nature biotechnology

Proteomics: a pragmatic perspective

Parag Mallick[1,2] & Bernhard Kuster[3,4]

The evolution of mass spectrometry–based proteomic technologies has advanced our understanding of the complex and dynamic nature of proteomes while concurrently revealing that no 'one-size-fits-all' proteomic strategy can be used to address all biological questions. Whereas some techniques, such as those for analyzing protein complexes, have matured and are broadly applied with great success, others, such as global quantitative protein expression profiling for biomarker discovery, are still confined to a few expert laboratories. In this Perspective, we attempt to distill the wide array of conceivable proteomic approaches into a compact canon of techniques suited to asking and answering specific types of biological questions. By discussing the relationship between the complexity of a biological sample and the difficulty of implementing the appropriate analysis approach, we contrast areas of proteomics broadly usable today with those that require significant technical and conceptual development. We hope to provide nonexperts with a guide for calibrating expectations of what can realistically be learned from a proteomics experiment and for gauging the planning and execution effort. We further provide a detailed supplement explaining the most common techniques in proteomics.

Proteomics[1] provides a complementary approach to genomics technologies by *en masse* interrogation of biological phenomena on the protein level. Two transforming technologies have been critical to the recent, rapid advance of proteomics: first, the emergence of new strategies for peptide sequencing using mass spectrometry (MS), including the development of soft ionization techniques, such as electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI); and second, the concurrent miniaturization and automation of liquid chromatography. Together these technologies enable the measurement and identification of peptides at a rate of thousands of sequences per day[2,3] with better than femtomole sensitivity (10^{-15} mol, or subnanogram)[4] in complex biological samples.

Early excitement about the potential for proteomics (**Supplementary Glossary**) to transform biological inquiry has been tempered by the discovery that the enormous molecular complexity and the dynamic nature of proteomes (**Supplementary Glossary**) pose much larger hurdles than encountered for either genome or transcriptome studies. In particular, issues related to splice variants, post-translational modifications (PTMs), dynamic ranges (**Supplementary Glossary**) of copy numbers spanning ten orders of magnitude, protein stability, transient protein associations and dependence on cell type or physiological state have limited our technical ability to characterize proteomes comprehensively and reproducibly in a reasonable time[5]. Despite the hurdles, after 15 years of evolution, proteomic technologies have significantly affected the life sciences and are an integral part of biological research endeavors (**Supplementary Fig. 1**).

At present, the field of proteomics spans diverse research topics, ranging from protein expression profiling to analyzing signaling pathways to developing protein biomarker assay systems. It is important to note that within each area, distinct scientific questions are being asked and, therefore, distinct proteomic approaches may have to be applied; these approaches vary widely in their versatility, technical maturity, difficulty and expense. Consequently, we must recognize that some biological questions are much harder to answer by proteomics than others. Here, we review biologically directed MS-based proteomics, focusing on which parts are routinely working, which applications are emerging and promising, and which paradigms still require significant future investment in technology development and study design.

Getting organized

The catalog of proteomics experiments contains a wide diversity of techniques and approaches. In this section, we clarify the naming of these approaches. Proteomics experiments are foremost divided by objective into either discovery or assay (**Fig. 1**). Both objectives have strong scientific rationale, but they come with very different study requirements and technical challenges. Proteomic assay experiments typically seek to quantify a small, predefined set of proteins or peptides, whereas discovery experiments aim to analyze larger, 'unbiased' sets of proteins (see **Supplementary Techniques**) for a deeper discussion of 'unbiased' proteomics). A typical example of an assay experiment would be the measurement of the levels of cardiac troponins in human plasma samples[6]. Such experiments are often called 'targeted', 'restrictive' or 'directed' proteomics' studies, and the analytical approach must typically address challenges such as data variation and sample throughput.

Within discovery proteomics, we distinguish among comprehensive, broad-scale and focused approaches because these distinctions have a large influence on how a biological question is approached technically. Comprehensive approaches are typically qualitative in nature and aimed at enumerating as many components of a biological system as possible. For example, the Human Proteome Organization (HUPO) Plasma Proteome Project (PPP) aims to identify every possible protein and peptide in human plasma. Such experiments can span years and require

[1] University of Southern California Center for Applied Molecular Medicine, Departments of Medicine and Biomedical Engineering, Los Angeles, California, USA. [2] Department of Chemistry & Biochemistry, Univeristy of California, Los Angeles, Los Angeles, California, USA. [3] Chair of Proteomics and Bioanalytics, Technische Universität München, Freising-Weihenstephan, Germany. [4] Center for Integrated Protein Science Munich, Munich, Germany. Correspondence should be addressed to P.M. (mallick@usc.edu) or B.K. (kuster@wzw.tum.de).

Published online 9 July 2010; doi:10.1038/nbt.1658

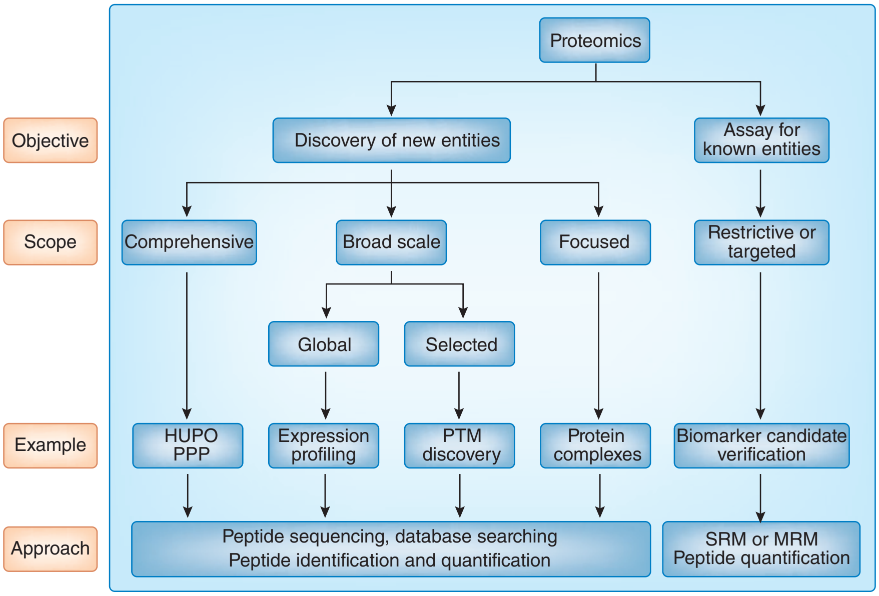


Figure 1 Conceptual organization of proteomic experiments. We broadly divide the objectives of proteomics into discovery and assay experiments. The scope of these experiments can range from very narrow (few proteins) to comprehensive (all proteins). A small set of examples is shown here, along with the technology used to study them.

input from many labs[7]. In contrast, broad-scale experiments attempt to globally or selectively sample a large, but not necessarily complete, fraction of the expressed proteome (for example, the phospho-proteome) and are commonly used as profiling tools to measure qualitative and quantitative changes in a system in response to perturbation or differences in genetic background[8,9]. The identification of several thousand proteins or phospho-peptides[10] may also require days to weeks of data acquisition and analysis time but can be shouldered by any well funded laboratory. Focused approaches, such as the identification of proteins present in a mammalian protein complex, restrict their scope from the start by copurifying relatively few interacting proteins. The challenge in these experiments is not complexity or dynamic range but the related challenges of either the detection sensitivity or the large-scale sample generation required to measure interaction partners, which may be of extremely low cellular abundance[11,12].

Many, but not all, conceivable biological questions can be approached through proteomic experiments. In **Figure 2**, we contrast the technical expertise required to implement and execute a proteomic inquiry with the sample complexity (that is, the complexity of the biological system being interrogated). Simply put, experiments at the upper left of the chart are straightforward; those at the bottom right are difficult or under development. This chart is critical for understanding the effort involved in planning and conducting a study using proteomics and for setting realistic expectations on likely results. Success in a proteomic study is enabled and confined by the biological system (for example, do the cells actually respond to stimulus?), the study design (for example, are all the appropriate controls and statistics in place?), the available technology (for example, does it deliver the required proteome coverage, sensitivity, accuracy (**Supplementary Glossary**)) and, finally, the ability to perform hypothesis-driven follow-up experiments required to transform proteomic information into biological knowledge. Shortcomings in any of these areas will significantly impair success, and clearly, expectations must be measured against what the study can actually accomplish. If, for example, the purpose of an experiment is to identify the components of a protein complex, it is unreasonable to expect that the analysis will also uncover the phosphorylation status of all proteins and their stoichiometries (**Supplementary Glossary**) at the same time.

The ability to conduct a successful and substantial proteomic study also heavily depends on the local or regional research infrastructure environment. Core facilities have been established in many places to give scientists access to mainstream proteomic technologies and applications (for example, protein identification). Even so, more sophisticated applications requiring specialized technologies or particular practical expertise (for example, top-down sequencing of intact proteins or ion mobility measurements of glycosylated protein isoforms) may only be available through collaboration with expert laboratories. In our view, much more effort needs to be expended in helping biologists understand proteomic technologies (and in helping technologists to understand more of the biology) so that the right experiment can be designed, meaningful conclusions can be drawn from the data, and the appropriate follow-up experiments can be initiated. Despite significant investments in people and infrastructure over the past decade, access to the technology and special expertise still constitutes a substantial bottleneck.

In this Perspective, we place biologically motivated proteomics in context by detailing components of each of the columns in **Figure 2**. As a comprehensive treatment of each topic is not possible, some topics are thoroughly discussed and the others only mentioned briefly. It is beyond the scope of this Perspective to cover aspects of structural biology that are often discussed in the context of proteomics. Instead, the interested reader may refer to reviews published on this topic[13,14]. The guiding thoughts within each section of this article are the following: given a biological question, what are the specific challenges and which proteomic methods may be able to address them; what methods are still experimental, but may emerge over the next decade; and what are reasonable expectations for the outcomes of a given experiment? A technical supplement to this Perspective (**Supplementary Techniques**) briefly explains the core proteomic technologies listed in **Figure 3**. In addition, definitions of important proteomics and MS terms (**Supplementary Glossary**), technical details of protein identification by MS (**Box 1**), and frequently asked questions (**Table 1**) provide more clarity and simplify reading. In **Figure 4**, we give a concrete example of a quantitative proteomics workflow drawing on elements from **Figure 3**.

Protein analysis

The classic tasks of characterizing the size, identity, presence of PTMs and purity of a single protein isolated from natural or recombinant sources draws on decades of experience in protein chemistry and is broadly accessible to scientists through core facilities or commercial service providers. Many of the tools developed for protein characterization are also frequently used on a broader scale in proteomic workflows. Thus, although previously described as 'protein characterization', some protein characterization techniques are now referred to as proteomics. We do not cover this area in detail, but instead touch on key points that also apply to later sections.

In protein characterization, what can and cannot be done depends primarily on technical factors, such as available sample amounts, purity, solubility and stability of the material. Using modern mass spectrometers (**Supplementary Glossary**), the mass of an intact protein can be determined with an accuracy (**Supplementary Glossary**) of better than 0.01% and can often be used to confirm the integrity of the isolated protein. MS can also be used to assess the purity of a protein preparation, as contaminating proteins can be detected at <5% abundance. This is important in the production of therapeutic proteins and in preparation of samples for structural studies by nuclear magnetic resonance (NMR) or X-ray crystallography. Very large (say, >150 kDa) and/or poorly soluble proteins can present a challenge because the detection efficiency of mass spectrometers rapidly degrades with increasing mass and the presence of detergents and salts can suppress the mass spectrometric signal or interfere with chromatography. In such cases, the identity of a protein can be confirmed by sequencing proteolytic fragments either by MS or by classical Edman degradation. Albeit far less sensitive than MS, the latter approach offers a simple route to determination of the sequence of the protein's N terminus.

The presence and sites of PTMs on a single protein can also be generally analyzed by MS-based proteomics because many of the >200 described PTMs alter the mass of a protein in a predictable fashion[15]. Even so, robust protocols are as yet available for relatively few low molecular weight PTMs, such as phosphorylation, acetylation and methylation[16]. Protein oxidation can also be readily detected by MS, but it is generally impossible to distinguish a biologically important oxidation event from an experimental artifact. Important PTMs such as ubiquitinylation[17] and glycosylation[18] are difficult to analyze, even on an isolated protein, because the modification may exist in multiple or combinatorial numbers and can lead to molecular branching of the otherwise linear protein sequence. This may require the application of a more specialized MS platform, such as electron transfer dissociation (ETD) and infrared multiphoton dissociation (IRMPD). Further challenges can arise from the necessity to cover the entire protein sequence to ensure that no potential site has been missed. This can often be addressed by using several alternative proteases to generate complementary protein fragments for analysis by MS, but a significant proportion of all proteins seem to be completely refractory to any of the tried approaches.

Determining the stoichiometry (**Supplementary Glossary**) of PTM at a given site is still challenging—even for a single isolated protein. The physicochemical properties of the modified and unmodified proteins or peptides are often vastly different, so that there is no unambiguous direct way to measure stoichiometry. Instead, one often must resort to indirect measures—for example, by chemically or enzymatically removing the PTM from the protein or peptide and then comparing the quantities of the unmodified peptide or protein before and after the transformation[19–21]. An alternative method for this purpose is the use of stable isotope (**Supplementary Glossary**) labeling with exogenously introduced analytical standards of precisely known quantities (absolute quantification, or AQUA[22]). Such standards have so far been generated for only very few PTMs (notably phosphorylation[23]) and, for economic reasons, are now mostly used to address specific questions rather than on a broad scale. A more fundamental factor that affects our ability to determine the quantity and stoichiometry of a PTM is the common occurrence of PTM microheterogeneity at a single site. An extreme example is human erythrocyte CD59, which carries more than 120 different asparagine-linked oligosaccharides at a single site[24]. The analytical task of PTM analysis becomes more complex still when multiple types of modifications are present at the same site or different sites of the protein. A prominent example is the extensive modification of the N-terminal tail of histones by acetylation, methylation and phosphorylation. Using highly specialized MS methods, including ETD and proton transfer reactions (PTR), 74 isoforms of histone H4 have been isolated from differentiating human embryonic stem cells and subsequently characterized[25]. However, these approaches are not yet routinely available in core facilities.

Generating comprehensive and quantitative information on protein modifications is a significant undertaking requiring several experimental approaches, significant amounts of pure starting material (mid-microgram range), special expertise and time. It should therefore only be undertaken if some functional hypothesis can be formulated or these data are required by regulatory agencies. A fundamental issue with the quantitative analysis of multiple PTMs present on a protein is that it is almost impossible to separate all existing protein isoforms (top-down proteomics; **Supplementary Glossary**), but this is required to estimate the amount of each isoform relative to the total protein amount. Electrophoretic and chromatographic methods in conjunction with high-resolution MS may resolve a substantial number of isoforms[26], but even then, identifying the site and stoichiometry of modification remains difficult. In practice, quantitative PTM analysis is mostly performed at the peptide level (bottom-up proteomics; **Supplementary Glossary**). Here, special care must be exercised because variations in protein digestion, peptide recovery and peptide detection may distort the quantification results, and measurement of total protein is often difficult. We therefore recommend using analytical protein and peptide standards whenever possible, to account for systematic bias, and confining the analysis to one PTM at a time[27].

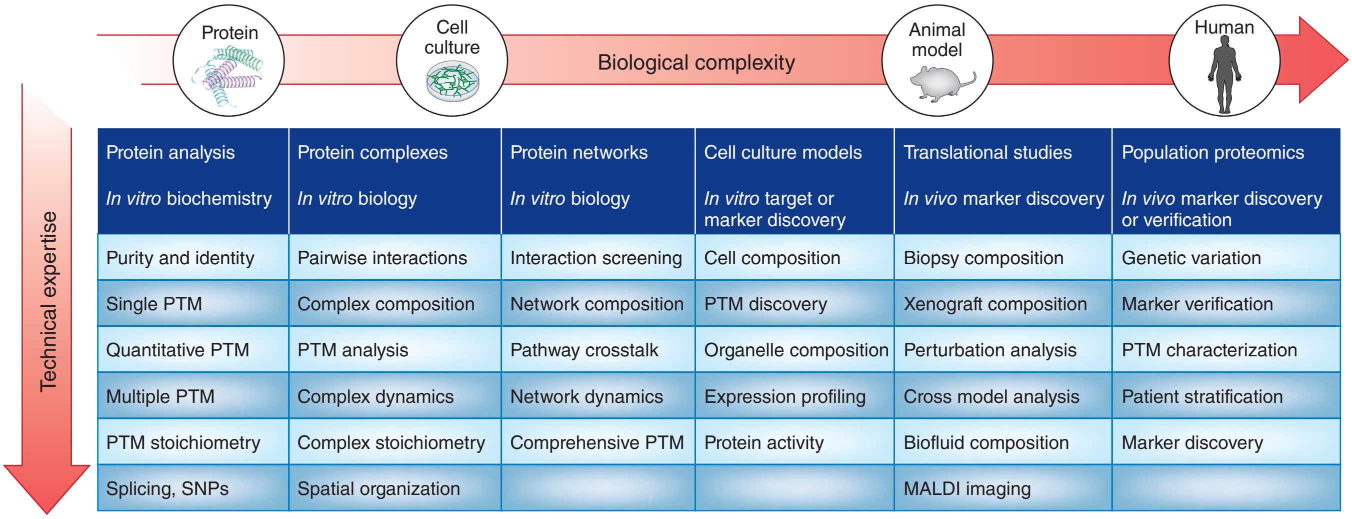
MS-based peptide sequencing can also be used to detect proteins resulting from splice variants and single-nucleotide polymorphisms[28]. This type of study has rarely been done systematically owing to the requirement for 100% sequence coverage and the difficulty of detection of low-abundance isoforms. With the advent of next-generation DNA sequencing techniques[29], we expect proteomics to play a lesser role in this area in the future.

Analysis of protein complexes

It is by now widely accepted that proteins exert their cellular functions as part of multiprotein complexes[30]. In the analysis of protein complexes, the contribution of proteomics has been nothing short of phenomenal. Since the groundbreaking mass spectrometric identification of the components of the yeast spliceosome in 1997 (ref. 31), the analysis of protein complexes has uncovered countless important specific as well as global biological phenomena. As quantitative MS methods, such as SILAC (stable isotope labeling in cell culture[32]; **Supplementary Glossary**), have been perfected, proteomics has provided a powerful means to distinguish true interactors from abundant contaminants[33].

Although proteomics has been very successful at determining the composition of complexes, the detailed study of binary protein interactions is still surprisingly difficult by proteomic methods. In part, this results from the general challenge of purifying protein pairs in the presence of other interacting proteins. *In vitro* surface plasmon resonance or chemical crosslinking experiments are often used, but these techniques suffer from the need for significant quantities of pure protein. As a result, binary protein interactions are still mostly identified by the yeast two-hybrid system, which can be readily automated to enable systematic studies of transient protein-protein interactions[34,35]. The yeast two-hybrid system is not without issues, however, as the interaction of two exogenous proteins in a yeast nucleus can lead to various artifacts.

In the analysis of the molecular composition of protein complexes, proteomics has several advantages. First, affinity purification typically yields moderately complex protein mixtures, a situation that



Protein analysis *In vitro* biochemistry	Protein complexes *In vitro* biology	Protein networks *In vitro* biology	Cell culture models *In vitro* target or marker discovery	Translational studies *In vivo* marker discovery	Population proteomics *In vivo* marker discovery or verification
Purity and identity	Pairwise interactions	Interaction screening	Cell composition	Biopsy composition	Genetic variation
Single PTM	Complex composition	Network composition	PTM discovery	Xenograft composition	Marker verification
Quantitative PTM	PTM analysis	Pathway crosstalk	Organelle composition	Perturbation analysis	PTM characterization
Multiple PTM	Complex dynamics	Network dynamics	Expression profiling	Cross model analysis	Patient stratification
PTM stoichiometry	Complex stoichiometry	Comprehensive PTM	Protein activity	Biofluid composition	Marker discovery
Splicing, SNPs	Spatial organization			MALDI imaging	

Figure 2 Applications of proteomic technologies. For the purpose of organizing the field of proteomics, it is instructive to compare and contrast the many conceivable applications on the basis of the complexity of the biological context versus the technical difficulty of implementing the appropriate technology. Each cell in the table shows an application of proteomics that is discussed in the main text.

is ideally matched by the capabilities of MS to identify proteins in mixtures. Second, interacting proteins can be purified under near physiological conditions from endogenous sources or from cell lines, limiting artifacts. Third, functionally important protein modifications, such as phosphorylation or acetylation, can often be determined in the same context. Finally, with few exceptions, 5–20 proteins are generally present in complexes and can usually be identified by LC-MS/MS after either a solution digest or a one-dimensional sodium dodecyl sulfate (SDS) gel.

Protein complexes can be purified in several ways[36,37]. One approach is to attach an affinity tag to the protein of interest, express it in a cell line and purify the interacting partners by virtue of the tag. The advantage of using tagged proteins is that the tag can be systematically applied to any number of proteins in a particular pathway, including proteins discovered to interact with a certain bait protein. To allow validation of the components found to be in the complex, a reciprocal tagging experiment can be performed. A newly identified interactor is tagged and in turn used for the purification of the same complex. If the same proteins are identified, the interactions are valid. As proteins may be part of more than one complex, results from this type of experiment depend on the abundance of the respective complexes. Disadvantages are that the tag modifies the protein, which may alter its activity. Issues may also arise from overexpression of the tagged protein, but this can often be overcome by tagging the endogenous gene locus[38,39] so that the endogenous promoter drives protein expression. The use of antibodies or other protein binders does not suffer from these shortcomings, as they purify the endogenous complex. High-quality antibodies are, however, available only for a limited set of proteins.

The biochemical approach aside, the ability to identify interacting proteins by MS depends on two main factors: the abundance of the protein complex and the affinity with which interacting proteins are held together. As modern mass spectrometers offer attomole sensitivity, the former issue can be overcome when sufficient quantities of starting material are used. The latter is harder to address, as the time required to perform an affinity purification biases the results toward submicromolar interactions. *In vivo* crosslinking with low concentrations of formaldehyde has been used to stabilize transient interactions before purification[40], but there are not enough examples in the literature to validate this approach as a generic strategy.

Because not all the proteins identified in the types of experiments mentioned above are genuine interactors, validation experiments at different levels are required. A common biochemical approach is to use coimmunoprecipitation of wild-type proteins at basal expression levels. Although coimmunoprecipitation is an independent approach, it suffers from the same issues of abundance and affinity. If the suspected interactor is nonspecifically copurified with a target protein, it will be detected by both coimmunoprecipitation and MS. A recent and elegant biochemical validation approach is a method called protein correlation profiling, in which the quantity of suspected interactors is compared across the different steps of the complex purification scheme[41]. Only those proteins that show an identical purification profile are genuine members of a complex, whereas all other proteins are (abundant) contaminants. As noted above, a reciprocal tagging experiment may also be used for validation. A common cell-biological approach is then to show cellular colocalization of the interacting proteins. Of course, none of these types of experiments demonstrates biological significance; this may come from experiments showing that the interaction takes place *in vivo* and is functional.

