nature biotechnology

MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification

Jürgen Cox & Matthias Mann

Efficient analysis of very large amounts of raw data for peptide identification and protein quantification is a principal challenge in mass spectrometry (MS)-based proteomics. Here we describe MaxQuant, an integrated suite of algorithms specifically developed for high-resolution, quantitative MS data. Using correlation analysis and graph theory, MaxQuant detects peaks, isotope clusters and stable amino acid isotope–labeled (SILAC) peptide pairs as three-dimensional objects in *m/z*, elution time and signal intensity space. By integrating multiple mass measurements and correcting for linear and nonlinear mass offsets, we achieve mass accuracy in the p.p.b. range, a sixfold increase over standard techniques. We increase the proportion of identified fragmentation spectra to 73% for SILAC peptide pairs via unambiguous assignment of isotope and missed-cleavage state and individual mass precision. MaxQuant automatically quantifies several hundred thousand peptides per SILAC-proteome experiment and allows statistically robust identification and quantification of >4,000 proteins in mammalian cell lysates.

Data analysis in MS-based proteomics is much more challenging than for other high-throughput technologies such as microarrays[1] and remains a principal bottleneck in proteomics[2,3]. In one popular format of MS-based proteomics, proteins are enzymatically digested to peptides, which are analyzed online by liquid chromatography (LC) coupled to electrospray and tandem MS (MS/MS)[4]. MS spectra contain peptide mass and intensity information, and the identity of the peptides is deduced by matching the MS/MS spectra against a sequence database[5,6]. Typically, peaks are extracted from raw data, the peptide mass is estimated from the scan from which the peak was 'picked' for sequencing and the peak files are sent to a search engine. Results consist of tables of identified proteins. In a quantitative proteomics experiment using stable isotopes, peptide and protein ratios are obtained by direct comparison of the signals of the 'light' and 'heavy' isotope in the same LC run[7,8].

There is already a substantial literature on 'computational proteomics' (reviewed in refs. 3,9–12). However, these efforts were usually not directed at high-resolution data of the type readily attainable today and they do not approach the quality of a skilled human expert. Here we describe a set of algorithms that efficiently and robustly extracts information from raw MS data and allows very high peptide identification rates as well as high-accuracy protein quantification for several thousand proteins in complex proteomes.

RESULTS

Analysis pipeline

MaxQuant incorporates all steps needed in a computational proteomics platform but currently uses Mascot[13] to generate peptide candidates for MS/MS spectra. Below, we describe the analysis framework and illustrate its performance with SILAC-treated HeLa cells that were stimulated for 2 h with epidermal growth factor (EGF)[14]. These data were obtained by triplicate analysis of 24 peptide fractions from isoelectric focusing using an LTQ Orbitrap mass spectrometer. We describe conceptual issues and computational analysis. A detailed explanation of algorithms is provided in **Supplementary Notes** online and their C# source code in **Supplementary Data** online.

Feature detection and quantification

The high resolution of modern mass spectrometers and the need for quantification in functional proteomics led us to start the data analysis with 'features' in the MS spectra (mass and intensity of the peptide peaks) rather than focus on the fragmentation spectra. This is already commonly done in MS-based biomarker discovery[9]. In MaxQuant, peaks are detected in each MS scan by fitting a gaussian peak shape to the three central raw data points and then assembled into three-dimensional (3D) peak hills over the *m/z*-retention time plane (**Fig. 1a–c**). Smoothed intensity profiles over retention time are split at significant local minima. From the centroid masses we obtain a high precision, intensity-weighted estimate of mass for the 3D peak (**Fig. 1d**). For each 3D peak an individual mass precision is calculated by bootstrap replication (**Supplementary Notes**).

Each of the 72 LC-MS runs of the HeLa proteome resulted in ~382,000 3D peaks, on average. It is not trivial to efficiently and reliably determine isotope patterns, and we employ a graph theoretical data structure to construct an undirected graph with the 3D peaks as vertices. An edge is inserted between two peaks when the difference in

Department for Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany. Correspondence should be addressed to J.C. (cox@biochem.mpg.de) or M.M. (mmann@biochem.mpg.de).

