JOURNAL OF PROTEOMICS 75 (2012) 3495–3513

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot

Review

Targeted proteome investigation via selected reaction monitoring mass spectrometry☆

Alessio Maiolica[a], *Martin A. Jünger*[a], *Iakes Ezkurdia*[b], *Ruedi Aebersold*[a,c,*]

[a]*Department of Biology, Institute of Molecular Systems Biology, Zurich, Switzerland*
[b]*Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre-(CNIO), Madrid, Spain*
[c]*Faculty of Science, University of Zurich, Zurich, Switzerland*

ARTICLE INFO

Article history:
Received 21 November 2011
Accepted 29 April 2012
Available online 8 May 2012

Keywords:
Selected Reaction Monitoring (SRM)
Proteomics
Antibodies
Mass spectrometry
Proteome mapping

ABSTRACT

Due to the enormous complexity of proteomes which constitute the entirety of protein species expressed by a certain cell or tissue, proteome-wide studies performed in discovery mode are still limited in their ability to reproducibly identify and quantify all proteins present in complex biological samples. Therefore, the targeted analysis of informative subsets of the proteome has been beneficial to generate reproducible data sets across multiple samples. Here we review the repertoire of antibody- and mass spectrometry (MS) -based analytical tools which is currently available for the directed analysis of predefined sets of proteins. The topics of emphasis for this review are Selected Reaction Monitoring (SRM) mass spectrometry, emerging tools to control error rates in targeted proteomic experiments, and some representative examples of applications. The ability to cost- and time-efficiently generate specific and quantitative assays for large numbers of proteins and posttranslational modifications has the potential to greatly expand the range of targeted proteomic coverage in biological studies. This article is part of a Special Section entitled: Understanding genome regulation and genetic diversity by mass spectrometry.

Contents

☆ This article is part of a Special Section entitled: Understanding genome regulation and genetic diversity by mass spectrometry.

* *Corresponding author at:* Institute of Molecular Systems Biology, ETH Zurich, Wolfgang-Pauli-Str. 16, 8093 Zurich, Switzerland. Tel.: +41 44 633 1071; fax: +41 44 633 1051.
E-mail address: aebersold@imsb.biol.ethz.ch (R. Aebersold).

1874-3919/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jprot.2012.04.048

1. The complexity of proteomes: the boundary condition for proteomics

In the 1990's, the cellular proteome was defined as the entirety of proteins expressed by a particular cell under specific conditions [1]. Although this is a generally accepted concept, there are still numerous knowledge gaps regarding the actual composition of proteomes. The number of protein species constituting a prokaryotic or eukaryotic proteome is dependent on at least three different cellular mechanisms: i) regulation of gene expression, ii) post-transcriptional events, iii) post-translational modifications of proteins. Determining the composition of cellular proteomes requires a comprehensive knowledge of those three mechanisms, without which attempts to define any proteome *a priori* will fail.

The gene-centric view expresses the simplest layer of proteomic complexity. It depends on experimental evidence that a certain open reading frame (ORF) predicted from the genome is indeed transcribed and translated into a protein. However, accurate gene identification in eukaryotic genome sequences is technically challenging, no single method has yet been able to exhaustively achieve it, and the number and type of gene prediction algorithms is still evolving, as is the annotation of the genomes [2–4]. For multicellular eukaryotic species, the number of proteins constituting the proteome is larger than the number of predicted ORFs because many multi-exon genes are able to produce at least two differently spliced mRNA transcripts by alternative splicing of pre-messenger RNA [5,6]. Exon rearrangement from primary transcripts [7] has the capacity to expand the cellular protein products with altered structure and biological functions [8–10]. In humans, at least 50,000 splice variants have been known to be transcribed, and this number is certain to grow as more advanced analytical methods become available. Further layers of complexity consist of single nucleotide polymorphisms (SNPs) that can lead to an amino acid mutation in the resulting protein. SNPs are the most common type of genetic variation among populations of individuals. Non-synonymous SNPs (nsSNPs) in protein-coding regions can explain half of the known genetic variations linked to human hereditary diseases [11] and can also influence post-translational modifications of proteins (PTMs) such as phosphorylation [12]. Large-scale polymorphism surveys have been recently published for *Saccharomyces cerevisiae* [13], *Arabidopsis thaliana* [14], and *Mus musculus* [15]. Initial results of the 1000 Genomes Project, an international project aiming to characterize human variation (1000 Genomes Project Consortium, 2010 [16]) have identified 15 million SNPs, 55% of which are novel, over 1 million indels and more than 20,000 structural variants.

An additional level of proteome complexity is provided by the post-translational modification (PTM) of proteins. PTMs are enzymatic processing events which can consist of either the proteolytic cleavage of the target protein or the chemical modification of a single or several amino acids. Frequently studied examples of PTMs include phosphorylation, glycosylation, methylation, acetylation, ubiquitination and lipidation [17]. Today we count at least 200 different types of PTMs, and conservative estimates suggest that each protein can be modified by about 10 different PTMs [18]. Only considering phosphorylation, the most extensively studied PTM, as an example, more than 500,000 sites are predicted to exist in the human proteome [19].