Although the identification of members of stable protein complexes of low cellular abundance is fairly routine, the analysis of PTMs at the protein complex level is possible but difficult[42]. Variations in biological conditions may lead to changes in the composition, PTM status and activity of protein complexes. To capture this dynamic behavior, the respective biological and proteomic experiments must be modified, and several controls must be performed to ensure that the data can be meaningfully interpreted. First, it must be demonstrated that the biological system from which the proteomic sample is derived actually responds to the stimulus with the expected kinetics, dose-responses or other appropriate criteria (as would be the case for any biologically motivated proteomic experiment). Second, a quantitative MS technique should be used so that the observed changes can be statistically measured rather than assessed by intuition. And third, functional assays should be in place to validate the observed changes. As with static protein complexes, one should only expect to identify relatively stable protein interactions as the time scale of the experiment generally does not permit the identification of transient interactions. Maybe not surprisingly, the dynamics of individual protein complexes are not often studied by proteomic approaches[43]; other biochemical and cell biological techniques are often more suitable for this purpose once the proteomic experiment has established the protein components of a complex.

One fundamental aspect of protein complex architecture is the stoichiometry of its constituents. Experiments to determine stoichiometry are technically very challenging and often require combinations of biophysical and proteomic approaches[44,45]. For stable protein complexes, gel filtration or centrifugation techniques can give indications of stoichiometry, but the larger the complex gets, the harder data become to interpret. Proteomic techniques are only beginning to be used to determine stoichiometry, but, given the sensitivity of MS, we anticipate that proteomics will be important in these types of analyses in the future. In the few published examples, stable isotope or fluorescently labeled reference standards of precisely known quantities have been used to determine the quantities of members of protein complexes[46–48]. The most rigorous controls must be used for this type of study because bias must be avoided in purification steps in order to arrive at meaningful numbers. Intact mass measurements of isolated protein complexes will be of utility, but very few laboratories now have the technical capability to perform these experiments[49,50].

The spatial organization of proteins in a complex is also of interest. Given that typical protein complexes are made up of up to 5–20 members[51], each protein in the supramolecular structure cannot physically contact all the other proteins. Supramolecular structure determination typically is the domain of biophysical techniques such as X-ray crystallography, NMR and cryo-electron microscopy. Proteomic approaches have not yet been prominent but may contribute in the future, given the comparatively small sample needed for MS. The general idea is to crosslink the complex and then to sequence the crosslinked peptides by MS to establish the nearest-neighbor relationships. Although conceptually simple, this is technically very demanding. Chemical crosslinking heavily modifies the proteins and may change the integrity of the complex. In addition, the yields of the crosslinking reactions are typically very low. Finally, the sequencing and identification of crosslinked peptides by MS is nontrivial because crosslinking generates branched peptides. Tandem mass spectra of such peptides often contain information about both of the sequences, but most database search algorithms are unable to process this information because they only consider the linear peptide sequences deposited in a protein sequence database. As a result of all these complications, the examples in the literature are mostly confined to binary protein interactions or very small protein complexes[52,53].

Analysis of protein pathways and networks

The next level of cellular organization is provided by pathways and networks in which proteins and protein complexes relay signals from the extracellular space into the cell or distribute information within a cell and its compartments. Much of what was said about protein complexes also applies to networks; however, many more proteins are involved in networks than in typical protein complexes. Charting a physical network is technically fairly straightforward, and analyzing dynamic behavior in a global sense by MS has become more doable as quantitative MS methods become more widely available. However, the functional validation of identified proteins is by no means trivial, as cross-talk between pathways can often render the results somewhat ambiguous.

Proteomic technologies have enabled the systematic charting of cellular pathways and networks in several model organisms[54–56]. In fact, two reports on large-scale protein interaction screens in yeast are among the five most highly cited papers in proteomics so far[51,57]. Technically, such interaction screens take advantage of affinity tagging of proteins using genetic or molecular biology techniques and the speed and sensitivity of MS. Use of affinity tags rather than antibodies

Box 1 Protein identification in mixtures by MS

Broadly, there are two strategies for protein identification in mixtures: first, mapping strategies that rely predominantly on the accurate mass, retention time, or both to infer the composition of a mixture; and second, tandem MS approaches, now the most common (for greater detail, see **Supplementary Techniques**). MS*n* refers to sequential MS/MS experiments, where *n* is the number of MS/MS experiments. For MS*n* approaches, peptides are first selected for fragmentation (in either a targeted or a data-directed manner) inside the mass spectrometer and then are fragmented by one of several methods (e.g., collision-induced dissociation (CID) or electron capture detection (ECD)); the mass spectrum of the peptide fragments is then recorded. It is most common to perform this step only once (that is, conventional MS/MS); however, some studies have shown value in multiple isolation and fragmentation steps (that is, MS*n*). Typically, the most intense ions are selected for fragmentation. Dynamic exclusion (**Supplementary Glossary**) and targeted inclusion lists are used to broaden the range of selected species.

Once ions have been selected and fragmented, three strategies are used to assign a peptide to the ion. The first is database searching (**Supplementary Glossary**). In this strategy, peptides are generated by an *in silico* digest of a proteome database and then a theoretical mass spectrum is predicted for each peptide. The theoretical spectrum is compared with the experimental spectrum and a peptide identity is inferred on the basis of the best match between the theoretical spectrum and the observed spectrum. In the second approach, *de novo* sequencing (**Supplementary Glossary**), peptide sequences are read out directly from fragment ion spectra. In hybrid techniques, short stretches of the peptides are sequenced, and then the rest of the spectrum is matched to existing data.

Though fragmentation-based methods are generally successful, there are several limitations. As noted in the main text, the largest limitation is the small number of peptides selected for sequencing. Many instruments are able to sequence only a subset of the hundreds of peaks present in each mass spectrum. In addition, relatively few peptides with fragmentation spectra give rise to high-confidence identifications. This low percentage can be attributed to several experimental and computational factors. Computationally, matching techniques are most successful with unmodified tryptic peptides. The inclusion of more modifications greatly increases the false discovery rate, and the larger size of the sample space complicates identification. In addition, gas phase chemistry or ion source effects can fragment or modify peptides. Finally, for the inference of protein identifications from peptide identifications, there is the issue that not all peptides are unique for a single protein, as close sequence homologs or proteins with similar domains can contain the same peptide sequence (so-called shared peptides). From this so-called peptide inference problem follows the requirement to ascertain whether protein identifications are made on the basis of unique or shared peptides. If only shared peptides are identified, a protein group rather than a single protein has been identified.

to purify network components means that the strategy is generic (that is, it can in principle be applied to any protein). Tags, such as the Flag peptide (DYKDDDDK or MDYKDDDDK), hemagglutinin, streptavidin, green fluorescent protein (GFP) and TAP (tandem affinity purification: a fusion cassette encoding calmodulin-binding peptide, a tobacco etch virus protease cleavage site and Protein A), and combinations thereof, have been used effectively. GFP is attractive because it enables both the monitoring of protein localization and complex purification. Although not technically demanding, systematic mapping of protein networks on a large or genome-wide scale requires significant technical resources. Thousands of samples must be analyzed by MS to produce a mostly static picture of the physical organization of cells into protein networks. Even larger numbers of samples will be required to capture the dynamic nature of protein networks or to extend analysis to different cell types. This means that genome-wide interaction studies can likely only be undertaken by substantially funded academic consortia or companies.

Proteomics has been important in identifying the component parts of smaller networks from all corners of biology. In the design of a proteomics experiment to evaluate a network, consideration should be given to the choice of initial bait proteins. Tagging scaffolding proteins or transcription factors has yielded particularly rich network coverage, whereas tagging of enzymes often results in disappointment because their interactions are generally too transient or too weak to be observed by proteomic methods. Thus, proteomic charting of networks typically provides a physical rather than functional view of a network. Because of the multitude of possible interactions within and between complexes, as well as the fact that many proteins present in a network have generic cellular function (say, maintaining cell homeostasis), the interpretation of network mapping data needs to be carefully controlled. The extent to which such controls may have to be applied is illustrated by a study in which the tumor necrosis factor-α (TNF-α)–nuclear factor-κB (NF-κB) pathway was mapped in human embryonic kidney (HEK293) cells using 32 TAP-tagged proteins[11]. The initial interaction map constructed from the mass spectrometric analysis of some 250 affinity purifications contained 680 proteins, only 130 of which were not identified in a counter-screen of 250 unrelated TAP purifications. This means that, even for relatively small protein networks, relatively large-scale proteomic analyses may be required for informed selection of new proteins for functional validation. Network mapping is most effective if carried out in a stepwise fashion in which one starts from proteins of well described biology to identify a small number of interaction partners that can be validated using functional assays established for the system under study.

In mapping protein interaction networks and pathways, one soon realizes that the pathways are interconnected at many different levels[58]. Such cross-talk is of great biological importance, as it offers a means to generate functional redundancy, diversity and compensating mechanisms should parts of a pathway become unavailable. To identify pathway cross-talk systematically, one would again start out from a well known protein interaction hub and map protein interactions in its

Table 1 Frequently posed questions in MS-based proteomics

Question	Answer
How do I prepare my sample for MS analysis?	High amounts of salts and detergents must be removed before MS analysis. There are many ways of accomplishing this, including protein precipitation, SDS-PAGE and ultrafiltration or dialysis. If in doubt, ask your analytical collaborator.
How much protein do I need for protein identification or quantification?	You can expect to identify and quantify: 1. 10s to 100s of proteins from nanograms of total protein 2. 100s to 1,000s of proteins from micrograms of total protein 3. 1,000 to 10,000 proteins from milligrams of total protein Results strongly depend on the complexity and dynamic expression range of samples. Typically, one-tenth as many proteins are identified from serum than from cell lines or tissues.
How much protein do I need for PTM analysis?	Systematic PTM analysis of a single protein requires microgram amounts of a reasonably pure protein. Proteome-wide shotgun (**Supplementary Glossary**) PTM analysis requires milligram amounts of protein. For very rare modifications, other requirements may apply.
What protein coverage can I expect to achieve?	This depends on (i) the complexity of the mixture, (ii) the amount of protein in the mixture and (iii) the MS/MS selection and dynamic exclusion criteria (**Supplementary Glossary**). There is a rough correlation between protein coverage and protein abundance; however, even for simple mixtures or for the most abundant proteins, it is rare to observe >60% coverage unless specific efforts are taken (for example, multiple digestion protocols) to increase coverage. In complex mixture experiments, many low-abundance proteins will be identified by only a single unique peptide.
What proteome coverage can I expect to achieve?	This depends on (i) the amount of protein used for the analysis and (ii) the degree of proteome fractionation. Coverage of 500–1,000 proteins may be achieved by direct LC-MS/MS of proteome digests. Coverage of 1,000–3,000 proteins requires at least one dimension of proteome fractionation on the peptide or protein level (for example, protein fractionation by one-dimensional SDS-PAGE followed by LC-MS/MS, or peptide fractionation by in-solution isoelectric focusing followed by LC-MS/MS). Coverage of >3,000 proteins usually requires multiple dimensions of fractionation on protein and/or peptide level. Note that typically, one-tenth as many proteins are identified from serum than from cell lines or tissues.
Which identifications can I trust?	Three general quality criteria (or combinations) can be applied: 1. Calculation of a global false discovery rate (FDR) for the list of identified proteins. FDRs of <1% are generally accepted. FDRs give information about the general quality of a data set. Most protein identification software packages provide FDR calculation tools. 2. Calculation of the probability that matching a tandem MS spectrum to a peptide sequence is a random event. Random matches of <1%–5% are generally accepted. Peptide probabilities give a quality assessment for each identified protein. Not all protein identification software can perform this probability calculation. 3. For publication in some journals, at least two peptide identifications are required. This is an *ad hoc* criterion and says very little about data quality.
How does the protein identification list correlate with protein amount?	As a rule of thumb, the abundance of a protein correlates with the number of tandem MS spectra that identify the peptides belonging to a protein. Proteins at the top of the list are therefore generally more abundant than proteins further down on the list. This is a very crude correlation as the relationship between detection efficiencies of different peptides in a proteomic workflow is complex and not well understood. Although it is fairly safe to compare the same protein across different experiments, it is more dangerous to make comparisons of different proteins in the same experiment.

(continues)

Table 1 Frequently posed questions in MS-based proteomics (continued)

Question	Answer
Where do I cut the list of identified proteins?	Physical presence of a protein may be judged by the criteria described above for protein identification. This does not automatically mean relevance for the experiment performed, as many of the identified proteins may be contaminants, either endogenous (for example, abundant housekeeping proteins) or exogenous (for example, keratins from human skin).
Which quantification approach should I choose?	This strongly depends on the experiment. Simple guides are the following: 1. Metabolic labeling (for example, SILAC ^{15}N) is best for small changes (10–50%) and cell culture systems. 2. Peptide labeling (for example, iTRAQ, TMT, dimethylation) is best for moderate changes (50%–200%), primary tissue protein sources and multiplex experiments (for example, time courses, dose responses). 3. Label-free methods using the MS detector response (for example, extracted ion chromatograms (**Supplementary Glossary**)) are best for moderate changes (20%–200%) and for comparison of many highly similar experiments. 4. Label-free methods using spectrum counts are best for large changes (>100%) and for comparison of many highly similar experiments. 5. Single or multiple reaction monitoring (SRM or MRM) in conjunction with spiked synthetic standards (AQUA) is best for determining the absolute quantity of a protein in a complex biological matrix (for example, serum).
What fold change can I trust in quantitative experiments?	Any observed change should bear a statistical measure of variance to define the changes that can be trusted. This may be computed for every protein on the basis of the number of available data points (for example, number of peptides per protein, amplitude of MS response, technical and biological replicates). Several free and commercial software packages have become available, but many proteomics laboratories still struggle with quantification statistics.
How reproducible are the results for protein identification?	Generally, reproducibility is a function of the complexity of a protein mixture and the number of upstream sample handling steps. For simple protein mixtures and short workflows (for example, immunoprecipitations), reproducibility should generally be better than 80%. For whole proteome analysis or complex proteome fractionation schemes, reproducibility may vary greatly, from 40 to 70%. It should be stressed that not identifying or quantifying a peptide or protein does not necessarily mean that the peptide or protein is not present in a mixture.
How reproducible are the results for protein quantification?	As for protein identification, sample complexity greatly affects reproducibility. Stable isotope labeling methods generally reproduce within 5%–25%, whereas spectrum counting typically shows larger variance.
How long will it take to get the results?	This depends largely on whether the work is done with a core facility or with a research lab. The following turnaround timesfrom sample submission to data reporting are typical for core facilities and research labs: 1. 5–10 working days for simple protein identification and quantification 2. 4–6 weeks for quantitative protein expression profiling 3. 2–6 months for PTM analysis
How much will this cost?	Proteomic analysis is not yet a commodity. Costs vary depending on the collaborator. For commercial and academic service providers, the costs scale with the requirements of time of personnel, cost for reagents and equipment and overheads. Typical figures would be as follows: 1. \$50–200 for simple protein identification 2. \$500–2,000 for simple PTM analysis 3. \$5,000–15,000 for complex PTM analysis 4. \$1,000–2,000 for quantitative protein expression profiling

close vicinity, rather than choosing biologically unconnected 'islands'. Technically, analysis of pathway cross-talk is no more demanding than mapping of protein interactions within a confined network. Even so, validation issues become more acute. For example, confirming the specificity of individual or even relatively few protein–protein interactions becomes a large-scale experiment because of the numbers of candidate proteins. In addition, the under-representation of enzymes in protein interaction studies makes direct functional validation of potential cross-talk events much more difficult. As a result, study of pathway cross-talk may be best approached by a battery of cell biological assays in combination with loss-of-function approaches, such as RNA interference, rather than by proteomics.

Clearly, pathways are dynamic, both in their physical makeup and their functional activity, although most published proteomics studies so far have provided static views. Going forward, the quantitative analysis of protein pathways and networks must include perturbation or stimulation experiments to learn about proteins moving in and out of complexes, changes in activation status, and the behavior of the network in general (rather than that of a single protein)[59,60]. Quantitative proteomic technology has advanced to enable a fairly accurate assessment of the relative changes between different cellular states. Although MS-based approaches are very successful in discovering the members of the network, measuring their dynamic behavior under a multitude of different conditions may be better served by normal or reversed protein arrays, owing to their inherent throughput[61–64]. Obviously, however, the effort involved in creating global or themed protein arrays is a significant up-front investment, and a protein array strictly speaking does not measure interactions occurring in a cell.

As mentioned before, the activity of a signaling pathway or network is often regulated by PTMs, and the techniques of PTM analysis can also be applied in the context of network analysis. Clearly, the comprehensive and simultaneous measurement of PTMs on many proteins is technically difficult, and the regulation mechanisms may be complex, so that analysis of PTM levels may not suffice to describe the behavior of the network or pathway. Still, MS today allows identification of thousands of phosphorylation sites in a quantitative manner and, as such, has made important contributions to our present knowledge of signaling pathways[65]. Nevertheless, our recommendation for a pathway-wide PTM study would be to focus on one particular PTM at a time and complement proteomic techniques with available PTM-specific antibodies if available.

Proteomic measurements are an important part of systems biology (**Supplementary Glossary**) data pipelines. The challenge here is to provide robust quantitative information so that mathematical models of the behavior of pathways and networks can be developed. So far, most proteomic studies have provided data on relative changes in protein abundance or PTM status in response to some form of biological perturbation. Although this often suffices to describe a pathway phenomenologically, information about the absolute numbers of molecules involved in a process is often required to compute a predictable outcome. Proteomics technologies based on MS are not now able to deliver such information routinely, even for one single pathway, let alone for the flux of information between pathways. But for focused applications (say, a small protein network), targeted analytical approaches such as the multiple reaction monitoring (MRM) technique hold considerable promise for the future[66].

Cell culture models

In vitro, prokaryotic[67], and eukaryotic[68,69] systems have been widely used to ask questions about the fundamental composition of proteomes and subproteomes (for example, phospho-proteome, mitochondrial proteome or cell-surface proteome) and how those proteomes are altered by genetic changes (for example, deletions or mutations), cell growth (for example, differentiation or cell state transition) or an intervention (for example, growth factor stimulation or drug treatment). Technically, qualitative protein expression profiling for thousands of proteins is no longer particularly difficult. The three principal challenges faced in system-scale analyses are sample purity, complexity and dynamic range. Consequently, the most profiling approaches aim to address all these in some shape or form. Sample purity is affected by contamination from other proteomes. For instance, the bovine or horse proteome from sera used in cell culture media may complicate secretome studies of human cell culture systems. Sample complexity refers to the number of different species within a sample being analyzed. Dynamic range refers to the range of protein concentrations from the least to the most abundant within a sample. Lastly, as very few protocols actually select for proteins, mixtures may contain a significant percentage of lipid, nucleic acid or small molecule contaminants that interfere with protein profiling.

One very common approach to reducing a sample's protein and peptide complexity is fractionation, such as by chromatographic methods. There are several key considerations when using chromatographic methods to partition a mixture before MS analysis. First is sample abundance. If this is severely limited, it may not be possible to use chromatographic methods. Next is analysis time. Chromatographic separation techniques can turn one sample into many and thus significantly increase analysis time and analysis cost. For reference, in a typical study of a cell lysate sample, ~400 proteins based on ~1,000 sequence-unique peptides (**Box 1**) can be confidently identified with a false discovery rate <5% within a 1–2 h gradient. Unfortunately, the relationship between number of chromatographic fractions and number of identified proteins is not linear[70]. For example, a typical 20-fraction experiment (requiring days rather than hours of instrument time) is likely to identify on the order of 3,000 proteins instead of the expected 8,000 (20 fractions with 400 proteins per fraction). This is because some analytes fall below the limit of detection of an instrument, but we may also be approaching the limit of expressed proteins in a biological system at a given time. Generally, chromatographic approaches can be applied at either the intact protein or peptide level, and it is not yet clear which fractionation strategy gives the best proteome coverage. A benefit of protein level fractionation (by one- or two-dimensional gels or column chromatography) is that proteins are separated both by mass and by other characteristics, which may distinguish among different protein isoforms. For example, glycosylated versions of a given protein will frequently segregate to different fractions than the parent protein. Another advantage of protein-level methods is the potential reduction in local dynamic range of a sample. However, many chromatographic separation techniques work better at the peptide level, providing better reproducibility and resolution. As a result, combinations of protein (for example, SDS-polyacrylamide gel electrophoresis; SDS-PAGE) and peptide separations (for example, the multidimensional protein identification technology, MUDPIT (**Supplementary Glossary**), which uses reverse-phase and strong cation exchange (SCX) columns in tandem, or peptide isoelectric focusing or hydrophilic interaction chromatography (HILIC) approaches) are frequently used to reduce proteome complexity and maximize proteome coverage[71].

Another possible explanation for the limitations imposed by sample complexity and dynamic range is ion suppression, a phenomenon wherein some analytes literally interfere with the ionization of other analytes so that they cannot be detected by the mass spectrometer, even though they are physically present. If one considers the process of ionization as containing a fixed amount of charge to be distributed, and that charge is distributed as a function of both abundance and ionizability, then high-abundance peptides that ionize well take most of the available charge, leaving only a small amount remaining. Probably more so, chemical noise present in a sample (for example, salts, detergents and solvent clusters) often limits dynamic range. The signal of low-abundance species (peptides) may be too weak to exceed the sometimes large signal of the chemical noise; or, conversely, highly abundant ions may saturate some detectors. Despite these issues, recent profiling approaches based on multidimensional chromatography and MS have demonstrated the ability to identify >5,000 proteins expressed over four orders of magnitude of cellular abundance in cellular proteomes[72,73], a tenfold increase over what was possible only five years ago[74].

In addition to proteome coverage, another key consideration in profiling experiment analyses is protein coverage. At first blush, it would seem that if a protein were present in a sample, all of its peptides should be readily observable. For several reasons that we have discussed elsewhere[75,76], this is not the case. Instead, in a typical proteomics experiment, only a single peptide is observed for many proteins; the median protein is identified by observation of only three peptides. This not only limits confidence in many of the identified proteins but also in their quantification (discussed further below). One can usually improve protein coverage by simplifying the composition of the mixture (for example, by the fractionation approaches described above). Alternatively, it has been shown that repeatedly analyzing the same mixture can improve coverage, at the expense of measurement time. Such resampling frequently not just increases a protein's peptide coverage but can also allow identification of 30% or more additional proteins[77]. A marked improvement can also be obtained by using multiple proteases with different cleavage specificities[78].

Analysis of PTMs in complex mixtures

In vitro cell culture models are also attractive systems for the broad-scale discovery of PTMs because one can subject cultured cells to a set of biologically well-controlled experiments. However, large-scale PTM expeditions require specialized upstream chromatography approaches that enrich or select for the PTM under investigation, highly sensitive mass spectrometers and sophisticated downstream software tools for the assignment of the modified peptide and the exact site of the modification.

