

# Proteomics: a pragmatic perspective

Parag Mallick<sup>1,2</sup> & Bernhard Kuster<sup>3,4</sup>

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<sup>1</sup> 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<sup>2,3</sup> with better than femtomole sensitivity ( $10^{-15}$  mol, or subnanogram)<sup>4</sup> 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<sup>5</sup>. 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.

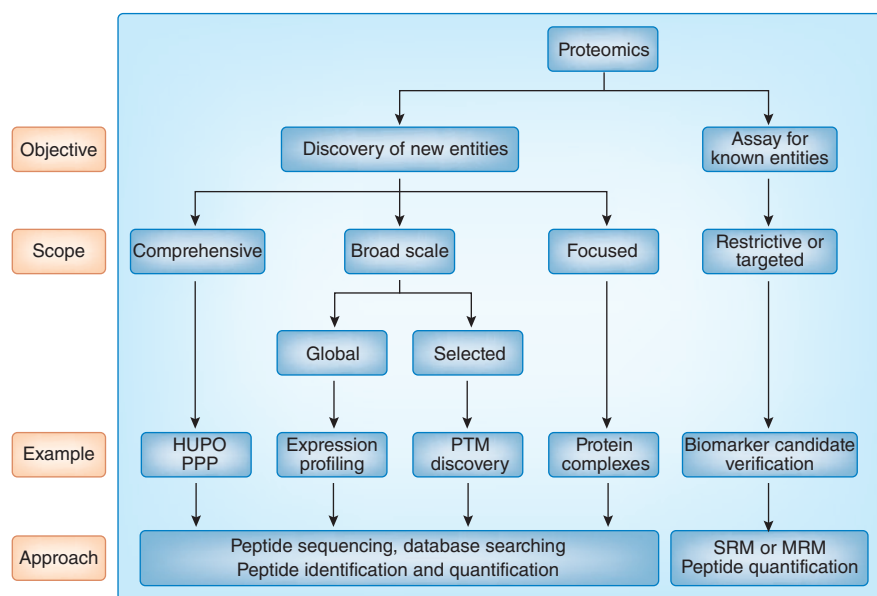
## Getting organized

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<sup>6</sup>. 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

<sup>1</sup>University of Southern California Center for Applied Molecular Medicine, Departments of Medicine and Biomedical Engineering, Los Angeles, California, USA. <sup>2</sup>Department of Chemistry & Biochemistry, University of California, Los Angeles, Los Angeles, California, USA. <sup>3</sup>Chair of Proteomics and Bioanalytics, Technische Universität München, Freising-Weihenstephan, Germany. <sup>4</sup>Center for Integrated Protein Science Munich, Munich, Germany. Correspondence should be addressed to P.M. (mallick@usc.edu) or B.K. (kuster@zww.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<sup>7</sup>. 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<sup>8,9</sup>. The identification of several thousand proteins or phospho-peptides<sup>10</sup> 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<sup>11,12</sup>.

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<sup>13,14</sup>. 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<sup>15</sup>. Even so, robust protocols are as yet available for relatively few low molecular weight PTMs, such as phosphorylation, acetylation and methylation<sup>16</sup>. 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<sup>17</sup> and glycosylation<sup>18</sup> 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<sup>19–21</sup>. 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<sup>22</sup>). Such standards have so far been generated for only very few PTMs (notably phosphorylation<sup>23</sup>) 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<sup>24</sup>. 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<sup>25</sup>. 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<sup>26</sup>, 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<sup>27</sup>.

