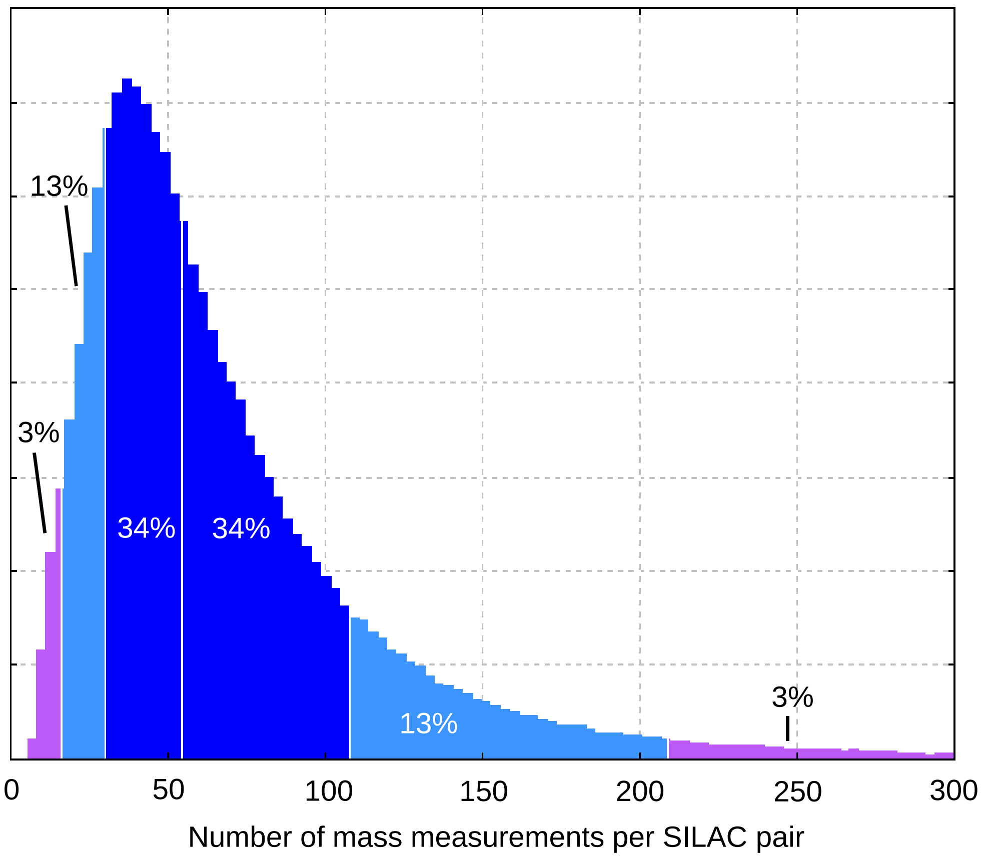
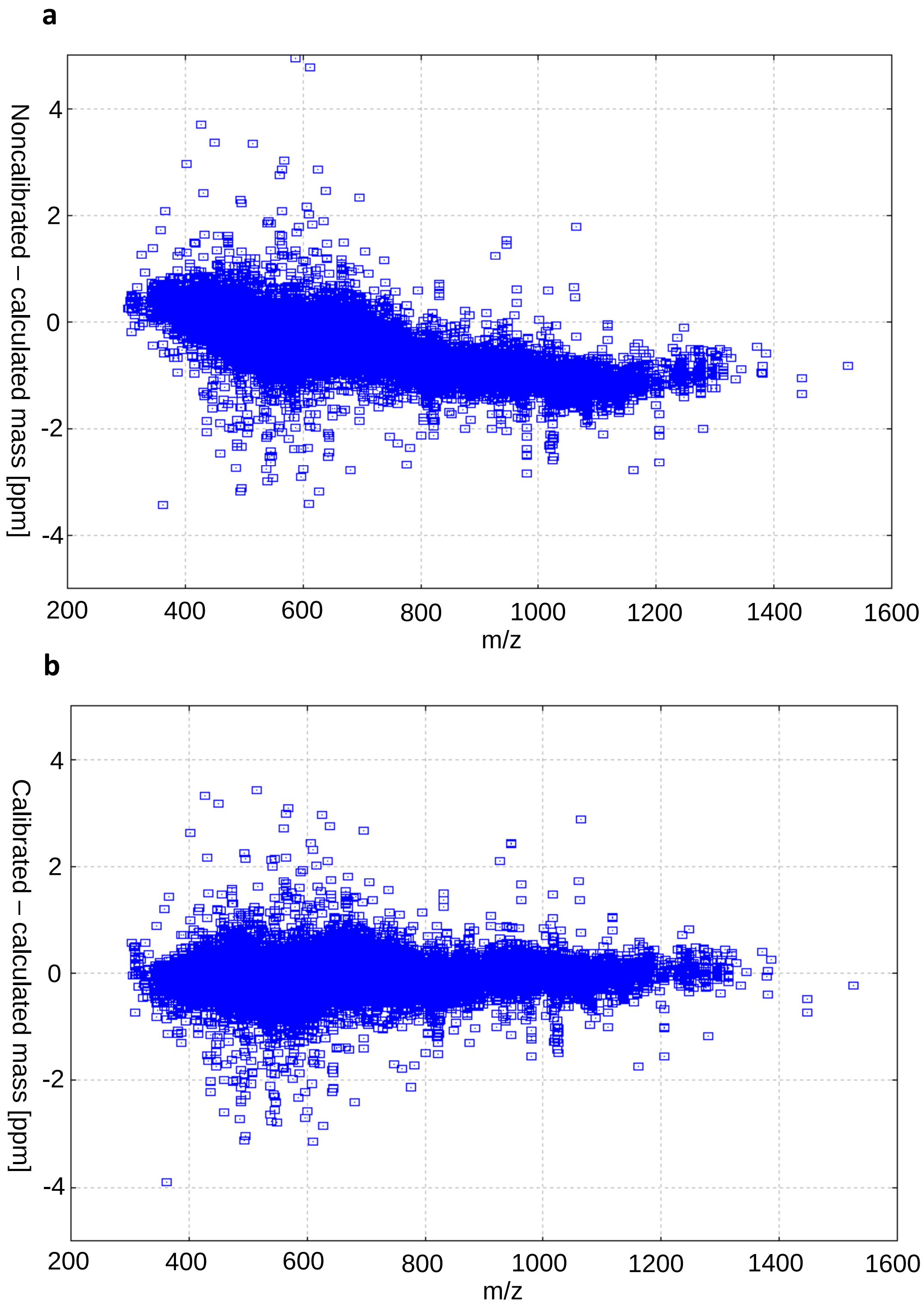
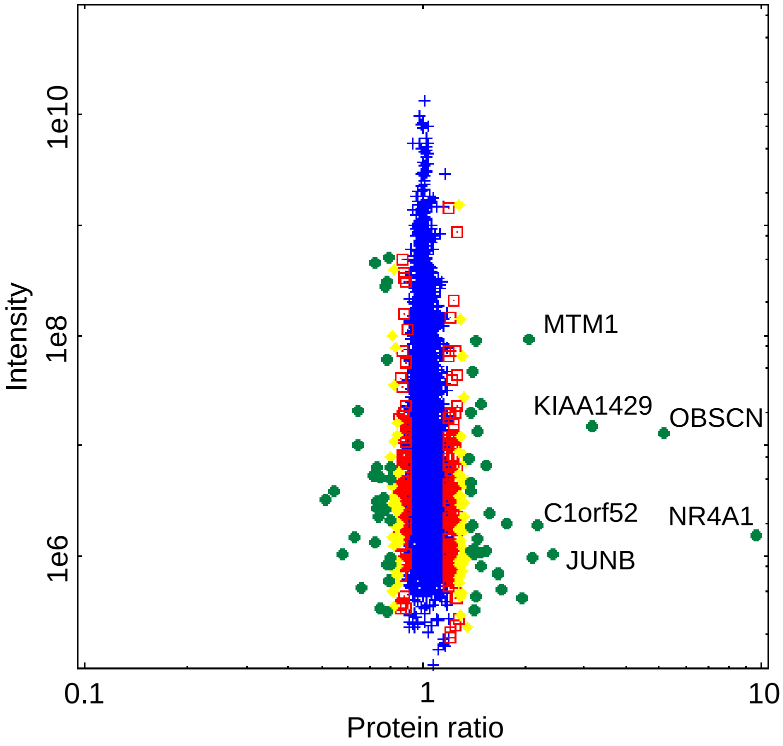
Supplementary Figure 1: Mass measurements per SILAC pair. Distribution of the number of mass measurements per SILAC pair from different scans, isotope peaks and members of the pair. A total of 50% of the SILAC pairs have more than 55 associated mass measurements.



