nature biotechnology

Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs

Jan Eriksson[1,3] & David Fenyö[2,3]

Truly comprehensive proteome analysis is highly desirable in systems biology and biomarker discovery efforts. But complete proteome characterization has been hindered by the dynamic range and detection sensitivity of experimental designs, which are not adequate to the very wide range of protein abundances. Experimental designs for comprehensive analytical efforts involve separation followed by mass spectrometry–based identification of digested proteins. Because results are generally reported as a collection of identifications with no information on the fraction of the proteome that was missed, they are difficult to evaluate and potentially misleading. Here we address this problem by taking a holistic view of the experimental design and using computer simulations to estimate the success rate for any given experiment. Our approach demonstrates that simple changes in typical experimental designs can enhance the success rate of proteome analysis by five- to tenfold.

Experimental design is critical for the success of a proteomics experiment. A good design must handle the complexity and the very wide range of protein abundances. Global abundance measurements in *Saccharomyces cerevisiae*—using an antibody against a tag engineered into *S. cerevisiae* genes, followed by quantitative western blot analysis—have revealed a bell-shaped distribution of proteins spanning about six orders of magnitude in abundance[1]. Estimates for body fluids indicate much larger ranges of protein abundances (10^{10} or higher)[2], with a distribution that is still unknown. In contrast, the dynamic range of experimental methods typically used in proteomics spans only a few orders of magnitude, hampering the identification of low-abundance proteins. State-of-the-art experimental designs in proteomics involve[3] (i) taking samples of proteins relevant to the biological hypothesis or phenomenon explored; (ii) protein separation by liquid chromatography (LC) and/or gel electrophoresis[4]; (iii) protein digestion using an enzyme of high specificity, followed by chromatographic[5] or electrophoretic separation[6] and mass spectrometric (MS) analysis[7] of the proteolytic peptides; and (iv) searching a protein-sequence collection to identify proteins based on the MS and tandem MS (MS/MS) information[8,9]. The numerous choices available for each step in the workflow make it prohibitive to fully optimize the design of the workflow experimentally. To address this issue, we have developed a simulation tool for evaluating the success of current designs and for predicting the performance of future, further optimized proteomics experimental designs. The simulation takes a holistic view of a general analytical experiment and aims at identifying pertinent factors that influence the success rate. Here, we assess the performance of this approach for predicting the success of proteome analyses of human tissue and body fluid that use various state-of-the-art experimental design principles.

The model of the common major steps of a proteomics experiment, including separation of proteins and peptides, is shown in **Figure 1a**. Several parameters are required to simulate the steps of a proteome analysis: (i) the distribution of protein amounts in the sample analyzed; (ii) the loss of analyte material and the maximal limit of the amount loaded in each step of sample manipulation (separation, digestion, chemical modification and so forth); (iii) the dynamic range, the detection limit and the losses associated with MS analysis. Depending on what experiment is modeled the detection limit employed in a simulation can represent either protein identification only (lower limit of identification) or protein identification with quantification (lower limit of quantification).

We have defined and studied two quantities as a function of the model parameters: the success rate (**Fig. 1b**) and the relative dynamic range (RDR; **Fig. 1c**). The success rate is simply the ratio between the number of distinct proteins detected (or quantified) and the total number of distinct proteins in the sample. The quantity RDR is defined as the ratio between the logarithm of the range of abundances of proteins detected (or quantified) and the logarithm of the entire range of protein abundances of a proteome. Hence, RDR is a measure of how far down in protein abundance we can reach for a given proteome. The success rate and RDR provide key pieces of information that so far have been neglected when reporting proteome analysis results. Appropriate assignment of the success rate and RDR for a proteomics study can simplify evaluation and prevent a misleading interpretation of results.

