RESEARCH ARTICLE

MapQuant: Open-source software for large-scale protein quantification

Kyriacos C. Leptos[1], David A. Sarracino[2], Jacob D. Jaffe[1], Bryan Krastins[2] and George M. Church[1]***

[1] Harvard Medical School, Department of Genetics, Boston, MA, USA
[2] Harvard Partners Center for Genetics and Genomics, Cambridge, MA, USA

Whole-cell protein quantification using MS has proven to be a challenging task. Detection efficiency varies significantly from peptide to peptide, molecular identities are not evident *a priori*, and peptides are dispersed unevenly throughout the multidimensional data space. To overcome these challenges we developed an open-source software package, MapQuant, to quantify comprehensively organic species detected in large MS datasets. MapQuant treats an LC/MS experiment as an image and utilizes standard image processing techniques to perform noise filtering, watershed segmentation, peak finding, peak fitting, peak clustering, charge-state determination and carbon-content estimation. MapQuant reports abundance values that respond linearly with the amount of sample analyzed on both low- and high-resolution instruments (over a 1000-fold dynamic range). Background noise added to a sample, either as a medium-complexity peptide mixture or as a high-complexity trypsinized proteome, exerts negligible effects on the abundance values reported by MapQuant and with coefficients of variance comparable to other methods. Finally, MapQuant's ability to define accurate mass and retention time features of isotopic clusters on a high-resolution mass spectrometer can increase protein sequence coverage by assigning sequence identities to observed isotopic clusters without corresponding MS/MS data.

Received: April 2, 2005
Revised: August 23, 2005
Accepted: September 6, 2005

Keywords:
Computer program / Mass spectrometry / Quantitative analysis

1 Introduction

Knowing the quantities of proteins in a biological system is crucial to understanding post-transcriptional events [1, 2] including translational efficiency, post-translational modifications, and turnover. Now that whole-cell proteome analysis has become routine [3, 4], the need for protein quantification software has become increasingly apparent. While earlier methods involved the quantification of excised 2-D protein gel spots [1, 5], current methods employ chromatographic separation methods coupled to MS. The latter are almost exclusively relative quantification methods and require simultaneous injections of the samples to be compared into the spectrometer and they involve some sort of stable-isotope labeling [6, 7]. Recent studies, however, have shown that relative quantification can be carried out as separate injections with spiked in standards [8]. Although chromatographic separation methods coupled to MS have proven to be more easily automated, the identification and quantification of the signal acquired from tryptic peptides of a small bacterial proteome, comprising $\sim 10^5$ isotopic clusters has proven to be a desirable, but problematic goal. This

Correspondence: Kyriacos C. Leptos, Room 238, The New Research Building, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA
E-mail: leptos@fas.harvard.edu
Fax: +1-617-432-6513

Abbreviations: UPZV, unique peptide charge-variants; ***xcorr***, cross-correlation

* Current address: The Broad Institute of MIT and Harvard, Cambridge, MA, USA
** Second corresponding author: Professor George M. Church, address details at http://arep.med.harvard.edu/gmc

is because the detection efficiency varies significantly from peptide to peptide, molecular identities are not evident *a priori*, and peptides are dispersed unevenly throughout the multidimensional separations. Currently available quantification software packages either are driven by sequencing data (http://msquant.sourceforge.net), or fail to approach quantification in a systematic manner that addresses overlapping peaks and intertwined isotopic clusters [9].

Accordingly, we have developed open-source software, MapQuant, which, given large amounts of MS datasets, is designed to quantify as many organic species in the sample as possible. In this study, we show that MapQuant can quantify tryptic peptides of single-protein, medium-complexity, and proteome-complexity samples. Issues of linear response, ionization suppression, protein coverage and variance across replicates are addressed.

2 Materials and methods

2.1 Data acquisition

The tryptic peptide samples used in this study were purchased from Michrom BioResources and included BSA (PTD/00001/15), chicken conalbumin (PTD/00001/21), bovine lactoperoxidase (PTD/00001/27), *Eschericchia coli* tryptophanase (PTD/00001/38), human acid glycoprotein (PTD/00001/41), rabbit aldolase (PTD/00001/45), and yeast phosphoglucose isomerase (PTD/00001/50). The angiotensin mixture used was also purchased from Michrom BioResources (910/00002/02). The tryptic digests were diluted in 95% water/5% ACN/0.1% formic acid.

For the LC/MS experiments performed on the low-resolution mass spectrometer, samples were subjected to nanoflow RP chromatography coupled to mass spectrometric detection. The HPLC system consisted of a gradient pump (ThermoElectron, Waltham, MA), and autosampler (LC Packings; San Francisco, CA) and an LCQ Deca XP+ IT mass spectrometer (ThermoElectron). The column was a laser-pulled fused silica capillary (75 μm id) packed in-house with 15 cm of Magic C_{18} (5 μm 200 Å AQ-type) resin. For the RP chromatography, buffer A was 0.1% formic acid in HPLC gradewater (Burdick and Jackson) and buffer B was 0.1% formic acid in HPLC/ACS grade ACN (Burdick and Jackson). The chromatographic gradient employed was linear: from 5–35% B over 130 min with a flow rate of 85 μL/min and from 35–95% B over 30 min with a flow rate of 125 μL/min. For the calibration dataset, 1, 3.3, 10, 33, 100, 333 and 1000 fmoles of BSA tryptic peptides were used. The dataset addressing ionization suppression was acquired on the same low-resolution mass spectrometer mentioned above; each data point contained 100 fmoles of BSA tryptic peptides with increasing amounts of a mixture containing tryptic peptides from chicken conalbumin, bovine lactoperoxidase, *E. coli* tryptophanase, human acid glycoprotein, rabbit aldolase, and yeast phosphoglucose isomerase, all in equimolar concentrations. In the low-resolution experiments, signal acquisition for each data point was carried out four times: three times in full profile mode (m/z interval = 0.067) with a signal acquisition method that included one MS/MS scan per MS scan, for quantification purposes (q-experiments), and once in centroid mode with a signal acquisition method that included five MS/MS scans per MS scan, for sequencing purposes (s-experiments). The LC/MS experiments on the low-resolution spectrometer gave rise to ~3500 MS scans of data acquisition per injection.

For the LC/MS experiments on the high-resolution mass spectrometer, samples were subjected to a similar procedure as above except that the chromatography column was changed to 125 μm id and packed with 18 cm of RP material and a steeper liner chromatography gradient was employed: from 9–33% B over 50 min with a flow rate of 90 μL/min and from 33–100% B over 1 min with a flow rate of 180 μL/min. Buffer A was 3% ACN/0.1% formic acid and buffer B was 95% ACN/0.1% formic acid. The spectrometer used was a hybrid linear IT/FTICR mass spectrometer (LTQ-FT; ThermoElectron). For the calibration dataset, 0.2, 0.66, 2, 6.6, 20, 66.6 and 200 fmoles of BSA tryptic peptides were used. The signal acquisition method for each data point was carried out in triplicate and included the acquisition of two MS/MS dependent scans in centroid mode for each MS scan in full profile mode (average m/z interval = 0.0048). The files acquired were used for both quantification and sequencing purposes (qs-experiments). Given the steeper chromatographic gradients employed, these LC/MS experiments gave rise to ~1300 MS scans of data acquisition per injection.

All calibration experiments were run from low to high concentrations to minimize carryover effects, and dynamic exclusion was employed in experiments where MS/MS scans were acquired to reduce redundancy in MS/MS data.

2.2 Extraction of LC/MS data

We used an application programming interface provided by the manufacturer of the mass spectrometer to extract the LC/MS data into a novel data structure that we termed OpenRaw. Briefly, an OpenRaw data structure contains archives of 'full scan' mass spectra (MS), higher order CID spectra (MS^n), and global information about the specific experiment. A more detailed description of this platform-independent file format can be found in Suppl. Text 1 and Suppl. Fig. 1.

