protein and peptide separation methods enables detection of low-abundance proteins in human plasma and serum proteomes. *Proteomics* **5,** 3329–3342
23. Listgarten, J., and Emili, A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **4,** 419–434
24. Hilario, M., Kalousis, A., Pellegrini, C., and Muller, M. (2006) Processing and classification of protein mass spectra. *Mass Spectrom. Rev.* **25,** 409–449
25. Strittmatter, E. F., Rodriguez, N., and Smith, R. D. (2003) High mass measurement accuracy determination for proteomics using multivariate regression fitting: application to electrospray ionization time-of-flight mass spectrometry. *Anal. Chem.* **75,** 460–468
26. Lipton, M. S., Pasa-Tolic, L., Anderson, G. A., Anderson, D. J., Auberry, D. L., Battista, J. R., Daly, M. J., Fredrickson, J., Hixson, K. K., Kostandarithes, H., Masselon, C., Markillie, L. M., Moore, R. J., Romine, M. F., Shen, Y., Stritmatter, E., Tolic, N., Udseth, H. R., Venkateswaran, A., Wong, K. K., Zhao, R., and Smith, R. D. (2002) Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 11049–11054
27. Prakash, A., Mallick, P., Whiteaker, J., Zhang, H., Paulovich, A., Flory, M., Lee, H., Aebersold, R., and Schwikowski, B. (2006) Signal maps for mass spectrometry-based comparative proteomics. *Mol. Cell. Proteomics* **5,** 423–432
28. Radulovic, D., Jelveh, S., Ryu, S., Hamilton, T. G., Foss, E., Mao, Y., and Emili, A. (2004) Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **3,** 984–997
29. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **78,** 779–787
30. Silva, J. C., Denny, R., Dorschel, C. A., Gorenstein, M., Kass, I. J., Li, G. Z., McKenna, T., Nold, M. J., Richardson, K., Young, P., and Geromanos, S. (2005) Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* **77,** 2187–2200
31. Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P., and Geromanos, S. J. (2006) Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **5,** 144–156
32. Aach, J., and Church, G. M. (2001) Aligning gene expression time series with time warping algorithms. *Bioinformatics* **17,** 495–508
33. Calvo, S., Jain, M., Xie, X., Sheth, S. A., Chang, B., Goldberger, O. A., Spinazzola, A., Zeviani, M., Carr, S. A., and Mootha, V. K. (2006) Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat. Genet.* **38,** 576–582
34. Mootha, V. K., Bunkenborg, J., Olsen, J. V., Hjerrild, M., Wisniewski, J. R., Stahl, E., Bolouri, M. S., Ray, H. N., Sihag, S., Kamal, M., Patterson, N., Lander, E. S., and Mann, M. (2003) Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* **115,** 629–640
35. Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4,** 1985–1988
36. Leptos, K. C., Sarracino, D. A., Jaffe, J. D., Krastins, B., and Church, G. M. (2006) MapQuant: open-source software for large-scale protein quantification. *Proteomics* **6,** 1770–1782
37. Banfield, J. D., and Raftery, A. E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49,** 803–821
38. Fraley, C., and Raftery, A. E. (1998) MCLUST: software for model-based cluster and discriminant analysis, in *Technical Report No. 342*, University of Washington, Seattle, WA
39. Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003) *Bayesian Data Analysis*, CRC Press, Boca Raton, FL
40. Kass, R. E., and Raftery, A. E. (1995) Bayes Factors. *Journal of the American Statistical Association* **90,** 773–795
41. R Development Core Team (2005) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria
42. Sokal, R. R., and Rohlf, F. J. (1995) *Biometry*, 3rd Ed., p. 887, W. H. Freeman and Co., New York
43. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537
44. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57,** 289–300

# Additions & Corrections

## Vol. 5 (2006) 1927–1941

### PEPPeR: A Platform for Experimental Proteomic Pattern Recognition

*Jacob D. Jaffe, D. R. Mani, Kyriacos C. Leptos, George M. Church, Michael A. Gillette, and Steven A. Carr*

On Page 1940: Acknowledgments: The grant information should be listed as follows: "This work was supported by grants from the National Institutes of Health (Grant NCI R01 CA126219 to D. R. M., as part of NCI Clinical Proteomic Technologies for Cancer Program), The Gates Foundation, and The Entertainment Industry Foundation (to S. A. C.), and the United States Department of Energy (DOE-GTL) (to G. M. C.)."

## Vol. 7 (2008) 1810–1823

### Proteomic Identification of Cyclophilin A as a Potential Prognostic Factor and Therapeutic Target in Endometrial Carcinoma

*Zhengyu Li, Xia Zhao, Shujun Bai, Zhi Wang, Lijuan Chen, Yuquan Wei, and Canhua Huang*

On Page 1810, the correct affiliation for the authors should be: The Department of Gynecology and Obstetrics, West China Second Hospital and the State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu 610041, China.

We suggest that subscribers photocopy these corrections and insert the photocopies at the appropriate places where the article to be corrected originally appeared. Authors are urged to introduce these corrections into any reprints they distribute. Secondary (abstract) services are urged to carry notice of these corrections as prominently as they carried the original abstracts.