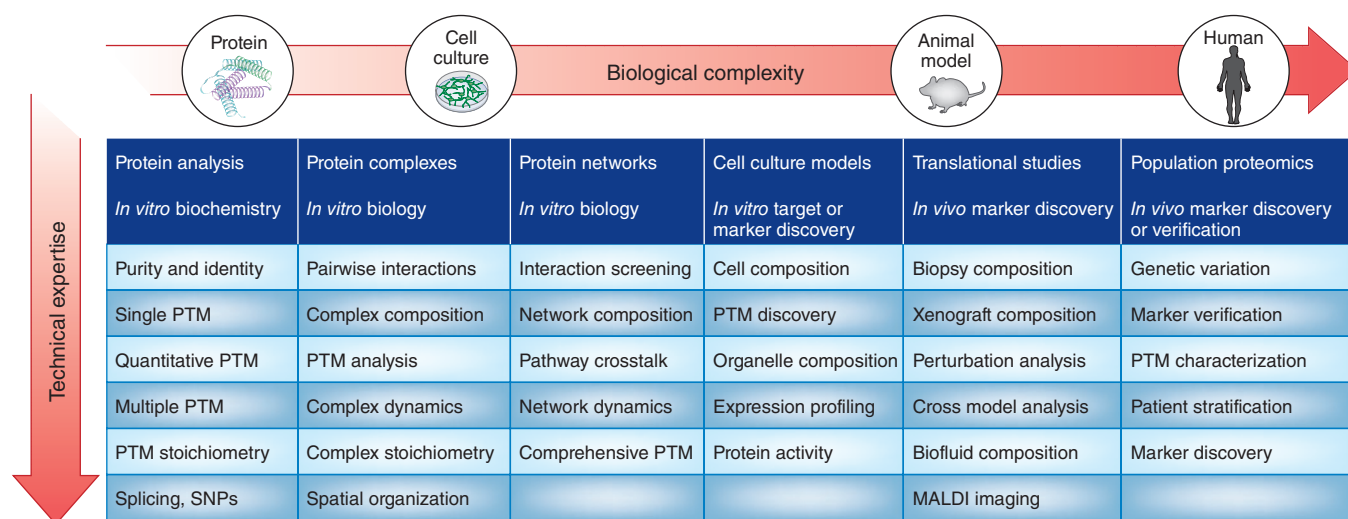
MS-based peptide sequencing can also be used to detect proteins resulting from splice variants and single-nucleotide polymorphisms<sup>28</sup>. This type of study has rarely been done systematically owing to the requirement for 100% sequence coverage and the difficulty of detection of low-abundance isoforms. With the advent of next-generation DNA sequencing techniques<sup>29</sup>, 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<sup>30</sup>. 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<sup>32</sup>; **Supplementary Glossary**), have been perfected, proteomics has provided a powerful means to distinguish true interactors from abundant contaminants<sup>33</sup>.

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<sup>34,35</sup>. 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



**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<sup>36,37</sup>. 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<sup>38,39</sup> 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<sup>40</sup>, 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<sup>41</sup>. 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<sup>42</sup>. 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<sup>43</sup>; 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<sup>44,45</sup>. 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<sup>46–48</sup>. 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<sup>49,50</sup>.

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<sup>51</sup>, 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<sup>52,53</sup>.