2.3 Data analysis using MapQuant

2.3.1 Overview of MapQuant

MapQuant is a program designed to isolate unique organic species and quantify their relative abundances from an LC/MS experiment. In this study, we propose a novel quantification method to perform this analysis. Data from an LC/MS experiment is analyzed after being formatted into a data structure called a 2-D map, analogous to a grayscale image. A

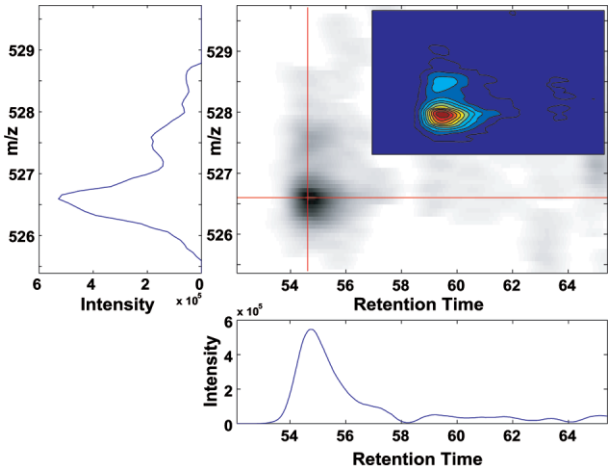

Figure 1. A detailed look at an isotopic cluster that was acquired on a low-resolution spectrometer (LCQ) as it is visualized as a 2-D map. The x-axis and y-axis of this noise-filtered 2-D map represent the chromatography and MS dimensions, respectively. Sections of the 2-D map, such as relevant mass spectrum and mass chromatogram are also shown (represented as cross-sections). A contour plot of the 2-D map is shown in the inset.

2-D map is stored and manipulated as a matrix whose rows and columns represent scans and *m/z* bins, respectively. The 2-D map of a tryptic peptide from BSA is shown in Fig. 1. The separation dimensions are considered orthogonal since they describe two independent properties of the peptides: mass and hydrophobicity. The advantage of this visualization method is that the experimentalist gets a global view of the species present in the sample. Although the concept of a 2-D map is not novel, it has only just started to be considered as the primary data structure for quantification [9] and visualization [10]. Commercial software packages such as MSView [8], Spectromania™ [11] and MosaiquesVisu [12] and the free software MSight [13] are available for dealing with large numbers of mass spectrometric data, however the methods employed are either not transparent enough or not open-source.

Given MS data in OpenRaw file format, MapQuant outputs a list of candidate organic species and their integrated signal abundances. A simple analogy is to traditional chromatography peak integration algorithms, except that MapQuant works in three dimensions (time, *m/z*, and intensity) and is designed specifically for the concerns unique to mass spectral data at various levels of resolution and accuracy. The following steps are implemented in order to achieve the above goal: (1) smoothing by convolution, (2) watershed segmentation, (3) peak finding and peak fitting, (4) peak clustering, and (5) peak refining (*i.e.* deconvolving by fitting and subtracting) and deisotoping.

Algorithms were implemented as MapQuant functions using ANSI C and a command-line user interface (*MQParser*) was developed using *bison* (http://www.gnu.org/software/bison/bison.html). MapQuant functions can be assembled into scripts, readable by the MQParser. All MQParser function syntax is documented at http://arep.med.harvard.edu/mqparser_functions.html. MapQuant also includes 2-D map, mass spectrum, and mass chromatogram visualization capabilities.

2.3.2 Definitions and data structures

Experiment is the data structure that holds information about a single LC/MS experiment, including the associated retention times of its constituent scans, and the mass spectrometer's *m/z* sampling capabilities.

Scan is the sampling unit in the chromatography dimension, referring to one mass spectrum. Each scan has an associated retention time.

Mass bin is the sampling unit of the mass spectrometer when measuring the *m/z* of the produced ions.

2-D map is a matrix whose rows and columns coincide with the intensity values of *mass spectra* and *mass chromatograms* of the LC/MS experiment at particular retention time values and *m/z* values, respectively (Fig. 2).

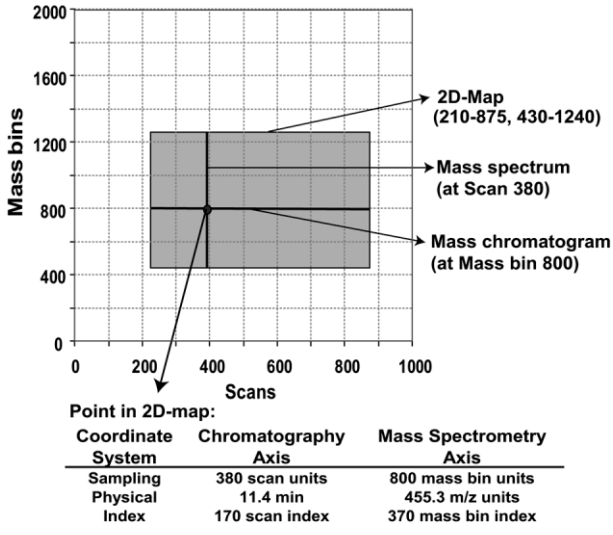

Figure 2. Illustration of the definitions surrounding the concept of a 2-D map. As seen in the figure, a 2-D map can also describe a fraction of an experiment, indicated by the shaded rectangle. A 2-D map is defined by scan boundaries (*e.g.* 210–875) and by mass bin boundaries (*e.g.* 430–1240). Any column of a 2-D map is defined as a mass spectrum at a particular scan, and any row is defined as a mass chromatogram at a particular mass bin. Positions of data points in a 2-D map can be addressed in three different ways: (a) Using sampling coordinates, where position is given in scan and mass bin units that refer to the experiment as a whole, (b) using physical coordinates, where position in the 2-D map is described in time units and *m/z* units, and (c) using index coordinates, where position is given as the indices of the matrix that describes the corresponding 2-D map. Sampling and index coordinates are important in the description of the implementation of the algorithms used.

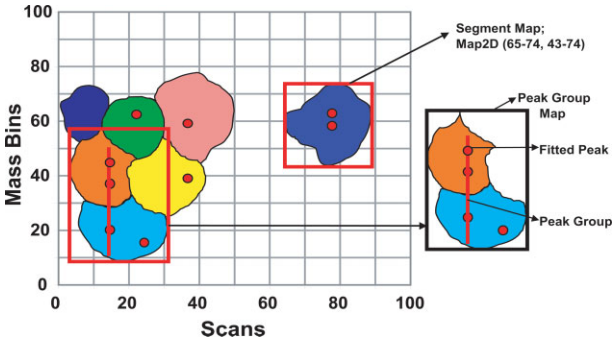

Figure 3. Illustration of data structures and concepts required for the understanding of the algorithms. A segment map is a 2-D map that contains all the data points belonging to a segment as a result of performing the operation of watershed segmentation on a parent 2-D map. A peak group is defined as a cluster of fitted peaks (centroids shown as red circles) that can represent candidate co-eluting isotopic clusters. A peak group map is the minimum 2-D map needed for fitting the estimated number of isotopic clusters that a peak group might contain. A peak group map may also contain "extra" peaks that do not belong to the corresponding peak group. This can happen if such fitted peaks (*e.g.* peak in the blue segment map and on the right of the peak group) slightly overlap with any of the peaks in the corresponding peak group and thus forced in the same segment. Moreover, peak groups may contain peaks that are unevenly spaced, as shown in the figure, indicating the presence of more than one isotopic cluster in the corresponding peak group.

Segment map is defined as a region of a parent 2-D map upon which the operation of segmentation was performed. Segmentation is performed to partition the map signal into tractable segments (Fig. 3).

Peak is a local maximum in the 2-D map.