We demonstrate that modeling and simulation can teach us how to improve an experimental design with a low success rate and low RDR (**Fig. 2a**,**b**, I). In this experimental design, a complex mixture of proteins from a tissue or body fluid sample are digested without

[1]Department of Chemistry, Swedish University of Agricultural Sciences, Box 7015, SE-750 07, Uppsala, Sweden. [2]The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA. [3]These authors contributed equally to this work. Correspondence should be addressed to J.E. (jan.eriksson@kemi.slu.se) or D.F. (fenyo@rockefeller.edu).

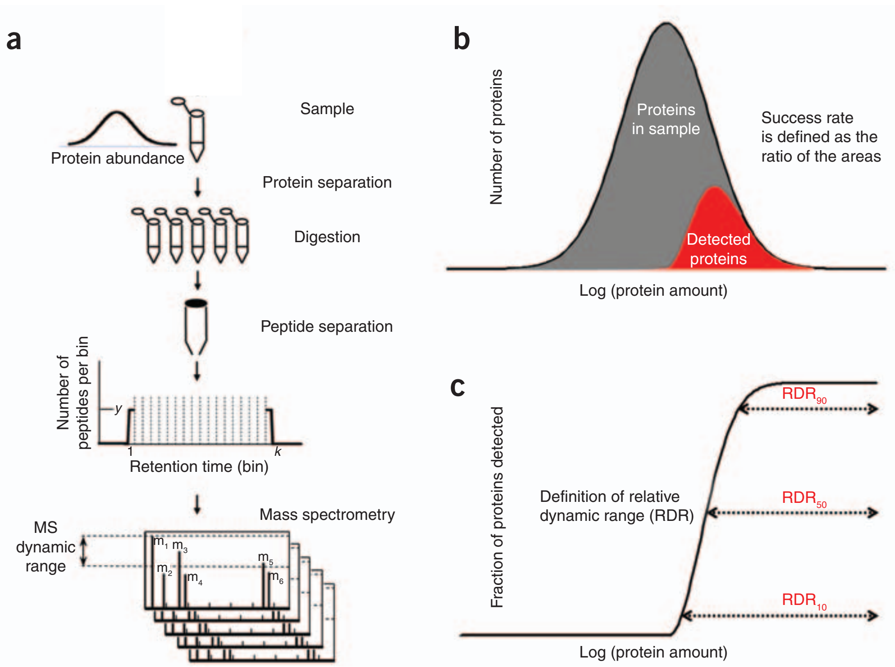

Figure 1 The proteomics workflow modeled and definitions of success rate and relative dynamic range, RDR. (**a**) Workflow model used for simulating proteome analysis. A protein sample is taken from an organism. The protein-abundance distribution is modeled according to the organism and type of sample (e.g., tissue or body fluid). The proteins are separated into fractions and each fraction is digested with a proteolytic enzyme of high specificity (typically trypsin). Each digested fraction is then subjected to peptide separation. The separated peptides are subjected to mass spectrometric analysis. In the simulations, two quantities—success rate and RDR—are studied as a function of experimental design. (**b**) The success rate of a proteomics experiment is the ratio between the area under the distribution of detected proteins and the area under the protein-abundance distribution of the proteome investigated. Alternatively, when the abundance distribution is unknown, the success rate can be calculated simply as the number of distinct proteins identified divided by the total number of distinct proteins in the proteome. The number of proteins in an organism is defined as the number of genes plus known splicing variants. (**c**) The relative dynamic range RDR_x is the logarithm value of the range of abundances where at least x% of the proteins are detected divided by the logarithm value of the entire range of protein abundances in the proteome studied. RDR_x is a measure of how far down in protein abundance we can reach for a given proteomic experiment. The relative dynamic range and the success rate provide appropriate measures of what fraction of the proteome an analysis is detecting and what fraction is missing due to experimental imperfection.

prior protein separation. The resulting proteolytic peptides are separated by reversed-phase chromatography (RPC) using a nanobore column (≈75 µm internal diameter (i.d.)) before MS analysis. Potential improvement of this experimental design is demonstrated by varying three fundamental experimental design parameters: the degree of protein separation, the amount of peptides loaded on the RPC column and the degree of peptide separation. The explicit effect of each respective fundamental design parameter is elucidated by assuming that losses are independent of other parameters.