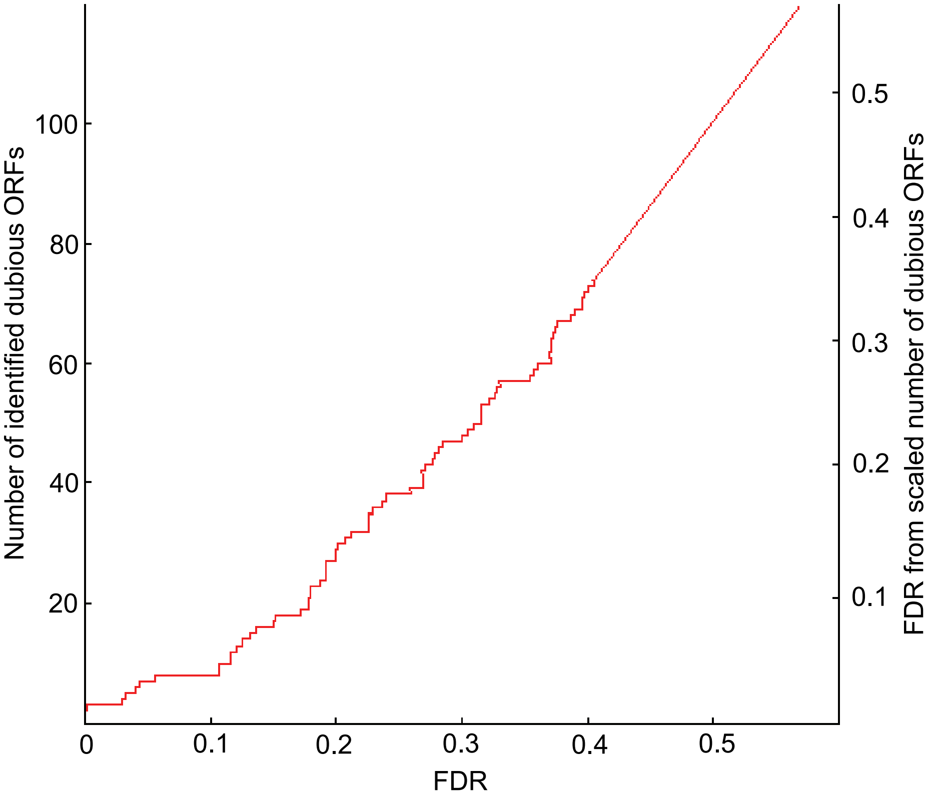
Supplementary Figure 2: Non-linear recalibration of the mass scale. Distribution of mass error (measured minus calculated mass) for identified SILAC pairs without (**a.**) and with (**b.**) non-linear recalibration using charge pairs.



Supplementary Figure 3: Proteome-wide accurate quantitation and significance. Same as **Figure 6** but colored by ‘significance A’.



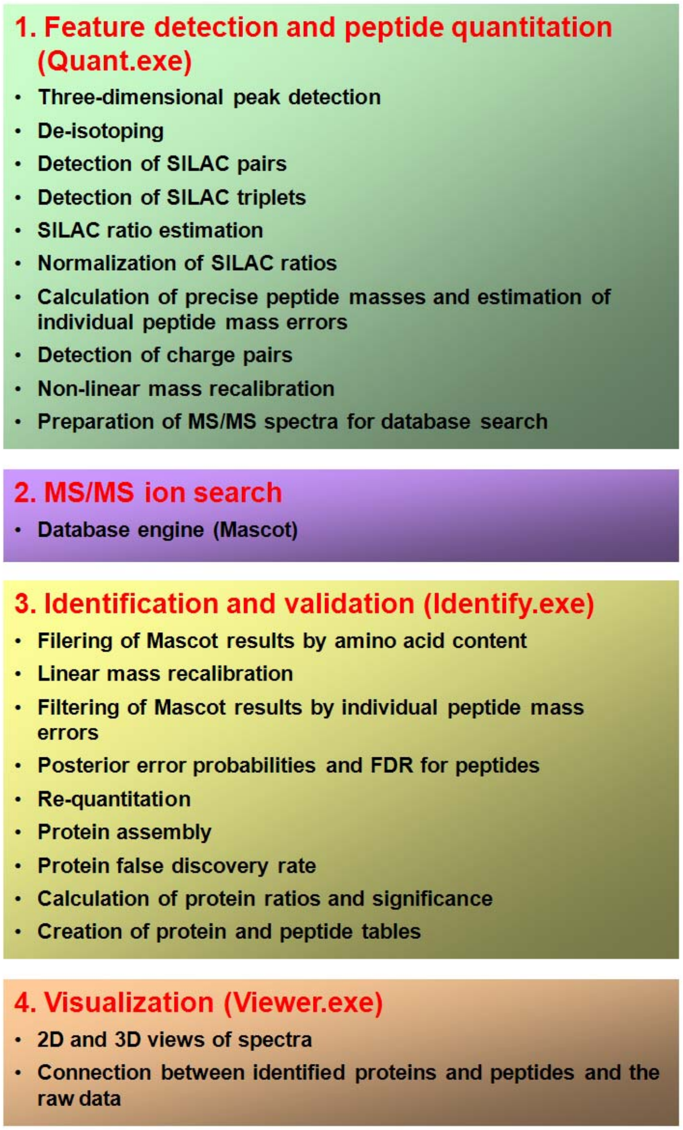
Supplementary Figure 4: FDR compared to identification rate of dubious ORFs in yeast. Number of dubious ORFs identified in the yeast proteome data from[34] plotted against FDR. An approximate linear relationship is observed. On the right axis the number of dubouos ORFs has been rescaled by a suitable factor to best reflect the FDR.