Received 27 May; accepted 31 October; published online 30 November 2008; doi:10.1038/nbt.1511

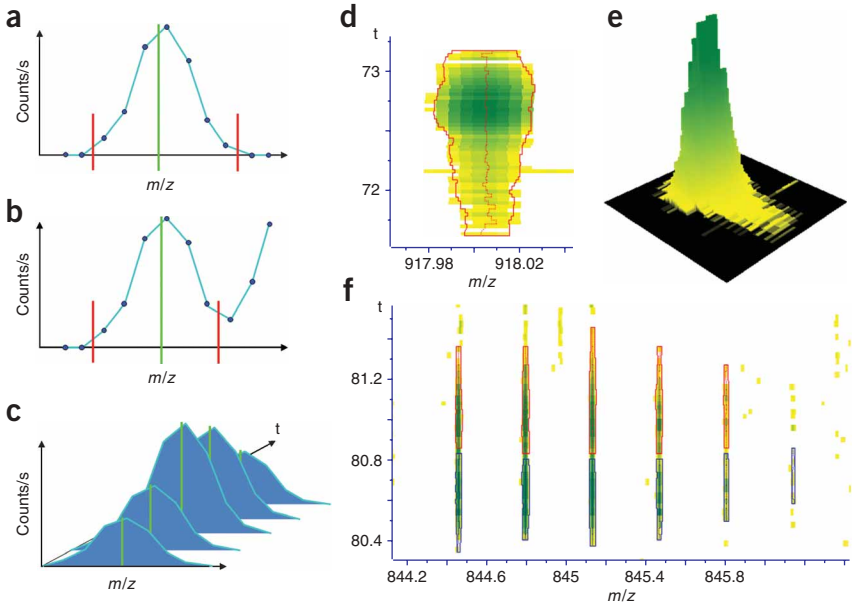

Figure 1 Three-dimensional peak detection. (**a**) Two-dimensional (2D) peaks whose intensity drops to zero on both sides. The centroid mass of a 2D peak is calculated as a fit of a gaussian peak shape to the three central raw data points. (**b**) Peaks are broken up at local intensity minima. (**c**) 2D peaks in adjacent MS scans are assembled to 3D peak hills over the *m/z*-retention time plane. Two peaks in neighboring scans are connected whenever their centroid *m/z* positions are sufficiently close. (**d**) 3D peak eluting over 1.5 min represented with color-coded intensity, decreasing from green over yellow to white, in the mass-retention time plane. Forty-nine centroids (dotted red line) have been joined to form this 3D peak. Note that fluctuations in mass become larger at low abundance. (**e**) 3D representation of the same peak. (**f**) Eleven 3D peaks forming two isotope patterns. The masses of the upper and lower isotope patterns are identical. The sixth peak of the lower isotope pattern has just been detected, whereas the sixth peak of the upper isotope pattern has just escaped detection.

mass equals the difference in isotope mass of an average amino acid ('averagine'[15]) within bootstrap errors, with an additional error tolerance due to unknown atomic composition and when the intensity profiles have a sufficiently high overlap in retention time. The resulting graph contains millions of edges, connecting 'pre-isotope' patterns, which, however, are not necessarily consistent in terms of charge state. We then iteratively determine the longest, consistent subgraphs. In the LC-MS run in **Figure 2**, the number of 3D peaks was 317,658, assembling into 31,806 isotope patterns. Thus, isotope patterns reduce data features tenfold and are a potent noise filter. A particularly dense region is greatly enlarged in **Figure 2c**. In this small region of a few *m/z* units, three overlapping isotope patterns are automatically and correctly assigned, despite the overlap caused by peptides of different charge states ($z = 5$ versus $z = 2$) and near-identical masses of two co-eluting peptides. This would have been difficult from the MS information alone, even for an expert human scientist (**Fig. 2d**).

In our example we carried out SILAC[16] using arginine and lysine. To detect heavy-light SILAC partners we consider all possible pairs of isotope patterns. Potential SILAC pairs are first required to have sufficient intensity correlation over elution time (allowing for some retention-time shift due to isotope effects) and to have equal charges. By default we assume at most three labeled amino acids per peptide. Therefore, pairs could contain lysine (K), arginine (R), KK, KR, RR, KKK, KKR, KRR and RRR. For each of these cases we convolute the two measured isotope patterns with the theoretical isotope patterns of