Fitted peak (*FPeak*) is a peak that has been fitted to a particular mathematical model, *e.g.* a 2-D Gaussian.

Isotopic cluster is a group of peaks that represents the isotopic variants of a molecular species.

Peak group is the cluster of fitted peaks that represents candidate co-eluting isotopic clusters (Fig. 3).

Peak group map is the minimal 2-D map needed to fit the estimated number of isotopic clusters that a peak group might contain (Fig. 3).

2.3.3 Algorithms

2.3.3.1 Smoothing by convolution and other noise reduction algorithms

Since MS data are usually quite noisy, especially in the chromatography dimension, noise filters were applied. More specifically, smoothing algorithms were applied to facilitate the detection of all local maxima (peaks) found in the 2-D map. Smoothing was implemented using convolution [14]. Preset convolution functions that can be applied by MapQuant include box-car, Gaussian, and Savitzky-Golay [14]. In this study for the low-resolution spectrometer a Gaussian filter was applied in the retention time dimension. The width of the Gaussian filter was chosen to have approximately the same SD as the average SD of the peaks observed. One way to estimate the average standard deviation of peaks is to run MapQuant in the interactive mode. The average SD of peaks can be safely assumed not to change significantly among LC/MS runs of the same dataset since it is highly dependent on the chromatographic gradient. Additionally, morphological image operations such as opening and closing [15] were also used for noise filtering.

2.3.3.2 Watershed segmentation

Because fitting all peaks in a 2-D map simultaneously is computationally too expensive, we segmented the map using the watershed segmentation algorithm [16]. The function implementing this algorithm returns a 2-D labeled non-grayscale map that has the form of a mosaic, which, along with the noise-filtered 2-D-map from the previous step and information about the rectangular circumscribed boundaries of the segment (Fig. 3), can be used to cut out a so-called segment map (Fig. 4). Peaks that are well resolved are confined into individual segment maps whereas overlapping peaks are confined into common segments. The latter is possible through a morphological opening operation of the noise-filtered 2-D-map prior to segmentation. Confining overlapping peaks to the same segment is required for downstream peak fitting since peaks are fitted simultaneously in order to get an accurate value on their abundance (Suppl. Fig. 2).

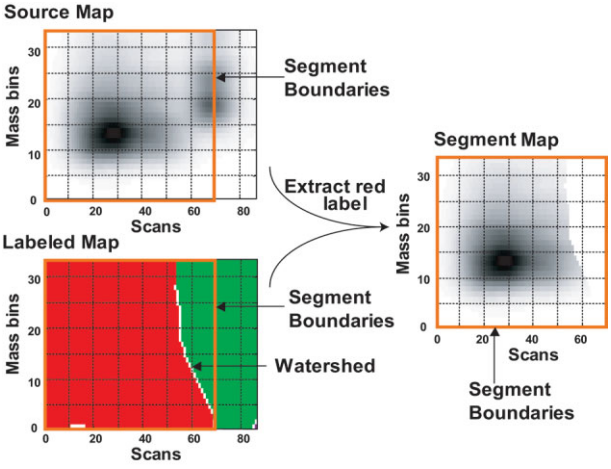

Figure 4. Operation of watershed segmentation on a 2-D map. This algorithm is utilized to divide the noise-filtered map in non-overlapping regions so that fitting individual peaks becomes less computationally intensive. The product of segmentation is a 2-D map called labeled map where each data point is given a segment number, which it belongs to (indicated by different shades). The labeled map can be used as a guiding mask to extract the data points needed for a particular segment, thus creating a segment map as described in Fig. 3.

2.3.3.3 Peak finding and peak fitting

After segmenting the global 2-D map, the goal of peak finding and peak fitting becomes computationally tractable. A peak detection algorithm described below was applied to find local maxima in every segment map. The positions of the local maxima were then used as seeds for the curve-fitting algorithm. The peak detection algorithm uses concepts from mathematical morphology such as the structuring element [15]. A structuring element can be considered a small binary image N that an image operator ^ can take as input along with the image of interest I, resulting in a binary image T as shown in Eq. 1.

$$T = I \hat{} N \tag{1}$$

In structuring element N, the sub-element N_i has a value of 1 if it is to be taken into account in deciding which points neighboring each data point in the 2-D map are to be included in the image operation.

A data point s_k in the 2-D map is considered a local maximum only if t_k is equal to 1 (Eq. 2).

$$t_k = \begin{cases} 1 \ if \left\{ \sum_i \Lambda(s_k, s_k N_i) \right\} = |N|, \\ 0 \ otherwise. \end{cases} \tag{2}$$

Where $\Lambda(p,q) = \begin{cases} 1 \ if \ p \geq q, \\ 0 \ otherwise. \end{cases}$ and $|N| = \sum_i N_i$. To avoid detecting pseudo-peaks due to noise, an abundance threshold was set for all points in the structuring element. MapQuant allows local determination of this threshold based on the mean or median and standard deviation of the 2-D map, with or without considering zero values. The abundance threshold for the datasets in this study was set to the median plus two or three average absolute deviations from the median depending on the spectrometer. Candidate peaks are compiled into a list and are fitted as a sum of curves described by a mathematical equation. In a segment map, if there are n candidate peaks, and if each peak is chosen to be fitted as curve C, then the whole segment-map would be fitted as $\sum_i^n C_i$. We chose to fit each curve with a 2-D Gaussian, referred to from now on as the gaussioid curve (Eq. 3), *i.e.* a bivariate function depending on retention time (r) and *m/z* (m) as described by the curve

$$f(m, r; A, r_0, m_0, \sigma_m, \sigma_r) = \frac{A}{2\pi\sigma_m\sigma_r} e^{-\frac{(r-r_0)^2}{2\sigma_r^2}} e^{-\frac{(m-m_0)^2}{2\sigma_m^2}} \tag{3}$$

There are five parameters to be fitted per peak: abundance (A), retention-time centroid (r_o), *m/z* centroid (m_o), the SD of the Gaussian in the retention time dimension (σ_r), and finally the SD of the Gaussian in the *m/z* dimension (σ_m). The method used for peak fitting is non-linear least squares method [14]; it is a minimization method using steepest descent. It requires knowledge of the first derivative for each of the parameters to be fitted.

Finally, to address asymmetric chromatographic peak profiles the choice of fitting peaks with the exponentially modified Gaussian (EMG) curve in the chromatographic dimension of the 2-D Gaussian was made available. The EMG curve was chosen as it has been shown to be the best analytical fit for asymmetric chromatographic peak profiles [17]. In the examples shown in this study only the simple 2-D Gaussian was chosen for reasons of simplicity since the chromatographic peak profiles were mostly symmetric (Fig. 5b and data not shown).

More refined peak finding and peak fitting was required on segment maps of datasets of the low-resolution spectrometer (LCQ), but not of the high-resolution spectrometer (LTQ-FT). The algorithm employed (Suppl. Text 2) involved an iteration of subtraction and residual fitting based on previously estimated peak widths. The above algorithm depends on knowledge of peak widths that we have calculated from isotopic clusters of known charge. Peak widths on the low-resolution spectrometer were found to be dependent on the charge state of the peptide rather than its *m/z* value. The following values were calculated and fed to the algorithm: For +1 peptides 0.22 ± 0.06, for +2 peptides 0.17 ± 0.03 and for +3 peptides 0.14 ± 0.03 *m/z* units. Peptides with a +4 charge could not be resolved by this method and could not be verified by SEQUEST.

2.3.3.4 Peak clustering

To assign fitted peaks to isotopic clusters we partitioned the peaks into data structures called peak groups using single linkage clustering. Peak groups represent co-eluting peaks and might consist of one or more possible isotopic clusters. The position of peaks belonging to an isotopic cluster is constrained, *i.e.* they must be co-eluting and cannot be separated by >1 *m/z* unit (maximum distance defined by peptides with charge +1). We use these constrains to call peak groups.