Protein separation improves the success rate and extends the RDR of the experiment, and most proteins can be detected when extensive protein separation is performed (**Fig. 2a,b**). However, the drawbacks of extensive protein separation include: longer analysis time, larger protein losses and a requirement for larger sample amounts (provided that the same amount of proteolytic peptides is loaded on the column for each protein fraction).

Increasing the amount of peptides loaded on the RPC column improves the success rate and the RDR (**Fig. 2c,d**). Unfortunately, many experimental designs are not focused on maximizing the amount of peptides loaded on the RPC column but are instead focused on minimizing sample handling and losses. For example, the popular nanobore columns allow flow rates that are suitable for online MS, but these narrow columns severely limit the amount of material that can be analyzed (maximum amount loaded $\propto$ i.d.^{2}). An alternative strategy to increasing the amount of peptides would be to improve the detection sensitivity of the mass spectrometer (**Supplementary Fig. 1** online).

Finally, improving the peptide separation leads to an additional increase in the success rate and the RDR (**Fig. 2e,f**). An alternative strategy to improve peptide separation would be enhancement of the dynamic range of the mass spectrometer (**Supplementary Fig. 2** online).

The combined effect of these improvements—limited protein separation, analyzing more material and better peptide separation—leads to an experimental design that can be used to sample a large portion of the proteins in tissue. In contrast, for the wider range of protein abundances of body-fluid samples, this experimental design fails to analyze anything but the most abundant proteins.

The relative increases in the success rate and RDR shown in **Figure 2** are obtained by improving protein separation, increasing the amount of peptides loaded on the column, and improving the peptide separation (I–IV). The impact on the success rate and RDR of each respective design parameter is dependent on the other design parameters. For example, the effect of improving the peptide separation with and without loading more material on the column employed for separating the proteolytic peptides is dramatically different (**Fig. 3**). The enhanced peptide separation *per se* does not lead to an improvement of the success rate and RDR (**Fig. 3**, II, III) unless the amount loaded on the column is increased (**Fig. 3**, III, IV).

The examples of **Figures 2** and **3** display the effects of changing the parameters related to peptide separation for a fixed level of protein separation. The effects of changing the parameters related to peptide separation for different levels of protein separation are displayed in **Supplementary Figures 3** and **4** online.

We conclude that the factor limiting the success of many proteomics experiments is the amount of material analyzed—often because experimental designs include a nanobore-column or because the detection limit of current mass spectrometers is too poor. The effect of this limiting factor is reduced when two dimensions are used for the peptide separation, but the improvement is hampered by the increased losses associated with the additional separation step (**Supplementary Fig. 5** online).

In **Figures 2** and **3** losses are modeled as being independent of other parameters, and the amount of peptide material analyzed is