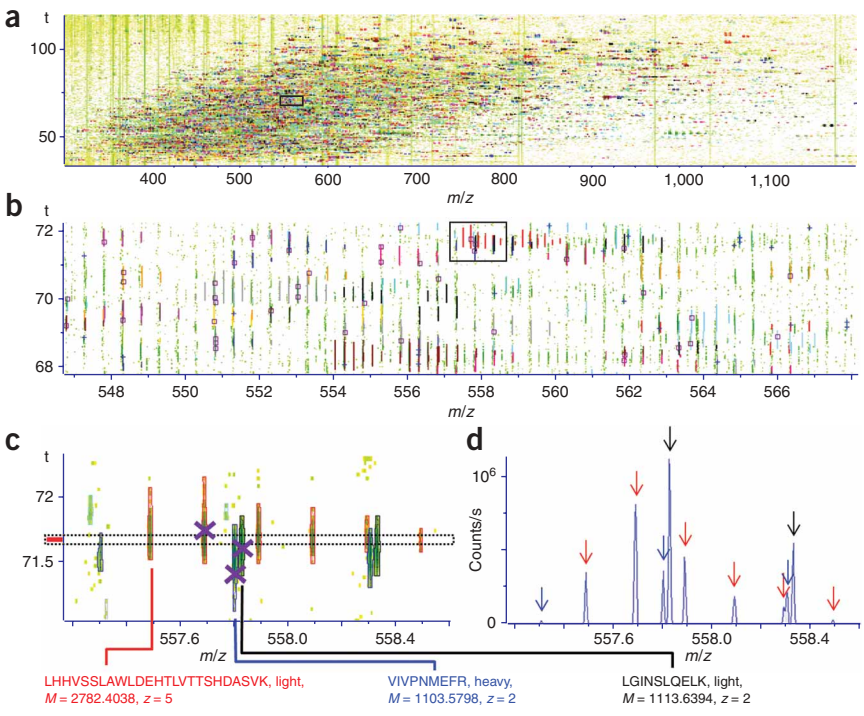

Figure 2 Automatic large-scale SILAC pair detection. (**a**) Overview of the part of the mass-retention time plane capturing most of the peptides in one LC-MS run of an OFFGEL fraction of HeLa cell lysate. 5,666 SILAC pairs have been detected in this run and are coded in different colors. (**b**) Zoom into the region indicated by the black rectangle in **a**. Several SILAC pairs can be seen with charges ranging up to five. MS/MS sequencing events are indicated either by squares, in case they led to a peptide identification, or by crosses. (**c**) Zoom into the region indicated by the black rectangle in **b** showing a challenging case for isotope pattern detection involving three peptides. Note that MaxQuant correctly assigned the monoisotopic mass, whereas the instrument software picked the C13 peak for sequencing. The heavy-labeled blue peptide has a small peak at the low-mass side of the monoisotopic peak because of the usual impurities of the commercially available heavy amino acids. (**d**) The mass spectrum corresponding to the dotted rectangle in **c**.

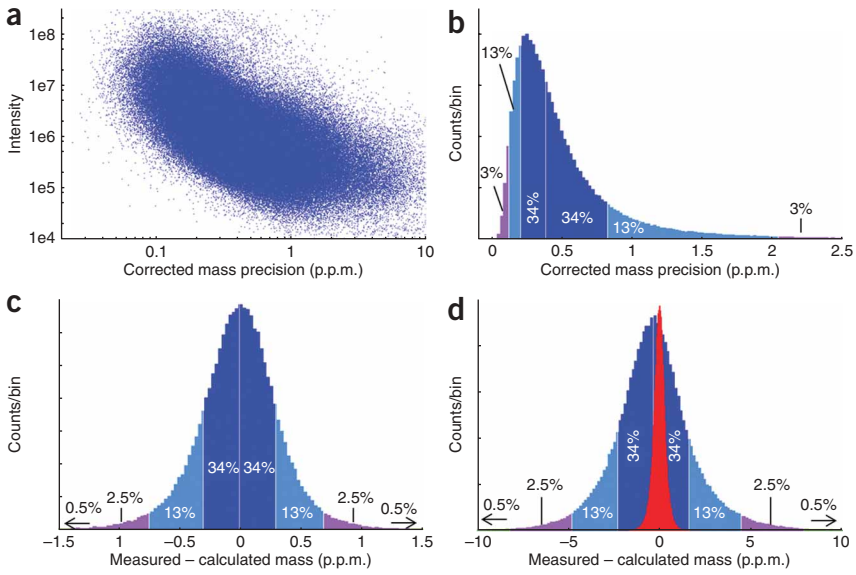

Figure 3 Accurate masses and individual peptide mass errors. (**a**) Mass precision corrected for autocorrelation of >477,000 SILAC pairs as a function of integrated signal intensity. Precision is inversely proportional to the square root of peptide intensity. (**b**) Same data as **a** but binned by corrected mass precision. (**c**) Mass deviation of all identified peptides. (**d**) Mass deviation without MaxQuant: precursor masses were taken directly from instrument software ('monoisotopic *M*/*Z*'). The scaled distribution from **c** is shown in red for comparison.

the difference atoms, that is, the atoms that have to be added so that both peptides would have the same atomic composition. If the mass differences are within the combined bootstrap error and if there is sufficient intensity correlation of the two isotope patterns in *m*/*z* dimension, the peaks are associated as a SILAC pair. **Figure 2** contains 5,666 SILAC pairs.

The resulting isotope patterns are then scaled to each other using all ratios, starting with a least-square solution and determining the best median fit iteratively by bisection. This yields the fold-change between the two SILAC peptides (**Supplementary Notes**). For triple-labeling SILAC experiments[17] more cases need to be considered but the procedures are very similar. In each LC-MS run, we normalize peptide ratios so that the median of their logarithms is zero, which corrects for unequal protein loading, assuming that the majority of proteins show no differential regulation.