While the proteome of a cell is complex, it is not static. The cell changes the proteome in response to stimuli, including a myriad of environmental factors, and therefore the expressed cellular proteome is difficult to predict. Both the complexity and transient nature of proteomes complicate their study. New technological developments continue to incrementally increase the fraction of a proteome that can be measured, but its character of constant change precludes the definition of a fixed endpoint of protein identification, implying that at present all proteomic studies must be considered incomplete [20]. It is therefore an important intermediate goal to at least strive for the conclusive and reproducible detection and quantification of predetermined sets of proteins or subproteomes, the selection of which is driven by biological and clinical questions.

Over the last two decades, the genomes of several eukaryotic organisms, including *Homo sapiens*, have been sequenced. From these sequences, the protein coding genes that are potentially transcribed and translated can be computationally predicted. The early gene prediction algorithms predicted 25,000–30,000 protein-coding genes in the human genome [21]. This number is significantly lower than the one estimated not long before the end of the genome sequencing project (80,000–140,000 genes) [22]. Furthermore, the initial estimate derived from the human genome sequence has been further revised downward. Current estimates assume that the human genome contains about 21,000 annotated protein-coding genes (Ensemble rel.64 [16]), the genome of the nematode *Caenorhabditis elegans* contains about

23,500 (Wormbase WS228 release [23]) and that of *Drosophila melanogaster* about 14,000 (BDGP rel. 5.25 [24]). These numbers are in the same range as for the human species. Therefore, it is apparent that the number of genes or the size of the genome cannot be used as parameters to judge the complexity of an organism [25]. Rather, the biological complexity seems to mainly depend on different mechanisms, including those that govern gene regulation and gene expression within species [26,27] and on the combinatorial association of individual proteins into functional protein complexes or modules [28]. A single gene can lead to a large number of different protein products, and this multiplicity is an additional mechanism to increase the pool of proteins generated by a genome. This is true both at the level of a population due to the prevalence of single nucleotide polymorphism (SNPs) and at the level of each individual member of a population, due to alternative transcript splicing and differential post-translational modifications of the proteins. In addition, the quantitative pattern of expressed proteins and their assembly into functional complexes, even in a simple unicellular organism such as a bacterium, changes continuously during its cell life but it is also heavily affected by a number of intra- and extracellular stimuli (i.e. temperature, availability of nutrients, cell density, etc.) [28]. The adaptation of the proteome to both different physiological and environmental conditions reflects the fact that proteins are the final products of the gene expression process and the actual actors that regulate most biological processes, including gene expression [29]. Therefore, while the number of protein-coding loci in the human genome and that of other species is by now reasonably well known, the composition of the proteome of a species and the acute proteome expressed in a tissue or cell at a specific state largely remains to be explored [20,30].