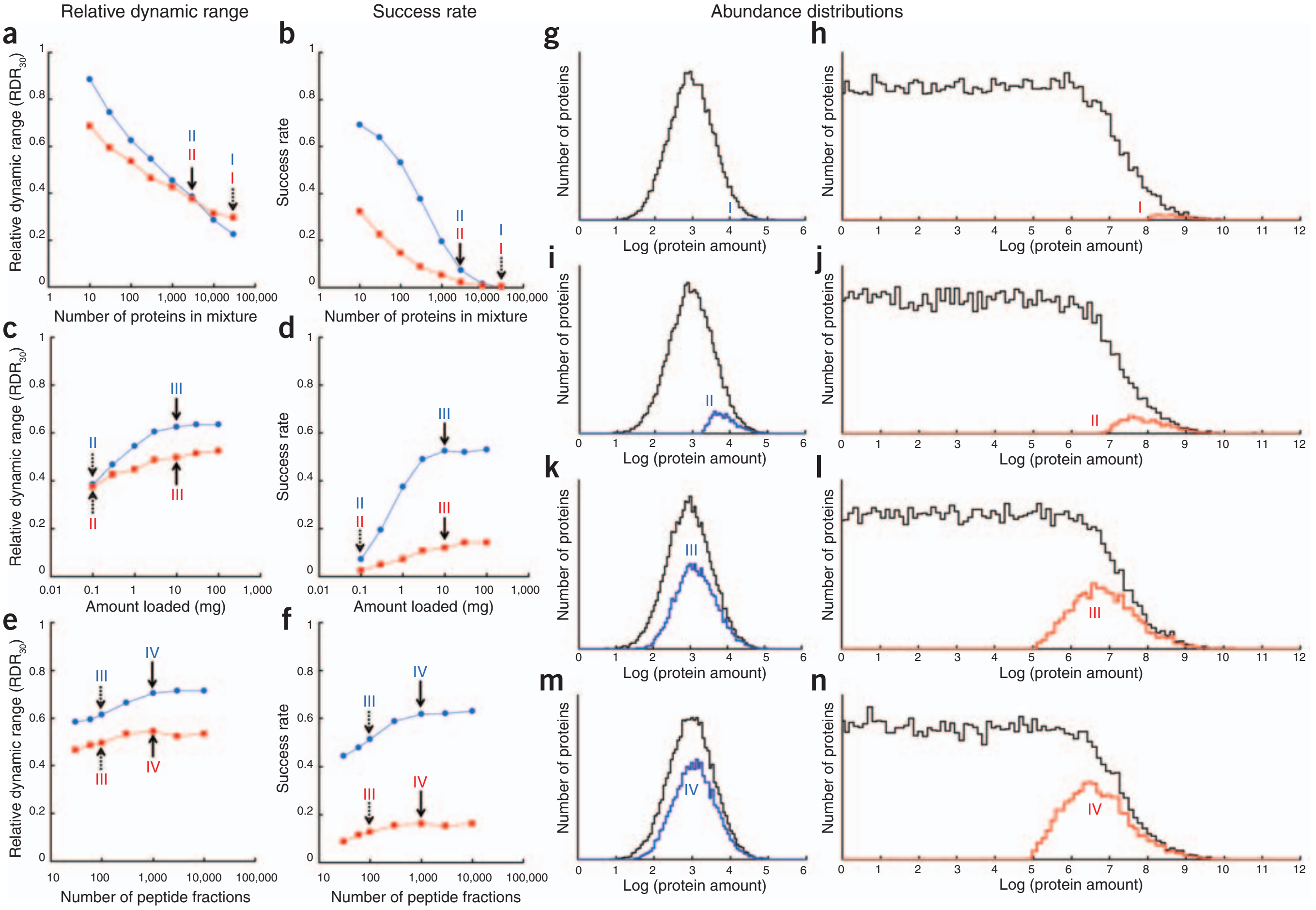

Figure 2 Simulations of proteome analysis of *Homo sapiens*. (**a**–**f**) Relative dynamic range, RDR_{30}, as a function of experimental design parameters (**a**,**c**,**e**). Success rate as a function of experimental design parameters (**b**,**d**,**f**). (**g**–**n**) The distributions of protein abundances (black) and the distributions of detected proteins (blue for tissue and red for body fluid) for selected experimental design parameter sets (I–IV). In the first selected experimental design parameter set (I) the sample contains 30,000 proteins (no protein separation) and the proteolytic peptides are separated into 100 fractions with an RPC column loaded with 0.1 μg of peptides. The influence of different degrees of protein separation is shown in **a** and **b**. It is seen that a change from a design that analyzes digests of 30,000 proteins (I), to a design that analyzes digests of 3,000 proteins (II) yields a valuable improvement of the RDR_{30} (**a**) and the success rate (**b**). The effect of increasing the amount of peptides loaded onto the column is shown in **c** and **d**. Important gains in the RDR_{30} (**c**) and the success rate (**d**) are obtained by increasing the amount loaded from 0.1 μg (II) to 10 μg (III). The effect of improving the peptide separation is shown in **e** and **f**. As the peptide separation is enhanced from yielding 100 fractions (III) to 1,000 fractions (IV) a substantial fraction of the human tissue proteome can be detected, whereas the detection from the body fluid still displays only moderate success. Based on the measurements on yeast in Ref. 1, we assumed that the shape of the protein-abundance distributions is Gaussian (σ = 0.6) for tissue and semi-Gaussian (σ = 1.2) for body fluid ranging six and twelve orders of magnitude (±5σ), respectively. The semi-Gaussian distribution takes into account that many different proteins can be secreted in low amounts into a body fluid. The total and pre-column survival probabilities were 10% and 90%, respectively. The MS detection sensitivity was 1 fmol and the MS dynamic range was 10^2.