### 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<sup>54–56</sup>. In fact, two reports on large-scale protein interaction screens in yeast are among the five most highly cited papers in proteomics so far<sup>51,57</sup>. 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- $\alpha$  (TNF- $\alpha$ )-nuclear factor- $\kappa$ B (NF- $\kappa$ B) pathway was mapped in human embryonic kidney (HEK293) cells using 32 TAP-tagged proteins<sup>11</sup>. 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<sup>58</sup>. 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: <ol style="list-style-type: none"> <li>1. 10s to 100s of proteins from nanograms of total protein</li> <li>2. 100s to 1,000s of proteins from micrograms of total protein</li> <li>3. 1,000 to 10,000 proteins from milligrams of total protein</li> </ol> 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 ( <b>Supplementary Glossary</b> ) 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 ( <b>Supplementary Glossary</b> ). 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: <ol style="list-style-type: none"> <li>1. Calculation of a global false discovery rate (FDR) for the list of identified proteins. FDRs of &lt;1% are generally accepted. FDRs give information about the general quality of a data set. Most protein identification software packages provide FDR calculation tools.</li> <li>2. Calculation of the probability that matching a tandem MS spectrum to a peptide sequence is a random event. Random matches of &lt;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.</li> <li>3. For publication in some journals, at least two peptide identifications are required. This is an <i>ad hoc</i> criterion and says very little about data quality.</li> </ol>
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: <ol style="list-style-type: none"> <li>1. Metabolic labeling (for example, SILAC <sup>15</sup>N) is best for small changes (10–50%) and cell culture systems.</li> <li>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).</li> <li>3. Label-free methods using the MS detector response (for example, extracted ion chromatograms (<b>Supplementary Glossary</b>)) are best for moderate changes (20%–200%) and for comparison of many highly similar experiments.</li> <li>4. Label-free methods using spectrum counts are best for large changes (&gt;100%) and for comparison of many highly similar experiments.</li> <li>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).</li> </ol>