Supplementary Notes

In the following, all sequential data analysis steps of the MaxQuant quantitative proteomics workflow are described in detail. Numerical values of parameters and thresholds are reported in their versions optimized for Thermo Fisher LTQ Orbitrap or FT Ultra mass spectrometers and might need further tuning for data produced by other instruments. We generated a table (to be found at the end of this document) listing all these parameters and giving the rationale for their choice. These parameter values have been determined by analyzing many datasets of very different types that were generated over months in our department and by having many users monitoring the performance of the algorithms on their own data. While the parameters should be optimized for different platforms, we do not expect any conceptual changes, as long as shotgun proteomics data of high sample complexity is used. The algorithms and general ideas should be applicable to any kind of high mass resolution data. The figure on the next page provides an overview of the MaxQuant processing pipeline. It consists of four blocks of computational tasks that are sequentially applied to the high precision mass spectrometry data. The first block, 'pre-processing and quantitation', contains all tasks that can be performed without knowing the identity of peptides. In particular, the assembly of isotope patterns into SILAC pairs is already done here, before the submission of data to an MS/MS search engine. Advanced three-dimensional peak detection and isotope pattern detection is a pre-requisite for the high peptide mass precision which subsequently leads to high peptide identification rates. In the second block of computational tasks we currently use Mascot as a database search engine[1] for generating peptide candidates. In 'identification and validation' prior knowledge in form of individual mass tolerances and partial amino acid composition (i.e. number of Arg and Lys) is applied to the tentative peptide identifications, thus increasing their accuracy and coverage. Peptides are assembled into proteins for which quantitative information in form of protein SILAC ratios and measures of significance is generated. The last module enables interactive two and three-dimensional visualization of the raw data in the context of the identification and quantitation results for proteins and peptides.

Please see accompanying **Supplementary Data** for source code of all the algorithms used. This zip file contains a C# .NET project that is compilable to a dynamic linked library. The classes in the subfolder ‘Tasks’ correspond to the algorithmic steps discussed below. The classes have self-explanatory names to link them to the specific numeric tasks.



1. Feature detection and peptide quantitation (Quant.exe)

- Three-dimensional peak detection
- De-isotoping
- Detection of SILAC pairs
- Detection of SILAC triplets
- SILAC ratio estimation
- Normalization of SILAC ratios
- Calculation of precise peptide masses and estimation of individual peptide mass errors
- Detection of charge pairs
- Non-linear mass recalibration
- Preparation of MS/MS spectra for database search

2. MS/MS ion search

- Database engine (Mascot)

3. Identification and validation (Identify.exe)

- Filering of Mascot results by amino acid content
- Linear mass recalibration
- Filtering of Mascot results by individual peptide mass errors
- Posterior error probabilities and FDR for peptides
- Re-quantitation
- Protein assembly
- Protein false discovery rate
- Calculation of protein ratios and significance
- Creation of protein and peptide tables

4. Visualization (Viewer.exe)

- 2D and 3D views of spectra
- Connection between identified proteins and peptides and the raw data

Three-dimensional (3D) peak detection

In each MS scan, peaks are detected in a conventional two-dimensional (2D) way by first searching for local maxima of the intensity as a function of *m/z*. The lower and upper limits of the *m/z* interval for a 2D peak (indicated by red vertical lines in Figure 1a/b in main text) are then determined by moving from the maximum position to smaller and larger m/z values, until either the intensity has dropped to zero (Figure 1a), or a local intensity minimum has been reached (Figure 1b, right flank). This straightforward approach of peak detection without any de-convolution, smoothing or de-noising is sufficient for MS data generated by modern high precision mass spectrometers such as LTQ FT or Orbitrap. The centroid position of a 2D peak is then determined in different ways, depending on the number of raw data points that are spanned by the peak. If the peak consists of only one raw data point, then the *m/z* value of that point is taken as the centroid position. If there are two raw data points in a peak, then the centroid position is defined as the average of the two raw *m/z* values, weighted by the raw intensities. These two cases are rarely found, since most of the time a peak consists of at least three raw data points. In this case, only the three central points are taken into account and they are fitted to a Gaussian peak shape. The center position is then given by

$$m = \frac{1}{2}\frac{(L_0 - L_1)m_{-1}^2 + (L_1 - L_{-1})m_0^2 + (L_{-1} - L_0)m_1^2}{(L_0 - L_1)m_{-1} + (L_1 - L_{-1})m_0 + (L_{-1} - L_0)m_1},$$

where $m_{-1,0,1}$ are the m/z positions of the three central raw data points, and $L_{-1,0,1}$ are the logarithms of the corresponding raw intensities. We also tried to determine the centroid position based on a Gaussian fit to the four, five, and seven central points if available which only changed the resulting mass accuracy by about 5%. We therefore kept the simple formula above based on three central points. Replacing the Gaussian fit by an intensity-weighted average of the three central masses did worsen the mass accuracy by 35%. As the intensity of a centroid we take the sum of the intensities of all contained raw data points. This is proportional to the area under the peak to very high precision, because the mass spacing between raw data points is locally nearly constant.

In the next step, the 2D peaks in adjacent MS spectra are assembled into 3D peak hills over the *m/z*-retention time plane as schematically displayed in Figure 1c. Two peaks in neighboring scans are connected whenever their centroid *m/z* positions differ by less than 7 ppm. If for a given centroid in MS scan *n* no matching centroid is found in scan *(n+1)* in the ±7 ppm mass window, then it is checked if there is a centroid in scan *(n+2)* in the same mass window to continue the peak in time. We adjusted the window size to 7 ppm by visual inspection of many very low abundant peaks. While this value might seem large compared to the mass precision achievable with Orbitrap mass spectrometers, it ensures that virtually all instances that a human expert would call a 3D peak are included. The number of 'wrong' connections is still low and most of them will be disconnected again in the next step by detecting intensity minima in the elution profile. A 3D peak is defined as the maximal chain of 2D peaks that results from connecting the centroids in time direction in the described way. At least two centroids have to be matched together to form a 3D peak, i.e. centroids that cannot be matched to centroids in the two previous or the two next scans are discarded. This way, a major fraction of the noise in form of peaks not reproducible in time is removed. Then, the intensity profiles are smoothed over retention time applying a window mean filter of ±1 scan width and checked for local minima. If a minimum is found whose value is 1/1.3 of the lower of the two local maxima the 3D peak is split into two at the minimum position. The threshold value for a significant minimum in combination with the smoothing was chosen such that random intensity fluctuations in the intensity elution profile, which are on our machines usually below 20%, would not introduce erroneous splits of peaks. Multiple peaks in a retention time profile at the same mass may for instance be produced by isomers, resulting from multiple possibilities for the positioning of a posttranslational modification, e.g. phosphorylation, but also by the coincidental 'collision' of two different peptides.