2.3.3.5 Deconvolve and fit isotopic carbon peaks

In this algorithm we assume that each peak group represents one or more isotopic clusters. Accordingly, we devised an algorithm for sequential prediction of each isotopic cluster by analyzing fitted peaks based on increasing *m/z*. The peaks that are candidates for an isotopic cluster are fitted using a binomially distributed sum of 2-D Gaussians as a bivariate function of retention time (r) and *m/z* (m).

$$f(m, r; A, r_0, m_0, \sigma_m, \sigma_r, c, z) = A \sum_i \frac{B(i; c, p)}{2\pi\sigma_m\sigma_r} e^{-\frac{(r-r_0)^2}{2\sigma_r^2}} e^{-\frac{\left(m-\left(m_0+\frac{i}{z}\right)\right)^2}{2\sigma_m^2}} \tag{4}$$

The function in Eq. 4 describes a binomially-distributed sum of 2-D Gaussians. The shape of the curve is defined by seven

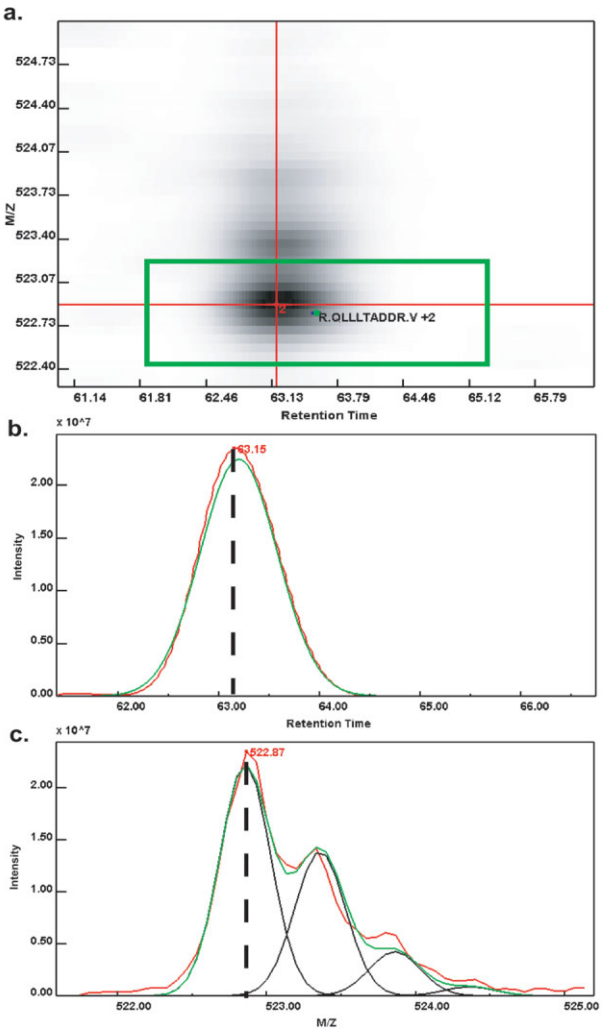

Figure 5. MapQuant view of a magnified area of a 2-D map (after noise filtering) (a) representing the peptide R.QLLLTADDR.V in its +2 state, along with the corresponding mass chromatogram (b) and mass spectrum (c) as defined by the crossing red lines. The isotopic clusters found by MapQuant are plotted as red points accompanied by a number representing their charge state. Sequence identification of the isotopic clusters is carried out by plotting the coordinates of sequenced MS/MS events, for example peptide R.QLLLTADDR.V in its +2 charge state. If there are more than one MS/MS event representing the same scoring sequence, the median retention time coordinate of the peptide is calculated (green point). Around this point a tolerance window (green rectangle) is used in the 2-D map to search for isotopic clusters found inside its boundaries that agree on the charge state and thus assigning an abundance value to a sequence. In the corresponding mass chromatogram (a) and mass spectrum (b) the fitted isotopic cluster is shown in green in contrast to the observed shown in red. The individual isotopic peaks composing the fitted isotopic envelope are shown in black.

parameters, the total abundance of the isotopic cluster (A), the retention-time centroid (r_o) and the *m/z* centroid (m_o) of the monoisotopic peak, the SD of the Gaussian in the retention time dimension (σ_r) and the *m/z* dimension (σ_m), both assumed to be common among the peaks of the isotopic cluster, the charge state of the peptide (z) and finally $B(i;c,p) = \binom{c}{i} p^{c-i}(1-p)^{i}$ which describes the binomial distribution. In the binomial distribution expression c is the total number of carbons in the molecule and p the natural isotopic abundance of carbon-13. The algorithm used in this step can be divided into two parts that are iterated until all fitted peaks are distributed into isotopic clusters:

1a. Guess the most likely subset of fitted peaks in a peak group that can form a potential isotopic cluster. This is the step where possible charge-states are determined (Suppl. Fig. 3).

1b. Substitute those peaks with a binomially distributed gaussioid curve by estimating its carbon content.

2. Refit the peak-group map with Eq. 5 to get a better estimate of the number of carbons and the total abundance of the isotopic cluster. In Eq. 5, m denotes the number of single gaussioid curves C (Eq. 3) that do not belong to any isotopic clusters and n is the number of binomially distributed gaussioid curves B (Eq 4).

$$\sum_{i}^{m} C_i + \sum_{i}^{n} B_i \qquad (5)$$

For high-resolution MS the second step of fitting can be omitted. The fitting of an isotopic cluster representing a BSA tryptic peptide is illustrated in Fig. 5c. Similar algorithms for carbon deconvolution have been reported in the literature [18], but ours uses a tree data structure that enables it to deconvolve isotopes of intertwined isotopic clusters (Suppl. Text 3 and Suppl. Fig. 3). Moreover, the reported (observed) number of carbons for each isotopic cluster was found not to be always the same when compared with the expected, but the deviation was nevertheless consistent (Suppl. Fig. 4).

2.4 Post-MapQuant analysis

To create a validated set of peptide identities, we used the well-established strategy of sequencing peptides by using commercially available software (SEQUEST [19]) on their MS/MS fragmentation pattern. If we had chosen to perform multiple MS/MS scans per MS scan in a single acquisition scheme, we would have limited the number of MS scans acquired, reducing the sampling of data points available for quantification. To circumvent this problem, we collected MS data with five MS/MS spectra per MS scan (s-experiments) and MS data with only one MS/MS spectrum per MS scan (q-experiments). This was not the case for data acquired on the hybrid LTQ-FT instrument, where MS/MS scans can be diverted to its linear IT, thus allowing simultaneous collection of MS and MS/MS scans without reduction in chromatographic sampling of the former. The work flow chart of how we linked the quantitative output from a q-experiment to the identification output of an s-experiment is outlined in Suppl. Fig. 5.