Specialized laboratories have identified thousands of phosphorylation sites in many model organisms, ranging from yeast to worms and flies to plants and mammals[10,79–83]. Such studies produce a rough PTM signature of a biological system rather than of an individual protein. To obtain a full picture of the phosphorylation status of a particular protein or group of proteins, more focused experiments have proven successful. For instance, immunoprecipitation of phosphotyrosine-containing proteins or peptides have led to interesting discoveries in diverse applications[84,85]. Despite substantial advances in this area, however, several issues remain for global PTM discovery. It remains difficult to study glycosylation (owing to heterogeneity of oligosaccharide types and structures), ubiquitinylation (owing to branching of the protein) and very transient PTM events.

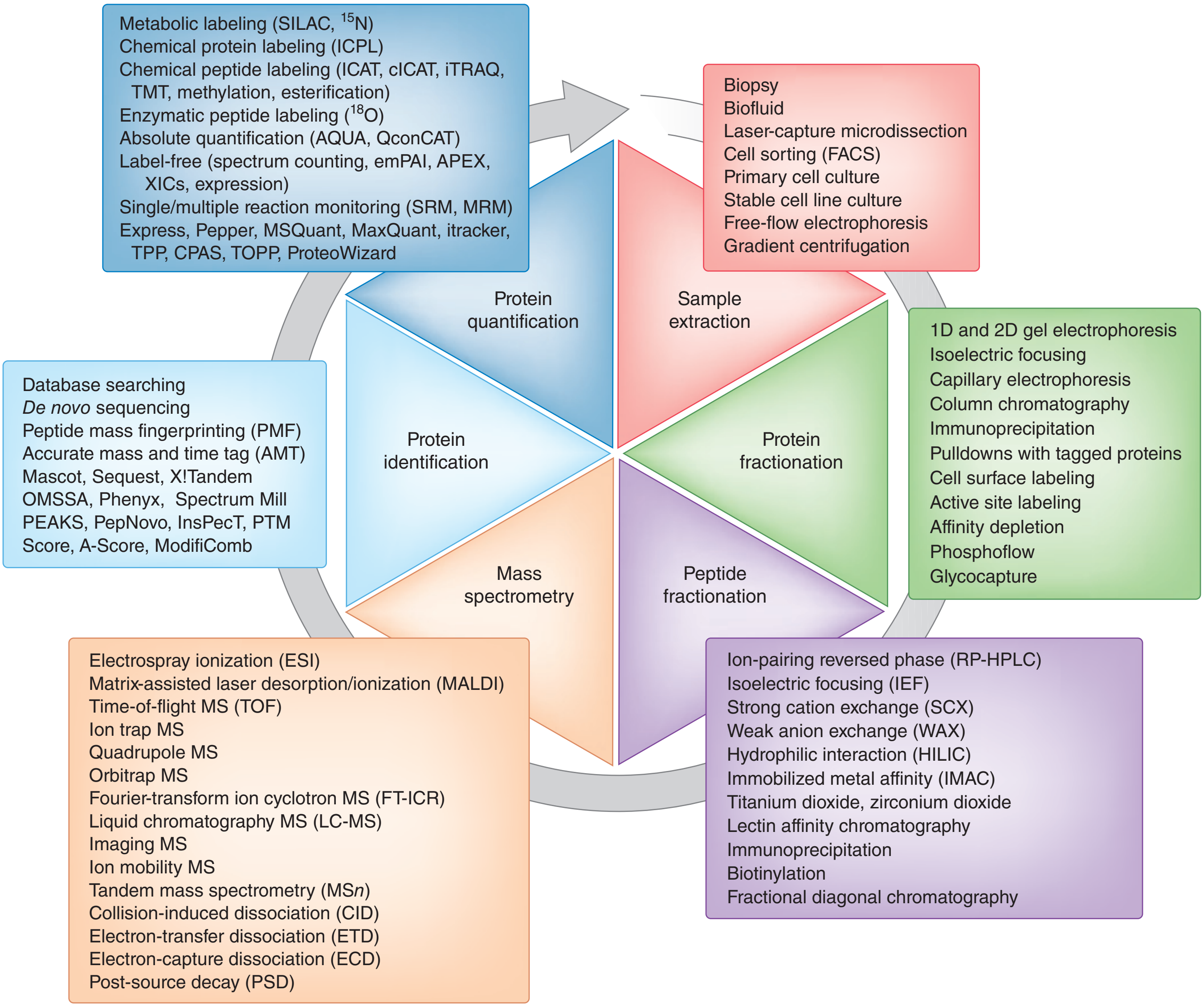

Metabolic labeling (SILAC, ^{15}N)
Chemical protein labeling (ICPL)
Chemical peptide labeling (ICAT, cICAT, iTRAQ, TMT, methylation, esterification)
Enzymatic peptide labeling (^{18}O)
Absolute quantification (AQUA, QconCAT)
Label-free (spectrum counting, emPAI, APEX, XICs, expression)
Single/multiple reaction monitoring (SRM, MRM)
Express, Pepper, MSQuant, MaxQuant, itracker, TPP, CPAS, TOPP, ProteoWizard
Biopsy
Biofluid
Laser-capture microdissection
Cell sorting (FACS)
Primary cell culture
Stable cell line culture
Free-flow electrophoresis
Gradient centrifugation
Protein quantification
Sample extraction
1D and 2D gel electrophoresis
Isoelectric focusing
Capillary electrophoresis
Column chromatography
Immunoprecipitation
Pulldowns with tagged proteins
Cell surface labeling
Active site labeling
Affinity depletion
Phosphoflow
Glycocapture
Database searching
De novo sequencing
Peptide mass fingerprinting (PMF)
Accurate mass and time tag (AMT)
Mascot, Sequest, X!Tandem
OMSSA, Phenyx, Spectrum Mill
PEAKS, PepNovo, InsPecT, PTM Score, A-Score, ModifiComb
Protein identification
Protein fractionation
Mass spectrometry
Peptide fractionation
Electrospray ionization (ESI)
Matrix-assisted laser desorption/ionization (MALDI)
Time-of-flight MS (TOF)
Ion trap MS
Quadrupole MS
Orbitrap MS
Fourier-transform ion cyclotron MS (FT-ICR)
Liquid chromatography MS (LC-MS)
Imaging MS
Ion mobility MS
Tandem mass spectrometry (MS*n*)
Collision-induced dissociation (CID)
Electron-transfer dissociation (ETD)
Electron-capture dissociation (ECD)
Post-source decay (PSD)
Ion-pairing reversed phase (RP-HPLC)
Isoelectric focusing (IEF)
Strong cation exchange (SCX)
Weak anion exchange (WAX)
Hydrophilic interaction (HILIC)
Immobilized metal affinity (IMAC)
Titanium dioxide, zirconium dioxide
Lectin affinity chromatography
Immunoprecipitation
Biotinylation
Fractional diagonal chromatography

Figure 3 Technologies for proteomics. This figure depicts the proteomic workflow from sample extraction to protein quantification. For each step in the workflow, the text boxes give examples of commonly used techniques, many of which may be combined in any one study. All featured techniques are discussed in detail in the **Supplementary Techniques**. Further details related to the terms database searching, *de novo* sequencing, peptide mass fingerprinting, electrospray ionization and matrix-assisted laser desorption/ionization can be found in the **Supplementary Glossary**. FACS, fluorescence-activated cell sorting; 1D, one-dimensional; 2D, two-dimensional.

Analysis of organellar protein compositions

A logical extension to the mapping the proteome of cells is the analysis of the protein complement of organelles or other large cellular structures. Organellar proteomics[86,87] links individual proteins to the functional context of a particular organelle (for example, drug receptors at the cell surface, transport mechanisms by vesicles or cell fate decisions at mitochondria). These experiments obviously depend on isolation of a particular organelle before identification of its protein constituents. The methods for the isolation of many organelles (mainly based on the sedimentation characteristics during centrifugation) are quite well established, and a combination of enzyme assays, western blotting and electron microscopy can be used to assess the enrichment and integrity of a preparation. Still, the field is plagued by controversies over whether or not certain proteins are genuine constituents of organelles or mere 'contaminants'. The use of stable isotope labeling and quantitative MS can offer insight here, as hundreds of proteins can be followed throughout the purification scheme and only those proteins showing the same quantitative behavior during purification are part of the same cellular structure[41,88]. Likewise, targeting the cell-surface proteome by protein chemistries specific for certain structures (for example, glycosylation) in combination with quantitative MS has led to determination of the proteomic content of this important organelle[89–91].

In vitro quantitative expression profiling

Although MS has been very successfully used in the analysis of proteins in complex mixtures, these studies have been so far dominated by qualitative results. This situation is slowly changing as quantitative measurement methods are becoming more widely available.

Quantitative expression profiling aims not just to identify the components of a proteome but also to compare two or more distinct proteomes to identify proteins with altered expression levels or post-translational forms in response to a given stimulus. Broadly, there are two primary approaches[92]: so-called label-free quantification methods and those that use stable isotope labeling of proteins or peptides. The former is attractive because one can in principle perform comparisons across many samples. The strength of the latter is its superior accuracy of quantification, albeit only for a small number of samples (up to eight). Each quantification approach compares the peptide signals observed in samples prepared under different conditions (for example, cells undergoing normal growth compared with cells treated with a therapeutic agent). Historically, proteomics has been most successful at relative quantification—determination of a ratio between a protein's concentration in one sample versus that in another. Absolute protein quantification approaches do exist, but they typically require the time-consuming and costly development of reference materials and assay conditions for each of the proteins of interest.

The simplest quantification techniques are the spectral counting (**Supplementary Glossary**) approaches (one variant of label-free quantification), which infer the abundance of a protein using the number of distinct peptides observed and/or the number of times a peptide from a protein is sequenced in an experiment. These approaches rely on the empirical observation that peptides from more abundant proteins are more likely to be sequenced and identified than peptides from less abundant proteins. Recently, counting approaches have demonstrated a dynamic range approaching six orders of magnitude[93], but several experimental conditions, including the selection criteria for picking a peptide for sequencing, can skew data. For example, MS acquisition regimes using 'inclusion lists', wherein only peptides from a predetermined list are sequenced (for example, in experiments probing a particular set of proteins from a pathway) are incompatible with counting approaches. In addition, if multiple 'dynamic exclusion' criteria (criteria by which peptides are excluded from further sequencing once they have been selected once by the mass spectrometer) are used, data sets may not be readily comparable. In addition, digestion artifacts and the variability of peptide ionization can make data unreliable. Furthermore, if only a few peptides are observed for a given protein (as is commonly the case), quantification accuracy decreases significantly. As with all label-free approaches, variation in sample handling can affect the reliability of estimates of protein relative abundance. Counting approaches are not exceptionally sensitive to small changes in abundance and cannot provide information about the change in abundance of a peptide relative to a protein, such as frequently arises by truncation or modification of a protein. When using the spectral counting technique, results can be computed in any of several ways. The simplest reports the average of ratios[94], and using an intensity threshold can help to minimize the noise-based bias[95]. More reliable results are achieved when computing the ratios on the basis of the intensity-weighted average or on the sum of all the observed spectra[94,96] or when using linear regression analysis[97].

An alternative label-free quantification technique compares the mass spectrometric intensity of each peptide in each of the experiments[98]. Peak intensity is a more direct measure of abundance than is the count of peptide identifications and thus offers some advantages (for example, linearity and accuracy). Unfortunately, this is as yet beyond the reach of most laboratories owing to the stringent requirements for MS quality assurance measures, as well as a lack of sophisticated software that can normalize for experimental variables introduced by peptide chromatographic drift between experiments. As more effort is put into building these software tools[99–106], this form of label-free quantification can be expected to become much more widely used.

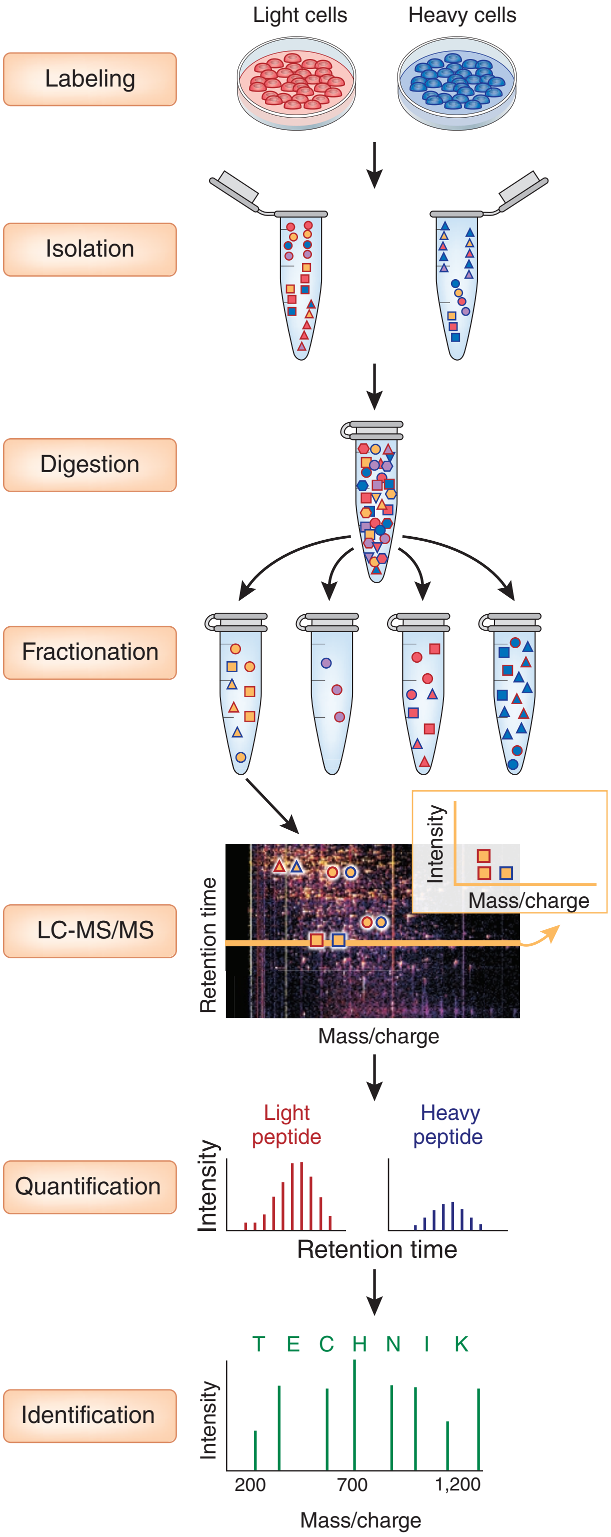
Use of isotopic labels, wherein samples are labeled either biosynthetically (as in SILAC) or, after isolation, chemically (as in isotope coded affinity tag (iCAT) or isobaric tags for relative and absolute quantification (iTRAQ) approaches) to create populations of peptides that are either isotopically light or isotopically heavy, provides more reliable quantification than label-free methods. When light and heavy samples are mixed and then measured in a mass spectrometer, the ratios of the intensities of the ions with slightly different masses, but the same chemical properties, can reliably be used for determining relative quantities. The addition of the label allows mixing of samples originating under different conditions for simultaneous analysis. When samples are mixed early in the workflow (that is, before a separation step), little bias is introduced during sample processing, resulting in high reproducibility. Therefore, methods that incorporate the stable isotope label at the protein level have generally higher reproducibility than those that introduce it at the peptide level. Label-based approaches have been shown to have excellent resolving power for quantifying small differences in protein abundance if combined with the appropriate mass spectrometer. For instance, the SILAC technique works best when using instruments with high resolving power, whereas the AQUA technique benefits from the large dynamic detection range of instruments capable of performing single or multiple reaction monitoring (SRM or MRM) experiments (discussed in more detail in ref. 107). As for protein identification, the dynamic range of protein quantification is often limited by the presence of chemical noise and the complexity of the analyzed peptide mixture. In practice, the linear dynamic range of quantification is often limited to 10- to 20-fold.

Several factors have to be considered when performing quantitative experiments. When choosing a stable isotope label, it must be determined whether the label alters the physicochemical properties of a peptide. For example, there is minimal impact when using ^{13}C, ^{15}N, or ^{18}O labeling[108], but deuterium labeling can be problematic because labeled and unlabeled peptides often differ in their retention time in reversed-phase high-performance liquid chromatography[109]. If retention times of labeled and unlabeled peptides differ, an extra signal integration step must be used to correct for this. The spectral quality also greatly affects accuracy. Data should be scrutinized when the signal is very low (close to the noise level) or very high (possibly resulting in detector saturation) because both will lead to distortion of the isotope envelope (**Supplementary Glossary**) intensity and result in inaccurate quantification. Accuracy also depends on the ability of the instrument to discriminate between interfering signals resulting from coeluting peptides of almost the same mass (a particular problem when using labels that are quantified in tandem mass spectra, such as in iTRAQ and TMT (tandem mass tags) techniques). This can be minimized by reducing the sample complexity through fractionation or by computational means[110]. A complicating factor is that analytes often do not elute in a narrow profile and sometimes even elute into two or more fractions in separate regions[111].

In vitro activity profiling

Activity- and affinity-based approaches are finding application in proteomics because they directly or indirectly focus on protein function and thus add a dimension that has mostly been missing in expression proteomics. Activity profiling was first demonstrated for serine hydrolases[112] but has since been applied to other enzyme classes, such as kinases, phosphatases and histone deacetylates[113,114]. In a typical activity-profiling approach, a small molecule inhibitor that can bind to members of an enzyme class is used as an affinity tool to purify these enzymes from a complex proteome before quantification by MS. This generates an enzyme class–specific expression profile of the underlying biological material that can be used to identify enzymes over- or

Figure 4 Protein identification and quantification by mass spectrometry. A typical proteomic workflow starts by extracting proteins from cells (here metabolically labeled), followed by proteome complexity reduction by fractionation techniques before MS is used to identify and quantify the proteins present in the original sample. Each element in the tubes represents a peptide, with its identically shaped elements originating from the same protein.



underexpressed in healthy versus pathological conditions. Activity profiling enables the identification of the targets of small molecule drugs in a proteome-wide fashion. We envisage activity proteomics playing a significant role in drug discovery as it becomes possible to profile the selectivity of drugs and their mechanisms of action systematically in relevant tissues.

An alternative to investigating enzymes is to look at their substrates. For kinases, this can be accomplished by techniques such as global and quantitative phosphopeptide profiling[85]. This is particularly attractive for studying cancer, as many individual tumor biologies arise from the dysregulation of signaling pathways. Global phosphorylation profiling therefore offers a route to classifying patients into groups on the basis of signaling pathways that underlie the development or progression of the disease[115]. An important future task will be to link quantitative phosphorylation profiles with the upstream kinases. This is not routinely possible now because the substrate specificities of most kinases are not precisely known. Even so, substrate trapping approaches[116] may make these studies possible. The ability to link enzymes and substrates is important, as it will reveal regulatory mechanisms as well as potential therapeutic targets. The analytical hurdles are often not very high for activity proteomics (unless accurate quantitative data is required) because the approach drastically reduces the complexity of the proteome by focusing on a class of proteins. The downside is that synthetic activity probes are often not available. This is because a fair amount of structural data on the catalytic site of an enzyme class is required to design a probe of suitable potency, and the catalytic site must be accessible for a generic inhibitor to purify a class of enzymes. Proteins with highly constrained binding sites will therefore be difficult to target. We believe that once organic chemistry is further integrated into proteomics research, activity-based approaches will become mainstream.

Translational studies

Above, we focused on the qualitative and quantitative characterization of *in vitro* systems. Here, we extend the discussion to proteomic characterization of *in vivo* systems. Studies in murine and human systems bear strong similarity to one another, and so the canon of techniques, except where noted, can typically apply to either. Specimen or sample extraction does add potential for introduction of bias that is largely irrelevant in *in vitro* studies. This is true for analysis of both body fluids and tissue biopsies. Furthermore, biological heterogeneity (genetic background, multiple cell types in organs and host/graft issues) poses significant technical and conceptual challenges. Unsurprisingly, as mice can be genetically identical and maintained on identical diets in near identical environments (for example, adjacent cages with similar temperature and light) their biological heterogeneity is much less than that of humans. The small size of mice can make sample extraction from tissues, such as ovaries, prostate or brain substructures, difficult and can lead to a sample-to-sample heterogeneity. Often, biological heterogeneity will require performing the biomarker discovery phase in a subgroup of proteins (to reduce the false discovery rate) or in cell culture models (molecular phenotyping), with subsequent corroboration in the relevant *in vivo* situation.

One typical study type is the characterization of the protein content of an organ or biopsy sample. Such studies are used to define organ proteomes (for example, the liver proteome) and to describe how such proteomes are altered by endogenous or exogenous perturbations. Studies have also been done to investigate the impact of wounds on the local proteome or the serum proteome. The challenge of these studies is primarily in sample extraction. For example, the extraction process itself can lead to inflammation and hypoxia, which significantly alters the proteome. In addition, contamination with vasculature, stroma and neighboring tissue (as is often encountered in tumor biopsies) may lead to quantitative differences between samples that are a function of differences in sample collection. Generally, replicate analysis of a given tissue can help distinguish biological variability from technical variability. Another approach, mostly used in the study of cancer, has been to use nearby 'normal' tissue as a control. Another option is to use cell sorting or tissue microdissection before proteome analysis[117–119]. However, the low amounts of material available from these techniques or the presence of fixation or crosslinking reagents can markedly limit the desired analysis[120].

One broad question still under debate is how best to handle biological heterogeneity in discovery experiments. Practically, this issue is often handled by analysis of pooled samples rather than samples from individuals. In pooled samples, effects are averaged. In addition, if resources are a concern, pooling may reduce the amount of instrument time required or allow a deeper fractionation approach and thus a broader look at the proteome. Pooling is sometimes necessary because material collected from a single subject is insufficient for a desired analysis. For example, a single tail bleed from a mouse provides only 50 µl of blood; after depletion, one is left with less than 10 µg of protein, which is insufficient for extensive fractionation. On the other hand, if subpopulations exist in a group of individuals, such signals are likely to be averaged by pooling. A commonly successful technique has been initial discovery in pooled samples to identify dominant effects and then verification and exploration of biodiversity in follow-up studies on individual samples[121,122].

Xenografts (**Supplementary Glossary**) or orthotopically implanted materials are extensively used in cancer research. One distinct benefit of these studies is their potential to differentiate proteins generated by the implant from host proteins, both in tissue and in the circulatory system. This benefit is also a complication, as sometimes 30% of tryptic peptides cannot be distinguished as murine or human by sequence. When performing quantification studies, such as regarding the growth of a tumor or its response to therapy, key experimental design questions must also be addressed. For example, how does one synchronize the size of samples extracted from multiple animals? What time points are most appropriate? At those time points, how does the disease or drug burden affect the intended results? For example, a common study comparing treated tumors to controls must consider the size and protein content of the tumor. If a tumor is smaller owing to treatment, what is the optimal way to normalize the samples before comparison?