In Figure 1d a 3D peak is shown that resulted from the procedure described above. It is the fourth peak in the isotope pattern of the doubly charged version of the peptide VGINYQPPTVVPGGDLAK found in isoelectric focusing fraction 13 with a heavy lysine and without modifications. It belongs, among others, to the protein with IPI identifier IPI00007750.1. Intensities are color coded, increasing from black through yellow to red. A total of 49 2D peaks have been joined to form this 3D peak. The dotted red line indicates the time trace of the centroids. From the centroid masses we obtain a high precision mass estimate for the 3D peak by taking the average of the masses of the *n* centroids in the 3D peak weighted by their intensities,

$$\overline{m} = \sum_{j=1}^{n} m_j I_j \Big/ \sum_{j=1}^{n} I_j \ .$$

As can be seen in Figure 1d, the fluctuation of the centroid position is smaller in the region of high intensities, reflecting the fact that highly intense peaks have a higher mass precision then low abundant ones. The weighting by intensity in the mass averaging takes this effect into account. Since the mass estimate $\bar{m}$ in the equation above is more complicated than just an average of the m_j, a standard deviation based estimate of the error would not be appropriate. Therefore we calculate the error as a bootstrap[2] estimate over B=150 bootstrap replications

$$\Delta\bar{m} = \sqrt{\sum_{b=1}^{B} (\bar{m}_b - \bar{m})^2 \Big/ (B-1)} \,,$$

where $\bar{m}_b$ is the result of evaluating $\bar{m}$ on the bth bootstrap replicate. A single bootstrap replicate consists of replacing the indices 1,…,n by n new indices that are uniformly and randomly drawn with replacement.

De-isotoping

On average we found about 386,000 3D peaks in each MS run, which may vary with the complexity of the sample. Many of these peaks will belong to isotope patterns of peptides. As a first step in determining which of the 3D peaks determined as described above belong together to form an isotope pattern, we construct an undirected graph with the 3D peaks as vertices. An edge is put between two peaks, whenever it is possible that they are neighbors in an isotope pattern, e.g. one might be the isotopic peak with all atoms in their lowest isotopic version, and the other has one atom in the next higher mass state, e.g. a ^{12}C atom has been replaced by a ^{13}C version. The peptide charge is for that purpose assumed to be within the range from one to six. The detailed criterion for putting an edge between two peaks is

$$\left| \Delta m - \frac{\Delta M}{z} \right| \leq \left[\left(\frac{\Delta S}{z} \right)^2 + \Delta m_1^2 + \Delta m_2^2 \right]^{1/2}$$

for any charge z, where Δm is the difference between the precise masses of the two 3D peaks, ΔM = 1.00286864 is the mass difference between the 13C peak and the monoisotopic peak in an averagine molecule of 1500 Da mass, which should represent a typical mass difference between two neighboring isotopic peaks. Δm_1 and Δm_2 are the bootstrap standard deviations of the peak masses times five, and

$$\Delta S = 2m(^{13}C) - 2m(^{12}C) - m(^{34}S) + m(^{32}S) = 0.0109135$$

is the maximal mass shift that the incorporation of a sulphur atom can cause. The latter term is included to take into account potential uncertainties in the mass difference due to the varying atomic composition of peptides.

In addition to the criterion above that the mass difference must fit, we also ensure that the intensity profiles have a sufficient overlap in retention time, by requiring them to have a cosine correlation (uncentered Pearson correlation) greater than 0.6:

$$\frac{\sum_{s=s\min}^{s\max} I_s J_s}{\sqrt{\sum_{s=s\min}^{s\max} I_s^2 \sum_{s=s\min}^{s\max} J_s^2}} \geq 0.6 .$$

This differs from the Pearson correlation by not subtracting the means. The sums run over MS scans from *s*min to *s*max such that all 2D peak centroids of the two 3D peaks being compared are contained in the interval $[s\min+1, s\max-1]$. I_s and J_s are the intensity profiles of the two 3D peaks. If for a given *s* the 3D peak has a 2D peak, then the intensity profile of this 3D peak and at this *s* is set to the corresponding 2D centroid intensity, and otherwise to zero. The rather permissive correlation threshold of 0.6 was chosen to also allow low abundant peaks with very short elution profile to pass. It is still sufficiently specific to reject peaks that have no or little time overlap.

For the creation of the graph, the two criteria above need to be checked on all pair-wise comparisons of 3D peaks, which results in a large number of arithmetic operations. Fortunately, there are two shortcuts, which apply in most of the cases. If two 3D peaks do not overlap in time, or are more than 1.01 Dalton apart, they will never be put together, because at least one of the criteria is not met.

Then, the graph is decomposed into connected sub-graphs, each of which represents a ‘pre-isotope’ pattern, in that each pair of peaks has a suitable mass difference and a sufficient correlation in retention time. However, these pre-isotope patterns may not be consistent because the pair-wise mass differences do not have to correspond to the same charge. Each pre-isotope pattern is now broken up into pieces such that each piece is a consistent isotope pattern for one single charge. To achieve this, each pre-isotope pattern is repeatedly split into two parts: a) the consistent isotope pattern contained in the pre-isotope pattern that has the highest number of peaks, and b) the remaining peaks. The former is taken over into the final list of isotope patterns, while the latter is recursively treated again as a pre-isotope pattern in the same way. We call an isotope pattern consistent if

$$\left| m - m_j - \frac{j \Delta M}{z} \right| \leq \left[\left(\frac{\Delta S}{z} \right)^2 + \Delta m^2 + \Delta m_j^2 \right]^{1/2}$$

where m is the mass of one of the peaks (central peak), and the index j numbers all other peaks relative to the chosen one, with negative/positive values to the left/right side. j has to run over consecutive values, meaning that 'holes' in the isotope pattern are not allowed. Furthermore, all peaks have to have a correlation of their retention time profile with the central peak of at least 0.6. Additionally, the correlation of the measured isotope pattern has to have a correlation of at least 0.6 with the averagine[3] isotope pattern of the same mass.

Detection of SILAC pairs

High precision quantitation of proteins can be achieved by using SILAC technology which is supported by MaxQuant in a fully automatic way. To be specific, we assume here that the cells were labeled in two forms, 'light' where all atoms have their natural isotope distribution, and 'heavy' where lysine and arginine have been incorporated in their heavy stable isotope labeled form - i.e. with $^{13}C_6{}^{15}N_2$ and $^{13}C_6{}^{15}N_4$ - leading to nominal mass shifts of eight and ten Daltons, respectively . The algorithms are not limited to this popular case. Which amino acids and which atoms in them are labeled can be freely configured. Also 'double triple' labeling[4] is supported, for which the additional algorithmic efforts are described in the next section. The algorithms do not assume that every labeling site coincides with an enzymatic cleavage site, though for the sake of efficiency of SILAC quantitation in general it is recommended that labels and enzyme are chosen in such a way that almost all labeled sites are also cleavage sites, for instance, that trypsin is used together with lysine and arginine as labeled amino acids.