Improving peptide mass accuracy

The peptide mass is calculated as the intensity-weighted average of all MS peak centroids in the 3D peaks within the isotope patterns belonging to a SILAC pair or triplet. The statistics of the number of mass measurements per SILAC peptide is given in **Supplementary Figure 1** online.

We use the several hundred SILAC charge pairs in every LC-MS run for recalibration without knowing their identity and minimize differences between mass estimates from different charge states. The resulting polynomial remaps experimental *m*/*z* values to their corrected values. Nonlinear mass corrections are about 1 p.p.m. for the LTQ Orbitrap mass spectrometer (**Supplementary Fig. 2** online).

We next use the two masses of peptide charge pairs to derive an estimate of the mass accuracy (deviation from the true value) from the estimate of the mass precision (repeatability of the measurement) by requiring that mass estimates are within the error range. We then scale the bootstrap errors by the required factor—between two to three in our data. As in similar cases[18], this factor is likely due to autocorrelation between the centroid determinations in subsequent spectra. To correct for global expansion or contraction of the mass scale we use well-identified peptides and minimize the mass deviation of these peptides weighed by their individual mass precisions.

We plotted the corrected mass precisions for the 477,511 SILAC pairs in our data set as a function of peptide signal (**Fig. 3a**). Mass precisions are extremely high (p.p.b.) and roughly proportional to one over the square root of the peptide signal. **Figure 3b** shows that 50% of the peaks have corrected mass precisions better than 393 p.p.b. In agreement with this, the actual mass deviations of all identified peptides (measured minus calculated mass) have a s.d. of 409 p.p.b. and average absolute mass deviation (average of the absolute value of the difference between measured and calculated masses) of 278 p.p.b. (**Fig. 3c**).

Peptide mass estimates are usually taken from the MS peak that leads to selecting the peptide for fragmentation (**Fig. 3d**). Average absolute mass accuracy in this standard approach is 1.8 p.p.m. and s.d. is 2.5 p.p.m. Thus mass accuracy measured as s.d., a key performance parameter in proteomics[19], improved sixfold using our computational approach. We suspect the improvement would have been even greater if we had not used the 'lock mass option'[20]. Worse, even including the lock mass, the normal approach would have necessitated a maximum allowed mass deviation of 10 p.p.m. for all peptides, whereas searches are performed with much tighter and individualized mass tolerances in MaxQuant.

Peptide and protein identification

Because the SILAC state of most isotope patterns is known beforehand, we can treat the label modifications as fixed in the database search. By counting the number of arginines and lysines, the SILAC state distinguishes limit tryptic peptides from incompletely cleaved ones. This a priori information decreases the search space about tenfold. For fragmentation spectra not associated with a SILAC pair, a conventional database search is performed. After a database search, the list of top ten sequences matching a fragmentation spectrum is sorted according to their peptide score or P-score[21] and filtered for consistency with a priori information, retaining the best scoring one. We allow a deviation between the measured and calculated mass of four s.d. of the individual bootstrap error for each peptide.

We use a database containing all true protein sequences, concatenated with reversed nonsense versions of these sequences[22,23]. To avoid spurious correlations because half of the reversed tryptic peptides have the same mass as the forward sequence, we also swapped every arginine and lysine with the preceding amino acid in the reversed sequences. This approach still retains the local amino acid relations—leading to the same length and mass distributions of peptides (**Supplementary Notes**).

To assess the likelihood of false identification we generate two lists of peptides, one for the hits in the forward sequences and one in the reversed sequences. We construct two histograms by gaussian kernel smoothing (**Fig. 4**). They can be interpreted as approximations to the