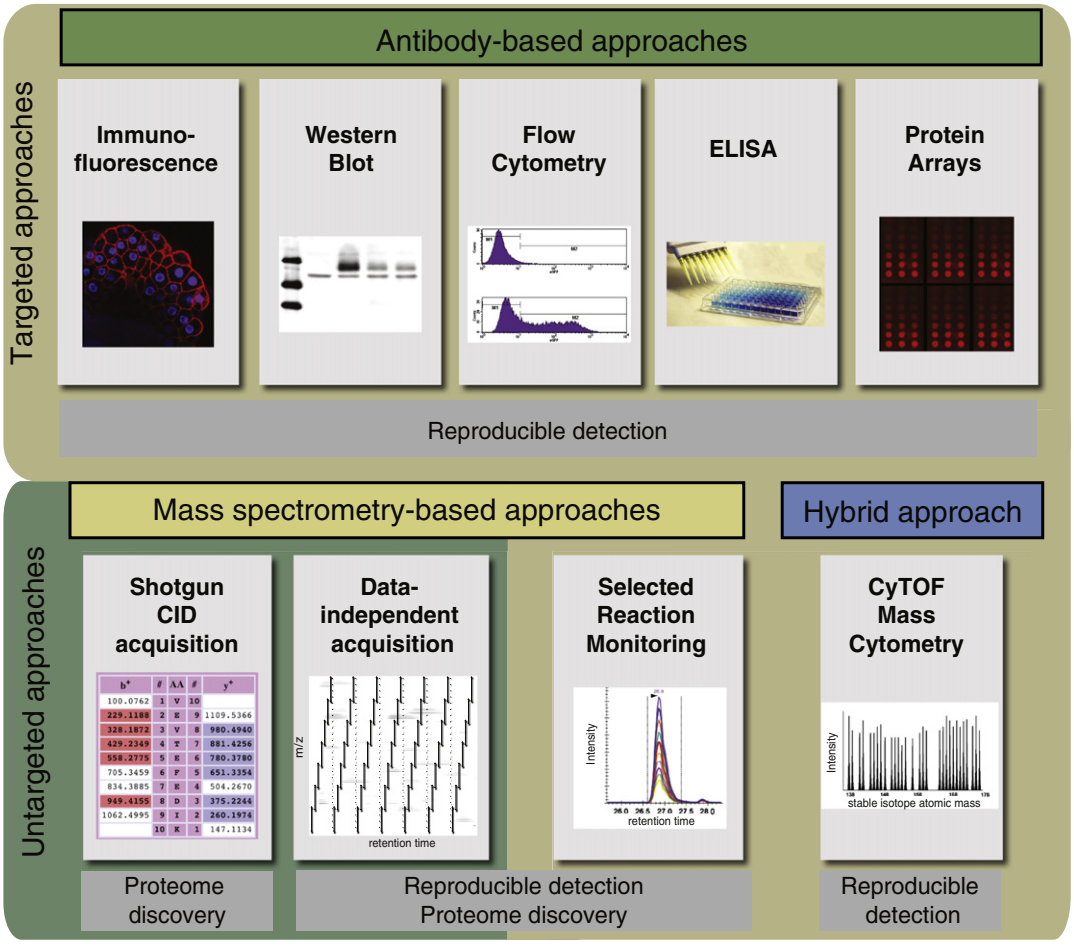

Fig. 1 – Overview of antibody- and mass spectrometry-based methods for protein detection and quantification. The approaches can be classified in targeted and untargeted and differ in the type of information they deliver. Immunofluorescence is the only method shown which can reveal *in situ* information such as the localization of proteins or PTMs within intact cells and tissues. Western blotting is semi-quantitative and additionally yields information about the molecular weight of the proteins analyzed. In contrast, the enzyme-linked immunosorbent assay (ELISA) does not confer molecular weight information but is more suited to protein quantification. Protein analysis in single cells, as opposed to extracts from pooled cell populations or tissues, is achievable with flow cytometry and mass cytometry. Protein arrays employ immobilized target proteins and is amenable for the analysis on a proteome-wide scale. Mass spectrometry-based approaches consent the identification and quantification of proteins in the sample. The concepts of CyTOF mass cytometry, Selected Reaction Monitoring, shotgun CID and Data-indipendent acquisition are explained in the text.

While the analysis of the proteome in its full complexity remains a daunting analytical challenge, there is no way around proteome studies within the framework of basic and applied life science research. This is mainly because of the biological and therapeutic importance of proteins, their organization into functional modules, pathways and networks that catalyze biochemical reactions and regulate essentially all biological processes and the fact that neither the presence or the abundance or activity of a protein can be predicted from genomic information. Therefore, the availability of tools for the accurate and comprehensive detection and quantification of cellular proteomes or defined subfractions thereof, is an essential target for life science research [31,32].

2. Methods used in proteome research

2.1. *Introduction to proteome analysis strategies*

The importance of proteome analysis in the life sciences accelerated the development of different analytical techniques for the detection and quantification of proteins. These analytical tools can be distinguished as i) methods for proteome discovery and ii) methods for the reproducible detection and quantification of subproteomes.

Affinity-based approaches and mass spectrometry-based techniques represent the preferred methods currently used for proteome investigations (Fig. 1). Traditionally, detection and quantification of proteins has been performed using antibody-based techniques such as Western blotting (WB), immunofluorescence (IF) or immunohistochemistry (IHC). The high specificity of antibodies and the sensitivity of the measurement methods enable the detection of low abundant protein species in complex backgrounds. With the advent of proteomics, the measurement of proteins by affinity-based methods has been scaled up. Using any one of several antibody array-based techniques, thousands of proteins can be detected and quantified with high reproducibility. Projects such as the Human Protein Atlas aim at creating a collection of antibodies against all human proteins [33]. Today the Human Protein Atlas database contains antibodies for the protein products of more than 11,000 genes and for a fraction of them also data regarding the localization of the antigen in cells or tissues [34]. At present, also new affinity based binding reagents based on proteins, peptides and nucleic acids are being developed that expand the range of available affinity reagents and their use [35]. However, despite the enormous success achieved with affinity reagent based-methods over the years, the definitive validation of these reagents and their routine implementation in specific assays usually requires months or in same cases, specifically in clinical research, even years [35–38]. Indeed, measurement of kinetic parameters such as association and dissociation rates, epitope specificity, selectivity and cross reactivity are essential to define the quality of the reagents, and so far there is no method with an appropriate throughput to measure any of those on a whole proteome scale [37].

Mass spectrometry-based proteomics has emerged as an alternative approach to affinity reagent-based methods for the detection and quantification of the components of a proteome [39]. Similar to the situation for proteomics based on affinity reagents, several mass spectrometry-based strategies have been developed, each one with its particular performance profile [40]. Today, the two major approaches used for proteome investigation are referred to as shotgun (or discovery) and targeted mass spectrometry, respectively [40]. The two methods are identical in the sample preparation steps upstream of injecting the samples into the mass spectrometer. First, the proteins, either in a gel or in solution, are digested with one or more proteases, and the resulting peptides are separated by liquid chromatography. The peptides eluting from the column are usually ionized by electrospray ionization (ESI) and injected into the mass spectrometer [41,42]. Less frequently, the peptides are spotted on the sample plate of a Matrix-Assisted Laser Desorption Ionization mass spectrometer (MALDI-MS), a method that is not further discussed here [43]. The resulting precursor ions are guided trough the mass spectrometer by electric or magnetic fields before they are detected. Despite those similarities in sample preparation, the two ESI-MS-based proteomic strategies differ in the manner in which the mass spectrometer operates. In a shotgun experiment, the mass-to-charge ratio of peptide ions is intermittently recorded during the course of the experiment to generate a mass spectrum often referred to as a survey scan. After each survey scan, shotgun instruments select and isolate peptide ions using a simple abundance-guided heuristic and subject them to fragmentation. This approach is known as product ion scanning and requires the employment of fast scanning analyzers, often ion traps, to rapidly complete each cycle and to aim for a comprehensive fragmentation of all peptides ions obtained by an enzymatic digest of cellular proteomes [40]. This measurement scheme is called data-dependent acquisition (DDA) because peptide fragmentation is guided by the abundance of detectable precursor ions. The recorded fragment ion spectra and the corresponding precursor masses are then used by protein database search algorithms to infer the sequence of the peptides and the identity of the protein from which they are derived (Fig. 2). Shotgun mass spectrometry is also typically referred to as a discovery method. This is because the selection of the fragmented peptide ions does not involve any prior knowledge about the composition of the sample.