Biofluid analysis

The analysis of biofluids is of interest for the discovery of serum- and plasma-borne markers. As with studies of tissue, biofluid analysis poses challenges in terms of sample collection and biological variability. So-called preanalytical variables have been confounding in serum studies because samples allowed to sit for varying amounts of time (from minutes to hours) have radically altered protein compositions. In addition, hemolysis, bacterial protein contamination and degradation are problems. There are also significant technical hurdles related to sample preparation, data reproducibility and protein dynamic range. A variety of chromatographic techniques for addressing sample complexity were described above, and each of these approaches has been used extensively for serum and plasma analysis. In addition, several chromatographic approaches have been developed specifically for serum and plasma[111,123]. Two techniques broadly used for improving the dynamic range of serum studies are targeted depletion of abundant proteins and selective enrichment of low-abundance proteins. Depletion techniques using antibodies (specific) or immobilized peptide bead libraries (nonspecific)[124] have been used to improve the dynamic range of proteomic analysis, as these techniques eliminate the most abundant proteins. Even though 99% of a target protein can be removed using these approaches, however, this may be insufficient when certain proteins (for example, albumin) are eight or ten orders of magnitude more abundant than proteins of interest. Both specific and nonspecific depletion techniques semirandomly deplete off-target proteins[125]. Consequently, some intersample differences in protein abundance may result from the depletion procedure itself[126]. Furthermore, proteins such as albumin are natural buffering and carrier agents. Consequently, depletion of these proteins can lead to adverse effects, such as precipitation.

In addition to targeted depletion, techniques to broadly or selectively enrich low abundance proteins are becoming popular. An example of broad enrichment is the capture of glycopeptides from serum proteome digests on hydrazide beads[91]. Recently, immunoprecipitation with antipeptide antibodies and MS have been used to quantify troponin I and interleukin-33 in serum[127,128]. The concept of using antibodies directed toward tryptic peptides is intriguing and has potential because it combines the advantages of the classic enzyme-linked immunosorbent assay (ELISA), namely selectivity, with the multiplexing capability of a mass spectrometer. There are challenges, including the questions of which peptide to choose for antibody generation, whether a single peptide is sufficient and representative and how one makes sure that plasma or serum digestion can be done reliably. Focusing on the glycosylated part of the proteome can be used in discovery projects, whereas selective enrichment through antibodies constitutes an assay for a particular protein of interest (**Fig. 1**).

As an alternative to discovering markers directly in samples from human subjects, several recent studies have first uncovered proteins in murine models and subsequently verified these findings in human clinical samples. Most of these studies used a combination of depletion and extensive fractionation to overcome dynamic range issues[129–134]. Such approaches are highly attractive because they can draw on the many murine models of human disease that have been established over the years. Obviously, the known limitations of using rodents as models of human disease apply.

Tissue imaging

One area of proteomics that is attracting increasing attention, particularly from pathologists, is imaging mass spectrometry (IMS)[135,136]. In this technique, a MALDI mass spectrometer records spectra from thin tissue sections to produce molecular weight–encoded 'images' of the distribution of constituent biomolecules. In contrast to conventional histological staining, IMS acts as a molecular microscope that records the distribution of hundreds of molecular species simultaneously without the need for a priori information about their molecular identity. IMS has proven utility in imaging of small molecules, such as lipids and drug metabolites (in this case, the molecule of interest is known)[137,138]; it is as yet unclear what the technique can deliver with respect to the discovery of protein biomarkers. This is because it is rarely possible to identify the molecular nature (that is, the protein identity) of a peak in an IMS spectrum. It is of paramount importance to overcome this hurdle as it is not even clear whether differential signals recorded for particular tissue areas indicate the underlying cellular structure or are artifacts of sample preparation such as cell or blood vessel damage. Despite these issues, one IMS protocol, in which HER2 status can be determined directly in breast cancer tissues, has been approved for diagnostic use[139].

Population proteomics

Population proteomics studies have proven difficult, and no protein biomarker discovered using proteomics has yet attained a level of validation accepted by regulatory agencies such as the US Food and Drug

Administration (FDA; Rockville, Maryland) or the European Medicines Agency (EMA; London). One challenge of population studies relates to the genetic variations among subjects. MS techniques have detected polymorphisms, truncations and splicing events in data repositories, such as peptideAtlas[7,140–142] and INSPECT[143–146], but the lack of complete coverage of proteins substantially impairs these studies, and genomics methods are at present far more powerful for these analyses. Although several pilot studies have been performed to quantify cardiovascular proteins and general serum proteins across a large number of people[6,128,147], the large inter-person variability of proteomes in the background of complex genetic diseases poses enormous challenges to study design, statistical significance and technical viability.

As mentioned before, it can be argued that before analyzing cohorts of individuals to discover biomarkers (**Supplementary Glossary**), the candidate list, regardless of which kind of molecular marker is sought (disease, diagnostic, prognostic, response, stratifying or other) and what molecular nature it may have (protein, peptide, modification), should be built from prior experiments in suitable and more controllable models. Once this information is available, emerging techniques such as MRM might be used to gather the many data points necessary for rigorous biomarker verification and validation. Proteomics is only one out of many pieces in the biomarker puzzle, and other techniques may be much more suitable for particular parts of the discovery process. The verification and clinical applications aside, proteomics for discovering biomarkers in human populations is in early development, and it will be some time before significant results can be expected. But several proof-of-principle studies have been performed, and some of these will hopefully develop into full clinical applications[148,149].

Conclusions

After 15 years of evolution, MS-based proteomics has measurably improved its robustness, sensitivity and usability and is now a routine part of biological inquiry workflows. MS-based proteomics is clearly a versatile tool and will become even more useful as currently novel proteomics approaches mature. Although proteomics technologies can now deliver very high quality data for basic biological research, their utility is most notable when the biological problem can be conceptually confined and experimentally approached in a focused fashion, with relevant discovery controls and extensive post-proteomic follow up. It is critical for the field's success that proteomics be treated as a component of broader biological studies. As with any experimental technique, the value of proteomics is not related to the price of the instrumentation being used, but instead to the rigor and thoroughness of the overall experimental design. As part of larger studies, there is no doubt that proteomics technology can help ask and answer important biological questions. For example, with the rapid pace of technological improvements, systems-wide profiling experiments are emerging as valuable additions to genomic technologies. Proteomics at the organism level, however, continues to pose significant conceptual and technical challenges. As our ability to deeply profile proteomes becomes more time and cost effective and the general understanding of biological systems is refined, biomarker candidates are likely to surface at increasing rates. For the full utility of proteomics experiments to be realized, improvement in productivity in the discovery phase must be complemented by more rapid and more globally applicable verification approaches than are now available. Though the gap between biologists' expectations of proteomics and what proteomics can deliver has historically often been wide, we fully anticipate that, through close collaboration, biologists and proteome scientists will be able to bridge this gap and use proteomic technologies to significantly contribute to our understanding of biological systems.

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

ACKNOWLEDGMENTS
The authors thank M. Bergman, C. Flinders, K. Kramer and S. Mumenthaler for comments on the manuscript. The work of P.M. was supported by NCI-1U54CA143907 and NCI-U54CA119367.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

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

1. Wasinger, V.C. *et al.* Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis* **16**, 1090–1094 (1995).
2. Ducret, A., Van Oostveen, I., Eng, J.K., Yates, J.R. III & Aebersold, R. High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry. *Protein Sci.* **7**, 706–719 (1998).
3. Washburn, M.P., Wolters, D. & Yates, J.R. III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001).
4. Wilm, M. *et al.* Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379**, 466–469 (1996).
5. Aebersold, R. Constellations in a cellular universe. *Nature* **422**, 115–116 (2003).
6. Keshishian, H. *et al.* Quantification of cardiovascular biomarkers in patient plasma by targeted mass spectrometry and stable isotope dilution. *Mol. Cell Proteomics* **8**, 2339–2349 (2009).
7. Omenn, G.S. *et al.* Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **5**, 3226–3245 (2005).
8. de Godoy, L.M. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1254 (2008).
9. Rush, J. *et al.* Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.* **23**, 94–101 (2005).
10. Olsen, J.V. *et al.* Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635–648 (2006).
11. Bouwmeester, T. *et al.* A physical and functional map of the human TNF-alpha/NF-κB signal transduction pathway. *Nat. Cell Biol.* **6**, 97–105 (2004).
12. Muzio, M. *et al.* FLICE, a novel FADD-homologous ICE/CED-3-like protease, is recruited to the CD95 (Fas/APO-1) death–inducing signaling complex. *Cell* **85**, 817–827 (1996).
13. Heck, A.J. Native mass spectrometry: a bridge between interactomics and structural biology. *Nat. Methods* **5**, 927–933 (2008).
14. Sharon, M. & Robinson, C.V. The role of mass spectrometry in structure elucidation of dynamic protein complexes. *Annu. Rev. Biochem.* **76**, 167–193 (2007).
15. Mann, M. & Jensen, O.N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **21**, 255–261 (2003).
16. Ong, S.E., Mittler, G. & Mann, M. Identifying and quantifying *in vivo* methylation sites by heavy methyl SILAC. *Nat. Methods* **1**, 119–126 (2004).
17. Denison, C., Kirkpatrick, D.S. & Gygi, S.P. Proteomic insights into ubiquitin and ubiquitin-like proteins. *Curr. Opin. Chem. Biol.* **9**, 69–75 (2005).
18. Zaia, J. Mass spectrometry of oligosaccharides. *Mass Spectrom. Rev.* **23**, 161–227 (2004).
19. Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W. & Gygi, S.P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. USA* **100**, 6940–6945 (2003).
20. Steen, H., Jebanathirajah, J.A., Springer, M. & Kirschner, M.W. Stable isotope-free relative and absolute quantitation of protein phosphorylation stoichiometry by MS. *Proc. Natl. Acad. Sci. USA* **102**, 3948–3953 (2005).
21. Zhang, X., Jin, Q.K., Carr, S.A. & Annan, R.S. N-terminal peptide labeling strategy for incorporation of isotopic tags: a method for the determination of site-specific absolute phosphorylation stoichiometry. *Rapid Commun. Mass Spectrom.* **16**, 2325–2332 (2002).
22. Kirkpatrick, D.S., Gerber, S.A. & Gygi, S.P. The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods* **35**, 265–273 (2005).
23. Gerber, S.A., Kettenbach, A.N., Rush, J. & Gygi, S.P. The absolute quantification strategy: application to phosphorylation profiling of human separase serine 1126. *Methods Mol. Biol.* **359**, 71–86 (2007).
24. Rudd, P.M. *et al.* The glycosylation of the complement regulatory protein, human erythrocyte CD59. *J. Biol. Chem.* **272**, 7229–7244 (1997).
25. Phanstiel, D. *et al.* Mass spectrometry identifies and quantifies 74 unique histone H4 isoforms in differentiating human embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **105**, 4093–4098 (2008).
26. Siuti, N. & Kelleher, N.L. Decoding protein modifications using top-down mass spectrometry. *Nat. Methods* **4**, 817–821 (2007).
27. Mayya, V., Rezual, K., Wu, L., Fong, M.B. & Han, D.K. Absolute quantification of multisite phosphorylation by selective reaction monitoring mass spectrometry: deter-

mination of inhibitory phosphorylation status of cyclin-dependent kinases. *Mol. Cell. Proteomics* **5**, 1146–1157 (2006).

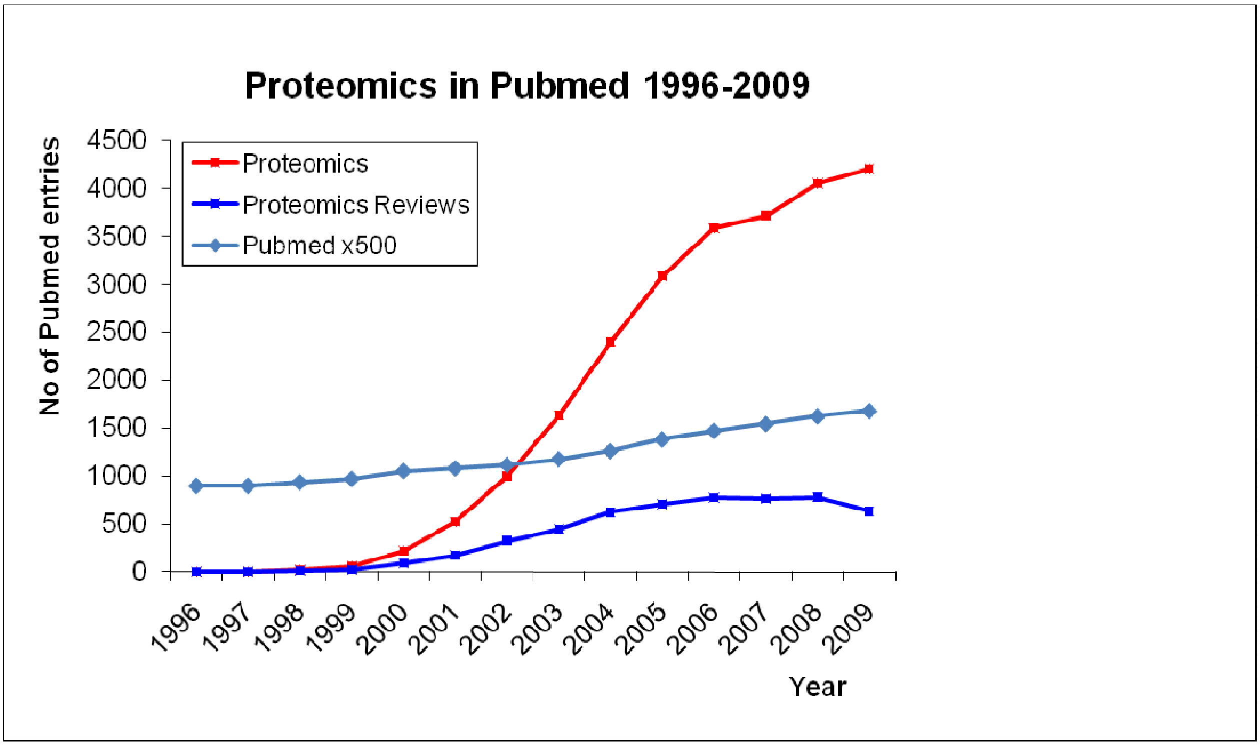
28. Desiere, F. *et al.* Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **6**, R9 (2005).
29. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
30. Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**, 291–294 (1998).
31. Neubauer, G. *et al.* Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc. Natl. Acad. Sci. USA* **94**, 385–390 (1997).
32. 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).
33. Blagoev, B. *et al.* A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat. Biotechnol.* **21**, 315–318 (2003).
34. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
35. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae. Nature* **403**, 623–627 (2000).
36. Bauer, A. & Kuster, B. Affinity purification-mass spectrometry. Powerful tools for the characterization of protein complexes. *Eur. J. Biochem.* **270**, 570–578 (2003).
37. Gingras, A.C., Gstaiger, M., Raught, B. & Aebersold, R. Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* **8**, 645–654 (2007).
38. Poser, I. *et al.* BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods* **5**, 409–415 (2008).
39. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17**, 1030–1032 (1999).
40. Schmitt-Ulms, G. *et al.* Time-controlled transcardiac perfusion cross-linking for the study of protein interactions in complex tissues. *Nat. Biotechnol.* **22**, 724–731 (2004).
41. Andersen, J.S. *et al.* Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426**, 570–574 (2003).
42. Pflieger, D. *et al.* Quantitative proteomic analysis of protein complexes: concurrent identification of interactors and their state of phosphorylation. *Mol. Cell. Proteomics* **7**, 326–346 (2008).
43. Andersen, J.S. *et al.* Nucleolar proteome dynamics. *Nature* **433**, 77–83 (2005).
44. Alber, F. *et al.* Determining the architectures of macromolecular assemblies. *Nature* **450**, 683–694 (2007).
45. Alber, F. *et al.* The molecular architecture of the nuclear pore complex. *Nature* **450**, 695–701 (2007).
46. Hochleitner, E.O., Sondermann, P. & Lottspeich, F. Determination of the stoichiometry of protein complexes using liquid chromatography with fluorescence and mass spectrometric detection of fluorescently labeled proteolytic peptides. *Proteomics* **4**, 669–676 (2004).
47. Menetret, J.F. *et al.* Single copies of Sec61 and TRAP associate with a nontranslating mammalian ribosome. *Structure* **16**, 1126–1137 (2008).
48. Nanavati, D., Gucek, M., Milne, J.L., Subramaniam, S. & Markey, S.P. Stoichiometry and absolute quantification of proteins with mass spectrometry using fluorescent and isotope-labeled concatenated peptide standards. *Mol. Cell. Proteomics* **7**, 442–447 (2008).
49. Hernandez, H. & Robinson, C.V. Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nat. Protoc.* **2**, 715–726 (2007).
50. Lorenzen, K., Olia, A.S., Uetrecht, C., Cingolani, G. & Heck, A.J. Determination of stoichiometry and conformational changes in the first step of the P22 tail assembly. *J. Mol. Biol.* **379**, 385–396 (2008).
51. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
52. Maiolica, A. *et al.* Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. *Mol. Cell. Proteomics* **6**, 2200–2211 (2007).
53. Sinz, A. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass Spectrom. Rev.* **25**, 663–682 (2006).
54. Ewing, R.M. *et al.* Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3**, 89 (2007).
55. Gavin, A.C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
56. Krogan, N.J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae. Nature* **440**, 637–643 (2006).
57. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
58. Kolch, W. Meaningful relationships: the regulation of the Ras/Raf/MEK/ERK pathway by protein interactions. *Biochem. J.* **351**, 289–305 (2000).
59. Schubert, P., Hoffman, M.D., Sniatynski, M.J. & Kast, J. Advances in the analysis of dynamic protein complexes by proteomics and data processing. *Anal. Bioanal. Chem.* **386**, 482–493 (2006).
60. White, F.M. Quantitative phosphoproteomic analysis of signaling network dynamics. *Curr. Opin. Biotechnol.* **19**, 404–409 (2008).
61. Kung, L.A. & Snyder, M. Proteome chips for whole-organism assays. *Nat. Rev. Mol. Cell Biol.* **7**, 617–622 (2006).
62. Paweletz, C.P. *et al.* Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene* **20**, 1981–1989 (2001).
63. Speer, R. *et al.* Molecular network analysis using reverse phase protein microarrays for patient tailored therapy. *Adv. Exp. Med. Biol.* **610**, 177–186 (2008).
64. Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105 (2001).
65. Huang, P.H. & White, F.M. Phosphoproteomics: unraveling the signaling web. *Mol. Cell* **31**, 777–781 (2008).
66. Picotti, P., Bodenmiller, B., Mueller, L.N., Domon, B. & Aebersold, R. Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. *Cell* **138**, 795–806 (2009).
67. Van, P.T. *et al. Halobacterium salinarum* NRC-1 PeptideAtlas: toward strategies for targeted proteomics and improved proteome coverage. *J. Proteome Res.* **7**, 3755–3764 (2008).
68. King, N.L. *et al.* Analysis of the *Saccharomyces cerevisiae* proteome with PeptideAtlas. *Genome Biol.* **7**, R106 (2006).
69. Chen, E.I., Hewel, J., Felding-Habermann, B. & Yates, J.R. III. Large scale protein profiling by combination of protein fractionation and multidimensional protein identification technology (MudPIT). *Mol. Cell. Proteomics* **5**, 53–56 (2006).
70. Malmstrom, J. *et al.* Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. *J. Proteome Res.* **5**, 2241–2249 (2006).
71. Hubner, N.C., Ren, S. & Mann, M. Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis. *Proteomics* **8**, 4862–4872 (2008).
72. Chen, E.I., McClatchy, D., Park, S.K. & Yates, J.R. III. Comparisons of mass spectrometry compatible surfactants for global analysis of the mammalian brain proteome. *Anal. Chem.* **80**, 8694–8701 (2008).
73. Graumann, J. *et al.* Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Mol. Cell. Proteomics* **7**, 672–683 (2008).
74. Schirle, M., Heurtier, M.A. & Kuster, B. Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **2**, 1297–1305 (2003).
75. Kuster, B., Schirle, M., Mallick, P. & Aebersold, R. Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* **6**, 577–583 (2005).
76. Mallick, P. *et al.* Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **25**, 125–131 (2007).
77. Tabb, D.L. et al. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **9**, 761–776 (2009).
78. Swaney, D.L., Wenger, C.D. & Coon, J.J. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J. Proteome Res.* **9**, 1323–1329 (2010).
79. Gruhler, A. *et al.* Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol. Cell. Proteomics* **4**, 310–327 (2005).
80. Pinkse, M.W. *et al.* Highly robust, automated, and sensitive online TiO_2-based phosphoproteomics applied to study endogenous phosphorylation in *Drosophila melanogaster. J. Proteome Res.* **7**, 687–697 (2008).
81. Reiland, S. *et al.* Large-scale *Arabidopsis* phosphoproteome profiling reveals novel chloroplast kinase substrates and phosphorylation networks. *Plant Physiol.* **150**, 889–903 (2009).
82. Villen, J., Beausoleil, S.A., Gerber, S.A. & Gygi, S.P. Large-scale phosphorylation analysis of mouse liver. *Proc. Natl. Acad. Sci. USA* **104**, 1488–1493 (2007).
83. Zielinska, D.F., Gnad, F., Jedrusik-Bode, M., Wisniewski, J.R. & Mann, M. *Caenorhabditis elegans* has a phosphoproteome atypical for metazoans that is enriched in developmental and sex determination proteins. *J. Proteome Res.* **8**, 4039–4049 (2009).
84. Lemeer, S. *et al.* Endogenous phosphotyrosine signaling in zebrafish embryos. *Mol. Cell. Proteomics* **6**, 2088–2099 (2007).
85. Zhang, Y. *et al.* Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. *Mol. Cell. Proteomics* **4**, 1240–1250 (2005).
86. Au, C.E. *et al.* Organellar proteomics to create the cell map. *Curr. Opin. Cell Biol.* **19**, 376–385 (2007).
87. Lilley, K.S. & Dupree, P. Plant organelle proteomics. *Curr. Opin. Plant Biol.* **10**, 594–599 (2007).
88. Dunkley, T.P., Watson, R., Griffin, J.L., Dupree, P. & Lilley, K.S. Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics* **3**, 1128–1134 (2004).
89. Jang, J.H. & Hanash, S. Profiling of the cell surface proteome. *Proteomics* **3**, 1947–1954 (2003).
90. Wollscheid, B. *et al.* Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat. Biotechnol.* **27**, 378–386 (2009).
91. Zhang, H., Li, X.J., Martin, D.B. & Aebersold, R. Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat. Biotechnol.* **21**, 660–666 (2003).
92. 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).
93. Griffin, N.M. *et al.* Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat. Biotechnol.* **28**, 83–89 (2009).