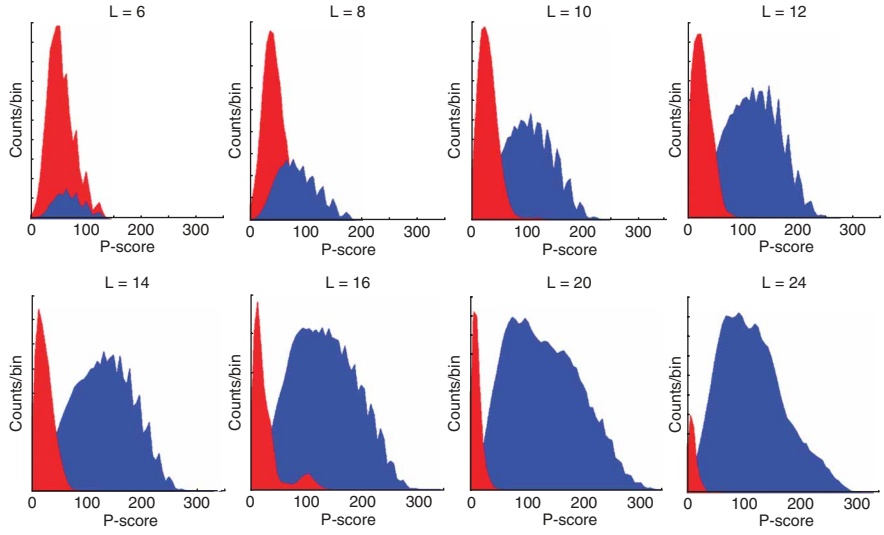

Figure 4 Peptide score (P-score) distributions. The panels show the distributions of scores in the forward (blue) and reverse (red) database with peptide length (L) as the parameter. MaxQuant filters potential hits by a priori information, which moves the reverse hit distribution far to the left. These distributions are used to calculate the false-positive rate for peptide identification as a function of peptide length.

total and the conditional probability densities

$$p(s, L) \text{ and } p(s, L|X = \text{false})$$

where the Boolean variable X indicates 'true or false' (forward) or 'false' (reverse) sequences. s is the peptide database score and L the peptide length. The probability of a false hit, given the peptide identification score and the length of the peptide is then

$$p(X = \text{false}|s, L) = \frac{p(s, L|X = \text{false})p(X = \text{false})}{p(s, L)}$$

the posterior error probability (PEP) of each individual peptide. We use the PEP only as input for calculating the false-discovery rate (FDR) below. The a priori probability $p(X{=}\text{false})$ is a constant with no effect on the final list of accepted peptides at a given FDR. Longer peptides, which are less frequent in the database, are automatically accepted with lower scores.

To determine a cutoff score for a specific FDR, we sort all peptide identifications, from the forward and the reverse database, by their PEP, starting with the best. Peptides are accepted until 1% of reverse hits/forward hits has accumulated. The fraction of wrong identifications in the forward database is then 1% as well.

In this run, 11,299 sequencing events led to 7,307 peptide identifications (identification rate of 64.7%, **Fig. 5**). Sequencing events associated with SILAC pairs have identification rates of 84.4%. Identifications (red squares) cluster in particular regions of the contour plot (**Fig. 5a**), with characteristic polymer patterns devoid of peptide identifications (**Fig. 5b**) and fragmentation events in peptide-rich regions almost uniformly identified (**Fig. 5c**). Note that many SILAC pairs were not targeted for sequencing at all (32.3% in this run).

We next assemble peptide hits into protein hits, a nontrivial step in shotgun proteomics[24]. Whenever the set of identified peptides in one protein is equal to or completely contained in the set of identified peptides of another protein these two proteins are joined in a protein group. Shared peptides are most parsimoniously associated with the group with the highest number of identified peptides ('razor' peptides[24]) but remain in all groups where they occur. Protein quantification may then be performed based only on unique peptides, including razor peptides, or using all peptides. By default we use unique and razor peptides as a compromise between unequivocal peptide assignment and most-accurate quantification.

We assign to each protein group a PEP by multiplying their peptide PEPs. Only peptides with distinct sequences and only the highest-scoring identified spectra are used to avoid bias due to dependent peptides. Similarly to the peptide PEP, the protein PEP serves to sort the list of hits from forward and reverse databases. Using a protein FDR of 1% and requiring that each protein group contain a unique peptide, we identified 4,149 proteins in the cell line proteome (**Supplementary Table 1** online).