Shotgun mass spectrometry is mostly used for discovery and large scale mapping of cellular proteomes and has been employed for proteome analysis of both prokaryotic and eukaryotic cells, and tissues [18,40]. Today modern mass spectrometers optimized experimental protocols and mature software tools for error-rate controlled data analysis enable the identification of large fractions of proteomes. However, the stochastic sampling of this technique affects the reproducible detection of proteins between different experiments.

In contrast, targeted mass spectrometry identifies only predefined set of peptides. At present, the most widely used targeted mass spectrometry technique is Selected Reaction Monitoring (SRM) [44]. It uses the capability of triple quadrupole mass spectrometers to act as ion filters. In an SRM experiment, a specific precursor ion is preselected in the first (Q1) analyzer, fragmented by collision-activated dissociation (CAD) in the second quadrupole (Q2) and one or several of its fragments are specifically measured by the second mass

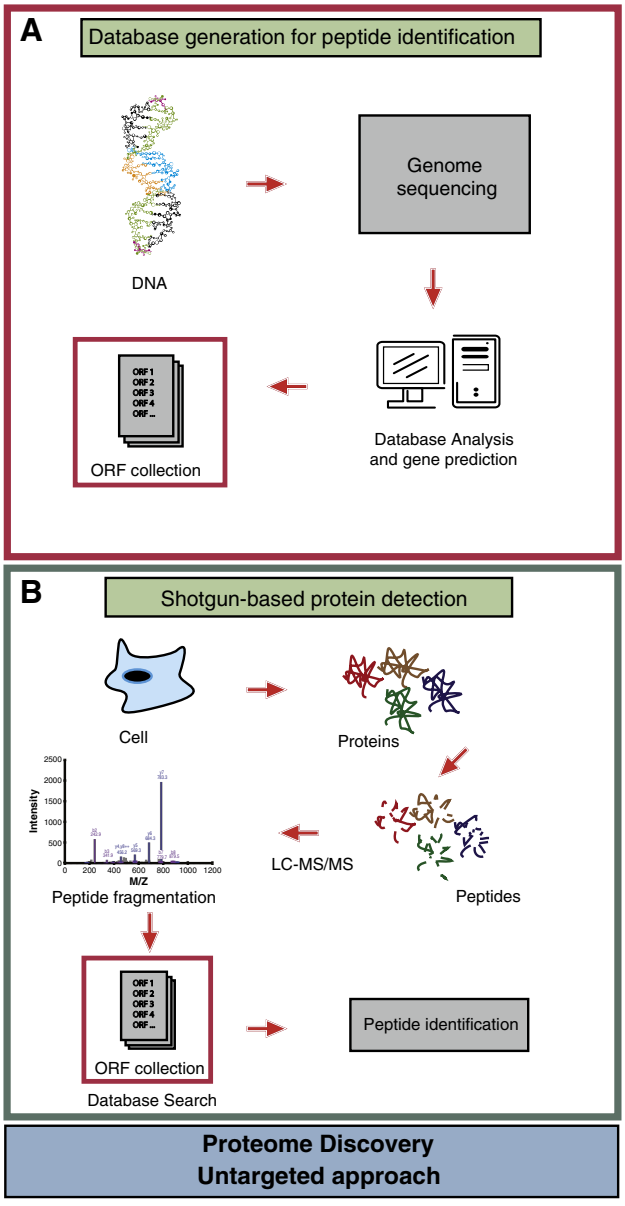

Fig. 2 – Untargeted shot-gun mass spectrometry approach for protein identification. Protein identification by mass spectrometry requires genome sequencing and ORF identification in the organism of interest (A). Proteins extracted from cells or tissues are digested with a protease, usually trypsin, and analyzed by LC-MS/MS. The peptide ion fragment spectra are searched against protein sequence databases for the identification of the peptides (B).