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 times from sample submission to data reporting are typical for core facilities and research labs: <ol style="list-style-type: none"> <li>1. 5–10 working days for simple protein identification and quantification</li> <li>2. 4–6 weeks for quantitative protein expression profiling</li> <li>3. 2–6 months for PTM analysis</li> </ol>
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: <ol style="list-style-type: none"> <li>1. \$50–200 for simple protein identification</li> <li>2. \$500–2,000 for simple PTM analysis</li> <li>3. \$5,000–15,000 for complex PTM analysis</li> <li>4. \$1,000–2,000 for quantitative protein expression profiling</li> </ol>

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)<sup>59,60</sup>. 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<sup>61–64</sup>. 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<sup>65</sup>. 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<sup>66</sup>.

### Cell culture models

*In vitro*, prokaryotic<sup>67</sup>, and eukaryotic<sup>68,69</sup> 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<sup>70</sup>. 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<sup>71</sup>.

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<sup>72,73</sup>, a tenfold increase over what was possible only five years ago<sup>74</sup>.

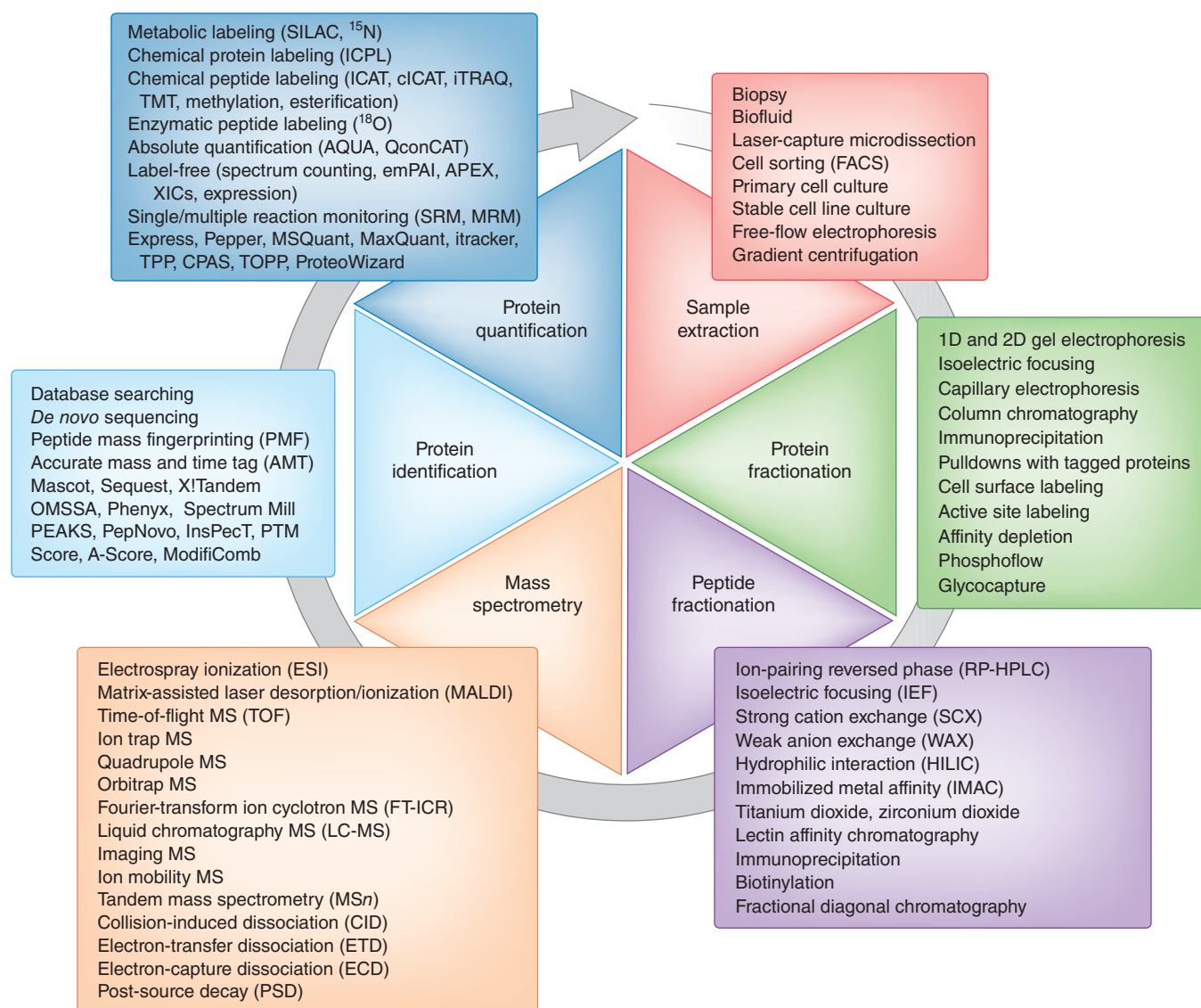
In addition to proteome coverage, another key consideration in profiling experiment analyses is protein coverage. At first blush, it would seem that if a protein were present in a sample, all of its peptides should be readily observable. For several reasons that we have discussed elsewhere<sup>75,76</sup>, 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<sup>77</sup>. A marked improvement can also be obtained by using multiple proteases with different cleavage specificities<sup>78</sup>.