94. Saito, A. *et al.* AYUMS: an algorithm for completely automatic quantitation based on LC-MS/MS proteome data and its application to the analysis of signal transduction. *BMC Bioinformatics* **8**, 15 (2007).
95. Wolf-Yadlin, A., Hautaniemi, S., Lauffenburger, D.A. & White, F.M. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc. Natl. Acad. Sci. USA* **104**, 5860–5865 (2007).
96. Ono, M. *et al.* Label-free quantitative proteomics using large peptide data sets generated by nanoflow liquid chromatography and mass spectrometry. *Mol. Cell. Proteomics* **5**, 1338–1347 (2006).
97. Parish, R. Comparison of linear regression methods when both variables contain error: relation to clinical studies. *Ann. Pharmacother.* **23**, 891–898 (1989).
98. Mueller, L.N. *et al.* SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **7**, 3470–3480 (2007).
99. Cox, J. *et al.* A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat. Protoc.* **4**, 698–705 (2009).
100. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
101. Han, D.K., Eng, J., Zhou, H. & Aebersold, R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* **19**, 946–951 (2001).
102. Faca, V. *et al.* Quantitative analysis of acrylamide labeled serum proteins by LC-MS/MS. *J. Proteome Res.* **5**, 2009–2018 (2006).
103. Rauch, A. *et al.* Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome Res.* **5**, 112–121 (2006).
104. Jaffe, J.D. *et al.* PEPPeR, a platform for experimental proteomic pattern recognition. *Mol. Cell. Proteomics* **5**, 1927–1941 (2006).
105. Park, S.K., Venable, J.D., Xu, T. & Yates, J.R. III. A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat. Methods* **5**, 319–322 (2008).
106. Du, X. *et al.* A computational strategy to analyze label-free temporal bottom-up proteomics data. *J. Proteome Res.* **7**, 2595–2604 (2008).
107. Domon, B. & Aebersold, R. Three strategies for quantitative proteomics and their use. *Nat. Biotechnol.* **28**, 710–721 (2010).
108. Zhang, R. & Regnier, F.E. Minimizing resolution of isotopically coded peptides in comparative proteomics. *J. Proteome Res.* **1**, 139–147 (2002).
109. Zhang, R., Sioma, C.S., Wang, S. & Regnier, F.E. Fractionation of isotopically labeled peptides in quantitative proteomics. *Anal. Chem.* **73**, 5142–5149 (2001).
110. Zhang, Y. *et al.* A robust error model for iTRAQ quantification reveals divergent signaling between oncogenic FLT3 mutants in acute myeloid leukemia. *Mol. Cell Proteomics* **7**, 780–790 (2009).
111. Faca, V. *et al.* Contribution of protein fractionation to depth of analysis of the serum and plasma proteomes. *J. Proteome Res.* **6**, 3558–3565 (2007).
112. Liu, Y., Patricelli, M.P. & Cravatt, B.F. Activity-based protein profiling: the serine hydrolases. *Proc. Natl. Acad. Sci. USA* **96**, 14694–14699 (1999).
113. Bantscheff, M. *et al.* Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nat. Biotechnol.* **25**, 1035–1044 (2007).
114. Cravatt, B.F., Wright, A.T. & Kozarich, J.W. Activity-based protein profiling: from enzyme chemistry to proteomic chemistry. *Annu. Rev. Biochem.* **77**, 383–414 (2008).
115. Rikova, K. *et al.* Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell* **131**, 1190–1203 (2007).
116. Blethrow, J.D., Glavy, J.S., Morgan, D.O. & Shokat, K.M. Covalent capture of kinase-specific phosphopeptides reveals Cdk1-cyclin B substrates. *Proc. Natl. Acad. Sci. USA* **105**, 1442–1447 (2008).
117. Emmert-Buck, M.R. *et al.* Laser capture microdissection. *Science* **274**, 998–1001 (1996).
118. Lu, Q. *et al.* Analysis of mouse brain microvascular endothelium using immuno-laser capture microdissection coupled to a hybrid linear ion trap with Fourier transform-mass spectrometry proteomics platform. *Electrophoresis* **29**, 2689–2695 (2008).
119. Johann, D.J. *et al.* Approaching solid tumor heterogeneity on a cellular basis by tissue proteomics using laser capture microdissection and biological mass spectrometry. *J. Proteome Res.* **8**, 2310–2318 (2009).
120. Reimel, B.A. *et al.* Proteomics on fixed tissue specimens - a review. *Curr. Proteomics* **6**, 63–69 (2009).
121. Faca, V.M. *et al.* A mouse to human search for plasma proteome changes associated with pancreatic tumor development. *PLoS Med.* **5**, e123 (2008).
122. Harsha, H.C. *et al.* A compendium of potential biomarkers of pancreatic cancer. *PLoS Med.* **6**, e1000046 (2009).
123. Bandhakavi, S., Stone, M.D., Onsongo, G., Van Riper, S.K. & Griffin, T.J. A dynamic range compression and three-dimensional peptide fractionation analysis platform expands proteome coverage and the diagnostic potential of whole saliva. *J. Proteome Res.* **8**, 5590–5600 (2009).
124. Righetti, P.G., Boschetti, E., Lomas, L. & Citterio, A. Protein equalizer technology: the quest for a "democratic proteome". *Proteomics* **6**, 3980–3992 (2006).
125. Brand, J., Haslberger, T., Zolg, W., Pestlin, G. & Palme, S. Depletion efficiency and recovery of trace markers from a multiparameter immunodepletion column. *Proteomics* **6**, 3236–3242 (2006).
126. Seam, N. *et al.* Quality control of serum albumin depletion for proteomic analysis. *Clin. Chem.* **53**, 1915–1920 (2007).
127. Anderson, N.L. *et al.* Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). *J. Proteome Res.* **3**, 235–244 (2004).
128. Kuhn, E. *et al.* Developing multiplexed assays for troponin I and interleukin-33 in plasma by peptide immunoaffinity enrichment and targeted mass spectrometry. *Clin. Chem.* **55**, 1108–1117 (2009).
129. Pitteri, S.J. *et al.* Integrated proteomic analysis of human cancer cells and plasma from tumor bearing mice for ovarian cancer biomarker discovery. *PLoS ONE* **4**, e7916 (2009).
130. Katayama, H. *et al.* Application of serum proteomics to the Women's Health Initiative conjugated equine estrogens trial reveals a multitude of effects relevant to clinical findings. *Genome Med.* **1**, 47 (2009).
131. Faça, V., Wang, H. & Hanash, S. Proteomic global profiling for cancer biomarker discovery. *Methods Mol. Biol.* **492**, 309–320 (2009).
132. Faça, V.M. & Hanash, S.M. In-depth proteomics to define the cell surface and secretome of ovarian cancer cells and processes of protein shedding. *Cancer Res.* **69**, 728–730 (2009).
133. Faça, V.M. *et al.* Proteomic analysis of ovarian cancer cells reveals dynamic processes of protein secretion and shedding of extra-cellular domains. *PLoS ONE* **3**, e2425 (2008).
134. Hanash, S.M., Pitteri, S.J. & Faça, V.M. Mining the plasma proteome for cancer biomarkers. *Nature* **452**, 571–579 (2008).
135. Caprioli, R.M., Farmer, T.B. & Gile, J. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Anal. Chem.* **69**, 4751–4760 (1997).
136. Cornett, D.S., Reyzer, M.L., Chaurand, P. & Caprioli, R.M. MALDI imaging mass spectrometry: molecular snapshots of biochemical systems. *Nat. Methods* **4**, 828–833 (2007).
137. Hsieh, Y., Chen, J. & Korfmacher, W.A. Mapping pharmaceuticals in tissues using MALDI imaging mass spectrometry. *J. Pharmacol. Toxicol. Methods* **55**, 193–200 (2007).
138. Woods, A.S. & Jackson, S.N. Brain tissue lipidomics: direct probing using matrix-assisted laser desorption/ionization mass spectrometry. *AAPS J.* **8**, E391–E395 (2006).
139. Taguchi, F. *et al.* Mass spectrometry to classify non-small-cell lung cancer patients for clinical outcome after treatment with epidermal growth factor receptor tyrosine kinase inhibitors: a multicohort cross-institutional study. *J. Natl. Cancer Inst.* **99**, 838–846 (2007).
140. Deutsch, E.W. *et al.* Human Plasma PeptideAtlas. *Proteomics* **5**, 3497–3500 (2005).
141. Deutsch, E.W., Lam, H. & Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* **9**, 429–434 (2008).
142. Zhang, Q. *et al.* A mouse plasma peptide atlas as a resource for disease proteomics. *Genome Biol.* **9**, R93 (2008).
143. Castellana, N.E. *et al.* Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc. Natl. Acad. Sci. USA* **105**, 21034–21038 (2008).
144. Gupta, N. *et al.* Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Res.* **18**, 1133–1142 (2008).
145. Gupta, N. *et al.* Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res.* **17**, 1362–1377 (2007).
146. Tanner, S. *et al.* InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **77**, 4626–4639 (2005).
147. Gerszten, R.E., Carr, S.A. & Sabatine, M. Integration of proteomic-based tools for improved biomarkers of myocardial injury. *Clin. Chem.* **56**, 194–201 (2010).
148. Kentsis, A. *et al.* Discovery and validation of urine markers of acute pediatric appendicitis using high-accuracy mass spectrometry. *Ann. Emerg. Med.* **55**, 62–70 (2010).
149. Rifai, N., Gillette, M.A. & Carr, S.A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* **24**, 971–983 (2006).

Supplementary Note:

Accompanying information to Mallick and Kuster: Proteomics: a pragmatic perspective

Table of contents



Supplementary Figure 1: Historical survey of publications in the field of proteomics. The number of publications in the field of proteomics has grown exponentially over the past decade. As in many new areas of science, the proportion of reviews was initially high and decreased over time as the field matured. The data shown was generated by searching PubMed for the term Proteomics (all fields). For comparison, the total number of manuscripts in PubMed is also plotted (PubMed x500).

1. Glossary

Basic terms:

Biomarker: Measurable molecular phenotypic parameters that characterize an organism's state of health or disease or response to a particular therapeutic intervention [1].

Bottom-up proteomics refers to the analytical approach of separating and analyzing peptides following proteolytic digestion of a sample. This mode of operation is attractive because of its sensitivity and proteome coverage but comes at the expense of lost information about protein isoforms and inability to distinguish close sequence homologs.

Proteogenomics refers to the use of proteomic information to annotate a genome.

Proteome refers to the entire complement of proteins expressed by a genome and present in a cell, tissue, biofluid or organism.

Proteomics is the large-scale study of the proteome, particularly with regard to expression, structure, function, modifications, interactions, and the changes therein in different environments and conditions.

Stoichiometry: The relative proportion of components in a mixture (e.g., proteins in protein complexes or the occupancy of a phosphorylation site).

Systems biology is the study of relationships and interactions between various components of a biological system (metabolic pathways, organelles, cells, and organisms) and the integration of this information to allow understanding of how biological systems function.

Top-down proteomics refers to the analytical approach of separating and analyzing intact proteins. This mode of operation retains information on protein isoforms (sequence and/or modification) but for technical reasons comes at the expense of sensitivity and proteome coverage.

Xenograft: A surgical graft of tissue or cells from one species to another. In cancer research, most xenografts are human cancer cell lines or human tumors that have been transplanted into immune-deficient rodents.

Terms related to mass spectrometry:

Atomic mass unit refers to a unit of mass defined by the convention that the ^{12}C isotope of carbon has a mass of exactly 12 u; the mass of 1 u is therefore 1.67 x 10^{-27} kg.

Ion Selection & Dynamic Exclusion are the processes by which an ion is selected for MS/MS analysis. Typically an instrument is set to data-dependent selection, wherein the instrument will select the most intense ions for MS/MS analysis. Significant variation in which proteins

Nature Biotechnology: doi:10.1038/nbt.1658

are identified across experiments can be attributed to differential ion selection. To enhance the diversity of selected ions dynamic exclusion criteria can be set to exclude an ion from selection if it has been recently fragmented.

Isotope: An isotope is one of two or more forms of an element all having the same chemical properties but with different numbers of neutrons, different relative atomic mass, and different nuclear properties; for example, carbon occurs as ^{12}C, ^{13}C, and ^{14}C. In a high resolution mass spectrometer one is able to distinguish amongst instances of a peptide composed entirely of 'light' (e.g., ^{12}C, ^{14}N) atoms from those that have 1, 2 or more 'heavy' atoms (e.g., ^{13}C, ^{15}N). For the purpose of this appendix, all isotopes shall be stable (i.e. non-radioactive) isotopes.

Isotope envelope (isotopomers): The collection of peaks (signals in a mass spectrum) for a peptide containing different numbers of heavy atoms is referred to as an isotope envelope. For a small peptide, the first (lightest) peak in the envelope is the most intense peak. This is not the case for larger molecules in which the incorporation of naturally occurring heavy isotopes becomes statistically more and more likely and hence the isotope envelope shifts to higher and higher masses.

LC-MS/MS is an online-coupled HPLC-tandem mass spectrometer configuration that is frequently used for the identification of proteins from simple to complex mixtures (e.g., generated by shotgun proteome digestion).

[M+H]+ ion: The singly protonated molecular ion species. Likewise multiply charged molecular ion species are denoted [M+2H]2+ and so forth. Other ionizing agents might also be present in spectra such as sodium, denoted e.g. [M+Na]+.

MALDI is the abbreviation for matrix-assisted laser desorption/ionization, a particular mass spectrometer configuration that is frequently used for the identification of single proteins (e.g., from 2D gel spots).

Mass defect: The amount by which the mass of an atomic nucleus is less than the sum of the masses of its constituent particles: (sum of masses of protons and neutrons) - (measured mass of nucleus). As an example, one atom of carbon 12, the isotope of carbon with six protons and six neutrons, has a mass of exactly 12 amu, whereas the mass of six protons and six neutrons is 12.0956 amu (the mass of a proton is 1.00728 amu and a neutron is 1.00866 amu). Therefore, the mass defect is 0.0956.

Mass accuracy refers to how close a mass measurement is to its true (theoretical, exact, or accurate) value. In mass spectrometry this is often expressed in parts-per-million and is calculated as (theoretical mass - measured mass) / (theoretical mass)) x 1^6.

Mass precision is the ability of a measurement to be reproduced consistently. High precision measurements may not necessarily be accurate.

Mass spectrometer refers to an instrument that ionizes sample molecules and separates the ions according to their mass-to-charge (m/z) ratio. A mass spectrum is a graphical plot of measured ion intensity vs. its m/z value.

Mass resolution refers to the ability of a mass spectrometer to resolve different molecular species with similar but distinct masses. Mass resolution is the dimensionless ratio of the m/z value of a peak divided by its width at half maximum intensity. Resolution of 1,000 would mean that the instrument can resolve two peptides that differ by 1 u at a mass of 1,000. Resolution depends on the type of mass spectrometer and typically ranges from 1,000 to 100,000.

Molecular mass of a peptide is computed by adding the atomic masses of its constituent elements. Three mass definitions are frequently used: The ***monoisotopic mass*** is the mass of a molecule containing just the most abundant isotopes for each element. For high resolution spectra of peptides this is the first (lightest) peak in an isotope envelope. For example, the monoisotopic mass of the peptide 'TECHNISCHE' is 1172.4459 u. The ***average mass*** considers all stable isotopes that exist in a molecule. It is a weighted average sum of all isotope masses of all elements in a molecule multiplied by their relative abundance. For low resolution spectra of peptides, this is the apex of the unresolved isotope envelope and is always higher than the monoisotopic mass. The average mass of the peptide 'TECHNISCHE' is 1173.2600 u. The ***nominal mass*** of a molecule is the sum of the integers of the average masses of all atoms in a molecule. It is a convenient figure but has no physical meaning and cannot be determined by mass measurements. The nominal mass of the peptide 'TECHNISCHE' is 1172. **The *precursor ion mass*** is the mass of an intact peptide ion prior to fragmentation; it includes the mass of the charging species (i.e., protons). Likewise, the ***fragment ion mass*** is the mass of a peptide fragment generated by tandem mass spectrometry.

Stable isotopes are the one or more stable (i.e., non radioactive) forms of an element with different numbers of neutrons and thus different atomic mass (e.g., ^{12}C, ^{13}C). Naturally occurring molecules incorporate all stable isotopes of each constituent element at the frequency of their natural abundance. A high-resolution mass spectrometer can resolve all the isotopomers that make up the isotope envelope of, for example, a peptide composed of 'light' (e.g., ^{12} C, ^{14}N) atoms only or incorporating 1, 2 or more 'heavy' atoms (e.g., ^{13}C, ^{15}N). For a small peptide (few heavy atoms), the first peak in the envelope is the most intense peak. This is not the case for larger molecules (many heavy atoms).

Proteotypic peptides are those that are repeatedly and consistently identified for any given protein. These peptides are good candidates for targeted proteomics studies.

Terms related to protein identification:

Database searching refers to the process of using mass spectrometric information (peptide masses, fragment masses) to identify a peptide or protein in a database of known protein sequences. The idea is that the peptide mass and fragments thereof are indicative of the underlying peptide sequence and that experimentally determined mass spectrum can be compared to the computed spectra of all proteins in a sequence database.

***De novo* sequencing:** *De novo* sequencing is the process in which amino acid sequences are derived from tandem mass spectra by the fact that series of adjacent signals in these spectra often differ by the mass of an amino acid. This technique is frequently used when no genome sequence is available and when post-translational modifications are validated.

False discovery rate (FDR) is a statistical measure used to estimate the rate of false positives among a list of identified proteins. This is often required to describe the overall quality of large-scale protein identification experiments. One simple way to estimate an FDR is to search the mass spectrometric data against a normal and an artificially randomised sequence database (target-decoy). Any assignment in the randomized database must be wrong by definition and one can assume that the same number of hits may also be obtained in the normal database by chance.

Peptide mass fingerprinting: Peptide mass fingerprinting (PMF) refers to the use of accurate peptide mass information to identify a protein based on comparison to a sequence database. It is based on the idea that while any one proteolytic peptide mass of a protein may be found in many protein sequences, it is unlikely that multiple peptides from one protein will match several protein sequences by chance.

Terms related to quantitative proteomics:

Dynamic range is the ratio between the concentration of the least abundant and the most abundant protein in a complex mixture. The linear dynamic range, refers to the concentration range over which the calibration curve of a quantitative measurement platform remains linear. Typical LC-MS platforms have dynamic ranges of 3-5 orders of magnitude.

Ion chromatograms: Ion chromatograms refer to the mass spectrometric signals recorded over the course of an LC-MS/MS experiment. The total ion chromatogram (TIC) corresponds to the total MS signal intensity recorded over time. It includes all ions measured by the mass spectrometer and is thus similar to the UV signal of ordinary HPLC separations. The base peak intensity chromatogram (BPI) displays the signal intensity of the most intense ion at any one time. Ion chromatograms are used to monitor the overall performance of a separation. The extracted ion chromatogram (XIC) displays the elution behaviour of one particular peptide. The area under this curve is used for peptide quantification.

Label-free quantification of peptides/proteins does not use stable isotopes or any other label. Instead, most of these techniques compare the intensities of peptide signals detected in two or more separate experiments.

Multiple reaction monitoring or selected reaction monitoring: In multiple reaction monitoring (MRM) or selected reaction monitoring (SRM), a triple quadrupole (or linear ion trap) instrument is used to quantify a particular peptide species. The first quadrupole 1 (Q1) is set for a distinct precursor mass (m/z value) that enters the collision chamber. In the collision chamber (Q2), the peptide is fragmented. Q3 is set for one or more distinct fragment masses (m/z values). In this manner the transition Q1 – Q3 is monitored and only ions with a specific parent and fragment ion are measured. In LC-MS runs of complex peptide mixtures, an MRM experiment generates a chromatogram showing only those peptide species that generate signals for the set precursor and fragment masses. The area under the signal is a measure for the quantity of such peptides.

Quantitative proteomics experiments attempt to determine how much (in relative or absolute terms) of each peptide/protein is present in a mixture. Because mass spectrometry is not inherently quantitative, the vast majority of proteomics experiments employ ***relative quantification*** and report ratios of protein changes between experiments. ***Absolute quantification*** (e.g., determination of molar concentration) is much more difficult but can be achieved by spiking an isotopically labeled reference standard of known quantity.

Spectral counting is a special type of label-free quantification that uses protein identification results and is based upon the observation that the number of spectra matched to peptides from a protein is correlated to a protein's abundance.

Stable isotope labeling is the process of introducing a stable isotope into a proteomic sample to selectively increase the mass of all peptides/proteins. By mixing an unlabeled experiment with a labeled control, the change in protein abundance can be measured by comparing the intensities of the light form to the heavy form of any peptide.

Terms related to proteomics sample preparation approaches:

GeLCMS: GeLCMS is a technique for separation of proteins using 1D SDS-PAGE followed by in gel digestion of gel segments to generate moderately complex peptide mixtures. An LC-MS/MS experiment is then performed to identify the proteins in each gel segment.

Glycocapture: Glycocapture is a technique for enriching glycopeptides within a complex mixture. It involves chemically capturing the sugar moiety of a peptide on a hydrazide bead followed by elution using a glycosidase. The formerly glycosylated peptides are then identified by MS.

Multi-dimensional protein identification technology: Multi-dimensional protein identification technology (MUDPIT) is a particular type of shotgun proteomics in which 2D liquid chromatography is used to separate and identify proteins in complex mixtures. The classical set-up comprises a strong cation exchange column in series with a reversed-phase column to separate peptides by charge and hydrophobicity.

Protein arrays: Protein arrays are measurement devices used in biomedical applications to determine the presence and/or amount proteins in biological samples. Diverse capture agents, most frequently monoclonal antibodies or recombinant proteins, are deposited on a chip surface (glass or silicon) in a miniature array and binding of a sample to individual capture agents is evaluated by optical or radioactive detection systems.

Shotgun proteomics is a method in which a complete proteome is digested and the resulting complex peptide mixtures are separated by one or more dimensions of liquid chromatography before tandem mass spectrometry is used to identify the peptides.

SILAC is the abbreviation for stable isotope labeling with amino acids in cell culture; it is an approach for in vivo incorporation of a heavy isotope label into proteins for mass spectrometry-based quantitative proteomics. When labeled analogs of an amino acid are supplied to cells in culture instead of the natural amino acid, the label is incorporated into all newly synthesized proteins. After a number of cell divisions, proteins are fully labeled with this particular labeled amino acid

Unbiased Proteomics: As is true of most experimental techniques – some proteins are naturally more amenable to analysis than others. In proteomics, these tend to be proteins that are soluble and are easily digested. Likewise, as noted[2], some peptides are more amenable to analysis than others. When referring to an 'unbiased' experiment – researchers typically mean that no *intentional*, protein or peptide selection has been performed.

2. Sample extraction

Any proteomic sample should be collected from a source that reflects the biology of the system under study and success or failure is intimately linked to this first step. The sources of protein may be highly diverse and many different techniques may apply to acquire a specimen but several of considerations apply fairly generally when planning a proteomic project.

Biological origin: Is the sample extracted from a cell culture system, biological fluid, or primary tissue? In cell culture, cell populations are generally very homogeneous but sorting techniques such as fluorescence activated cell sorting (FACS) may be needed to select a particular sub-population. Homogeneity makes linking of the result of a proteomic experiment to the biology under study often straightforward.