analyzer (Q3) as a function of time. Detection of peptide/fragment ion pairs over time produces chromatographic peaks indicative of the presence of a specific peptide. As the fragment ions of the targeted precursors are monitored constantly during the course of the experiments, the method is not data-dependent. It has an improved limit of detection and extended dynamic range of analysis compared to discovery methods, in part because the targeted precursor ion does not have to be explicitly detected above the noise. However, the technique focuses on the detection of a predefined set of peptide candidates for which the charge states, chromatographic retention time and the relative product ion intensities need to be known prior to the measurement [40,44]. SRM has the ability to systematically detect predefined sets of proteins in complex samples with high reproducibility [20,40]. It is therefore frequently used in systems biology and biomedical applications where the same set of proteins has to be reproducibly measured in multiple samples. In addition to DDA and SRM, there are several data-independent acquisition (DIA) schemes, analytical strategies that can broaden the scope of proteomic experiments as well. These techniques are conceptually similar to precursor ion scanning and fragment the entire population of present precursor ions either completely without selection or by cycling through the whole precursor m/z range in discrete increments [45–48]. Because the link between precursor m/z and fragment ion spectrum is lost in these acquisition schemes, algorithms for peptide identification have to rely on the chromatographic co-elution of signals stemming from the same peptide. Alternatively, as exemplified by the recently introduced SWATH-MS technique, fragment ion maps generated by DIA methods can be searched by *in silico* targeted data analysis strategies [46]. We will not discuss these strategies further here and instead focus on SRM as a targeted proteomic method.

2.2. *Discovery proteomics — shotgun MS as a tool for proteome discovery*

In the early nineties, the introduction of shotgun mass spectrometers in proteomic research and the rapid development of quantitative approaches sparked enthusiasm in the scientific community. The prospects were exciting: rapid experiments, systems-wide analysis of proteins and PTMs, accurate quantification in different cellular states. For a significant period of time, shotgun proteomic studies aimed to increase throughput and a cataloging phase soon began that focused on identifying the largest possible number of proteins and PTMs in microorganisms, cell lines and tissues. Soon after their publication, the first high-throughput studies attracted the critical attention of the scientific community with respect to the confidence of the identifications of both peptides and proteins. Different statistical methods were developed to evaluate the false discovery rate (FDR) of mass spectrometry-based proteomic measurements. Several groups used a reverse database search strategy to derive a global statistic measurement of the quality of peptide assignments for a given dataset [49,50]. In this case, the MS/MS spectra are searched against a composite database composed of the correct protein sequences together with the reversed sequences, which do not correspond to naturally occurring proteins and therefore cannot be present in the sample. The assumptions on which the false discovery rates are estimated are: a) the match of a MS/MS spectrum to a reverse peptide is by definition incorrect, and b) the frequency of incorrect assignments for a certain dataset is the same for both the direct or reversed protein sequences in the database [51]. Other methods to estimate the rate of false peptide and protein identifications in proteomic experiments such as the one implemented in the PeptideProphet software

use statistical models to assign a probability of correctness for each peptide identification [52]. In most proteomic strategies, researchers are interested in the identification of proteins in their samples rather than just the MS-based peptide identifications. To infer protein identities, the individual identified peptides have to be grouped according to their corresponding protein and a new FDR has to be computed at the protein level [53–55]. Today, proteome analysis tools for mass spectrometry-based approaches have reached a high level of maturity, and the continuing advances in the field are constantly improving the quality of the published data [56].

2.3. *Present status of proteome maps*

The advent of proteomics created the basis for studying the systems-wide effects of cellular perturbations such as kinase inhibition or drug treatment in cellular systems [32]. The enormous technological advances of discovery proteomics over the last decade enabled the detection of a large fraction of all the predicted proteins for different organisms, tissues or cells, even though such measurements represent a significant investment of time, money and require special instrumentation and computational infrastructure. Despite significant proteome identification success, comprehensive detection of all the proteins in an organism remains a challenge, first for inability to *a priori* know or predict the proteome of any cell [20], and second, for technical limitations. It is expected that tryptic digestion of a proteome can produce up to 1 million different peptide species and sample complexity may be further artificially increased by the sample preparation protocols used prior to mass spectrometry analysis. It has been demonstrated that the actual number of peptides generated by tryptic digestion of a protein sample exceeds the predicted number of fully tryptic peptides by at least a factor of 10, mostly because of partially tryptic peptides and missed cleavages [57]. The limited scan speed of shotgun mass spectrometers prevents the sequential fragmentation of the very large number of different peptides typically obtained by the digestion of entire cellular proteomes. This makes shotgun analyses incomplete as they typically cover only a part of the proteome with a bias towards abundant peptides. To counteract this problem, scientists usually perform extensive fractionation of proteomes to reduce sample complexity. LC-MS/MS analysis of each fraction then cumulatively achieves a more comprehensive detection of proteins that may still not be sufficient to cover the whole proteome [18]. While this strategy has proven very successful for increasing proteome coverage, it also raises specific issues of error propagation. We have recently shown that the combination of many LC-MS/MS-derived datasets of peptide identifications bears the danger of inflating the number of identified proteins unless special precautions are being taken to control the FDR of protein identification [53,54].