varied over a wide range. This is straightforward to implement experimentally if the MS analysis is performed off-line, which allows, for example, the peptide amount to be increased by using a larger bore column. For online analysis it is less straightforward to load more material. The ideal flow rate and the maximum amount loaded scale in the same way with column i.d., thereby keeping the concentration of eluting peptides constant and cancelling the effect of loading more material. However, a gain in RDR and success rate for online experiments can still be obtained by improving the detection limit.

Another consideration in online MS is that a finite limit of the rate of data acquisition (sampling) yields losses that depend on the complexity of the sample analyzed. The finite rate of acquisition will lead to losses of proteins that are detectable (according to the detection sensitivity and dynamic range). The effect of sampling losses can be modeled and is demonstrated in **Supplementary Figure 6** online, where it is seen that when the acquisition rate is limiting, improved separation yields detection of a larger number of high-abundance proteins and therefore improves the success rate but not the RDR. An alternative way to partially overcome sampling losses is to perform repeated online analysis. The repeated analysis will yield a higher overall success rate than a single experiment (**Supplementary Fig. 7** online), but with negligible impact on RDR.

Much method development in proteomics has been focused on improving peptide separation in online experiments to improve detection of low-abundance proteins. Contrary to this belief, the modeling and simulations indicate that improving RDR is not possible by improving the peptide separation alone. We validated this finding experimentally. Whole-cell lysate of *S. cerevisiae* was proteolytically digested, and the resulting peptide mixture was analyzed by LC-MS/MS using different degrees of separation (two different gradient