Body fluids are relatively easy to obtain by standard clinical procedures, although special care must be taken in order not to contaminate a sample with proteins from the surrounding tissue (e.g., blood or lymph vessels, epidermis, muscle). In contrast to cell culture systems, body fluids are inhomogeneous mixtures of different cell types and soluble proteins from all regions of the organism. Separation of cellular from soluble components is often necessary. Samples obtained from tissues by surgical procedures (e.g., during an operation or by needle biopsy) are a particular challenge because these samples are almost invariably mixtures of different cell types and contain varying quantities of blood. In addition, there often is a considerable time between sample collection (e.g., in an operating theatre) and protein extraction; any delay may compromise the integrity of a sample. It is quite common to snap-freeze or, more recently snap-heat [3], specimens immediately after resection in order to preserve the *in vivo* state of a proteome as much as possible. This is much easier said than done as surgical routines in hospitals typically do not readily accommodate this step and ressected tissue must often first be inspected by a pathologist. In addition, most biopsies will contain considerable amounts of blood that may interfere with the analysis or mask the presence of less abundant proteins. Removing this blood from the specimen is difficult as perfusion of blood vessels is often not possible and passive diffusion is inefficient and carries the risk of undesired proteome changes. A common technique used to deal with cell heterogeneity in tissues is laser capture microdissection (LCM) [4] in which frozen or paraffin embedded tissue sections are covered with a plastic film and a laser is used to fuse cells of interest to the plastic backing. This allows the retrieval of these cells from the section for subsequent analysis. Although LCM is very effective, it is also very laborious and proteins extracted in this way are mostly inactive. An alternative to LCM is isolation of certain cells by more classical approaches and establishment of a relevant primary cell culture system, which is then propagated for a small number of passages prior to the proteomic experiment or analysis.

Protein localization and solubility: Proteins of interest may be extra- or intracellularly localized and many proteins reside inside organelles or membranes. As a result, extraction methods for proteins are very diverse and often require careful optimization of buffer systems (ionic strength, detergents, etc). Most techniques for organelle purifications rely on differential and/or gradient centrifugation, which takes advantage of the fact that many cellular components have different densities and sedimentation coefficients. Alternatively, electrophoretic methods such as free-flow electrophoresis are often successfully used for this purpose [5].

Protein activity: Depending on the goal of a study, protein activity may have to be preserved or destroyed. Most proteomic studies measure protein expression rather than activity. Therefore, most approaches focus on rapidly destroying protein activity by use of inhibitors, heat, chaotropic agents, detergents or combinations thereof. However, for many functional (e.g., analysis of protein-protein interactions) and chemical (e.g., protein-drug interaction analyses) proteomics experiments, it is important to keep proteins intact, properly folded, and active for the duration of the experiment. This limits the range of extraction methods that can be applied and poses the concomitant risk of rapid changes (e.g., protein degradation, stress response) that adversely influence the outcome of an experiment.

Protein quantity: It is often practically impossible or ethically unacceptable to collect large quantities of protein for proteomic studies. This is typically less of a problem when cell culture systems are studied but almost invariably challenges the analysis of primary material. Should this be the case, most experimental strategies aim to miniaturize and shorten all procedures to avoid losses and use the most sensitive detection systems available.

3. Protein fractionation

Because proteomes are very complex mixtures, some degree of protein fractionation is often employed prior to analysis. This is often referred to as ‘top-down proteomics’. The most common techniques for this purpose are one- or two-dimensional gel electrophoresis or affinity chromatography. Classical column chromatography (e.g., size exclusion, ion exchange, reversed-phase) can also be used but this is rarely done, mostly because of rather limited resolution (one would rarely collect more than 20 fractions) and unavoidable sample dilution that results. If and when column chromatography is used, the proteins are typically denatured prior to separation. Capillary electrophoresis of proteins is a somewhat higher resolution technique but is also rarely used because it cannot easily be used in a small-scale preparative fashion.

Gel electrophoresis: This class of protein separation techniques exploits the size and charge of native or denatured proteins as parameters for separation; proteins are essentially driven by an electric current across a gel matrix. The most common method is the classical SDS-PAGE. In general, the gel matrix is made up of cross-linked acrylamide that yields a porous network in which the ratio of acrylamide and bis-acrylamide concentrations affects the pore size. Prior to SDS-PAGE, proteins are loaded with SDS which produces a uniform charge to mass ratio for all proteins so that separation occurs based solely on size. Denatured proteins are loaded into wells at the top of the gel and when the electric field is on, proteins move toward the anode. Gels are usually sectioned into two steps to achieve better resolution. In the first step (stacking), proteins are concentrated by an isotachophoresis process. In the second step, proteins are separated based on their ability to navigate the matrix. An alternative to this classic configuration is to enhance resolution by the use of gradient gels in which the pore size of the gel gets progressively smaller toward the bottom. After electrophoresis, proteins are visualized using common stains such as Coomassie brilliant blue or silver. Depending on gel size and resolution, SDS-PAGE enables separation of proteins into about 10-50 fractions that are recovered by excision and digested into peptides for sequencing by MS. The technique is simple and almost universal applicable. Not only will most proteins (including membrane proteins) run on these gels, it also is a ‘safe container’ for proteins that removes buffer components (salt, detergent) that would interfere with subsequent chromatographic and mass spectrometric analysis. Still, the technique is not without issues: Significant sample loss may occur as the extraction of peptides from gels may be inefficient due to size or solubility. Another issue is that some proteins, such as glycoproteins and very basic proteins, have poor detergent binding and the high pH conditions of this electrophoresis method may remove some naturally occurring PTMs or introduce artificial protein modifications. One-dimensional SDS-PAGE also generally does not allow separation of proteins with different PTMs because of insufficient resolution.

To increase resolution, two-dimensional gel electrophoresis has been used for several decades in proteomics. These separations include an isoelectric focusing (IEF) step prior to SDS-PAGE. IEF separates proteins based on isoelectric points. Separation is achieved by placing proteins in a pH gradient, in which their movement is driven by an electric current. Once proteins reach their isoelectric point zone, their net charge becomes zero and they cease movement. Like chromatofocusing, this technique is able to resolve proteins that differ only slightly in pI value (by as little as 0.1 pH unit). IEF is a useful tool for separating proteins and their post-translationally modified forms and, in conjunction with SDS-PAGE, results in high resolution separations (up to several thousand visible spots). Over the years, the downsides of 2D gels have also become apparent, however: Integral membrane proteins are very difficult to display on 2D gels, many proteins turn insoluble at their pI resulting in losses and large proteins (>100 kDa) are underrepresented in these separations. Nevertheless, 2D gels are still a valuable component of many proteome analysis strategies.

Affinity chromatography: An altogether different protein separation is affinity chromatography. Here, specific proteins, complexes, protein classes or proteins residing in parts of the cell (e.g., cell surface) are selectively captured by a matrix-immobilized ligand (e.g., antibody for immunoprecipitation; tagged protein for pull down; small molecule for active site labeling; lectin/hydrazide for glycocapture; chelator for phospho-capture, etc.) and all non-binding proteins are washed away. The ligand-protein complex is then destabilized by salt, pH shift, addition of competing ligand, denaturing agents or enzymatic digestion and proteins of interest are recovered for analysis. Many of the above techniques (and others not listed) have been successfully used in proteomics, particularly in the sub-discipline of ‘functional proteomics’ that typically focuses on a sub-proteome in which the proteins are functionally linked (e.g., signaling pathways). In proteomics, affinity chromatography is also used in the reversed sense to deplete highly abundant proteins and thereby reduce the complexity of a sample. This strategy is frequently utilized in proteomic studies of serum, plasma, and other body fluids to enhance the detection of low abundance proteins and achieve broader proteome coverage. In human plasma, the 22 most abundant proteins account for ~99% of the total protein mass and if not removed, these proteins mask the detection of low abundance proteins of interest[6]. Depletion may however also lead to losses of low abundance proteins as these are often associated with the more abundant proteins that are depleted. Commercial kits exist that deplete a number of high abundance proteins and the concept of depletion has been generalized by the introduction of so-called 'equalizer beads' that carry a highly diverse immobilized peptide library to specifically bind and recover proteins in comparable amounts from any protein source [7]. This is a fairly recent development and time will tell whether this technology will become widely applicable. For all types and purposes of affinity chromatography, one has to bear in mind that this step introduces a bias into the sample and, therefore, the quantitative relationship of proteins before and after affinity chromatography are not the same (and mostly unknown).

4. Peptide fractionation

Many proteomic projects do not employ any protein separation but instead digest the entire proteome into peptides that are subsequently fractionated and identified by MS. This approach is called 'shotgun proteomics' and is most frequently used in quantitative protein expression profiling experiments because it is thought to introduce the least quantitative bias into a biological sample. Like in protein fractionation, electrophoretic and affinity methods can also be used on the peptide level for the separation of complex mixtures. Column chromatography plays a major role in peptide analysis because the resolution and peak capacity of the various available techniques (see below) is typically much higher on the peptide than on the protein level.

Ion exchange chromatography: In any chromatographic separation, peptides interact with a stationary phase and are then released by displacement into the mobile (liquid) phase using a more strongly interacting mobile phase component. In ion exchange chromatography, peptides are separated based on their electrical charge. Peptides bind to the solid phase through electrostatic interactions between the opposite charges found on the peptides and those of chromatographic phase. The mobile phase composition (ionic strength, pH) during binding is chosen such that (almost) all peptides bind to the stationary phase. For cation exchange chromatography, peptides are loaded at low pH (all peptides protonated, positive net charge), for anion exchange chromatography peptides are loaded at high pH (all peptides deprotonated, negative net charge). Peptides are then eluted by gradually increasing the ionic strength (salt concentration) of the mobile phase. The most common ion exchange chromatography techniques are strong cation exchange (SCX) and weak anion exchange (WAX) separations. Peak capacities for ion exchange chromatography of tryptic peptides range from 20 to >100. In practice however, the actual separation power of the technique is almost always underutilized because it is mostly employed as the first dimension of a 2D chromatographic strategy (the second dimension in proteomics is almost exclusively reversed-phase chromatography, see below) and it is often impractical to collect and analyze more than 20-30 fractions. For SCX of tryptic peptides in particular, one observes a strong correlation of net peptide charge with peptide retention time such that peptides with few net positive charges (e.g., phosphopeptides, N-terminally acetylated peptides) elute much earlier than non-modified peptides.

Isoelectric focusing: IEF of peptides is emerging as another important technique for proteomics. IEF separates peptides by their isoelectric points [8]. As for proteins, peptides migrate in a strong electrical field along an immobilized pH gradient. Once they reach their pI, the net charge on the peptide is zero and migration stops. The majority of pI values of tryptic peptides range from ~3 to ~10 and the resolution of the technique is <0.1 pI units, making IEF a powerful approach. However, as for ion exchange chromatography, IEF is mostly used as the first dimension of a 2D peptide separation approach and thus, one rarely collects more than 12-24 fractions.

Hydrophobicity: A third important parameter used in peptide fractionation is hydrophobicity. Ion-pairing reversed-phase high-performance liquid chromatography (RP-HPLC) separations are based on this property. A non-polar stationary phase (hydrocarbons of varying chain length) and an aqueous (polar) mobile phase containing an ion pairing reagent (such as formic acid or other organic acids) are used. Peptides bind to the non-polar stationary phase by two mechanisms: *i)* through direct hydrophobic interactions of the stationary phase with hydrophobic side chains of amino acids in the peptide and *ii)* through interactions of the stationary phase with polar parts of the peptide mediated by the amphiphilic ion pair molecule. By increasing the non-polar character of the mobile phase (i.e., increasing the percentage of organic solvent such as acetonitrile or methanol) adsorbed peptides are eluted. Peptides that are more non-polar have a longer retention time than polar ones. RP-HPLC of peptides offers excellent resolution (peak capacity 100 to > 500) and the solvent composition is directly compatible with mass spectrometry. Therefore, in proteomics, RP-HPLC is almost always directly (LC-ESI-MS) or indirectly (LC-MALDI-MS) coupled to a mass spectrometer. In conjunction with SCX or IEF as a first peptide separation dimension, up to 10,000 peptides (equivalent to ~3,000 proteins) can be identified in a single sample.

Chromatography of phosphopeptides: The above techniques address the separation needs for global proteome profiling. Analysis of PTMs require additional steps, notably the enrichment of particular peptides. Although many useful approaches exist for the analysis of phosphorylated peptides, other important areas of PTM analysis are much less developed. A classical way to enrich for phosphorylated peptides utilizes the immobilized metal affinity chromatography (IMAC) technique [9]. In this technique, a metal ion (iron or gallium) is linked to iminodiacetic acid resin. This trivalent cation complex binds phosphopeptides (and highly acidic peptides) by chelation and elution is accomplished with solvent systems containing competing molecules such as immidazole or phosphate or by raising the pH. To minimize non-specific binding from peptides rich in carboxylate groups, tryptic peptides can be converted to methyl esters using methanolic HCl prior to enrichment [10]. Strong cation exchange chromatography (see above) is also useful for phosphopeptide enrichment because phosphorylated tryptic peptides tend to elute in early fractions (lower net positive charge compared to unmodified tryptic peptides). Conversely, phosphopeptides are strongly retained on hydrophilic interaction chromatography (HILIC), which enriches for phosphopeptides in late fractions [11]. Today, metal oxides such as TiO_2 and ZrO_2 [12] are increasingly used for the enrichment of phosphopeptides. These transition metal oxides appear to bind phosphopeptides by coordinating the oxygen of the phosphate moiety into free d-orbitals of the transition metal. This results in very good selectivity for phosphopeptides over acidic peptides. These columns are typically used in a convenient two-step process: In the first step, a digested sample is passed over the TiO_2 column and only the phosphopeptides bind. In the second step, TiO_2-bound phosphopeptides are eluted by raising the pH of the mobile phase to 10.5. This enables separation of phosphorylated and non-phosphorylated peptides into two pools that can subsequently be analyzed. The simplicity of this approach makes it particularly amenable to automation. A technique for the enrichment of tyrosine phosphorylated peptides is immunoprecipitation (IP) using generic pY antibodies (4G10, pY100 and others). Phospho-tyrosine is a strong epitope so these IPs work on the peptide level [13]. Tyrosine phosphorylation is comparatively rare (approx 1% of serine/threonine cases) and the IP boosts sensitivity and selectivity. Immunoprecipitation with pY antibodies are usually performed on full proteome digests. This ensures that the quantitative relationship of pY levels in different samples can be compared with little or no bias; this is not the case on the protein level as it is unclear how changing levels of Tyr phosphorylation affect the amount of immunoprecipitated protein.

Chromatography of other modified peptides: Separations for PTMs other than phosphorylation are far less well developed and currently demand a case by case crafted strategy. For example, chromatographies showing strong retention of polar molecules such as HILIC and graphitized carbon [14] are useful for glycopeptides and the interaction of sugars with lectins can also be exploited for this purpose [15]. Functional groups of peptides such as the N-terminus or cysteine side chains can be targeted by special chemistries. This in turn allows for their selective retrieval by the appropriate chromatography. One example is the isolation of cysteine-containing peptides by biotinylation. A cysteine-selective chemical reagent that is coupled to biotin is used to chemically modify Cys-containing peptides and these peptides are selectively retrieved by virtue of the interaction of biotin with avidin (or derivatives thereof). Another interesting example is the isolation of N-terminal peptides by the Cofradic technique (combined fractional diagonal chromatography). It consists of two consecutive identical chromatographic separations with a chemical modification step between the two separations that targets a subset of peptides (say those with a free N-terminus). The chemically modified peptides display different chromatographic properties (typically a radical retention time shift) and therefore are segregated from the bulk of unaltered peptides in the second run. This essentially sorts the N-terminal peptides out of the entire proteome and offers an elegant way to study the substrate repertoire of proteases [16]. Despite the above choices, most PTMs, including such important ones as ubiquitinylation, are still very difficult to study systematically.

5. Mass spectrometry of peptides

Mass spectrometry is the key analytical technique in proteomics as the information provided by the technique is used for the identification of proteins and, increasingly, for measuring the quantities of proteins. It is important to note that mass spectrometers do not measure the mass of a molecule but the mass to charge ratio of ions in the gas phase. Therefore, in any mass spectrometric technique, peptides must first be transferred into the gas phase, then ionized so that they can then be analyzed for the mass to charge ratio (m/z). Essentially, a mass spectrometer consists of three parts: *i)* an ion source in which analytes (say peptides) are transferred into the gas phase and ionized, *ii)* a mass analyzer in which (peptide) ions are separated by their mass to charge ratio, and *iii)* a detection system that registers the individual m/z separated analyte ions.

Methods for peptide ionization: For proteomics, the two relevant techniques for ionization are matrix assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI). Both are so-called ‘soft’ ionization techniques because unlike traditional approaches, they are able to ionize large and polar molecules without physically destroying them. For MALDI, a peptide mixture (say from a protein digest) is first co-crystallized with a large excess of a UV-light absorbing organic molecule (the matrix) and transferred into the vacuum of the ion source of a mass spectrometer. A UV laser pulse is then used to irradiate the co-crystal, which leads to sublimation of the solid matrix along with peptide molecules into the gas phase. The energy deposited into the material leads to ionization of the matrix molecules, which in turn ionize the peptide molecules by proton transfer reactions. As a result, peptide molecules pick up a proton and become singly charged peptide ions (denoted as $[M+H]^+$).

In ESI, the peptide mixture is dissolved in a liquid solvent system, which makes the technique suitable for direct coupling to HPLC separations (see above). The liquid is passed through a thin capillary to which an electrostatic potential is applied. At the tip of the capillary, this potential causes the liquid to disperse into a mist of multiply charged droplets. The peptide molecules are dissolved in these droplets. As the droplets travel towards the vacuum system of the mass spectrometer, the solvent evaporates, which leads to an increased surface charge of the droplets. Ionization occurs by two mechanisms: *i)* when peptide molecules accumulate at the surface of the droplet, they are extracted and ionized from the droplet by field desorption or *ii)* when all solvent is evaporated, the excess charge of the droplets remains on the peptides. In both cases, peptides pick up multiple protons (roughly equivalent to one proton per basic site in the peptide) and form multiply charged ions (denoted as $[M+nH]^{n+}$). Most tryptic peptides carry two or three charges.

Methods for mass-to-charge determination: Ions can be manipulated by electrostatic and/or magnetic fields and the motion of (peptide) ions in such field is often a function of their m/z ratio. There are many technical ways to measure this ratio and thus, many different mass analyzers have been devised. For proteomics, several types of mass analyzers are found in a multitude of commercial instruments. In time-of-flight (TOF) mass analyzers, the m/z value of peptides is measured by measuring the time peptide ions require to travel over a fixed distance inside the mass analyzer. This time is a function of the square root of the m/z value. Peptides of low m/z values travel faster down a flight tube than peptides of high m/z. The flight times of each peptide ion can be converted to m/z values using a calibration function built on the flight times of standards of known m/z. Modern time-of-flight analyzers offer mass resolution of >30,000, mass accuracy of <5 ppm and sub-femtomol sensitivity.

Another important mass analyzer is the quadrupole. It functions as a mass filter and works on the principle that ions of different m/z values have different (stable) oscillations (trajectories) in a quadrupolar field. The quadrupolar field is generated by applying an oscillating electrostatic field to two pairs of parallel rods. This arrangement forces ions onto a 'spiraling' trajectory through the quadrupole rods; at any particular setting of field frequencies and amplitudes, only species of a single m/z value can travel through the device and be recorded at the detector. By 'scanning' the amplitudes of the applied fields, each m/z value can be focused on the detector in turn to create a full mass spectrum. Quadrupoles do not offer very high resolution and mass accuracy but are very sensitive and thus useful in quantitative measurements (see MRM technique below).

A third class of mass analyzer is the family of ion traps. As the name implies, these devices can trap ions in a (quadrupolar) field. Similar to quadrupoles, the stability of ions in the trap is a function of the applied electrostatic field parameters. By systematically varying the field parameters, ions of different m/z values can be ejected from the trap and focused onto the detector. Ion traps offer a significant advantage over other mass analyzers in that they can accumulate ions of interest and thus offer very good sensitivity. Their downside is that they only offer moderate mass accuracy (~0.1-0.5 amu) and resolution (500-2000) in the most commonly used measurement modes.

Yet another group of mass analyzers are Orbitraps and Fourier transform ion cyclotron resonance (FT-ICR) mass analyzers. The former separates ions of different m/z values based on their differences in oscillation frequencies along a central spindle inside a particularly shaped ion trap chamber. Each ion induces an oscillating image current and deconvolution of the many different frequencies observed for peptide mixtures is carried out by Fourier transformation. Similarly, in FT-ICR devices, ions of different m/z values display different 3D motions in a strong homogeneous magnetic field and their in-phase oscillation frequencies are recorded by an induced current. Again, overlapping frequencies of image currents can be deconvoluted by Fourier transformation. Both Orbitrap and FT-ICR instruments offer superior mass resolution (>100,000) and mass accuracy (<1 ppm), which has important implications for the quality of protein identification (see below).

It should be noted that ionization and m/z analysis are two independent steps in the analysis process. In fact, it is possible to combine virtually any ionization technique with any type of mass analyzer. In practice however, not all possible combinations are actually realized because several combinations turn out to be particularly fruitful. Classic combinations for proteomics are MALDI-TOF (e.g. for 2D gel analysis) or ESI-ion trap (for LC-MS) but other combinations are also commonly found (such as ESI-FT, ESI-Orbitrap, and ESI-TOF).

Tandem mass spectrometry: Peptide mass determination is the first step toward the identification of a peptide sequence. The second step is to generate partial amino acid sequence information for a particular peptide. There are a number of MS-based techniques for this purpose. The general idea is that a peptide is first isolated from a mixture of peptides inside the mass spectrometer. In a second step, the peptide is supplied with sufficient energy to undergo fragmentation. The m/z values of the fragments are then recorded in a so-called tandem mass spectrum. Tandem mass spectrometers simply contain two mass analyzers (tandem in space) or perform the experiment sequentially inside the same mass analyzer (tandem in time). Each configuration has certain advantages and disadvantages but both concepts are widely used. Examples for the former type are quadrupole-time-of-flight (QTOF), triple quadrupole (QQQ), and TOF-TOF instruments in which the first mass analyzer is used for peptide isolation and the second is used to determine the fragment m/z values. Examples for the latter are ion traps and FT-ICR instruments. There are also instruments that combine features of both approaches such as the ion trap-Orbitrap configuration.