In Table 1 we list several organisms for which proteomes were investigated in depth. The table reports proteome coverage as the fraction of the identified proteins over the predicted ORFs. While proteome exploration in unicellular organisms almost reached saturation, the exhaustive proteome mapping in multicellular organisms in which different cell types express partially overlapping segment of the proteome is more challenging. In two recent publications, the proteome of two different mammalian cell lines was extensively investigated by shotgun mass spectrometry [58,59]. The two studies employed different protein extraction protocols, protein fractionation methods and data analysis tools [54,60,61] and both reached almost identical proteome coverage. About 10,000 proteins were identified in two different human cell lines with a similar false discovery rate (FDR), suggesting that the two studies independently reached the maximum number of detectable proteins with currently available technologies. The limitation of current technologies for proteome investigation appear even more dramatic when we consider the number of splice variants and SNPs that could be detected at the mRNA but not the protein level. The difficulties of detecting SNPs and protein isoforms are due to several factors. First, such protein species could be only expressed in certain tissues at certain stages of development or under specific cellular conditions, some of them may not be transcribed or have a shorter half-life. Second, they could be only expressed in low quantities, and the detection of proteins at low levels remains a challenge [20,62]. Third, the identification of SNPs and protein isoforms by database search algorithms suffers still of higher occurrence of false-positive hits, which is especially problematic due to search space inflation when large protein databases derived from multi-frame translations of nucleic acid sequences are used.

Different groups have developed approaches for the detection of protein sequence polymorphisms in shotgun datasets. The strategies applied, include both the addition of known SNPs to the sequence database used for protein database searching [63,64] and more recently, filtering of high-quality spectra with iterative searching of unassigned spectra [65]. Bunger et al. [66] used a novel peptide prediction algorithm and a decoy database based on random peptide substitutions to find protein-coding non-synonymous SNPs. The authors identified a total of 629 nsSNPs from shotgun data of highly fractionated samples of human breast cancer cell lines. Recently, two research groups detected alternative protein isoforms for 53 genes in mouse [67] and for 150 genes in human samples [68], again by searching large shotgun datasets. Similar approaches have been used to identify splice isoforms in *Arabidopsis* [69] *D. melanogaster* [70] and *Aspergillus flavus* [71]. Recently Ning and Nesvizhskii [72] examined the feasibility of using MS data for the identification of novel alternatively spliced forms by searching MS/MS spectra, using publicly available mouse tissue proteomic and RNA-Seq datasets. The authors demonstrated a correlation between the likelihood of identifying a peptide from MS/MS data and the number of reads in RNA-Seq data for the same gene. However, the number of novel peptides that were actually identified from MS/MS spectra was substantially lower than the number expected based on *in silico* prediction. Similar approaches towards enhanced proteome coverage include parallel RNA expression profiling [73,74] and the use of RNA-Seq-derived customized databases for protein identification to better approximate the proteome space actually present in the samples under investigation [75]. The use of such sample-specific databases can improve the proteome coverage by enabling the identification of protein splice variants that may not be covered by the context-independent protein sequence databases which are derived from genome sequencing

Table 1 – Distribution of proteome coverage and predictions for several organisms.

Annotated protein coding genes for *Homo sapiens*, *Mus musculus*, *Drosophila* melanogaster and Arabidopsis thaliana include sequences whith at least one mach in sequence repositories external to Ensembl. Protein coverage is calculated as the percentage of proteins identified in proteomics experiments over annotated protein coding genes. Annotated protein transcripts include all experimentally derived protein-coding and splice-variants. Prediction of the protein coding genes was performed using the following algorithms: Glimmer [137] for *Mycoplasma pneumoniae*, *Thermoplasma acidophilum*, *Staphylococcus aureus* and *Leptospira interrogans serovar Copenhageni*; Genescan [99] for *Drosophila melanogaster*, Mus *musculus* and *Homo sapiens*, FGENESH [137] for *Arabidopsis thaliana* and finally GeneFinder (P. Green, unpubl.) for *Pristionchus pacificus* and *Caenorhabditis elegans*. Non-synonymous protein coding single nucleotide polymorphisms were derived from Ensembl Variation datasets. When several cellular strains were available all were considered. Abbreviations used : CMR, Comprehensive Microbial Resource, SGD, Saccharomyces Genome Database.

Organism	Proteins identified	Annotated protein coding genes	Proteome coverage	Annotated protein transcripts	Predicted protein coding genes	nsSNPs (source database)	Genomic annotations source
Bacteria and archaea							
Mycoplasma pneumoniae M129	620 Ref. [138]	688	90%	–	779 (GLIMMER)	–	CMR Ref. [139]
Thermoplasma acidophilum	1025 Ref. [140]	1478	69%	–	1630 (GLIMMER)	–	
Staphylococcus aureus COL	1703 Ref. [141]	2681	63%	–	2716 (GLIMMER)	–	
Leptospira interrogans serovar Copenhageni	2221 Ref. [141]	3660	61%	–	4475 (GLIMMER)	–	
Eukaryotes							
Saccharomyces cerevisiae	4399 Ref. [142]	6696	66%	7130	7940 (GLIMMER)	64124 (SGRP)	SGD rel. 64.10 Ref. [143]
Drosophila melanogaster	9263 Ref. [113]	13781	67%	23017	19437 (GENSCAN)	459982 (BDGP 5)	Ensembl rel. BDGP 5.25 Ref. [16]
Pristionchus pacificus	4029 Ref. [144]	5211	77%	5211	24217 (GeneFinder)	–	Wormbase rel. W228 [23]
Caenorhabditis elegans	6779 Ref. [113]	23358	29%	29872	25391 (GeneFinder)	35483 (WS220)	Wormbase rel. W220 [23]
Arabidopsis thaliana	13029 Ref. [145]	27299	48%	34183	23868 (FGENESH)	896537 (TAIR10)	Ensembl rel. 64.10 Ref. [16]
Mus musculus	7686 Ref. [113]	22234	35%	44337	46375 (GENSCAN)	29373 (dbSNP128)	Ensembl rel. 64.37 Ref. [16]
Homo sapiens	12141 Ref. [113]	20996	58%	72065	47019 (GENSCAN)	271732 (dbSNP135)	Ensembl rel. 64.3 Ref. [16]