To detect SILAC pairs we loop over all possible pairs of isotope patterns that have been determined in the previous step. First, the correlation of the intensity profiles over retention time is checked and required to be at least 0.5. This works technically in the same way as the correlation of 3D peaks in the determination of isotope patterns, with the exception that the intensity profiles are now summed up over the isotope peaks, and it is checked whether the correlation improves if one of the isotope patterns is shifted in retention time by plus or minus one scan. This is done because the labeling, especially with deuterium, might alter the chemical properties slightly, thus altering the retention time by a small amount. For this reason the correlation threshold is also lowered slightly in comparison to the isotope pattern assembly. For all pairs of isotope patterns that pass this retention time correlation test, have equal charges, and are close enough in mass, it is further investigated whether they may belong together to form a SILAC pair. Per default it is assumed that at most three labeled amino acids per peptide are possible. That means in the chosen example that we have to consider the following nine combinations of labeled amino acids: K, R, KK, KR, RR, KKK, KKR, KRR, and RRR. These label contents would correspond to the nominal mass differences 8, 10, 16, 18, 20, 24, 26, 28, and 30, respectively. For each of the nine hypothetical label contents we convolute the two measured isotope patterns with theoretical isotope patterns of the atoms one has to add so that both peptides would have the same atomic composition. For instance, when checking if the peptide contains one K and no R, we have to convolute the light isotope pattern with $^{13}C_6{}^{15}N_2$ and the heavy isotope pattern with C_6N_2 where C and N have the natural isotope distribution. As a result, the light isotope pattern will only be shifted by $6\left(m(^{13}C)-m(^{12}C)\right)+2\left(m(^{15}N)-m(^{14}N)\right)$ relative to the heavy pattern, while for the heavy pattern the relative peak heights will also be altered. The resulting isotope pattern should be exactly identical up to a global factor which is the intensity ratio between the heavy and light labeled form of the peptide. It is then checked if for the most intense peak in the light isotope pattern there is a corresponding peak in the heavy isotope pattern (and vice versa) such that

$$\Delta m \leq \mathrm{Max}\left(\left[\Delta m_1^2 + \Delta m_2^2\right]^{1/2}, 0.002\ \mathrm{Da}\right).$$

At this point we do not filter on the individual mass errors below 0.002 Da, because the correction for autocorrelation as well as the nonlinear recalibration have not been applied yet. This might otherwise lead to erroneous rejection especially of abundant and therefore very precisely measured SILAC pairs. In case the masses match it is checked if the intensity correlation of the two isotope patterns is at least 0.5. If up to here all tests have been passed, the two isotope patterns are associated in a SILAC pair.

Detection of SILAC triplets

Another popular application of SILAC labeling which allows comparing three cellular states in one experiment is 'double triple' labeling. In the 'light' form all atoms occur in their natural isotope distribution, in the 'medium' form lysine and arginine are labeled with $^{2}H_4$ and $^{13}C_6$ atoms, and in the 'heavy' form with $^{13}C_6{}^{15}N_2$ and $^{13}C_6{}^{15}N_4$. Also here the labeling of the medium and heavy peptides are not fixed to this choice but can be configured arbitrarily. The triplet case is reduced to the doublet algorithm by first finding light/medium, light/heavy and medium/heavy pairs. SILAC triplets are then detected by searching for triplets of isotope patterns that occur consistently in all three lists of pairs; in particular the labeled amino acid content has to be the same. Then, also triplets are taken whose isotope patterns occur consistently in two out of the three pair lists, e.g. the light/medium and the light/heavy pair have been detected, but the corresponding medium/heavy pair was not in the list. The remaining pairs that cannot be connected to a full triplet are discarded.

SILAC ratio estimation

As described in the section on the detection of SILAC pairs, the measured isotope patterns of the heavy and light peptides are convoluted with the theoretical isotope patterns of the missing atoms to obtain molecules of equal atomic composition. The resulting isotope patterns should only differ by a global factor which is the ratio between the heavy and light peptide. To determine this ratio we make a list of all corresponding 2D centroid intensities that are present in the heavy and light form. To these intensity pairs we fit a straight line through the origin, whose slope is the desired ratio. The linear fit is done in a robust way, taking the least squares solution as initial value and then solving the best median fit equation iteratively by bisection[5].

Normalization of SILAC ratios

To correct for mixing errors of total protein amount the SILAC ratios determined in the previous section are normalized so that the median of logarithmized ratios is at zero. This normalization is done in intensity bins, similarly as described in the section on the calculation of protein ratios and significance. It is done separately for lysine and arginine labeled peptides to compensate for any possible label-specific bias. This peptide ratio normalization is done for each LC-MS run separately, allowing for different protein mixing ratios in different runs.

Calculation of precise peptide masses and estimation of individual peptide mass errors

The mass of a peptide is calculated as a weighted average of all the 2D peak centroids in the 3D peaks within the isotope patterns belonging to a SILAC pair. As a weight the intensity of the 2D peak is taken. A peptide mass estimate from a single 2D peak is

$$m_{peptide} = z(m_z - m_P) - \Delta M_{iso} - \Delta M_{SILAC} \ .$$

m_z is the measured mass value, m_P the proton mass, z the charge, ΔM_{iso} is the mass difference in an averagine peptide between the isotopic peak that the 2D peak belongs to and the monoisotopic peak, and ΔM_{SILAC} is the mass difference of the monoisotopic peak of the SILAC pattern to the monoisotopic peak of the unlabeled peptide. The standard error of $m_{peptide}$ is again calculated by bootstrap resampling on the 2D peaks that go into the calculation, similar to the estimation of the mass error of the 3D peak masses described in the section on 3D peak detection.

Detection of charge pairs

Depending on the type of experiment, we want to find peptides that have been measured in multiple charge states, among the SILAC pairs. Among other things, these peptides can be used for non-linear mass re-calibration. Differently charged versions of the same peptide should show exactly the same retention time profile, since ionization happens after the chromatographic separation. Two SILAC pairs are detected as a charge pair if the retention time correlation, calculated in the same way as for the SILAC pair detection, exceeds 0.6, and the two peptide mass estimates are the same within seven ppm.