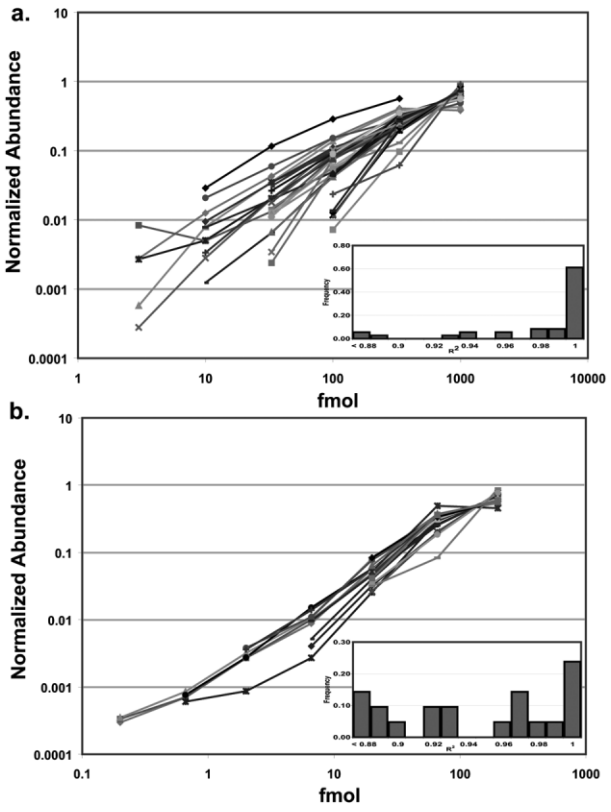

Figure 6. The linear response graph along with insets of the corresponding R^2 distributions are shown for the low-resolution mass spectrometer (a) and the high-resolution mass spectrometer (b). Different data points at each concentration correspond to different peptides, and data points from the same peptide are connected with a line. Only peptides whose sequences were mapped to abundances for three or more data points of the calibration curve were used. The above peptides amount to 36 with a mean R^2 of 0.97 for the low-resolution and 21 with mean R^2 of 0.92 for the high-resolution mass spectrometer. The R^2 values refer to the linear regressions applied. The data are plotted on a logarithmic scale solely for visualization purposes.

To associate a sequenced MS/MS event to a MapQuant isotopic cluster, we exploited the fact that every entry, representing a peptide charge variant, in the SEQUEST summary file (Suppl. Fig. 5) reflects a sequencing event (MS/MS scan) with a unique retention time and *m/z* coordinates in a 2-D map. Sequencing events assigned to the same peptide charge variants are pooled together into a list of unique peptide charge-variants (UPZV) whose new retention time value is calculated as the median of the retention time values of its constituent sequencing events (green point in Fig. 5a). Isotopic clusters that were identified using MapQuant also have centroids, represented by red points with assigned charge states (Fig. 5a). For each median-point calculated for a group of sequencings events, a rectangular area (shown in green in Fig. 5a), called a tolerance window, was searched for possible MapQuant isotopic clusters that matched the charge of the UPZV sequence it represents.

This strategy assumes alignment between q- and s-experiments. To align LC/MS runs we utilized the common SEQUEST-verified peptide identities between the q- and s-experiments. The alignment was achieved by performing either linear or quadratic regression of retention-time values as shown in Suppl. Fig. 6. Regression coefficients can then be fed to the *assignmq*/*assignsq* program (Suppl. Fig. 5). Runs on the high-resolution mass spectrometer (qs-experiments) could be treated both as q-experiments and as s-experiments. For example, for the high-resolution mass spectrometer an initial tolerance window of half size equal to 2 min and 20 ppm in retention time and *m/z* dimensions, respectively, was used to match UPZV from sequencing MS/MS scans of the same q-experiment. The UPZV that matched a single MapQuant isotopic cluster were used to calculate more statistically significant windows in both dimensions that were later used in the assignment of UPZV from other aligned s-experiments.

3 Results

3.1 General remarks

MapQuant is a program that, given raw MS data in profile mode, outputs the features of as many as possible organic species in the sample. Programs and scripts outside MapQuant are used to compile the processed data into tables (Suppl. Fig. 5). The four datasets described in Section 2 were used to develop benchmarks in order to assess MapQuant performance.

3.2 BSA coverage

The BSA sequence used in the study was identified by subjecting MS/MS data acquired to a SEQUEST search against a database composed of nine BSA sequences present in the NCBI nr database (Suppl. Table 1). From the peptides that scored well [cross-correlation (*xcorr*) >2.0], it was evident that the 24-amino acid leading peptide was not present in the mature form of the BSA used in the experiment, implying a protein of 583 amino acids in length. The sequence is shown in Suppl. Fig. 7 and referred to from now on as mBSA-A214T.

SEQUEST was re-run using a protein database composed of mBSA-A214T as well as 27 trypsin sequences and 306 keratin sequences. No proteolytic enzyme specificity was set for database searching, and therefore assignments of peptide sequences that were fully tryptic in nature could be considered with increased confidence despite the increased background posed by the "no enzyme" search. MS/MS spectra were extracted as +1, +2 and +3 charged variants because the scan modalities employed did not allow for precursor charge state determination *a priori*. Moreover, it was set to take into account amino acid modifications such as lysine and arginine carbamylation, methionine oxidation,

Table 1. The distribution of UPZV found by SEQUEST[a)]

1a. UPZV observed in the 1000 fmol data point of the calibration performed on a low-resolution mass spectrometer

Charge	+1	+2	+3	Total
Fully tryptic	15	64	12	91
Partially tryptic	0	23	3	26
Non-tryptic	0	0	0	0
Total	15	87	15	**117**

1b. UPZV observed in the 200 fmol data point of the calibration performed on a high-resolution mass spectrometer

Charge	+1	+2	+3	Total
Fully tryptic	0	43	13	56
Partially tryptic	0	6	0	6
Non-tryptic	0	0	0	0
Total	0	49	13	**62**

a) The SEQUEST results were filtered using the charge and cleavage-specific *xcorr* cutoff values cited in the text.

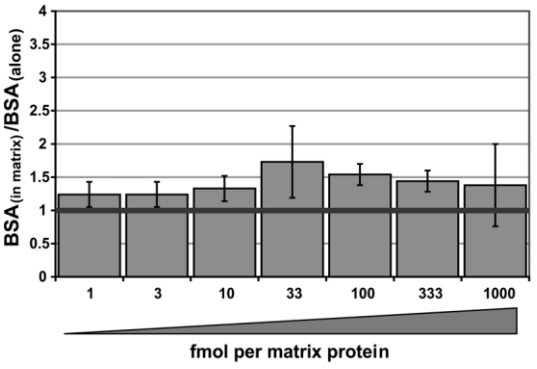

Figure 7. The means and standard deviations of the abundance ratio of BSA peptides in the presence of matrix peptides to BSA peptides alone. The mean ratios for the seven concentrations of protein mixture are 1.24, 1.24, 1.33, 1.73, 1.54, 1.44 and 1.38 for 1, 3.3, 10, 33, 100, 333 and 1000 fmoles of each matrix protein, respectively. A line marking the ratio value of 1 is shown.

and loss of ammonia from N-terminal glutamine. The *xcorr* cutoff values set were charge-specific and protease cleavage site-specific as described in previous studies on the evaluation of the *xcorr* based on false positive rates [20].

For the calibration experiment conducted on the low-resolution mass spectrometer (1–1000 fmol), MapQuant identified 117 UPZV for the 1000 fmol s-experiment with corresponding SEQUEST assignments surpassing the *xcorr* score thresholds. From these, 91 UPZV were fully tryptic at both termini. By observing sequences of non-tryptic peptides, we concluded that an enzyme with chymotrypsin activity was present in the digestion mixture, as 9 of the 26 partially tryptic peptides had phenylalanine, tyrosine or leucine at the C terminus of their cleavage site [21]. However we hypothesize that chymotrypsin activity was attributed to an enzyme that was co-purified with trypsin in a lesser amount (personal communication with vendor). The 117 UPZV cover 80.3% of the amino-acid residues of BSA. With respect to the calibration experiment conducted on the high-resolution mass spectrometer (0.2–200 fmol), the number of UPZV with corresponding SEQUEST assignments was only 62 for the first technical replicate of the 200 fmol data point. These UPZV cover 71% of the total 583 amino-acid residues of BSA.

The large difference in the number of UPZV found between the two calibration experiments can be attributed to the duration of the elution gradient used (4 *vs.* 1 h), the number of MS/MS spectra obtained (many fewer on the LTQ-FT) and the total amount of peptides used (1000 fmol *vs.* 200 fmol). The charge state distributions of the peptides from both calibration experiments are shown in Table 1.