projects. In several cases, large MS datasets were also instrumental in improving genome annotation, specifically the definition of ORFs by confirming annotated protein-coding genes, correcting protein-coding gene annotations and identifying novel protein-coding genes [76–84]. Overall, while the existence of many differently spliced mRNA transcripts is supported by several pieces of evidence such as cDNA, EST sequence and by microarray data [85], the proof of their translation in proteins by mass spectrometry has generally remained difficult [10,69–72]. We think that the analysis of splice variants and mutations affecting amino acid sequences on the protein level is important, and that is not sufficient to rely solely on genomic and transcriptomic data. Only the protein analysis reveals whether the predicted mutant proteins and isoforms are indeed present at detectable levels, and also allows quantitative comparisons of expression levels of these variants which may have very diverse biological functions.

The same limitations that exist for the detection of SNPs and splice variants also challenge the comprehensive characterization of PTMs on a proteome-wide level. PTMs are often substoichiometric and therefore challenging to detect. Mass spectrometry-based analysis of PTMs requires several steps to be successful: first, the enrichment of the modified subproteome consisting of the peptides carrying the specific modification of interest; second, the detection and identification of the modified peptides in proteolytic digests; and third, the unambiguous identification of the modified residue(s) within the peptide/protein sequence. Advances in both, instrumentation and enrichment protocols, resulted in the identification of thousands of PTMs such as phosphorylation, acetylation and methylation sites in different organism and tissues. Today, enormous amounts of data are collected and represented in high quality publicly available databases such as PhosphositePlus [86], Phosida [87] Phosphopep [88] or Uniprot [89]. The PhosphositePlus database, for example, integrates both low- and high-throughput data sources into a single comprehensive resource. Today it contains more than 1,000,000 phosphorylation sites identified in different organisms. The enormous amount of accumulated data stimulated the development of several PTM predictor tools [90]. To name a few, Scansite [91] predicts phosphorylation sites by searching amino acid patterns corresponding to kinase consensus motives in the protein databases, Phosida [87] offers predictor tools for both phosphorylation and lysine acetylation, and UbPred [92] is instead used to identify ubiquitination sites on proteins. The overarching problem in PTM site predictions from protein sequence is a high FDR that prevents us from making accurate predictions on the extent of PTMs in a cell. However, such predictors have proven to be useful tools for the identification of the precise PTMs site when independent experimental evidence is available.

The complexity of proteomes arising from cellular mechanisms of protein expression and modification pose challenges for the comprehensive detection of cellular proteomes. Very often improvements in instrument performances, sample preparation protocols and data analysis workflows enabled the detection of increasing numbers of proteins. In this context, approaches for the fast and targeted detection of newly discovered proteins are becoming more and more important as these will offer new tools to biologists for their functional characterization in biological processes.

2.4. *How to expand the proteome maps by targeted analysis*

Most large-scale proteomic studies have operated in a continuous discovery mode, whereby overlapping segments identified in proteome fractions have provided a bird eye's view of the proteome expressed by different cells and organisms. A different or complementary approach to increase the detectable fraction of a proteome is to specifically target protein species expected to be expressed by cell or organisms. The cellular proteome could thus be "extrapolated" either by prediction of the expressed genes or by experimental determination of gene expression, e.g. by deep sequencing of mRNA transcripts. Gene finding is relatively straightforward in prokaryotes because their gene structure is simple, genomes are compact and void of introns, and regulatory regions are well defined. Even for these simple organisms though, shadow open reading frames (ORFs) exist that can overlap with annotated genes in another reading frame and thus escape detection. Gene prediction is more challenging in eukaryotes because their gene structures are more complex, mainly because of the presence of introns and exons. Probabilistic algorithms such as GENESCAN [93], GlimmerHMM [94], GeneMark [95] partition sequence segments into introns, exons and intergenic regions. Although these methods can quite successfully predict protein coding regions and individual exons, accurate prediction of the genes structure still is a remaining challenge. Next generation sequencing technology now provides efficient tools for comprehensive exome, transcriptome and genome analyses of uni- and multicellular organisms. RNA-Seq analysis and the mapping of short sequence fragments onto the reference genome identify introns, exons and their boundaries in a DNA gene sequence, identify SNPs and recently it provided evidence of novel RNA editing mechanisms [96].

While prediction of genes and splice variants and innovative sequencing technologies are now able to reveal the sequences of the different proteins potentially expressed in a cell, the real challenge is now to use proteomics to prove the value of these predictions at the protein level. Targeted methods based on either affinity reagents or mass spectrometric measurements can be used for the detection of predicted proteins or protein isoforms. This experimental set-up differs from untargeted approaches for proteome discovery since it is driven by the hypothesis that predicted populations of proteins are expressed in defined cellular conditions, even if they were never detected before probably due to the sensitivity limits of protein detection approaches used.