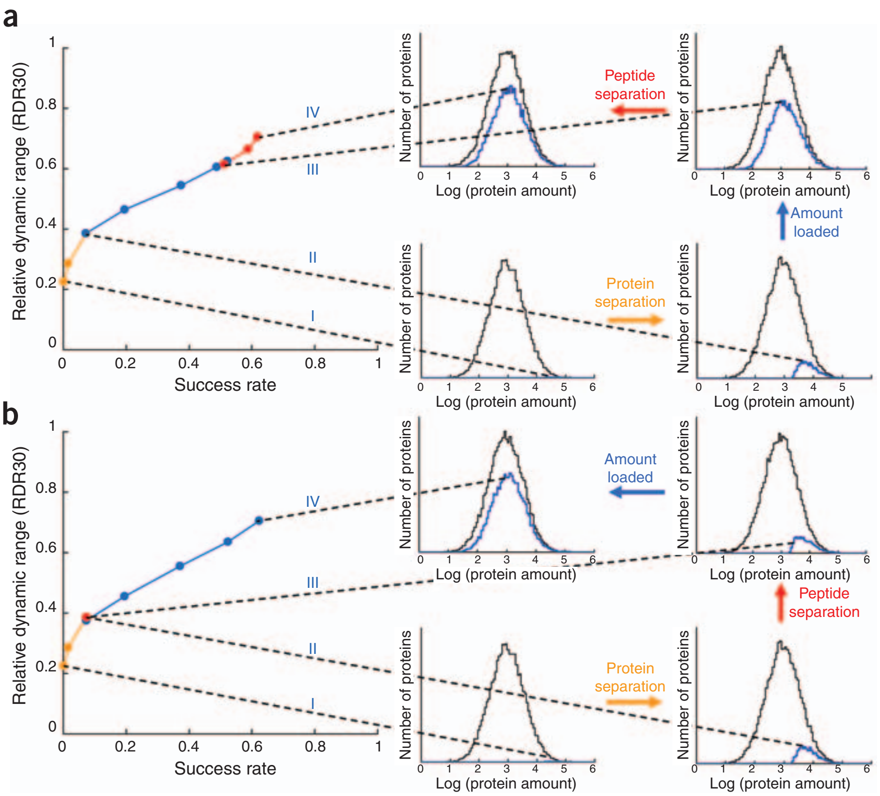

Figure 3 Simulation results revealing the influence of the order by which various design parameters are varied on success rate and RDR_{30}. (**a**) The RDR_{30} as a function of the success rate when first improving the protein separation (I–II), followed by increasing the amount of tryptic peptides loaded on the column (II–III), and finally enhancing the tryptic peptide separation (III–IV). (**b**) The improved protein separation (I–II) is followed by enhanced peptide separation (II–III), which leads to no improvement of the success rate and RDR_{30} until the amount loaded on the column is increased (III–IV).

lengths) and with a fixed low amount of material loaded on the RPC column. The results from these experiments show that improving the separation increases the success rate but has no influence on RDR (**Supplementary Fig. 8a**,**b** online). Repeated online analysis is an alternative strategy for improving the success rate, but again there is no improvement of RDR (**Supplementary Fig. 8c**,**d**). The gain in success rate when improving the separation or repeating the analysis is entirely due to reduced sampling losses of high-abundance proteins, whereas low-abundance proteins remain undetected—in agreement with the simulation results (**Supplementary Figs. 6** and **7**).

Our results clearly show that to reach both a high RDR and a high success rate some protein separation should be performed even at the expense of analysis time. After one has settled on a level of protein separation, the focus should be to find out how to analyze more material or alternatively improve the detection sensitivity of the mass spectrometer—depending on whether the MS analysis will be performed off-line or online. Only after this has been done should the focus be on improving the peptide separation or, alternatively, the dynamic range of the mass spectrometer. Even slight modifications of common proteomics techniques can raise the success rate for proteome analysis of human tissue from 1–10% up to 50–70%.

Computer simulations are ideal for optimizing experimental designs and identifying bottlenecks, because a large parameter space can be investigated even with modest computational resources. The exact values of many of the parameters needed to describe the experiment are uncertain, including the losses of proteins and peptides as they travel through the system. By investigating a wide range of parameters (e.g., protein-abundance distributions, losses of peptides and separation models), we have shown that the identification of bottlenecks and the general trends of influences on the performance of various experimental designs are not very sensitive to the precise values of the design parameters (**Supplementary Figs. 1–5** and **9–14** online).

We foresee that in the near future computer simulations using detailed models of the many analytical steps will be used to design better proteomics experiments. These simulations will refocus proteomics efforts toward experimental designs that have higher chances of success and will limit the current waste of resources. Higher chances of success are instrumental to the completeness of the analyses needed for developing systems biology and for discovering low-abundance protein biomarkers. We envision that proteomic studies will be reported with accurate assignment of the statistical significance[10,11] of each result and with information on what fraction of the proteome was actually covered by the experiment. The success rate can always be assigned when there is an estimation of the total number of proteins in the proteome. The RDR provides more refined information on what fraction of a proteome has been detected and can be assigned once the protein abundances of a proteome are known.