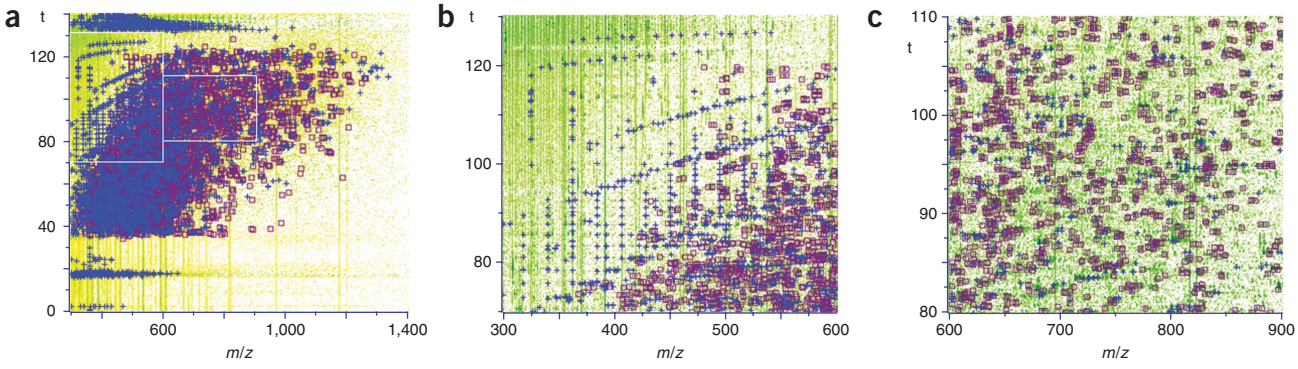

Figure 5 High rate of identified MS/MS spectra. MS/MS sequencing events are indicated in the mass-retention time plane (contour plot). Identified and unidentified MS/MS spectra are represented by red squares and blue crosses, respectively. (**a**) Peptides elute between 40 and 120 min and peptide identifications are shifted to higher *m/z* values at later points in the gradient. (**b**) Left rectangle of **a**. In this region, characteristic polymer patterns that do not lead to peptide identifications are prevalent. (**c**) In contrast, in a peptide-rich region of the contour plot (right rectangle in **a**), almost all fragmentation events lead to successful peptide identification.

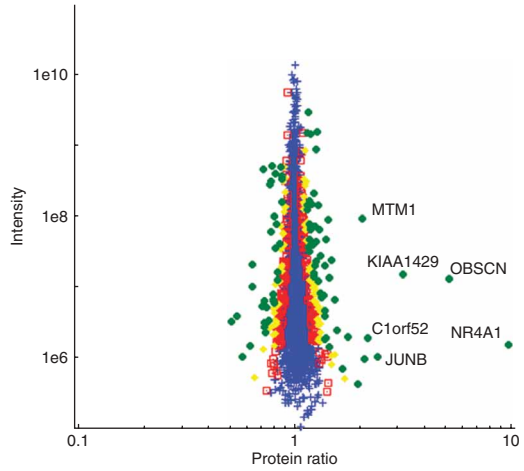

Figure 6 Proteome-wide accurate quantification and significance. Normalized protein ratios are plotted against summed peptide intensities. The spread of the cloud is lower at high abundance, indicating that quantification is more precise. The data points are colored by their 'significance B', with blue crosses having values >0.05, red squares between 0.05 and 0.01, yellow diamonds between 0.01 and 0.001 and green circles <0.001.

Protein quantification

Many of the isotope patterns that have not been assembled into SILAC pairs are nevertheless identified by database search. For these peptides the *m*/*z*-elution time shapes of the 3D peaks belonging to the identified SILAC version are translated to the location of the missing SILAC partner and after integration of intensities, ratios are calculated in the same way as for SILAC pairs that were detected before identification.

Protein ratios are calculated as the median of all SILAC peptide ratios, minimizing the effect of outliers. We normalize the protein ratios to correct for unequal protein amounts.

We next calculate an outlier significance score for log protein ratios (significance A). To create a robust and asymmetrical estimate of the s.d. of the main distribution we calculate the 15.87, 50 and 84.13 percentiles r_{-1}, r_0, and r_1. r_1-r_0 and r_0-r_{-1} are right- and left-sided robust s.d. For a normal distribution, these would be equal to each other and to the conventional definition of an s.d. A suitable measure for a ratio $r > r_0$ being significantly far away from the main distribution is the distance to r_0 measured in terms of the right s.d.

$$z = \frac{r - r_0}{r_1 - r_0}$$

As a *P*-value for detection of significant outlier ratios we define

$$\text{significance A} = \frac{1}{2}\,\text{erfc}\left(\frac{z}{\sqrt{2}}\right) = \frac{1}{\sqrt{2\pi}}\int_z^\infty e^{-t^2/2}dt$$

which is the probability of obtaining a log-ratio of at least this magnitude under the null hypothesis that the distribution of log-ratios has normal upper and lower tails (**Supplementary Fig. 3** online).

For highly abundant proteins the statistical spread of unregulated proteins is much more focused than for low abundance ones[8] (**Fig. 6**). To capture this effect, we define another quantity, significance B, which is calculated only on the protein subsets obtained by intensity binning. We define bins of equal occupancy such that each contains at least 300 proteins.

We quantified 4,100 proteins, comparable to the number of significant messages in a microarray experiment on the same cell type[14] (**Supplementary Table 1**). If a minimum of three quantification events (three SILAC pairs) is required, quantification becomes very reliable[25,26] because an outlier ratio has no effect on the median. Strikingly, 99.3% of proteins were within 50% of the one-to-one ratio. This implies excellent SILAC partner identification as wrongly partnered peptides would have ratios strongly deviating from 1. We found 48 proteins to be significantly upregulated based on significance B with a Benjamini-Hochberg[27] FDR < 5% (**Supplementary Table 2** online).