Non-linear mass recalibration

We use the SILAC pairs that occur in more than one charge state to do non-linear mass re-calibration. Usually, several hundreds of these pairs exist in every LC-MS run. They allow for a non-linear re-calibration without knowing the identity of any peptide. To determine the best re-calibration function, we minimize the quantity

$$\chi^2 = \sum_{pairs} \frac{(m_1 - m_2)^2}{\Delta m_1^2 + \Delta m_2^2}$$

where the sum is over all charge pairs. m_1 and m_2 are the peptide masses calculated from the two differently charged SILAC pairs, and Δm_1^2 and Δm_2^2 are the corresponding bootstrap errors. The dependence on parameters that are to be determined during minimization enters by replacing m_z by its re-calibrated version

$$m_z \to m_z\left(1 + 10^{-6}\left(P(u) + Q(v)\right)\right)$$

where

$$u = \frac{100}{\sqrt{m_z}} - \frac{100}{\sqrt{m_0}} \text{ with } m_0 = 445.120025$$

is the parameter through which the non-linear mass dependence enters and

$$v = \ln(I) - \ln(10^7)$$

is the logarithmized intensity up to an additive constants. $P(u)$ and $Q(v)$ are polynomials whose coefficients are to be determined in the minimization,

$$P(u) = p_1 u + p_2 u^2 + p_3 u^3 + p_4 u^4 + p_5 u^5 \text{ and } Q(v) = q_1 v \text{ .}$$

The constants in the definitions of u and v are not strictly necessary and their purpose is to have convenient number ranges for the parameters. m_0 is arbitrarily chosen as the mass of one of the compounds used as a lock mass for the LTQ Orbitrap. The dependence on

$1/\sqrt{m_z}$ is inspired by the relations between mass and frequency domains for FT and Orbitrap mass spectrometers. Note that there is no constant term present in either of the polynomials. The reason for this is that the method of comparing masses of charge pairs, while being very powerful for correcting for nonlinearities, is blind towards a constant ppm mass shift. This global shift will be determined later in the workflow after the peptide identities are known, as described in the section on linear mass recalibration. The minimization of χ^2 is done numerically with the Levenberg-Marquardt method[5].

Re-adjustment of individual peptide mass errors

The individual peptide mass errors estimated by the bootstrap method give us information about the precision of the mass measurements in that this is the variability of repeated measurements of the same mass. It would be advantageous to promote this error estimate to a bound on the mass accuracy, meaning the deviation from the measured to the true value. While this is impossible to achieve without knowing the identity of the peptide, we can try to do the next best thing, by requiring that mass estimates for the same peptide coming from different regions of the *m/z* scale should coincide within the quoted errors. Again we can use the charge pairs for this purpose. If the individual peptide mass errors should be used as bounds on mass accuracies, then at least the difference in the peptide mass estimates from differently charged versions of a peptide should be comparable to the estimated individual mass errors. The quantity

$$\delta = \frac{m_1 - m_2}{\sqrt{\Delta m_1^2 + \Delta m_2^2}}$$

can be calculated for every charge pair and is the difference in the independent mass estimates normalized by its expected error. The standard deviation of this quantity carries information about possible systematic over- or underestimation of errors. It would be expected to equal one for perfectly consistent error estimates. We often observe values of the standard deviation of δ ranging from two to three, indicating that the bootstrap errors systematically underestimate the true error. This is likely caused by autocorrelation[6] of the 2D centroid masses, meaning that the difference of mass estimates in immediately neighboring scans is systematically smaller than the difference of mass estimates in next to nearest scans. In any case, we take this kind of systematic effect into account by multiplying all error estimates by the standard deviation of δ which effectively normalizes all errors by the appropriate autocorrelation time.

Preparation of MS/MS spectra for database search

MS/MS spectra have been recorded in centroid mode. They are filtered so that the six most intense peaks per 100 Dalton intervals pass through. The processed MS/MS spectra are submitted to a Mascot server for MS/MS ion searches against digested sequence databases.

Since the SILAC state of many isotope patterns is known beforehand, we can treat the label modifications as fixed modifications in the Mascot search. In the example of lysine and arginine labeled with $^{13}C_6{}^{15}N_2$ and $^{13}C_6{}^{15}N_4$ respectively, we can sort the MS/MS spectra into three classes: (1) MS/MS spectra that were triggered on an MS isotope pattern detected as being the light form of a SILAC pair. These spectra can be searched by Mascot without any modifications related to SILAC labeling. (2) MS/MS spectra on the heavy part of a SILAC pair. For these spectra *'Arginine-13C615N4 (R-full)'* and *'Lysine-13C615N2 (K-full)'* can be taken as fixed modifications, since all arginines and lysines should occur in their heavy forms. (3) MS/MS spectra not associated with a SILAC pair. For these spectra *'Arginine-13C615N4 (R-full)'* and *'Lysine-13C615N2 (K-full)'* have to be taken as variable modifications. These Mascot searches are done with an initial peptide mass tolerance of 7 ppm, and with 0.5 Dalton tolerance for MS/MS peaks.

Filtering of Mascot results by amino acid content

In the .dat files containing all results of a Mascot search, for each submitted MS/MS spectrum not only the highest scoring sequence is reported, but a list of up to ten candidate peptides ranked by their score. We found that re-ordering the list of candidates according to the probabilistic p-score[7] slightly improves identification rates compared to keeping the order of candidates suggested by the Mascot score. Since for MS/MS spectra originating from a known SILAC pair the number of labeled amino acids (the number of lysines and arginines in the example above) is known, the candidate list can be filtered by removing all sequences that are incompatible with this information. Also, for the MS/MS spectra that were searched with the labelings as variable modifications, it is ensured that either all or no arginines and lysines are labeled, by removing those peptide candidates that do not fulfill this criterion.

Linear mass recalibration

While non-linear calibration problems have already been flattened out before peptide identification, there may still remain a global ppm shift that needs to be taken care of. Also here, we want the mass deviation of well measured peptides to enter with more weight into the averaging. So we calculate the Δppm global mass shift as the weighted mean of the ppm shifts of all identified SILAC pairs/triplets with a Mascot score of at least 30,

$$\Delta ppm = \frac{\sum_j \frac{m_j^2}{\Delta m_j^2} \Delta ppm_j}{\sum_j \frac{m_j^2}{\Delta m_j^2}} \ .$$

This Δppm is subtracted subsequently from all measured peptide masses.

Filtering of Mascot results by individual peptide mass errors

After re-calibration of the peptide masses we can remove all candidates suggested by Mascot whose difference between measured and calculated mass exceeds the tolerance as previously described calculated by the bootstrap method. We discard all candidates that are more than four standard deviations away from the calculated mass.

Posterior error probabilities and false discovery rates for peptides identified by MS/MS ion search

We perform the Mascot search against a concatenated database containing all true protein sequences, e.g. all proteins in the IPI database for a given species, plus all sequences in their reversed version. This decoy approach[8] allows for a straightforward assessment of the likelihood of false positive identifications. However, a problem with the reverse database approach is that it creates peptides that have precisely the same composition as the peptides in the forward database. This happens in half the cases with tryptic peptides and almost all peptides in the case of LysC peptides. Thus the reverse database overestimates the number of random hits, especially for very high mass accuracy data. Several remedies have been suggested, for example randomizing the sequences. However, these procedures change the local relationships between amino acids and may lead to a different length distribution. Here we avoid both problems by constructing the decoy database in two steps. First we reverse all sequences as before. In a second step, we swap each arginine and each lysine with the preceding amino acids (or only each lysine in the case of LysC). The decoy peptide database constructed in this way has the same mass and amino acid distribution but avoids the spurious repetition of the exact same mass values that are in the forward database. Ensuring that peptide masses are not repeated by design in the reverse database was not important until recently, because most of the information leading to the identification of a peptide came from the fragment ions in the MS/MS spectrum. However, when making maximum use of the high precursor mass accuracy, then this repetition does make a difference and the decoy model must be built in such a way that detrimental correlation artifacts are avoided. Indeed, we found that, in accordance with our expectations, the new procedure noticeably increased the number of positive identifications with FDR fixed at 1%.