The most common fragmentation techniques for tandem mass spectrometry today are collision induced dissociation (CID) and electron transfer dissociation (ETD) followed by electron capture dissociation (ECD) and post source decay (PSD). In CID, a peptide is forced to undergo multiple collisions with inert gas molecules (such as He, N2, Ar) during which the ions accumulate vibrational energy until chemical bonds are broken. Most of the bond cleavages involve the peptide bond (generating so called b- and y- fragment ions that contain the N- or C-terminus of the peptide respectively), which makes the technique highly valuable for peptide sequence determination. In CID, the weakest bonds in a molecule break preferentially. This often leads to problems if peptides contain labile modifications such as glycosylation or phosphorylation because such spectra often contain only few sequence-specific fragment ions. For a related reason, fragmentation of large peptides often leads to poor spectra because they tend to be dominated by few intense signals arising from cleavage of a few weak bonds. Both these issues can in principle be addressed by the alternative fragmentation techniques ETD and ECD. In ETD, peptide ions (positively charged) are reacted inside the mass spectrometer with an electron donor (such as fluoranthene). The transfer of an extra electron to the peptide leads to a charge-reduced ion that contains an unpaired electron. Unpaired electron configurations are highly instable and thus lead to rapid bond cleavage. Cleavages occur along the peptide backbone but, unlike those observed in CID, most ETD fragment ions result from the cleavage of the Cα-C’ bond (generationg so-called c- and z-fragment ions). ETD is advantageous for modified or large peptides because there is little cleavage bias. However, the dominant effect of ETD is often charge reduction rather than fragmentation which can impact negatively on sensitivity. The ECD technique results in spectra quite similar to those generated by ETD but differs in the way electrons are supplied to the peptide. Whereas ETD does so by chemical means and is implemented on many MS platforms, ECD uses low energy electrons generated by a heated filament and is primarily implemented on FT-ICR instruments. PSD is a particular fragmentation technique that can be used only on MALDI-TOF instruments. In a MALDI-TOF experiment, peptide ions are accelerated through a strong electrostatic field to acquire kinetic energy. Some of this kinetic energy is spontaneously converted into vibrational energy resulting in bond cleavage. Because intact peptides and fragments have different kinetic energies (while having the same velocity) fragment ions can be separated by the reflectron field that is implemented in all modern TOF instruments. Alternatively, the intact peptide and fragment ions can be elevated to the same kinetic energy by a post-source acceleration step (also known as LIFT) and analyzed by the ordinary time-of-flight effect.

LC-MS/MS: Even though tandem mass spectrometers are capable of analyzing peptide mixtures, the complexity of peptide mixtures generated in a proteomic experiment is often so great that chromatographic separations (see above) prior to tandem MS are required. These so called LC-MS/MS instruments are very commonly used today. Whereas many different techniques for peptide chromatography exist (see above), the coupling of LC to MS almost exclusively employs ion-pair reversed-phase chromatography. This is because the mobile phase of the LC separation consists only of water, acetonitrile, and an organic acid such as formic acid. All these components are volatile (i.e., they simply evaporate in the vacuum of the instrument) so that there is no need to remove interfering substances such as salts prior to MS analysis. A second common feature is that modern LC-MS/MS couplings employ so-called nanoHPLC systems. These are characterized by very small column diameters (e.g., 75 µm) and operate at low flow rates (e.g. ,200 nL/min). This is advantageous because the quantities of samples in proteomics are often limiting and so miniaturization is required. Depending on which ionization technique is used, LC-MS/MS couplings can be realized either 'on-line' or 'off-line'. In the on-line configuration, the (nano) HPLC is coupled directly to the ESI source of a tandem mass spectrometer. This coupling is convenient because both HPLC and ESI operate in the liquid phase so no sample collection step is needed, avoiding losses. A second strong point is that peptides can be sequenced with high throughput (typically 1-5 peptides per second) because peptides are analyzed immediately after they elute from the LC column into the mass spectrometer. Coupling of LC separations with MALDI ionization can only be achieved in an off-line configuration. Here, the eluting peptides are deposited onto the sample plate of a MALDI instrument and mixed with the matrix to form crystals. Following completion of the LC run, all collected fractions on the MALDI plate (hundreds to thousands) are transferred to the mass spectrometer for sequencing. The advantage of this configuration is that the chromatographic separation is 'frozen in space', which allows for in-depth analysis of any fraction. However, this comes at the expense of time as MS analysis of a one-hour LC separation may take many hours of MS time.

Selection of MS peaks for sequencing: It is critical to understand the process of how MS peaks are selected for sequencing when interpreting proteomics results. The most common selection paradigm is 'data directed' sequencing (also called 'data depenent') wherein the most intense signals in a spectrum are sequenced. Typically, exclusion criteria are put in place to minimize repeatedly sequencing the same entity. In addition, tools have been developed to intelligently determine the elution profile of a peptide in order to best pick a time to sequence the peptide. One challenge with data directed sequencing is its quite stochastic nature. Replicate experiments are highly likely to identify different peptides and proteins, because of MS/MS selection rather than biological variation. The more complex the analysed peptide mixture is, the more pronounced this effect is. Additional strategies for determining which ions to sequence include the use of an 'inclusion list', which is a set of masses that specifically trigger sequencing events. Experiments with 'inclusion lists' and 'exclusion lists' have demonstrated that intelligently targeted sequencing events can dramatically increase numbers of proteins identified.

Specialties: In proteomics, mass spectrometry can be used to analyze proteolytic peptides and also intact and in-vivo processed proteins from body fluids or the extracellular space of tissues. For example, the so-called MALDI imaging technique records MALDI-TOF spectra of tissue sections and correlates the measured m/z values with morphological features of the section. This approach has become popular in pathology departments because the mass spectrometer acts as a 'molecular microscope' that records hundreds to thousands of molecular features in sections simultaneously. The idea behind this approach is that some of these features might turn into molecular biomarkers indicative of a particular physiological state of a tissue. A similar approach can be taken to identify certain bacteria based on their MALDI-TOF signature (biotyping). Another specialty is ion mobility mass spectrometry: In this technique one measures the time it takes for ions to travel down a uniform electrostatic field (drift tube). In the field, the ions interact with neutral molecules and are separated according to parameters such as mass, size, and shape. When coupled to a mass spectrometer, isobaric peptides (e.g., structural isoforms of a glycopeptide) can often be differentiated, which adds another separation dimension to the classical mass spectrometer.

6. Peptide and protein identification by MS

For protein identification using mass spectrometry, the most common approaches all start by digesting proteins (or mixtures thereof) with a sequence-specific protease. Trypsin is most commonly used for this purpose but other sequence-specific proteases such as LysC, LysN, and GluC are also frequently employed because they produce peptides of precisely predictable mass. The resulting peptide mixtures are then subjected to mass spectrometric analysis. The information generated can be used in multiple ways to identify the underlying amino acid sequence.

Peptide mass fingerprinting: In peptide mass fingerprinting (PMF), a mass spectrometer records peptide m/z values of a protein digest from which the peptide masses can be calculated. These measured peptide masses are then searched against a database of known protein sequences of proteins for which the theoretically computed masses of peptides produced by an *in silico* digest matches to the observed peptide masses. As each measured m/z value should be indicative of a particular peptide of that protein, the basic underlying idea of protein identification by PMF is that while any single measured m/z value may match to many (actually thousands) protein sequences by chance, it becomes rapidly less likely that several (or many) measured peptide masses exist in the same protein by chance [17-19]. PMF is considered one of the fastest methods for identifying proteins but is only effective if the protein in question has actually been sequenced and is in a database. In addition, protein mixtures constitute a problem because the number of possible matches rises very quickly with the number of proteins in the mixture. Similarly, small proteins are difficult to identify by PMF because they often generate too few tryptic peptides for a unique database match. The very high mass accuracy afforded by modern mass spectrometers (low ppm) helps to deal with these issues. Together with the fact that each PMF analysis takes only a few seconds instrument time, it is a valuable protein identification approach, particularly for the analysis of the hundreds to thousands of spots present on a 2D gel.

Accurate mass and time tag: A conceptually related but alternative MS-based identification approach is 'accurate mass and time tag' (AMT) identification. In addition to peptide masses, it incorporates data on the retention times of peptides in an LC-MS experiment. Hence it is significantly more complex than other identification approaches and requires a database of previously identified peptides and their LC-MS 'coordinates'. Simply put, these approaches hypothesize that a peptide's LC-MS coordinate is conserved across experiments. This idea is quite promising in light of the very high mass accuracy data produced by today's MS instruments. Challenges include sample complexity overwhelming the resolution of the peptide inference at any one retention time and retention time drifts between LC-MS runs. The former might be improved by yet higher resolution mass spectrometers and the second by computational approaches that 'align' the different LC runs according to 'features' that are present in both experiments. Unfortunately, such computational methods are not yet widely available. More generally speaking, MS-only approaches are challenged by the presence of post-translational modifications as they can exponentially increase the size of lookup databases and thus make it challenging to confidently assign peptide identities.

Protein identification using tandem mass spectra: The information contained in spectra of fragmented peptide ions can be used in one of three strategies to identify the underlying peptide: *i)* in database searching, or spectral library searching, peptides are identified based on theoretical spectra predicted for that sequence or based on spectra from previous experiments, *ii)* in *de novo* sequencing, peptide sequences are read out directly from fragment ion spectra; and *iii)* in hybrid techniques, short stretches of the peptides are sequenced and then the rest of the spectrum is matched to existing data. Publically available tools for MS/MS based peptide identification have been comprehensively reviewed [20] and thus, we only cover the basic principles here.

In database searching, the spectrum of a peptide (m/z values and optionally intensity) is scored against theoretical fragmentation patterns constructed for peptides contained in the searched databases. The peptides queried are restricted to investigator specified criteria (e.g., proteolytic enzyme and post-translational modifications allowed). Once a spectrum is matched against the database, a list of ranked peptides (scored according to the parameters set by investigator) is returned. Discerning a true match from a false match is critical but far from trivial in proteomic data analysis. As a rule of thumb, the higher the score, the more confident the investigator can be that the peptide is a positive match (or at least that the match is not a random event). There are a number of scoring schemes used in different tools: spectral correlation functions (e.g., SEQUEST) [21], shared fragment counts and dot products (e.g., TANDEM, OMSSA, MASCOT) [22-24], empirically observed rules (e.g., Spectrum Mill), and fragmentation frequencies (e.g., PHENYX) [25]. Any of these scores can be converted into an expectation value (E value), which is the expected number of peptides with scores equal to or better than the observed score under the assumption that peptides match the experimental spectrum by random chance (e.g., OMSAA, TANDEM and MASCOT). Despite the success of database and spectral matching searches, false peptide assignments occur for a number of reasons, including the use of oversimplified scoring algorithms, presence of non-peptide contaminants, low quality spectra, simultaneous fragmentation of multiple peptide ions, presence of homologous peptides, incorrectly determined charge state or peptide mass, restricted/limited database search, sequence variants, and new peptides without entries in the database[20].

Obviously, the generation of high-confidence identifications is the goal in proteomics, but scoring is software/tool-dependent. The score distribution depends on mass spectrometer performance (e.g., spectrum quality, resolution, accuracy), the instrument settings and data acquisition methods, the quality of the sample (e.g., organismic purity), and the size (i.e., comprehensiveness) of the database. The global quality of an identification score may be improved by approaches such as target-decoy searching and use of empirical Bayes methods [26-28]. In the target-decoy method, each tandem MS spectrum in question is searched against the database of peptides in forward (target) and reverse (decoy) order (or with ‘shuffled’ sequences). Any hit in the decoy search indicates a false positive assignment. Subsequently, peptides are filtered with score cut-offs below which the number of false positive hits is above an arbitrary threshold (say 1%). This method is useful for globally ‘weeding’ out false positives and does so in a conservative fashion; it has the disadvantages of lower sensitivity (some true positives are lost) and of requiring twice the computing time. Programs such as PeptideProphet employ empirical Bayes approaches to validate peptide assignments made by database search programs. From each dataset, the software learns distributions of search scores and peptide properties among correct and incorrect peptides and uses those distributions to compute probabilities that assignments are correct.

In *de novo* sequencing, amino acids in the peptide are directly read from the fragment ion spectrum utilizing the mass differences between fragment ions that correspond to amino acid masses. This spectrum interpretation process can be facilitated by tools such as PepNovo and PEAKS [29, 30]. Direct sequencing is helpful in cases where no genome information is available, when peptides are modified in an unclear fashion, or when there are sequence polymorphisms. When limited genome information is available, *de novo* sequencing or a hybrid approach must be used (e.g., database searching against EST collections). One important issue in bottom up shotgun proteomics, and in fact in virtually all approaches utilizing mass spectrometric protein identification methods, is the so-called protein inference problem [31]. This problem arises from the fact that peptides rather than proteins are identified. Because the same peptide sequences may occur in different proteins, it is sometimes unclear which of the proteins has actually been identified (no unique peptide identified). For proteins that are represented by many sequenced peptides, this is often not a problem. But it becomes increasingly difficult to make unambiguous calls if very few (say one) peptide is available for identification or when proteins are very similar in their overall sequence. It has therefore become common (and necessary) to report identified protein groups rather than individual proteins. This problem is also relevant for protein quantification as discussed below.

PTM identification using tandem mass spectra: Although high-throughput peptide identification is quite straightforward, the identification of PTMs is not trivial. Known or suspected PTMs can be searched for by the aforementioned database searching techniques. However, the optional presence of PTMs (say phosphorylation on Y, S and T residues) greatly increases the number of possible matches and thus false assignments. In addition, the spectra of phosphorylated peptides in particular are often difficult to interpret by standard database search programs. Manual inspection of the tandem mass spectra is still considered to be the gold standard for the assignment of PTMs but this is becoming increasingly unrealistic as the sequencing speed of mass spectrometers increases. Specialized software tools like the A-Score [32] and PTM Score [33] have been developed but these are restricted to protein phosphorylation and are not readily available to most laboratories as algorithms have been published but no implementation provided. Unknown PTMs are best identified through direct *de novo* sequencing. Combinations of database searches and partial *de novo* sequencing have also been developed to identify PTMs. Particularly interesting examples are the ModifiComb [34] and InsPecT methods [35] that either systematically filter or examine tandem spectra for the presence of mass shifts indicative of a PTM.

7. Peptide or protein quantification by MS

Mass spectrometric methods are increasingly used for relative and absolute protein quantification. The many methods available for this purpose have been reviewed comprehensively elsewhere [36]. Hence, in the following section, we focus on the main principles and methods. Broadly speaking, quantitative mass spectrometry can be divided into analysis methods that incorporate stable isotope labeling and label-free methods. Stable isotope labeling takes advantage of the fact that the physicochemical properties of labeled and unlabeled versions of a molecule (protein/peptide) are identical and thus they behave identically during sample preparation (e.g., electrophoresis), sample separation (e.g., chromatography) and sample analysis (e.g., ionization, intensity). The only difference between labeled and unlabeled peptides/proteins is the mass increase provided by the label. Therefore, two (or more) experiments can be combined into the same analysis and the mass increase provided by the stable isotope labeled can be measured by a mass spectrometer. The relative signal intensities of the labeled and unlabeled forms of the analyte can be used to determine relative quantities. Stable isotopes labeling methods can be further distinguished by how and when the label is incorporated into the analysis. Labeling can be performed metabolically, chemically, or enzymatically at the level of the intact protein or at the level of proteolytic peptides. In the label-free methods, the mass spectrometric response or a derivative thereof (e.g., signal intensity, ion chromatogram, number of acquired spectra sequence coverage) is used directly to compare two (or more) separate analysis. All these methods have advantages and disadvantages.

Metabolic incorporation of labels: *In vivo* labeling takes advantage of cellular metabolism to effectively incorporate a stable isotope into proteins via the process of translation during cell growth and division. There are two approaches: the less practical global labeling of proteins by growing cells by in ^{15}N-supplemented cell culture medium [37 38] and the widely used selective labeling of amino acids. The most popular approach is the stable isotope labeling by amino acids in cell culture (SILAC) [39]. The stable isotope is incorporated by supplementing the cell growth medium with $^{13}C_6$-arginine and $^{13}C_6$-lysine (or other labeled amino acid), while growing an identical sample in a label free environment. This approach guarantees that the resulting peptides from the tryptic cleavage of a protein (excluding its C-terminus) contain no less than one labeled amino acid per peptide (heavy) with a constant mass increase as compared to the non-labeled corresponding peptides (light). Two cell populations grown in 'heavy' and 'light' conditions are pooled, lysed, and proteins are isolated, denatured, reduced and digested. The peptides are then quantified by MS. Protein identification is determined from either the 'heavy' or 'light' peptide spectrum and relative quantification is achieved by taking the ratios of the intensities of the two isotopes of the specific peptide in the MS spectrum. The advantage of SILAC over full metabolic protein labeling by ^{15}N is more straightforward data analysis due to the fact that the labels in SILAC are specifically incorporated and not peptide-sequence dependent.

Although, global ^{15}N metabolic protein labeling of higher organisms *in vivo* is feasible and has been performed for *C. elegans*, *Drosophila melanogaster*[40], rat[41], and plants [42], it is not widely used and is mostly restricted to the labeling of bacteria. On the other hand, SILAC is most widely used for metabolic labeling of higher eukaryotes. Near complete incorporation of labels typically occurs after five to 10 doubling of cells grown in SILAC media [39]. One of the main advantages of SILAC is the subsequent accuracy in quantitative MS-based methods. This is due to labeling and sample mixing prior to digestion and separation. In other methods, irreproducibility is introduced by labeling and mixing later in the process. SILAC is extremely useful in determining small variations in protein levels as well as post-translational modifications [33, 43, 44]. For the latter, it should be noted though, that quantification on the peptide level is far from trivial because all information is derived from a single or a few observations. There are challenges to the SILAC technology. For instance, certain cell lines readily form proline from excess arginine; this issue can be alleviated by supplementing medium with limited amounts of arginine [45] or supplying extra proline to the medium. Some cell lines or primary cells do not grow well in SILAC media and therefore cannot be labeled. Another limitation to the SILAC technology is that only certain isotopically labeled amino acids are available. As a consequence, in a single experiment only up to three conditions can be compared (e.g., unlabeled, $^{13}C_6$, and $^{13}C_6$ $^{15}N_4$-labeled arginine). In all proteomic experiments, sample extraction is considered to be the most critical step as any mistake made here will negatively impact the results. For metabolic labeling, quantification is actually not the last step in the experiment (Figure 3) but begins before sample extraction. This is relevant as the quality of the labeling experiment directly affects the quality of the results.

Post-biosynthetic labeling: Isotopic labeling of extracted proteins and peptides can be carried out *in vitro* either chemically or enzymatically. A stable isotope label can be incorporated into peptides enzymatically either during proteolytic digestion or in a separate step after proteolysis. Hence enzymatic labeling can be very specific and is often preferred over chemical labeling. For example, two ^{18}O isotope labels can be incorporated into the C-termini of peptides by either trypsin or Glu-C during or following protein digestion [46, 47]. This results in a 4 Da (2 Da/^{18}O) mass shift that can be utilized for discrimination of two samples. Other enzymes such as Lys-N introduce only one ^{18}O isotope [48] but this mass shift is not useful. Since ^{18}O labels are stable at low pH but can be lost at high pH values [49], this type of label is suitable for the mild acidic conditions typically utilized for ESI- and MALDI-MS. A drawback of enzymatic labeling is that the often-observed incomplete incorporation of isotopes and the differential incorporation efficiency of labels between peptides can lead to difficult data interpretation [50, 51].

Post-biosynthetic labeling at the protein level is possible and attractive as it would be compatible with proteomic experiments that rely on protein fractionation. The commercial IPCL reagent (isotope coded protein label) has been used successfully [52]. It targets the side chain of lysine residues and the N-terminus. It is very difficult to achieve full labeling and blocking lysine residues with the label renders trypsin ineffective at these sites resulting in a reduced number of peptides available for quantification. Many post-biosynthetic labeling strategies are more effective on digested peptides than on intact proteins, thus interfering with pipelines that rely heavily on intact protein fractionation. Since the isotope label should be introduced as early in a workflow as possible to minimize experimental variation, peptide labels are best utilized when the biochemical workflow is short (e.g., shotgun proteome digestion).

The chemical labeling approach mainly utilizes stable isotope–carrying chemical reagents to target reactive sites on peptides or proteins. The main targets for these reagents are the side chains of lysine and cysteine. The isotope-coded affinity tag (ICAT) was developed by Aebersold and co-workers[37] to modify cysteine residues and link them to a biotin tag by a polyether linker, which contains either eight (heavy) deuteriums or only hydrogen (light). The biotin tag is used for affinity purification and recovery of the labeled peptides. The ICAT experiment is performed on two isolated populations of proteins that are reduced and tagged with light and heavy ICAT reagents. The proteins are then pooled, digested, and the tagged peptides are recovered by affinity chromatography and quantified by MS. ICAT-generated samples are low complexity since only cysteine-containing peptides are present and cysteine is a rare amino acid. One of the shortcomings of the ICAT approach is that some proteins of interest containing only one or no cysteines. In addition, the biotin tag negatively affects the fragmentation spectra of peptides and there are elution time differences of light and heavy peptides during reversed-phase chromatography. These limitations have been overcome by recent technology advances, such as replacing the linker by a cleavable version[46, 47, 48]. A method similar to ICAT makes use of a 2-thiopyridyl disulfide group to react with cysteines, a deuterium-labeled alanine, a His$_6$-tag for affinity purification, and a tryptic cleavage site to limit the size of the tag[49]. Despite their drawbacks, ICAT and similar methods are valuable tools for a host of expansive, human plasma, or targeted analyses.

Chemical labeling can also be achieved by a group of reagents that modify the N-terminus of the peptide and the epsilon-amino group of lysine residues. The most common and specific reagents are the *N*-hydroxysuccinimide (NHS) and other active esters and acid anhydrides. This group includes isotope tags for relative and absolute quantification (iTRAQ) [53], the aforementioned isotope-coded protein label (ICPL) [52], tandem mass tags (TMT) [54], and acetic/succinic anhydride [55-58]. Less commonly used are isocyanates and isothiocyanates [59, 60], reagents for formaldehyde methylation of lysine residues, and cyanoborohydride [61-63]. Isobaric tagging of peptides[54] results in peptides that co-elute in liquid chromatography, leading to reduced variability. The different tags are then distinguished by the mass spectrometer after fragmentation occurs. For instance, in single MS mode, the same peptides with different labels are identical in mass. However, in tandem MS mode, where the peptides are fragmented, each tag generates a unique reporter ion. Protein quantification is then achieved by taking the ratio of the intensities of the reporter ions relative to each other in the MS/MS spectra. This approach allows the simultaneous determination of both identity and relative abundance of peptide pairs. The main advantage of the iTRAQ and TMT reagents is that they allow multiplexed quantification of up to eight samples at the same time, thereby reducing the amount of mass spectrometry time needed for analysis. These methods also allow monitoring of several time points or dose response measurements. Another type of chemical isotopic labeling involves esterification by deuterated alcohols of carboxylic acids in glutamic and aspartic acid and of the C-termini [64, 65]. This approach is useful for quantification studies of phosphorylated peptides, since it reduces the cross-reaction with IMAC that is often a required enrichment step [66]. β-elimination of phosphoric acid followed by Michael addition has also been used for quantification of phosphorylated peptides [67-70]. Quantitative studies of glycosylated peptides have been achieved by hydrazide chemistry, which replaces the carbohydrate with an isotopically labeled tag [71].