Notably, two of the most heavily upregulated proteins after 2 h of EGF stimulation were the transcription factors JunB and the orphan nuclear receptor NR4A1, also termed early-response protein NAK1 (**Fig. 6**). Both are known to be regulated by growth stimuli. Among the most upregulated proteins in **Figure 6** there is a conserved dual-specificity tyrosine-serine phosphatase (MTM1), widely studied in relation to myotubular myopathy[28] and, like PTEN, a lipid phosphatase[29]. The completely uncharacterized protein C1orf52 is tightly associated with the tumor suppressor BCL10 and therefore also called BAG for BCL10-associated gene. Neither of these proteins was known to be induced upon EGF stimulation. Many of the other significantly regulated proteins also have potential connection to growth factor signaling (**Supplementary Table 1**). Proteins encoded by genes having regulatory binding sites for SREBP-1 are shown to be significantly upregulated when analyzed by TRANSFAC[30]. SREPB-1 likely mediates the effects of EGF stimulation on cancer-relevant proteins like FAS[31].

DISCUSSION

We have introduced a set of computational proteomics algorithms with several useful features. Efficient extraction of mass information allows us to search protein databases with maximum allowed mass deviations that adjust themselves to the precision with which the peptide is measured. The mass accuracies achieved here are the highest yet reported in large-scale proteomics[32] and sharply limit the number of candidate peptides in database searches. With low-resolution data, only a few percent of fragmentation events lead to successful identification[33], whereas the mass accuracy and feature extraction in MaxQuant allow 73% of the fragmentation events associated with SILAC peptide pairs to be identified. Thus, standard ion trap fragmentation is extremely information rich, and nontryptic and modified peptides do not constitute the majority of fragmented peptides. The MaxQuant algorithms recently enabled comprehensive quantification of the yeast proteome[34]. Although we identified essentially the complete proteome, we found only three (<1%) of the 814 'dubious' open reading frames (ORFs) (http://www.yeastgenome.org/), which are not expected to be expressed from evidence such as comparative genome sequencing. This provides independent evidence that our FDR estimates of peptide and protein identifications are very stringent (**Supplementary Fig. 4** online). Much higher identification rates among dubious ORFs (3%) were found in genome-wide tagging experiments[35,36]. Likewise, aggregate data from yeast proteome resources cover 12% of these dubious ORFs[37], the same percentage as their occurrence in the genome.

We have already applied MaxQuant to quantify >5,000 proteins in the mouse stem cell proteome[38] and several other proteomes in similar depth. We conclude that the computational tools for proteome-wide quantification are now in hand. With further advances in instrumentation, particularly in the dynamic range of measurements[39,40], proteomics should be suitable for routine 'functional genomics' experiments, for which microarrays have so far been the only option.

METHODS

Software development and availability of MaxQuant. MaxQuant is developed for the .NET framework and written in the C# language. The interactive 3D data viewer was developed on the basis of DirectX. MaxQuant executables are available via http://www.maxquant.org/, whereas the source code of algorithms is available in **Supplementary Data**. It runs on Windows desktop computers and is compatible with XP and Vista. Processing time is currently about 20 min per raw file and per processing core. Detailed description of the algorithms used in MaxQuant can be found in **Supplementary Notes**.

Data processing. The Mascot program version 2.2.04 was used to generate up to ten peptide sequence candidates per fragmentation spectrum (Matrix Science), and International Protein Index (IPI) version 3.48 was searched. The database search is done with an initial maximum allowed mass deviation of 7 p.p.m. for the peptide mass and 0.5 *m/z* units for fragmentation peaks, which is optimal for linear ion trap data[41].

Gene Ontology, Pfam domain and TRANSFAC overrepresentation analysis. *P*-values for overrepresentation in regulated proteins were calculated with the Wilcoxon-Mann-Whitney test on the continuous significance B values calculated by the MaxQuant software.

Data used in analysis. The data used in this analysis have been published in reference 14. SILAC was performed as described[42]. Briefly, HeLa cells were stimulated with EGF for 2 h and mass spectrometric analysis performed as described[20]. 'Heavy' (EGF stimulated) and 'light' (control) SILAC cell populations were combined and lysed. Proteins were digested in solution with trypsin, and the resulting peptides were separated by isoelectric focusing into 24 fractions with an Agilent 3100 OFFGEL Fractionator. Each fraction was purified with StageTips[43] and analyzed by liquid chromatography combined with electrospray tandem mass spectrometry on a Thermo Scientific LTQ Orbitrap mass spectrometer with lock mass calibration[20]. The experiment was performed in triplicate.

Raw mass spectrometric data files and evidence tables containing peptide and protein data can be downloaded from Tranche at http://tranche.proteomecommons.org/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank all the other members of the Proteomics and Signal Transduction group for help with the development of MaxQuant. Shubin Ren helped in developing the 3D data viewer used in MaxQuant. Nina Hubner measured the data used in this analysis. This work was supported by the Max-Planck Society and by the 6th Framework Program of the European Union (Interaction Proteome LSHG-CT-2003-505520 and HEROIC LSHG-CT-2005-018883).