3.3 MapQuant performance

To evaluate MapQuant's performance we estimated the percentage of SEQUEST hits that could be assigned to a MapQuant isotopic cluster. Table 2 shows the percentage coverage of the total SEQUEST-identified peptides in all 21 q-experiments of the low-resolution calibration dataset. Corresponding results for the 21 q-experiments of the high-resolution calibration dataset are shown in Table 3.

Table 2. The number (and percentage) of observed isotopic clusters found by MapQuant out of the total UPZV that can be positively verified by SEQUEST, for the 21 q-experiments of the low-resolution BSA calibration dataset. The table refers to the total number of UPZV observed across all concentrations throughout the dataset

Charge	+1	+2	+3	Total
Fully tryptic	22/27	236/364	3/17	261/408
Partially tryptic	0/0	53/72	0/2	53/74
Non-tryptic	0/0	0/0	0/0	0/0
Total	22/27 (81%)	289/436 (66%)	3/19 (16%)	314/482 (65%)

Table 3. The number (and percentage) of observed isotopic clusters found by MapQuant out of the total UPZV that can be positively verified by SEQUEST, for the for the 21 q-experiments of the high-resolution BSA calibration dataset. The table refers the total number of UPZV observed across concentrations throughout the dataset

Charge	+1	+2	+3	Total
Fully tryptic	6/8	278/336	53/55	337/399
Partially tryptic	0/0	25/26	1/1	26/27
Non-tryptic	0/0	0/0	0/0	0/0
Total	6/8 (75%)	303/362 (84%)	54/56 (96%)	363/426 (85%)

A comparison between the two calibration experiments demonstrates that MapQuant performs better on data from high-resolution instruments since the total number of SEQUEST peptides mapped to MapQuant isotopic clusters is significantly higher (85 *vs.* 65%). It should be noted that in the low-resolution calibration experiment, the program performs better with +2 and +1 peptides (66 and 81%, respectively) than +3 peptides. This bias is likely due to the low resolution of the LCQ spectrometer in profile mode. For example, the average *m/z* bin size was about 1/15 (0.067) *m/z* units wide, meaning that peaks belonging to an isotopic cluster of a +3 peptide would be only 5 mass bins apart given an average peak width of 0.14 *m/z* (2.1 bins). This makes it extremely difficult for any algorithm to resolve these peaks.

It should be noted that the lack of finding an isotopic cluster with the correct charge does not imply that MapQuant did not find any peaks that were in the vicinity of the MS/MS event, since it might have misassigned the charge.

3.4 Additional non-SEQUEST peptides and amino acid modifications

MapQuant is designed to search for and report as many isotopic clusters as possible in a 2-D map. We exploited this comprehensive approach in an attempt to identify peptides based solely on the *m/z*, charge, retention-time, and number of carbons reported by the program. We used the 200 fmol data point from the high-resolution calibration experiment as a case study.

Our strategy taken is outlined in detail in Suppl. Fig. 8. We made use of the program *massfilter* to identify BSA peptides by matching observed *m/z* and *z* values of MapQuant isotopic clusters to all possible values accounting for all possible partially tryptic peptides in BSA including tryptic peptides from known contaminant proteins such as trypsins and keratins. The ppm-tolerance window used was calculated from known peptides and its standard deviation was found to be 2.52 ppm. Furthermore, to increase the confidence of identification by *m/z* and *z* alone a false-discovery rate study was carried out for SEQUEST-verified BSA peptides, the results of which are shown in Suppl. Fig. 9. We thereby identified a total of 381 BSA UPZV for the 200 fmol data point (Suppl. Table 2), a number far greater than obtained using SEQUEST alone, 117 and 62 in the 4-h low-resolution and 1-h high-resolution runs respectively (Table 4). There are two possible reasons for isotopic clusters present on a 2-D map not being identified successfully by SEQUEST. One reason could be that an MS/MS spectrum corresponding to a peptide isotopic cluster is not interpretable by the program. Another reason could be the complete absence of MS/MS spectra for a peptide isotopic cluster due to the difficulty of acquiring MS/MS spectra for a 2-D map too densely populated by peaks, especially when dealing with short run times.

These 381 UPZV increased the sequence coverage of BSA to 98%. Table 5 shows peptides that were found using the above method that share a common N-terminal sequence. The above example demonstrates the ability of MapQuant to identify peptides that were missed by SEQUEST as shown in column 3 of Table 5. The symbol # in this Table represents S-carboxymethylation of cysteine residues and % represents neutral loss of ammonia from N-terminal glutamine, consistent with the formation of pyroglutamate reported in the literature [22]. MapQuant was able to identify the next longer tryptic peptide K.Q%EPERNEC#FLSHKDDSPDLPK.L in its +4 charge state (rows 7); a charge state that cannot be identified by standard sequencing programs, but that is the most abundant of all charge states observed. Moreover we see that the two tryptic peptides

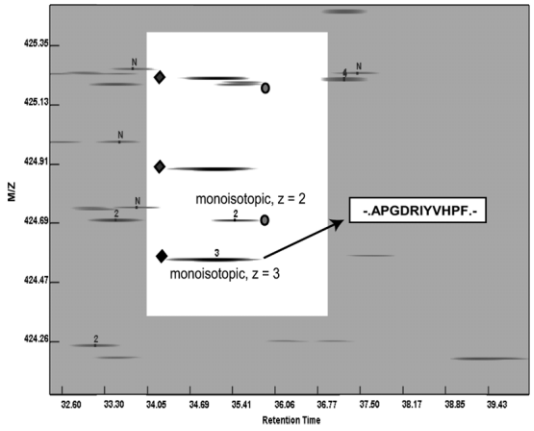

Figure 8. *Needle in a haystack*: finding and quantitating angiotensin peptides among the trypsinized proteome of *Prochlorococcus marinus* MED4 on an FT-ICR. This figure depicts a snapshot of MapQuant zooming in the 2-D map position (after noise filtering) of peptide APGDRIYVHPF at *z* = 3 (region highlighted for clarity). Note that MapQuant accurately derives the charge state and monoisotopic member of entwined isotopic clusters, as demonstrated by its ability to resolve the angiotensin isotopic cluster (marked with diamonds) from the isotopic cluster of another co-eluting species (marked with circles) of similar *m/z* but with *z* = 2.

Table 4. Amino acid protein coverage as calculated using different methods and experimental set up. The length of the chromatography run plays a role to the number of peptides being sequenced as shown by the first two rows of the table. Moreover, using a ppm window calculated from known SEQUEST hits as *m/z* metric for identifying BSA peptides solely on their *m/z*, the sequence percentage coverage could be increased to 98% all the partial tryptic peptides are taken into account

Experiment	Method of Identification	Number of unique peptide-charge-variants	Amino-acid sequence coverage
1000 fmol – 4 h	SEQUEST	117	80%
200 fmol – 1 h	SEQUEST	62	71%
200 fmol – 1 h	MASS-FILTER/SEQUEST	381	98%

Table 5. The diversity of peptides that share a common N terminus and are found by MapQuant among replicates of the 200-fmol BSA calibration data point