The absolute quantification strategy: The absolute quantification of proteins (AQUA) [72] and determination of their modification states can be accomplished by spiking the sample with modified peptides as internal standards. The synthetic peptides can contain stable isotopes to be used as internal standards similar to the endogenous proteolytic peptides. These peptides can also be synthesized with covalent attachments to mimic protein post-translational modifications such as phosphorylation, methylation, and acetylation. Data analysis is performed by comparing the signal of the synthetic peptide to the native peptide in the sample from the MS spectrum. The AQUA strategy is limited to the quantification of only a small subset of any sample, whereas other labeling techniques cover the whole sample. This strategy is nonetheless very useful if the aim of the study is focused on one or few proteins. For example, Gerber et al. [72] were able to measure the cell cycle-dependent modification of the human separase protein using this method. To alleviate the limitations of this strategy, a *de novo* gene design was developed in which artificial proteins that are concatemers of tryptic peptides (QconCAT) of one or multiple proteins are expressed [73]. In addition to the increased coverage, bias potential is reduced and accuracy is improved due to the introduction of the peptides early in the process. This strategy was applied successfully in the absolute quantification of the components of the eIF2B-eIF2 protein complex [74]. It should, however, be noted that the design and production of synthetic proteins is time consuming and expensive. Other limitations of AQUA and similar strategies are underscored by the inherit narrow dynamic detection range of present mass spectrometry, which is compounded by the complexity of the tryptic digests of entire proteomes. The amount of labeled standard required is rather difficult to determine since proteins of interest can be expressed differentially under different conditions. Also the specificity of the added standards is potentially problematic if they result in multiple isobaric peptides.

The use of a particular mass spectrometric strategy called selected or multiple reaction monitoring (SRM/MRM; discussed below) [75] [76] can alleviate both of these limitations. The SRM/MRM technique is discussed in much more detail by Domon and Aebersold in this issue, but briefly, in the SRM/MRM technique, a triple quadrupole mass spectrometer is used to assay the presence and the quantities of endogenous and AQUA peptides by focusing the first quadrupole on one particular peptide of interest followed by fragmentation of this peptide inside the second quadrupole and collection of one (SRM) or a few (MRM) particular fragment ions in the third quadrupole. One then generates a chromatogram of this peptide to fragment transition. By comparing the area under this chromatographic peak with the same experiment performed for the spiked AQUA peptide of known quantity, it is possible to calculate the amount of the endogenous peptide. It has been shown that this technique delivers quantitative information across 4-5 orders of magnitude, so one does not have to guess the amount of AQUA peptide needed. It is also very sensitive because the mass spectrometer is focused on a particular peptide of interest (other peptides present in the mixture are not detected). MRM experiments are also very fast; the mass spectrometer only needs milliseconds to record the MS signal, which allows one to measure hundreds of peptides in an LC-MS experiment. This experiment can also be very selective as the combination of peptide precursor and fragment mass are very characteristic features of peptides. Still, this selectivity is not absolute and more and more interferences are encountered as the complexity of a sample increases. Therefore, several transitions per peptide are usually measured. Although absolute quantification (with or without the use of the MRM technique) is attractive, some principle issues remain. Despite the ability to calculate protein amounts (in mol or gram) from an AQUA experiment, there are still questions as to how accurate these values are as any sample manipulation prior to adding the synthetic standard may bias the results (due to losses or enrichment). Consequently, the amount of a protein in an experiment determined by AQUA may not reflect the true expression levels of this protein in a cell and, quite often, the result of an AQUA experiment is relative rather than absolute quantification.

Label-free quantification: Proteomic quantification can also be achieved without artificially labeling parts of the sample. There are two very different approaches to accomplish such label-free quantification: extraction of peptide ion intensities [77-81] and spectral counting [82, 83]. The first approach is similar to HPLC quantification methods and is based on comparing integrated areas under the curve of extracted peptide ion intensity chromatograms [84]. Comparison of the same peptide signal between two (or more) experiments is a measure for the relative quantities of that peptide in the two experiments. The accuracy of this method is quite good but limited by the mass accuracy of the mass spectrometer and the reproducibility of the chromatography. To achieve high accuracy, one should minimize the MS signal overlap by utilizing a high mass accuracy spectrometer. Also, LC alignment software can optimize the chromatographic profiles of peptides [85-88], in turn enhancing reproducibility. These types of experiments are attractive because they (in principle) enable comparison of data from many experiments but in practice require an immense amount of data acquisition and analysis time. Therefore, a compromise has to be made between depth of identification and quality of quantification. As a consequence, better quantification accuracy is achieved at the expense of coverage and vice versa.

A very different alternative to using the direct signal response of the mass spectrometer is the spectral counting approach, which depends on high-throughput data acquisition for both identification and quantification. The spectral count approach is relatively new and relates the number of mass spectra identified for a protein to the protein's abundance [89-91]. The simple rational is that the more of a protein there is in a sample, the higher the number of tandem mass spectra the mass spectrometer will acquire for this protein (similar to the next generation RNA sequencing methods). Therefore, a direct comparison of two or more (similar) runs will allow the relative quantification of the protein of interest. The minimum number of spectral counts required to see a significant change was determined by Old et al. [92]; they observed that the relationship between protein quantity and spectrum counts is not linear, but rather exponential, and is different for every protein. They concluded that four spectra were sufficient to see three-fold protein changes, but up to fifteen spectra were needed to observe a two-fold change. They also showed that the spectral counting method yields reliable results as compared with extraction of peptide ion intensities, but both methods are less sensitive than isotopic labeling [93]. To achieve better accuracy and more reliable quantification, an exponentially modified protein abundance index (emPAI) [94] is utilized, which is proportional to concentration of proteins in a sample. In addition, further improvements in quantification are achieved through use of computational tools (APEX, Sieve) that select peptides in advance for detection by the mass spectrometer [95-98] and as such maximize the reproducibility of peptide detection between experiments. Although this approach simultaneously identifies and quantifies proteins and is conceptually simple to perform, it suffers a major drawback. Quantification relies on generating a very high number of MS/MS peptide identifications and is at the same time greatly dependent on the quality of these identifications, since errors in peptide identification will lead to inaccurate protein quantification [99, 100]. Although both label-free methods have their advantages and can be applied for global quantification studies, both require extensive MS and computational resources. Hence, only a handful of labs are currently able to take full advantage of these methods.

Informatics for quantitative proteomics: Software tools are required for quantitative proteomic measurements in order as thousands of data points are generated in any one experiment. Several commercial systems are available for quantification, often as part of data analysis packages that come with a particular mass spectrometer. Other commercial tools such as Progenesis, ScaffoldQ, Mascot, and Rosetta work independently of instrument platform or data format and thus are of interest for laboratories operating a mixed vendor proteomics platform. Numerous free tools have been developed, including MaxQuant, XPRESS, MSQuant, Census, i-tracker, Quant, and Skyline. It is beyond the scope of this review to discuss the features of any of these programs but in one shape or form, they all use peak detection algorithms to extract quantitative information (intensity, area) from peptide (intact or tandem) mass spectra. The one exception to this rule is the spectral counting approach that does not use MS intensity information for quantification (see above). In a distinct step, this quantitative information is merged with peptide identification information typically generated by tandem mass spectrometry. In another step, intensity values of particular peptides from, say, experiment vs. control are compared in order to learn if there were any differences. As in the case of peptide/protein identification, the quantitative information is initially generated at the peptide level and has to be transformed into protein quantification. There are several important considerations at this stage: *i)* Reliable protein quantification can only be achieved by using peptides that are unique to a particular protein (the protein inference problem discussed earlier). *ii)* Not all peptide spectra that correspond to the same protein are of the same quality. Therefore, simple averaging of quantification data points is often not useful (e.g., an intensity weighted average or median might be more appropriate). *iii)* Finally, quantification of PTMs must be treated with extra care as these might either change in parallel with the protein (indicating no particular biological significance of the PTM) or show a differential behavior without apparent changes in protein expression (often observed for signaling pathways). Thus, PTMs must be evaluated on a case by case basis. Many of the methods discussed here and shown in Figure 3 have been put into data processing pipelines that can be managed by informatics tools such as the transproteomic pipeline (TPP), the openMS proteomics pipeline (TOPP), the Platform for Experimental Proteomic Pattern Recognition (PEPPeR), ProteoWizard, and the Computational Proteomics Analysis System (CPAS). All these are quite useful but require significant help from the local systems administration team for installation and configuration.

An area currently underdeveloped in proteomics is how one stores and mines the end results (protein IDs and quantitative experiments) of the hundreds to thousands of experiments generated via the course of one or many projects. There are some public resources that are being developed for this purpose such as PRIDE [101], the global proteome machine (GPS) [102], and Tranche (https://trancheproject.org/). Currently, these are mainly useful for bioinformaticians but will hopefully become more broadly useful for the proteomic community in the future.

References

1. Negm, R.S., Verma, M. & Srivastava, S. The promise of biomarkers in cancer screening and detection. *Trends Mol Med* **8**, 288-293 (2002).
2. Mallick, P. et al. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol* **25**, 125-131 (2007).
3. Svensson, M. et al. Heat stabilization of the tissue proteome: a new technology for improved proteomics. *J Proteome Res* **8**, 974-981 (2009).
4. Emmert-Buck, M.R. et al. Laser capture microdissection. *Science* **274**, 998-1001 (1996).
5. Weber, G. & Wildgruber, R. Free-flow electrophoresis system for proteomics applications. *Methods Mol Biol* **384**, 703-716 (2008).
6. Anderson, N.L. & Anderson, N.G. The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* **1**, 845-867 (2002).
7. Thulasiraman, V. et al. Reduction of the concentration difference of proteins in biological liquids using a library of combinatorial ligands. *Electrophoresis* **26**, 3561-3571 (2005).
8. Hubner, N.C., Ren, S. & Mann, M. Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis. *Proteomics* **8**, 4862-4872 (2008).
9. Porath, J. Immobilized metal ion affinity chromatography. *Protein Expr Purif* **3**, 263-281 (1992).
10. Ficarro, S.B. et al. Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat Biotechnol* **20**, 301-305 (2002).
11. McNulty, D.E. & Annan, R.S. Hydrophilic interaction chromatography reduces the complexity of the phosphoproteome and improves global phosphopeptide isolation and detection. *Mol Cell Proteomics* **7**, 971-980 (2008).
12. Pinkse, M.W., Uitto, P.M., Hilhorst, M.J., Ooms, B. & Heck, A.J. Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns. *Anal Chem* **76**, 3935-3943 (2004).
13. Rush, J. et al. Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat Biotechnol* **23**, 94-101 (2005).
14. Karlsson, N.G. et al. Negative ion graphitised carbon nano-liquid chromatography/mass spectrometry increases sensitivity for glycoprotein oligosaccharide analysis. *Rapid Commun Mass Spectrom* **18**, 2282-2292 (2004).

15. Chalkley, R.J., Thalhammer, A., Schoepfer, R. & Burlingame, A.L. Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proc Natl Acad Sci U S A* **106**, 8894-8899 (2009).
16. Gevaert, K. et al. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat Biotechnol* **21**, 566-569 (2003).
17. James, P., Quadroni, M., Carafoli, E. & Gonnet, G. Protein identification by mass profile fingerprinting. *Biochem Biophys Res Commun* **195**, 58-64 (1993).
18. Mann, M., Hojrup, P. & Roepstorff, P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol Mass Spectrom* **22**, 338-345 (1993).
19. Pappin, D.J., Hojrup, P. & Bleasby, A.J. Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol* **3**, 327-332 (1993).
20. Nesvizhskii, A.I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* **4**, 787-797 (2007).
21. Eng, J.K., McCormack, A.L. & Yates, J.R. An approach to correlate tandem mass spectral data of peptides with amino-acid-sequence in a protein database. *J Am Soc Mass Spectrom* **5**, 976-989 (1994).
22. Craig, R. & Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466-1467 (2004).
23. Geer, L.Y. et al. Open mass spectrometry search algorithm. *Journal of proteome research* **3**, 958-964 (2004).
24. 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).
25. Colinge, J., Masselot, A., Giron, M., Dessingy, T. & Magnin, J. OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* **3**, 1454-1463 (2003).
26. 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).
27. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **74**, 5383-5392 (2002).
28. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440-9445 (2003).
29. Frank, A. & Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* **77**, 964-973 (2005).
30. Ma, B. et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* **17**, 2337-2342 (2003).
31. Nesvizhskii, A.I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **4**, 1419-1440 (2005).
32. Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J. & Gygi, S.P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* **24**, 1285-1292 (2006).
33. Olsen, J.V. et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635-648 (2006).
34. Savitski, M.M., Nielsen, M.L. & Zubarev, R.A. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of

modifications, and fingerprinting complex protein mixtures. *Mol Cell Proteomics* **5**, 935-948 (2006).

35. Tanner, S. et al. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* **77**, 4626-4639 (2005).

36. 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).

37. Gygi, S.P. et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**, 994-999 (1999).

38. Oda, Y., Huang, K., Cross, F.R., Cowburn, D. & Chait, B.T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci U S A* **96**, 6591-6596 (1999).

39. 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).

40. Krijgsveld, J. et al. Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics. *Nat Biotechnol* **21**, 927-931 (2003).

41. Wu, C.C., MacCoss, M.J., Howell, K.E., Matthews, D.E. & Yates, J.R., 3rd Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis. *Anal Chem* **76**, 4951-4959 (2004).

42. Gruhler, A., Schulze, W.X., Matthiesen, R., Mann, M. & Jensen, O.N. Stable isotope labeling of Arabidopsis thaliana cells and quantitative proteomics by mass spectrometry. *Mol Cell Proteomics* **4**, 1697-1709 (2005).

43. 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).

44. Park, K.S., Mohapatra, D.P., Misonou, H. & Trimmer, J.S. Graded regulation of the Kv2.1 potassium channel by variable phosphorylation. *Science* **313**, 976-979 (2006).

45. Ong, S.E., Kratchmarova, I. & Mann, M. Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC). *J Proteome Res* **2**, 173-181 (2003).

46. Reynolds, K.J., Yao, X. & Fenselau, C. Proteolytic 18O labeling for comparative proteomics: evaluation of endoprotease Glu-C as the catalytic agent. *J Proteome Res* **1**, 27-33 (2002).

47. Yao, X., Freas, A., Ramirez, J., Demirev, P.A. & Fenselau, C. Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem* **73**, 2836-2842 (2001).

48. Rao, K.C., Carruth, R.T. & Miyagi, M. Proteolytic 18O labeling by peptidyl-Lys metalloendopeptidase for comparative proteomics. *J Proteome Res* **4**, 507-514 (2005).

49. Schnolzer, M., Jedrzejewski, P. & Lehmann, W.D. Protease-catalyzed incorporation of 18O into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry. *Electrophoresis* **17**, 945-953 (1996).

50. Johnson, K.L. & Muddiman, D.C. A method for calculating 16O/18O peptide ion ratios for the relative quantification of proteomes. *J Am Soc Mass Spectrom* **15**, 437-445 (2004).

51. Ramos-Fernandez, A., Lopez-Ferrer, D. & Vazquez, J. Improved method for differential expression proteomics using trypsin-catalyzed 18O labeling with a correction for labeling efficiency. *Mol Cell Proteomics* **6**, 1274-1286 (2007).

52. Schmidt, A., Kellermann, J. & Lottspeich, F. A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics* **5**, 4-15 (2005).

53. Ross, P.L. et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **3**, 1154-1169 (2004).
54. Thompson, A. et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**, 1895-1904 (2003).
55. Che, F.Y. & Fricker, L.D. Quantitation of neuropeptides in Cpe(fat)/Cpe(fat) mice using differential isotopic tags and mass spectrometry. *Anal Chem* **74**, 3190-3198 (2002).
56. Glocker, M.O., Borchers, C., Fiedler, W., Suckau, D. & Przybylski, M. Molecular characterization of surface topology in protein tertiary structures by amino-acylation and mass spectrometric peptide mapping. *Bioconjug Chem* **5**, 583-590 (1994).
57. Ji, J. et al. Strategy for qualitative and quantitative analysis in proteomics based on signature peptides. *J Chromatogr B Biomed Sci Appl* **745**, 197-210 (2000).
58. Zhang, X., Jin, Q.K., Carr, S.A. & Annan, R.S. N-Terminal peptide labeling strategy for incorporation of isotopic tags: a method for the determination of site-specific absolute phosphorylation stoichiometry. *Rapid Commun Mass Spectrom* **16**, 2325-2332 (2002).
59. Lee, Y.H., Han, H., Chang, S.B. & Lee, S.W. Isotope-coded N-terminal sulfonation of peptides allows quantitative proteomic analysis with increased de novo peptide sequencing capability. *Rapid Commun Mass Spectrom* **18**, 3019-3027 (2004).
60. Mason, D.E. & Liebler, D.C. Quantitative analysis of modified proteins by LC-MS/MS of peptides labeled with phenyl isocyanate. *J Proteome Res* **2**, 265-272 (2003).
61. Hsu, J.L., Huang, S.Y. & Chen, S.H. Dimethyl multiplexed labeling combined with microcolumn separation and MS analysis for time course study in proteomics. *Electrophoresis* **27**, 3652-3660 (2006).
62. Hsu, J.L., Huang, S.Y., Chow, N.H. & Chen, S.H. Stable-isotope dimethyl labeling for quantitative proteomics. *Anal Chem* **75**, 6843-6852 (2003).
63. Ji, C., Guo, N. & Li, L. Differential dimethyl labeling of N-termini of peptides after guanidination for proteome analysis. *J Proteome Res* **4**, 2099-2108 (2005).
64. Goodlett, D.R. et al. Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation. *Rapid Commun Mass Spectrom* **15**, 1214-1221 (2001).
65. Syka, J.E. et al. Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J Proteome Res* **3**, 621-626 (2004).
66. Salomon, A.R. et al. Profiling of tyrosine phosphorylation pathways in human cells using mass spectrometry. *Proc Natl Acad Sci U S A* **100**, 443-448 (2003).
67. Goshe, M.B. et al. Phosphoprotein isotope-coded affinity tag approach for isolating and quantitating phosphopeptides in proteome-wide analyses. *Anal Chem* **73**, 2578-2586 (2001).
68. Goshe, M.B. et al. Phosphoprotein isotope-coded affinity tags: application to the enrichment and identification of low-abundance phosphoproteins. *Anal Chem* **74**, 607-616 (2002).
69. Qian, W.J. et al. Phosphoprotein isotope-coded solid-phase tag approach for enrichment and quantitative analysis of phosphopeptides from complex mixtures. *Anal Chem* **75**, 5441-5450 (2003).
70. Tao, W.A. et al. Quantitative phosphoproteome analysis using a dendrimer conjugation chemistry and tandem mass spectrometry. *Nat Methods* **2**, 591-598 (2005).

71. Zhang, H., Li, X.J., Martin, D.B. & Aebersold, R. Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat Biotechnol* **21**, 660-666 (2003).
72. Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W. & Gygi, S.P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* **100**, 6940-6945 (2003).
73. Beynon, R.J., Doherty, M.K., Pratt, J.M. & Gaskell, S.J. Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat Methods* **2**, 587-589 (2005).
74. Kito, K., Ota, K., Fujita, T. & Ito, T. A synthetic protein approach toward accurate mass spectrometric quantification of component stoichiometry of multiprotein complexes. *J Proteome Res* **6**, 792-800 (2007).
75. Kirkpatrick, D.S., Gerber, S.A. & Gygi, S.P. The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods* **35**, 265-273 (2005).
76. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol* **4**, 222 (2008).
77. Bondarenko, P.V., Chelius, D. & Shaler, T.A. Identification and Relative Quantitation of Protein Mixtures by Enzymatic Digestion Followed by Capillary Reversed-Phase Liquid Chromatography&#x2212;Tandem Mass Spectrometry. *Analytical Chemistry* **74**, 4741-4749 (2002).
78. Chelius, D. & Bondarenko, P.V. Quantitative Profiling of Proteins in Complex Mixtures Using Liquid Chromatography and Mass Spectrometry. *Journal of Proteome Research* **1**, 317-323 (2002).
79. Chelius, D., Zhang, T., Wang, G. & Shen, R.-F. Global Protein Identification and Quantification Technology Using Two-Dimensional Liquid Chromatography Nanospray Mass Spectrometry. *Analytical Chemistry* **75**, 6658-6665 (2003).
80. Wang, W. et al. Quantification of Proteins and Metabolites by Mass Spectrometry without Isotopic Labeling or Spiked Standards. *Analytical Chemistry* **75**, 4818-4826 (2003).
81. Li, X.-j., Yi, E.C., Kemp, C.J., Zhang, H. & Aebersold, R. A Software Suite for the Generation and Comparison of Peptide Arrays from Sets of Data Collected by Liquid Chromatography-Mass Spectrometry. *Mol Cell Proteomics* **4**, 1328-1340 (2005).
82. Gao, J., Opiteck, G.J., Friedrichs, M.S., Dongre, A.R. & Hefta, S.A. Changes in the Protein Expression of Yeast as a Function of Carbon Source. *Journal of Proteome Research* **2**, 643-649 (2003).
83. Liu, H., Sadygov, R.G. & Yates, J.R. A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Analytical Chemistry* **76**, 4193-4201 (2004).
84. Higgs, R.E., Knierman, M.D., Gelfanova, V., Butler, J.P. & Hale, J.E. Comprehensive Label-Free Method for the Relative Quantification of Proteins from Biological Samples. *Journal of Proteome Research* **4**, 1442-1450 (2005).
85. Strittmatter, E.F., Ferguson, P.L., Tang, K. & Smith, R.D. Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry. *J Am Soc Mass Spectrom* **14**, 980-991 (2003).
86. Wang, P. et al. A statistical method for chromatographic alignment of LC-MS data. *Biostat* **8**, 357-367 (2007).
87. Jaitly, N. et al. Robust Algorithm for Alignment of Liquid Chromatography&#x2212;Mass Spectrometry Analyses in an Accurate Mass and Time Tag Data Analysis Pipeline. *Analytical Chemistry* **78**, 7397-7409 (2006).

88. Bylund, D., Danielsson, R., Malmquist, G. & Markides, K.E. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *Journal of Chromatography A* **961**, 237-244 (2002).
89. Gilchrist, A. et al. Quantitative proteomics analysis of the secretory pathway. *Cell* **127**, 1265-1281 (2006).
90. Liu, H., Sadygov, R.G. & Yates, J.R., 3rd A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76**, 4193-4201 (2004).
91. Gao, J., Opiteck, G.J., Friedrichs, M.S., Dongre, A.R. & Hefta, S.A. Changes in the protein expression of yeast as a function of carbon source. *J Proteome Res* **2**, 643-649 (2003).
92. Old, W.M. et al. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics* **4**, 1487-1502 (2005).
93. Old, W.M. et al. Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics. *Mol Cell Proteomics* **4**, 1487-1502 (2005).
94. Ishihama, Y. et al. Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. *Mol Cell Proteomics* **4**, 1265-1272 (2005).
95. Robertson Craig, J.P.C., Ronald C. Beavis, The use of proteotypic peptide libraries for protein identification. *Rapid Communications in Mass Spectrometry* **19**, 1844-1850 (2005).
96. Tang, H. et al. A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics* **22**, e481-488 (2006).
97. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E.M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotech* **25**, 117-124 (2007).
98. Mallick, P. et al. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotech* **25**, 125-131 (2007).
99. Olsen, J.V. et al. Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. *Cell* **127**, 635-648 (2006).
100. Li, J., Steen, H. & Gygi, S.P. Protein Profiling with Cleavable Isotope-coded Affinity Tag (cICAT) Reagents: The Yeast Salinity Stress Response. *Mol Cell Proteomics* **2**, 1198-1204 (2003).
101. Vizcaino, J.A. et al. A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* **9**, 4276-4283 (2009).
102. Craig, R., Cortens, J.P. & Beavis, R.C. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* **3**, 1234-1242 (2004).