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Allison, D.B., Cui, X., Page, G.P. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.* **7**, 55–65 (2006).
2. Patterson, S.D. & Aebersold, R.H. Proteomics: the first decade and beyond. *Nat. Genet.* **33** Suppl, 311–323 (2003).
3. Nesvizhskii, A.I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4**, 787–797 (2007).
4. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
5. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5**, 699–711 (2004).
6. Sadygov, R.G., Cociorva, D. & Yates, J.R. III. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **1**, 195–202 (2004).
7. Ong, S.E. & Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **1**, 252–262 (2005).
8. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **389**, 1017–1031 (2007).
9. Listgarten, J. & Emili, A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **4**, 419–434 (2005).
10. Colinge, J. & Bennett, K.L. Introduction to computational proteomics. *PLOS Comput. Biol.* **3**, e114 (2007).
11. Matthiesen, R. Methods, algorithms and tools in computational proteomics: a practical point of view. *Proteomics* **7**, 2815–2832 (2007).
12. Mead, J.A., Shadforth, I.P. & Bessant, C. Public proteomic MS repositories and pipelines: available tools and biological applications. *Proteomics* **7**, 2769–2786 (2007).
13. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
14. Cox, J. & Mann, M. Is proteomics the new genomics? *Cell* **130**, 395–398 (2007).
15. Senko, M.W., Beu, S.C. & McLafferty, F.W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **6**, 229–233 (1995).
16. Ong, S.E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).
17. Blagoev, B., Ong, S.E., Kratchmarova, I. & Mann, M. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nat. Biotechnol.* **22**, 1139–1145 (2004).
18. Sokal, A.D. *Monte Carlo Methods in Statistical Physics: Foundations and New Algorithms* (Lausanne, Switzerland, 1996).
19. Zubarev, R. & Mann, M. On the proper use of mass accuracy in proteomics. *Mol. Cell. Proteomics* **6**, 377–381 (2007).
20. Olsen, J.V. *et al.* Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **4**, 2010–2021 (2005).
21. Olsen, J.V. & Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. USA* **101**, 13417–13422 (2004).
22. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
23. Käll, L., Storey, J.D., MacCoss, M.J. & Nobel, W.S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **7**, 29–34 (2008).
24. Nesvizhskii, A.I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **4**, 1419–1440 (2005).
25. Selbach, M. *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58–63 (2008).
26. Bonaldi, T. *et al.* Combined use of RNAi and quantitative proteomics to study gene function in *Drosophila*. *Mol. Cell* **31**, 762–772 (2008).
27. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289–300 (1995).
28. Laporte, J. *et al.* MTM1 mutations in X-linked myotubular myopathy. *Hum. Mutat.* **15**, 393–409 (2000).
29. Wishart, M.J. & Dixon, J.E. PTEN and myotubularin phosphatases: from 3-phosphoinositide dephosphorylation to disease. *Trends Cell Biol.* **12**, 579–585 (2002).
30. Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374–378 (2003).
31. Swinnen, J.V. *et al.* Stimulation of tumor-associated fatty acid synthase expression by growth factor activation of the sterol regulatory element-binding protein pathway. *Oncogene* **19**, 5173–5181 (2000).
32. Liu, T., Belov, M.E., Jaitly, N., Qian, W.J. & Smith, R.D. Accurate mass measurements in proteomics. *Chem. Rev.* **107**, 3621–3653 (2007).
33. Kuster, B., Schirle, M., Mallick, P. & Aebersold, R. Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* **6**, 577–583 (2005).
34. de Godoy, L.M. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1254 (2008).
35. Huh, W.K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
36. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
37. King, N.L. *et al.* Analysis of the *Saccharomyces cerevisiae* proteome with PeptideAtlas. *Genome Biol.* **7**, R106 (2006).
38. Graumann, J. *et al.* SILAC-labeling and proteome quantitation of mouse embryonic stem cells to a depth of 5111 proteins. *Mol. Cell Proteomics* **7**, 672–683 (2008).
39. Eriksson, J. & Fenyo, D. Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs. *Nat. Biotechnol.* **25**, 651–655 (2007).
40. Mann, M. & Kelleher, N.L. Special feature: precision proteomics: The case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. USA.* published online, doi: 10.1073/pnas.0800788105 (25 September 2008).
41. Cox, J., Hubner, N.C. & Mann, M. How much peptide sequence information is contained in ion trap tandem mass spectra? *J. Am. Soc. Mass. Spectrom.* published online, doi:10.1016/j.jasms.2008.07.024 (7 August 2008).
42. Ong, S.E. & Mann, M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat. Protocols* **1**, 2650–2660 (2006).
43. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protocols* **2**, 1896–1906 (2007).