	Sequence	Charge	SEQUEST identification[a)]	Retention time	Abundance	Peptide type
1	K.Q%EPERNEC#FLSHK.D	2	3/3	26.45 ± 0.27	85.42 ± 26.50	Fully tryptic
2	K.Q%EPERNEC#FLSHK.D	3	3/3	26.45 ± 0.27	262.83 ± 91.66	Fully tryptic
3	K.Q%EPERNEC#FLSHKD.D	3	0/3	26.53 ± 0.26	6.30 ± 1.61	Tryptic/aspartate
4	K.Q%EPERNEC#FLSHKDD.S	3	0/3	26.72 ± 0.18	22.67 ± 4.94	Tryptic/aspartate
5	K.Q%EPERNEC#FLSHKDDSPD.L	3	0/2	27.44 ± 0.09	11.78 ± 1.83	Tryptic/aspartate
6	K.Q%EPERNEC#FLSHKDDSPDLPK.L	3	3/3	31.06 ± 0.16	496.97 ± 83.48	Fully tryptic
7	K.Q%EPERNEC#FLSHKDDSPDLPK.L	4	0/3	31.05 ± 0.14	865.77 ± 218.58	Fully tryptic
8	K.QEPERNEC#FLSHK.D	3	0/3	22.96 ± 0.12	13.68 ± 4.44	Fully tryptic
9	K.QEPERNEC#FLSHK.D	4	0/3	22.95 ± 0.12	2.73 ± 0.44	Fully tryptic
10	K.QEPERNEC#FLSHKDDSPDLPK.L	3	0/2	27.46 ± 0.08	11.74 ± 3.59	Fully tryptic
11	K.QEPERNEC#FLSHKDDSPDLPK.L	4	0/3	27.48 ± 0.07	29.22 ± 6.48	Fully Tryptic
12	K.QEPERNEC#FLSHKDDSPDLPK.L	5	0/3	27.49 ± 0.09	6.74 ± 1.01	Fully Tryptic

a) Number of replicates in which SEQUEST identified the corresponding UPZV in comparison with MapQuant and massfilter.

mentioned above are also found in their non-pyroglutamate forms (rows 8, 9 and 10–12, respectively), albeit in a much lower abundance. Another interesting feature of the sequences in Table 5 is the presence of non-tryptic peptides (rows 3–5); these peptides maintain an N-terminal tryptic cleavage site but all have an aspartate C-terminal cleavage site reflecting either caspase activity or aspartate-activated autoproteolysis [23]. The above observation is corroborated by the observation of other SEQUEST-identified peptides having non-tryptic cleavage sites of the form D.X, where X is any amino-acid residue.

We were also interested in discovering possible peptide modifications. Among the 381 peptide charge-variants we focused on the following modifications: S-carboxymethylation of cysteines (due to preparation in iodoacetic acid), oxidation of methionine and histidine, carbamylation of lysine and arginine, and the neutral loss of ammonia (Table 5). With regard to carbamylation, SEQUEST indicated (Table 6, row 1) that BSA can be carbamylated at lysine-211. Table 6 also provides further information that this carbamylation site, indicated by *, is corroborated by carbamylated peptides identified by *massfilter* that contain sequences that run both upstream (rows 2, 3) and downstream (row 4) of the SEQUEST-identified peptide. Finally, SEQUEST results indicated that lysine-548 was carbamylated. Lysine-548 is also known to be glycated [24], indicating a sequence hot spot for attack by acidic molecules in the blood stream.

3.5 Linear response

We assessed the range of linear response for the two kinds of spectrometers. Although the results pertain to the particular instruments used in this study, our long-term goal is to be able to use the BSA tryptic peptide mix as a calibration standard, either internal or external, for all studies.

We used a linear model to fit the data points for the two calibration series (1–1000 fmol and 0.2–200 fmol). We used the equation $y = Ax + b$, where y is the median abundance of the isotopic clusters found by MapQuant, and x is the corresponding amount in fmoles injected into the mass spectrometer. We calculated the correlation coefficient R^2 for the linear response of each peptide that had at least three data points mapped to MapQuant isotopic clusters (Fig. 6). The correlation coefficient R^2 was chosen as a linearity metric because it is sensitive to outliers when only a few data points are available. The number of data points was limited by

Table 6. Peptides found by MapQuant corroborate the carbamylation site of lysine-211 detected by the UPZV found by SEQUEST (row 1)

	Sequence	Charge	SEQUEST identification	Retention time	Abundance
1	R.EK*VLTSSAR.Q	2	1/1	21.54	2.72
2	R.EK*VLTSSARQR.L	2	0/3	31.88 ± 0.50	68.08±21.92
3	R.EK*VLTSSARQR.L	3	0/3	31.88 ± 0.51	5.42± 0.44
4	K.IETMREK*.V	2	0/3	30.45 ± 0.34	4.23± 0.20

MapQuant performance, the number of MS/MS spectra acquired and the differential ionization efficiency of the BSA peptides. The distributions of R^2 are shown in the insets of Fig. 6. The mean R^2 for the peptides detected on the low-resolution mass spectrometer was 0.97 ($n = 36$) and the high-resolution mass spectrometer was 0.92 ($n = 21$). Any deviations from linearity can be attributed to saturation related effects.

3.6 Ionization suppression in a medium-complexity matrix

To address the issue of matrix effects on the tryptic peptides of a single protein, we used MapQuant to quantify SEQUEST-identified peptides of BSA in the presence of varying concentrations of tryptic peptides from six other proteins, referred to as matrix peptides (see Section 2). We chose 100 fmoles of BSA because it gave a reasonable number of sequence identities for comparison between different matrix conditions. Figure 7 shows the medians and median absolute deviations from the median of the abundance ratios of BSA peptides in the presence of matrix peptides to BSA peptides alone ($BSA_{(in\ matrix)}/BSA_{(alone)}$). The median ratios for the seven concentrations of protein mixture are 1.24, 1.24, 1.33, 1.73, 1.54, 1.44 and 1.38 for 1, 3.3, 10, 33, 100, 333 and 1000 fmoles of each matrix protein respectively. Thus, we do not observe any strong ionization suppression effects, although we can hypothesize that deviations of ratios above the value of 1 can be attributed to limitations in accurate volume transfer. The abundance of a few BSA and matrix peptides across different concentrations can be found in Suppl. Fig. 10.

3.7 MapQuant performance on a proteomic sample

3.7.1 Identification of minor components

To assess the performance of MapQuant at finding and identifying peptides in a complex mixture, we collected data of triplicate injections on a linear IT/FTICR mass spectrometer of the trypsinized proteome of the cyanobacterium *Prochlorococcus marinus* MED4, sampled at 25 different time points during its daily life cycle. An estimate of 5.7 μg of peptides from the trypsinized proteome of *P. marinus* were mixed with five angiotensin peptides (0.12 ng each) that were spiked in to each sample at a constant level across all samples. In this study we present the findings of the analysis for six time points across the cell cycle as summarized in Suppl. Table 3. MapQuant was able to find 340/375 (91%) expected isotopic clusters of angiotensins. The *m/z* metric for identification used was the ppm window calculated from the BSA study on the same instrument. In each LC/MS experiment MapQuant identified between 15 000 and 20 000 isotopic clusters, yet was easily able to reproducibly identify and quantify these five peptides despite the fact they accounted for less than 0.1% of all of the isotopic clusters

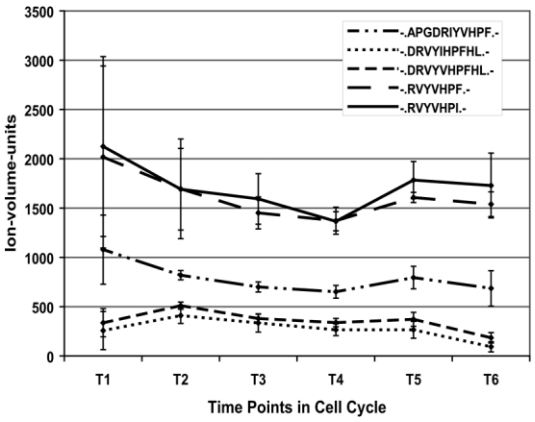

Figure 9. The quantification of the five angiotensin peptides that were spiked in the trypsinized proteome of *P. marinus* MED4 at six different time points of its daily cycle (T1 = 6 h, T2 = 14 h, T3 = 28 h, T4 = 30 h, T5 = 40 h, T6 = 44 h). The injections were done in triplicate and the mean and SD are shown in the graph.

detected. MapQuant can deconvolve peptides whose isotopic envelopes are intertwined, as shown in the area left of the angiotensin peptide -.APGDRIYVHPF.- in Fig. 8. The mean and the standard deviation of the abundances of the five angiotensin peptides are plotted for six time points of the cell cycle in Fig. 9.