First we generate two lists of number pairs. For every MS/MS spectrum submitted to Mascot whose candidate list of peptide sequences has at least one entry after the above filtering we take the Mascot score and sequence length from the top-scoring candidate and put it in one of the two lists. All instances where the top-scoring peptide was a digestion product from one of the original (forward) proteins, the number pair is put into the first list, otherwise, if it originates from a reverse sequence, it is put in the second list. From these lists of number pairs we generate two histograms by Gaussian kernel smoothing. These can be interpreted as approximations to the probability densities

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

where the Boolean variable X indicates whether the peptides originates from a correct protein hit or not. *s* denotes the Mascot score, while *L* is the length of the peptide. We chose Gaussian kernel smoothing for the determination of the functional dependence of these probabilities on the parameters *s* and *L*, because it is completely hypothesis-free and expected to be very precise in the regions of parameter space where both forward and reverse hits are present. Aiming for high accuracy in the tail regions where there are no decoy hits is not as important for the problem of determination of the true positive identifications. Peptides that are located in this region are by definition very likely to be correctly identified. Methods that attempt to precisely estimate PEP in the far tails have to rely on assumptions due to the lack of data and their inherent extrapolative nature. Therefore, we concentrated instead on the twilight zone where the decision is not as clear-cut and where an accurate discrimination is valuable to retrieve the largest number of possible identifications.

Using Bayes theorem, one can invert the conditional probability to obtain the probability of a false hit, given the Mascot score and the length of the peptide,

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

This quantity, which is a property of an individual peptide, is called posterior error probability (PEP) in the remainder. We assume that the a priori probability $p\left(X=\text{false}\right)$, which is just a constant overall factor, is 0.5. The smaller the PEP, the more certain is the identification of a peptide. The dependence on the peptide length will cause longer peptides to be accepted with a lower Mascot score than shorter peptides, for their comparatively higher uniqueness based on the parent mass alone. The functional dependence of the PEP on the Mascot score and the sequence length is not determined on individual LC-MS runs, but on the whole ensemble of files that are analyzed together, to ensure good statistics and make an estimate that is as precise as possible. PEP calculations are done separately for MS/MS spectra that were a priori detected as the light or heavy form of a SILAC pair, and for those where this information was not available, since they are assumed to have different statistical significance.

There is the option in MaxQuant to filter peptides by PEP, meaning that all peptides above a certain PEP threshold value, e.g. 0.01, would be discarded. The other option is to filter by false discovery rate (FDR). For this, all peptide identifications are sorted by their PEP, starting with the best. Then successively peptides are accepted along the list, until a specified fraction, e.g. 1%, of reverse peptides has accumulated. Then one can assume that among the forward peptides the fraction of wrong identifications is 1% as well. The FDR thresholding is done on individual raw files using the PEP formula generated by the whole ensemble of raw files. For the datasets analyzed in this publication a FDR rate of 1% was taken, while no additional threshold on the PEP was applied. Another option in MaxQuant allows accepting all those MS/MS spectra as identified when its top-scoring peptide sequence has already been identified in another spectrum. While having no influence on the number of identified peptides and proteins this will be valuable when comparing protein ratios across several experiments.

Re-quantitation

Many of the isotope patterns that have not been assembled into SILAC pairs will have an associated MS/MS spectrum which leads to the identification in the conventional Mascot search with the label mass differences as variable modifications. For these isotope patterns we know then whether it is the peptide in the light or in the heavy labeled form. Since we know the amino acid composition as well, we can calculate at which mass the potential missing SILAC partner is expected. The m/z-time shapes of the 3D peaks belonging to the identified SILAC version are then translated to the place where the SILAC partner is supposed to be located. If at least three 2D peak centroids fall into these shapes, ratios are calculated in the same way as described above for SILAC pairs that were detected before identification.

Protein assembly

To reduce the redundancy of the protein sequences that contain any of the identified peptides we first make all pair-wise comparisons between proteins. 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 together in a protein group. The proteins in each resulting protein group are then sorted in decreasing order by the number of identified peptides that they contain. This way the first protein sequence in a protein group contains all the peptides of the other proteins and is therefore able to explain the occurrence of these peptides by itself. A peptide is called unique to a protein group if it only occurs in this group and nowhere else in the proteome. To make a tentative assignment of the peptides that are shared between more than one protein group to one of these, we consider the total number of peptides in each group as a criterion. The non-unique peptide is assigned to the protein group with the highest number of peptides as so-called ‘razor’ peptide, alluding to Occam’s razor, since this would be the simplest explanation of the peptide.

Protein false discovery rate

We assign to each protein group a PEP by multiplying the contained peptide sequences’ PEPs. Here, each distinct peptide sequence contributes only one factor, the PEP of the MS/MS spectrum for a given peptide sequence with the lowest (best) PEP value. That means, the collection of all MS/MS spectra associated with a peptide sequence which might consist of re-sequencing events on the same peak, sequencing on different isotopic peaks in the same isotope pattern, sequencing of different charge states, different SILAC or modification states are all entering the protein PEP only with one single peptide PEP.

In this way we avoid an overly optimistic estimation due to violation of statistical independence. In any case, a statistical dependence as well as lack of multiple hypothesis testing corrections would generate a conservative list of identified proteins, since the protein PEP is only used for sorting the list of proteins as input for applying a protein FDR threshold based on the target-decoy approach. We can roughly estimate the correctness of statistical independence by considering the PEP calculated this way for the reverse protein with the smallest (best) PEP in the list. For our dataset, this PEP is 2.7e-5. This number should be compared with the best p-value one would expect to obtain in a multiple hypothesis testing context, if one would perform 4,506 tests, which equals the number of protein groups and also the number of times a protein PEP has been calculated, and if the null hypothesis would always be true, i.e. no real effect is present. In the spirit of the Bonferroni correction this best p-value obtained by chance would roughly be 1/4,506 which coincides within an order of magnitude with the best PEP for a reverse protein.

Protein groups are then sorted by the PEP and a given protein false discovery rate is ensured by terminating the list of proteins so that a given percentage of reverse proteins are contained. We used as a standard a protein FDR of 1%. Following this procedure ensures that at most one percent of the proteins are wrongly identified. Finally all proteins without unique peptides are removed. Proteins with only one unique peptide may be discarded as well or they may undergo a manual validation procedure, depending on the stringency that is desired in the project.