### 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<sup>10,79–83</sup>. 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<sup>84,85</sup>. 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.





**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<sup>86,87</sup> 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<sup>41,88</sup>. 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<sup>89–91</sup>.

### *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<sup>92</sup>: 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<sup>93</sup>, 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<sup>94</sup>, and using an intensity threshold can help to minimize the noise-based bias<sup>95</sup>. 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<sup>94,96</sup> or when using linear regression analysis<sup>97</sup>.

An alternative label-free quantification technique compares the mass spectrometric intensity of each peptide in each of the experiments<sup>98</sup>. 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<sup>99–106</sup>, 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 <sup>13</sup>C, <sup>15</sup>N, or <sup>18</sup>O labeling<sup>108</sup>, but deuterium labeling can be problematic because labeled and unlabeled peptides often differ in their retention time in reversed-phase high-performance liquid chromatography<sup>109</sup>. 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<sup>110</sup>. 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<sup>111</sup>.

### ***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<sup>112</sup> but has since been applied to other enzyme classes, such as kinases, phosphatases and histone deacetylases<sup>113,114</sup>. 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-

**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.

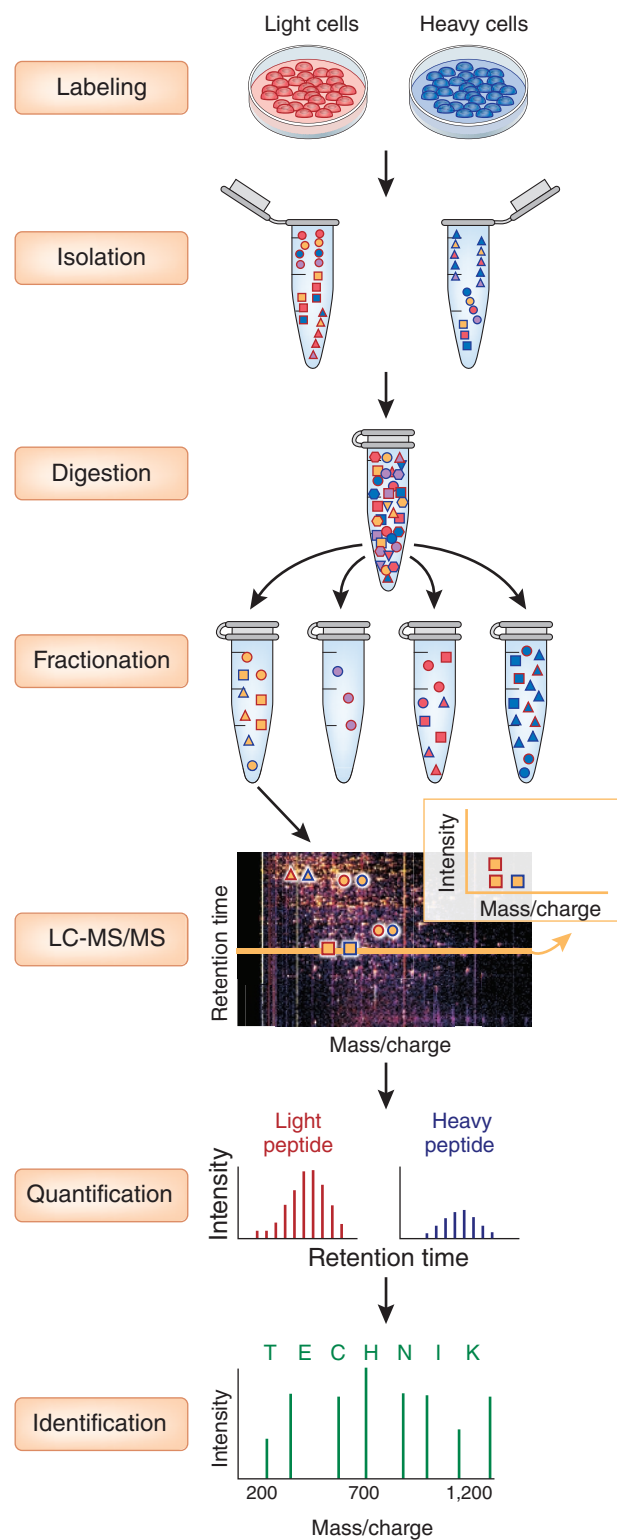
underexpressed in healthy versus pathological conditions. Activity profiling enables the identification of the targets of small molecule drugs in a proteome-wide fashion. We envisage activity proteomics playing a significant role in drug discovery as it becomes possible to profile the selectivity of 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<sup>85</sup>. 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<sup>115</sup>. 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<sup>116</sup> 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<sup>117–119</sup>. However, the low amounts of material available from these techniques or the presence of fixation or crosslinking reagents can markedly limit the desired analysis<sup>120</sup>.

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  $\mu$ l of blood; after depletion, one is left with less than 10  $\mu$ g 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<sup>121,122</sup>.

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<sup>111,123</sup>. 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 (non-specific)<sup>124</sup> 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<sup>125</sup>. Consequently, some inter-sample differences in protein abundance may result from the depletion procedure itself<sup>126</sup>. 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<sup>91</sup>. Recently, immunoprecipitation with anti-peptide antibodies and MS have been used to quantify troponin I and interleukin-33 in serum<sup>127,128</sup>. 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<sup>129–134</sup>. 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)<sup>135,136</sup>. 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)<sup>137,138</sup>; 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<sup>139</sup>.

### 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<sup>7,140–142</sup> and INSPECT<sup>143–146</sup>, 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<sup>6,128,147</sup>, 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<sup>148,149</sup>.

## 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 HUPPO 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- $\kappa$ B 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. Maya, V., Rezuall, 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 transcatheter 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<sub>2</sub>-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. Wollschlaedl, 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.* SuperHim - 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).