3.7.2 Reproducibility of quantification

To assess the reproducibility of quantification in a proteomic-scale sample we calculated coefficients of variance (CV) for 1663 SEQUEST-identified *P. marinus* peptides for the second time point shown in Fig. 9 (T = 14 h). The CV values of these peptides are shown in Suppl. Fig. 11 as a function of their corresponding abundance. The mean and median of the distribution of CV were calculated to be 0.37 and 0.16, respectively. These values were also comparable with the ones cited in the literature [9].

4 Discussion

We developed the software package MapQuant in order to quantify the relative abundances of thousands of peptides (or any organic species) in parallel across multiple experimental conditions by using mass spectrometric techniques. MapQuant mainly addresses the problem of isotopic cluster "feature detection" in MS/LC data. Through use of image processing techniques we are able to simplify this process and make it extremely reliable. The MapQuant architecture allows for user-defined combinations of operations that give an analyst total control of the feature detection steps performed; we have outlined a basic framework that we believe will be useful for a generic LC/MS experi-

ment. While we expect that MapQuant will be most useful with high-resolution/high-mass accuracy mass spectrometers (*i.e.* LTQ-FT), we have demonstrated that is useful for more common forms of instrumentation, and we are currently extending MapQuant's capabilities in dealing with "centroid" data acquired on intermediate-resolution mass spectrometers.

Using MapQuant, we have demonstrated that LC/MS should be considered as a valid platform for massively parallel quantification of peptides in a proteome-scale sample. We have demonstrated the linearity of response, within the dynamic range of at least two common LC/MS platforms (quadrupole IT and FTICR). We have shown that the potential problem of ion suppression in complex sample matrices is actually relatively negligible, although we have not ruled out that specific peptides might be severely affected by matrix conditions. We have also demonstrated the ability to reproducibly detect and accurately quantify minor constituents in complex sample mixtures.

We believe that MapQuant represents an excellent beginning of promoting standardization of quantitative MS tools, as it combines features of currently available programs (*e.g.* smoothing, peak-detection, deconvolution of isotopes [9] and visualization abilities [10]) into one package and at the same time it offers new algorithms such as 2-D watershed segmentation, 2-D peak fitting, peak clustering, and isotope deconvolution of intertwined isotopic clusters. It also provides a dedicated scripting language that allows for automated analysis methods by giving the user control over how the data are processed.

Importantly, MapQuant is completely open-source and independent of specific instrument vendor's proprietary data format if raw data can be translated, for the time being, into OpenRaw format. We feel strongly that MapQuant should be open-source, so that it could leverage expertise in the greater MS community at large for its continued improvement as well as for future support of emerging data standards such as mzXML [25], mzData, and hmsXML (http://arep.med.harvard.edu/hmsXML/; Nguyen *et al.* manuscript in preparation). MapQuant can be compiled and run on both Windows Visual C++ and Linux platforms – a feature that existing quantification software does not provide. MapQuant can be downloaded from http://arep.med.harvard.edu/mapquant.html through an open-source compatible Harvard University agreement.

We are currently using MapQuant to perform relative quantification of proteins in the proteome of cyanobacterium *P. marinus* MED4 over the course of infection by phage and during its 24-h diel division cycle (Lindell *et al.*, manuscript in preparation; Leptos *et al.*, manuscript in preparation). Integration of prior efforts and future developments in accurate mass, retention time, charge, and carbon-content feature assignments will enable comprehensive whole-proteome expression analysis where many isotopic cluster features can be simultaneously identified and quantified [26].

We would like to thank Jay McPhee and Brent Martin for the maintenance of the clusters where MapQuant was developed and run on. Moreover, we would like to thank professors Fritz Roth, Steve Buratowski and Steve Gygi for their advice and the latter for help with SEQUEST. Patrik D'Haeseleer and John Aach for thoughtful discussions and comments during the development of MapQuant, Nathan Walsh for troubleshooting the software and its documentation, as well as Nikos Reppas for his help on the decision of the name of the software and for his constructive comments on the manuscript. We would also like to thank Sallie Chisholm, Erik Zinser and Debbie Lindell for providing us with the Prochlorococcus marinus samples. This work was supported by the US Department of Energy: GTL.

5 References

[1] Gygi, S. P., Rochon, Y., Franza, B. R., Aebersold, R., *Mol. Cell Biol.* 1999, *19*, 1720–1730.

[2] Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S. *et al.*, *Mol. Cell Biol.* 1999, *19*, 7357–7368.

[3] Lipton, M. S., Pasa-Tolic, L., Anderson, G. A., Anderson, D. J. *et al.*, *Proc. Natl. Acad. Sci. USA* 2002, *99*, 11049–11054.

[4] Jaffe, J. D., Berg, H. C., Church, G. M., *Proteomics* 2004, *4*, 59–77.

[5] Pandey, A., Mann, M., *Nature* 2000, *405*, 837–846.

[6] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F. *et al.*, *Nat. Biotechnol.* 1999, *17*, 994–949.

[7] Heller, M., Mattou, H., Menzel, C., Yao, X., *J. Am. Soc. Mass Spectrom.* 2003, *14*, 704–718.

[8] Wang, W., Zhou, H., Lin, H., Roy, S. *et al.*, *Anal Chem*, 2003, *75*, 4818–4826.

[9] MacCoss, M. J., Wu, C. C., Liu, H., Sadygov, R. *et al.*, *Anal Chem.* 2003, *75*, 6912–6921.

[10] Li, X. J., Pedrioli, P. G., Eng, J., Martin, D. *et al.*, *Anal. Chem.* 2004, *76*, 3856–3860.

[11] Tammen, H., Kreipe, H., Hess, R., Kellmann, M. *et al.*, *Breast Cancer Res. Treat.* 2003, *79*, 83–93.

[12] Wittke, S., Kaiser, T., Mischak, H., *J Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 2004, *803*, 17–26.

[13] Palagi, P. M., Walther, D., Quadroni, M., Catherinet, S. *et al.*, *Proteomics* 2005, *5*, 2381–2384.

[14] Press, W. H., Teukolsky, S. A., Flannery, B. P., Vetterling, W. T., *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge, New York 1992, *13*: 538–541, *14*: 650–655, *15*: 681–688.

[15] Ritter, G. X., Wilson, J. N., *Handbook of Computer Vision Algorithms in Image Algebra*, CRC Press, Boca Raton 2001, 417.

[16] Vincent, L., Soille, P., *Pattern Analysis and Machine Intelligence*, IEEE Transactions on 1991, *13*, 583–598.

[17] Olivé, J., Grimalt, J., *J. Chromatogr. Sci.* 1995, *33*, 194–203.

[18] Wehofsky, M., Hoffmann, R., Hubert, M., Spengel, B., *Eur. J. Mass Spectrom.*, 2001, *7*, 39–46.

[19] Eng, J., McCormack, A. L., Yates, J. R., III., *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[20] Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. *et al.*, *J. Proteome Res.* 2003, *2*, 43–50.

[21] Antal, J., Pal, G., Asboth, B., Buzas, Z. *et al.*, *Anal. Biochem.* 2001, *288*, 156–167.

[22] Baldwin, M. A., Falick, A. M., Gibson, B. W., Prusiner, S. B. *et al.*, *J. Am. Soc. Mass Spectrom.* 1990, *1*, 258–264.

[23] Qian, X., Guan, C., Guo, H. C., *Structure (Camb)*, 2003, *11*, 997–1003.

[24] Wada, Y., *J Mass Spectrom*, 1996, *31*, 263–266.

[25] Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M. *et al.*, *Nat. Biotechnol.* 2004, *22*, 1459–1466.

[26] Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L. *et al.*, *Proteomics* 2002, *2*, 513–523.