Calculation of protein ratios and significance

Protein ratios are calculated as the median of all SILAC pair ratios that belong to peptides contained in this protein. By using this robust kind of averaging, the presence of outlier ratios will leave the protein ratio largely unaffected. A version of protein ratios that is corrected for unequal total protein amounts is calculated by taking the median of the normalized versions of SILAC pair ratios, which are calculated as described in a previous section.

The problem of finding significant responders can be treated in two different ways. In the microarray field, to find differentially regulated transcripts between two cellular states, one usually performs a number of replicate measurements of both, and then filters for 'responders' by a t-test like approach[9]. There are many variations on this approach but the general idea is always to compare 'within replicate group variance' with 'between replicate group variance'. The objective in proteomics, in contrast, is usually to determine outlier protein ratios in comparison to the distribution of all protein ratios. Our scores implement and substantially improve this tradition e.g. by taking the signal intensity into account.

To assess the significance of outlier ratios we propose two measures of deviation from the bulk of the distribution, which we call significance A and B. We consider all normalized protein ratios as obtained in the previous steps and take their natural logarithm. The base of the logarithm is irrelevant for our calculation. By taking the logarithm, an equal treatment of up- and down-regulation is ensured. To make a robust and asymmetrical estimate of the standard deviation of the main distribution we calculate the 15.87, 50, and 84.13 percentiles r_{-1}, r_0, and r_1. We define r_1-r_0 and r_0-r_{-1} as the right- and left-sided robust standard deviations. For a normal distribution, these would be equal to each other and to the conventional definition of a standard deviation. A suitable measure for a ratio $r>r_0$ of being significantly far away from the main distribution would be the distance to r_0 measured in terms of the right standard deviation

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

Similarly, for a ratio $r<r_0$ one would take

$$z = \frac{r_0 - r}{r_0 - r_{-1}}$$

for the same purpose. Under the null hypothesis of normal tails of the log ratio distribution the probability of obtaining a value this large or larger is

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

The assumption that the ratio distribution is normal in a case where no differential regulation takes place is reasonable, since the protein ratios are obtained as averages of many SILAC peptide ratios. In the limit of a large number of SILAC peptides per protein a normal distribution can be assumed due to the central limit theorem.

In Figure 6 it can be seen that the width of the bulk distribution of logarithmic ratios depends on the protein intensity. For highly abundant proteins the statistical spread of unregulated proteins is much more focused than for low abundant ones. Because of this, a protein that shows, for instance, a ratio of two should be very significant when it is highly abundant, while at very low abundance it should only be marginally significant. To capture this effect we define another quantity, called significance B, which is calculated in the same way as significance A, but on protein subsets obtained by grouping them into intensity bins. We divide the proteins into bins of equal occupancy such that each bin contains at least 300 proteins. The above calculation for significance A is then repeated in each bin to obtain significance B. Significance A and B can both be corrected for multiple hypothesis testing, e.g. with the method proposed by Benjamini and Hochberg[10]. For our dataset we report significant ratios based on significance B with a Benjamini-Hochberg corrected p-value threshold of 0.05.

References

1. 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).
2. Efron, B. & Tibshirani, R. An Introduction to the Bootstrap. (Chapman & Hall/CRC, 1993).
3. Senko, M.W., Beru, 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).
4. 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).
5. Press, W.H., Teukolky, S.A., Vetterling, W.T. & Flannery, B.P. Numerical Recipes in C, Second Edition. (Cambridge University Press, 1992).
6. Sokal, A.D. Monte Carlo Methods in Statistical Physics: Foundations and New Algorithms. (Lausanne; 1996).
7. Olsen, J.V. & Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc Natl Acad Sci U S A* **101**, 13417-13422 (2004).
8. 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).
9. Cui, X. & Churchill, G.A. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* **4**, 210 (2003).
10. 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).

Table of parameters

Step	Parameter	Comment
3D peak detection	Two peaks in neighboring scans are connected whenever their centroid *m/z* positions differ by less than 7 ppm.	Should scale linearly with mass precision. Window is relatively wide in order not to lose low abundance peaks at this early stage.
3D peak detection	Intensity profiles are smoothed over retention time applying a window mean filter of ±1 scan width and checked for local minima. If a minimum is found whose value is 1/1.3 of the lower of the two local maxima the 3D peak is split at the minimum position.	Has to be set sufficiently high to avoid peak splitting due to random intensity variations between neighboring scans. This may depend on instrument type and even the setup of the individual instrument.
3D peak detection; De-isotoping; SILAC pair detection	150 bootstrap replicates for determination of mass errors	Strikes a balance between accuracy of error and computation time. It should be largely independent of the instrumentation.
De-isotoping	We ensure that the intensity profiles have a sufficient overlap in retention time, by requiring them to have a cosine correlation (uncentered Pearson correlation) greater than 0.6. Additionally, the correlation of the measured isotope pattern has to have a correlation of at least 0.6 with the averagine isotope pattern of the same mass.	The correlation threshold should on the one hand guarantee a minimum overlap, but on the other hand not be too restrictive since low abundant peaks will not reach high correlations even when perfectly co-eluting.
SILAC pair detection	To detect SILAC pairs, the correlation of the intensity profiles between isotope patterns over retention time is checked and required to be at least 0.5. In case the masses match it is checked if the intensity correlation of the two isotope patterns is at least 0.5.	It turned out to be beneficial to have a slightly less stringent correlation threshold compared the isotope pattern assembly to include high ratio, and thus low abundant signals. This is possible since there are fewer false positives in SILAC pair assembly due to the restrictive mass difference criterion.

Charge pair detection	Two SILAC pairs are detected as a charge pair if the retention time correlation, calculated in the same way as for the SILAC pair detection, exceeds 0.5, and the two peptide mass estimates are the same within seven ppm.	Again, the correlation threshold should be low enough not to discard low abundance peptides. The mass window should be sufficiently large, because up to this point the mass re-calibration has not been performed.
Preparation of MS/MS spectra	MS/MS spectra have been recorded in centroid mode. They are filtered so that the six most intense peaks per 100 Dalton intervals pass through	For CID one would expect a singly charged y and a b ion and possibly some multiply charged ions per average amino acid mass interval. Therefore six peaks per 100 Dalton is generally a good choice.
Filtering of peptide candidates	We discard all candidates that are more than five standard deviations away from the calculated mass.	Only candidates that are clearly inconsistent with the individual mass errors are discarded.
PEP calculation	The prior probability $p\left(X = \text{false}\right)$ is set to 0.5.	This is just a constant factor independent of peptide identification score and peptide length. Its value has no effect on the final list of accepted peptides at a given FDR.
FDR control of identifications	Peptide FDR < 1%; Protein FDR < 1%	These are rather conservative identification thresholds, which could be loosened if desired, especially on the peptide level. However, we found that a restrictive FDR control also at the peptide level significantly improves the quality of protein quantitation.
Protein quantitation	Significance B based Benjamini-Hochberg FDR < 5%	Guarantees that at most 5% of the reported responders are in fact non-responding. This threshold might be adjusted according to personal preferences.