METHODS

Simulations. The simulations were performed by randomly selecting a mixture of proteins from the human proteome. Each protein in the mixture was randomly assigned a different amount based on the distribution of protein amounts in the sample. The mixture of proteins was digested and the resulting proteolytic peptides were randomly selected based on the assumed precolumn survival probability. The proteolytic peptides surviving the precolumn process were separated into fractions, either randomly or according to a separation model[12]. The separated peptides were randomly selected based on the assumed total survival probability. Finally, the surviving peptides were considered detected by MS if their amount was above the detection limit and their peak intensity was within the dynamic range of the mass spectrometer. The entire process was repeated to obtain sufficient statistics (at least 30,000 proteins selected). The simulation tool is accessible at http://prowl.rockefeller.edu/modeling/.

Distribution of protein amounts in the sample. Based on published measurements on yeast[1], we assumed that the shape of the protein-abundance distributions is Gaussian for tissue and semi-Gaussian for body fluids (**Figs. 2** and **3** and **Supplementary Figs. 1–5** and **11–14**). The semi-Gaussian distribution takes into account that many different proteins can be secreted in low amounts into a body fluid. The parameter σ defining the Gaussian shape of the distributions was scaled with the estimated range of protein amounts in tissue and body fluid, 6 (ref. 1) and 12 (ref. 2) orders of magnitude, respectively: $\sigma_{\text{tissue}} = 0.60$ and $\sigma_{\text{Body Fluid}} = 1.20$ (**Figs. 2** and **3**). The effect of changing the σ-value for the protein-abundance distribution for tissue was tested ($\sigma_{\text{tissue}} = 1.00$, **Supplementary Fig. 11**). The effect of alternative protein-abundance distributions for tissue was tested: semi-Gaussian (**Supplementary Fig. 9**) and constant over the entire range of protein amounts (**Supplementary Fig. 10**).

Sample handling: losses. Peptides can be lost by not being released during digestion of the sample, sticking to the walls of tubes and capillaries, not being eluted off the column, not ionizing properly in the ion source of the mass spectrometer, not being selected for fragmentation or not fragmenting well enough to provide sufficient information to identify the peptide with high statistical significance. All these potential losses were modeled using two parameters: (i) the total survival probability, that is, the probability that the peptide survives from digestion through identification; and (ii) the precolumn survival probability, that is, the probability that the peptide survives until it is bound to the column. **Supplementary Figure 5** shows the effect of changing the total survival probability.

Sample handling: amount limit. At each sample-handling step the maximum amount that can be used is potentially limiting. For the experiments typically performed in proteomics analysis, the limiting step is the RPC separation of the proteolytic peptides using a nanobore column (≈75 μm i.d.). We studied the effect of changing the amount limit for the peptide separation step (**Figs. 2** and **3** and **Supplementary Figs. 1–5** and **9–14**).

Sample handling: separation. The protein separation was assumed to distribute the proteins randomly and uniformly among the fractions resulting in an equal number of proteins in each fraction (10, 30, 100, 300, 1,000, 3,000, 10,000 and 30,000 proteins). It was assumed that the peptide separation results in the peptides being uniformly and randomly distributed among the fractions (30, 60, 100, 300, 1,000 or 10,000) (**Figs. 2** and **3**). This approach was compared with using a retention time model for RPC separation[12] (**Supplementary Figs. 12** and **13**). The different degrees of fractionation can be envisioned as a variation of the difference between the centroids of the peaks of eluting peptides without variation of peak duration. In many proteomics experiments the separated peptides are directly analyzed by MS and fractions are not collected. When comparing the simulated separation with these types of experiments, the number of fractions in the experiment can be estimated from the ratio between the effective length of the separation and the width of a peptide peak.

MS analysis. The MS detection limit (10, 100 and 1,000 amol) defines the minimum amount of a peptide that is required for detection in the mass spectrometer (**Supplementary Fig. 1**). The MS dynamic range (2, 3, 4 and 5 orders of magnitude) is the range of peak intensities that can be detected simultaneously—that is, it is the ratio between the largest and the smallest peaks that can be observed at the same time (**Supplementary Fig. 2**). The peak intensities were assumed to be proportional to the peptide amount with a proportionality factor that was selected from a random uniform distribution (± 1 order of magnitude). The effect of varying the range of the peak intensity variation is shown in **Supplementary Figure 14**. The finite MS data acquisition rate can limit the success rate and RDR. The finite acquisition rate was modeled by selecting the *x* most abundant peptides in each fraction (**Supplementary Fig. 6**). Losses due to acquisition-rate limitations can potentially be overcome by repeated analysis. The effect of repeated analysis was studied by repeated simulation of the MS detection step for the same randomly selected protein mixture (**Supplementary Fig. 7**).

Experimental validation. *S. cerevisiae*, grown to mid-log phase, was harvested by centrifugation and frozen as pellets in liquid nitrogen and disrupted with a Retsch MM301 mixer mill that was maintained at liquid nitrogen temperature, and stored at –80 °C. The proteins were precipitated with trichloroacetic acid, resuspended in 6 M guanidine HCl, and quantified using the BCA (bicinchoninic acid) Protein Assay kit (Pierce). The protein mixture was diluted with 100 mM ammonium bicarbonate and digested using two enzymes: first, Endoproteinase Lys-C (Sigma) for 6 h in 2 M guanidine HCl and second, trypsin (Promega) for 24 h in 0.5 M guanidine HCl. We loaded 0.6 μg of the resulting peptide mixture on to a Zorbax 300-SB C18 300 μm i.d. × 150 mm column (Agilent) and analyzed the results by LC-MS/MS using a SMART System (GE Healthcare) coupled to a Finnigan LTQ linear ion trap mass spectrometer (Thermo Fisher). The MS/MS spectra were searched with X! Tandem (http://www.thegpm.org/) and proteins with at least one matching peptide with $e<10^{-3}$ were included in the results.

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

ACKNOWLEDGMENTS
We thank R.C. Beavis, B.T. Chait, I. Cristea, S. Gohil, T. Jovanovic, J.C. Padovan and B.M. Ueberheide for helpful advice, discussions and assistance with the experiments.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

1. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
2. Anderson, N.L. & Anderson, N.G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1**, 845–867 (2002).
3. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
4. Wang, H. *et al.* Intact-protein-based high-resolution three-dimensional quantitative analysis system for proteome profiling of biological fluids. *Mol. Cell. Proteomics* **4**, 618–625 (2005).
5. Ishihama, Y. Proteomic LC-MS systems using nanoscale liquid chromatography with tandem mass spectrometry. *J. Chromatogr. A.* **1067**, 73–83 (2005).
6. Cargile, B.J., Bundy, J.L., Freeman, T.W. & Stephenson, J.L., Jr. Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. *J. Proteome Res.* **3**, 112–119 (2004).
7. Coon, J.J., Syka, J.E., Shabanowitz, J. & Hunt, D.F. Tandem mass spectrometry for peptide and protein sequence analysis. *Biotechniques* **38**, 519–523 (2005).
8. Fenyo, D. Identifying the proteome: software tools. *Curr. Opin. Biotechnol.* **11**, 391–395 (2000).
9. Johnson, R.S., Davis, M.T., Taylor, J.A. & Patterson, S.D. Informatics for protein identification by mass spectrometry. *Methods* **35**, 223–236 (2005).
10. Eriksson, J., Chait, B.T. & Fenyo, D. A statistical basis for testing the significance of mass spectrometric protein identification results. *Anal. Chem.* **72**, 999–1005 (2000).
11. Eriksson, J. & Fenyo, D. Probity: a protein identification algorithm with accurate assignment of the statistical significance of the results. *J. Proteome Res.* **3**, 32–36 (2004).
12. Krokhin, O.V. *et al.* An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS. *Mol. Cell. Proteomics* **3**, 908–919 (2004).