The important parameters which impact the success and dissemination of targeted detection of proteins on the proteome scale, are sensitivity, specificity, and throughput. Furthermore, the time needed for reagent generation, and the ease of use of the instrumentation, the costs involved, and the required infrastructure are critical parameters to consider.

Methods for protein localization analysis in a cell or in a tissue currently depend on the use of affinity-based reagents. However, we believe that targeted approaches based on mass

spectrometry have adequate performance and multiplexing capability to allow the detection and quantification of a large number of proteins in crude cell or tissue extracts. Such approaches are therefore becoming valuable tools that can be used to expand the borders of the observable proteome space.

For example, protein forms still escaping detection can be *in silico* digested with one or more proteases and the peptides with favorable MS detection properties can be predicted [97]. Such peptides can then be chemically sensitized and used to efficiently develop assays required for their detection and quantification in digested proteomes (Section 3.2).

3. Reproducible quantification of predefined subset of the proteome

Studying the dynamics of protein expression, protein complex assembly and PTMs in cells in different states is important for the understanding of biological processes. Quantitative proteomic measurements performed with targeted methods critically depend on the ability to quantify the specific protein sets relevant for a defined biological process [98]. Examples for such functionally defined proteome subsets are signal transduction pathways, organellar proteomes, protein complexes and enzymes involved in common or interconnected metabolic pathways. The two main strategies to quantify predefined protein sets in biological samples are, as discussed above, based on affinity reagents and on targeted mass spectrometry (Fig. 1). The development of affinity reagents other than antibodies has been recently reviewed elsewhere [35,37,99]. We will focus our discussion of affinity-based measurements on those which employ antibodies. Both affinity- and MS-based approaches have their advantages and drawbacks and differ in the type of information they can yield, as will be discussed below.

3.1. *Protein detection approaches employing antibodies*

Antibody-based techniques are among the most widely used tools for targeted protein detection. Their enormous success in biological and biomedical studies is mainly derived from the often exquisite sensitivity towards their targets and their simplicity of use. Antibodies are specialized proteins of the vertebrate immune system that evolved to recognize foreign molecules. Antibodies recognize defined regions, termed epitopes, of a molecule termed antigen. Protein epitopes usually cover 8–11 amino acids and the number of instances where these sequences occur in the proteome define the upper limit of the specificity of an antibody. The capability of antibodies to specifically recognize proteins and other biomolecules has been exploited by biologists for protein analysis in a variety of contexts such as the detection of proteins immobilized on a support, usually a membrane (Western Blot, ELISA), in the context of preserved cellular structure (e.g. immunofluorescence, flow cytometry), or even in dissected or sectioned tissues where multiple cell types are concomitantly present (e.g. immunohistochemistry). Antibodies are usually generated by the immunization of animals with proteins or peptides (Fig. 3). Immunization of animals with modified (e.g. phosphorylated, acetylated etc.) peptides can yield modification-specific antibodies which allow detection of the target protein only if it is subjected to the respective post-translational modification in the context of a peptide sequence. Probably the most widely used reagents of this type are phospho-specific antibodies. These have been instrumental in monitoring the activity of phosphorylation-based signaling pathways. Du et al. have demonstrated the successful application of antibody-based reagents to the simultaneous profiling of tyrosine kinase activity in different cancer cells [100]. They designed an immuno-sandwich assay in which specific antibodies against 62 of the 90 human tyrosine kinases are used to isolate these from cell lysates of 130 human cancer cell lines, and a fluorescently labeled phosphotyrosine antibody subsequently detects tyrosine phosphorylation levels of the kinases, an event which usually correlates with their activation. To extend antibody-based protein measurements to an ideally whole-proteome level, efforts are ongoing to build a centralized antibody database with tested performance of the individual affinity reagents. The Human ProteinAtlas project, initiated in 2005, aims at systematically investigating the human proteome using affinity-purified antibodies [34]. Today about 14,500 antibodies targeting epitopes on more than 11,000 human genes are deposited in the ProteinAtlas database. Technological advances enabled the immobilization of thousands of antibodies on miniaturized arrays often referred to as antibodies chips (in the mm^2 to cm^2 range) which facilitate the generation of protein expression profiles of defined sets of proteins in unfractionated samples such as plasma, cell extracts or tissues. Recently, the Uhlen group performed a comparative study of both the transcriptomes and the proteomes of three different human cell lines of functionally different origins [74]. Using array technology, the authors profiled about 4000 proteins by immunofluorescence, quantified more than 5000 proteins by shotgun mass spectrometry and more than 14,000 transcripts by RNA sequencing in the three cell lines. The authors found that the majority of the proteins are expressed in all three cell lines, displaying a good correlation with transcript levels, but only 30% of the quantified proteins showed high expression level differences between the individual cell lines. This suggests that the amount of expressed proteins or different regulatory mechanisms, which influence protein levels, define cellular function rather than differentially expressed genes. However, it is important to mention that the proteins that are still escaping detection by these methods may play an important role in defining cellular functions, and therefore expansion of the experimentally observable proteome to these unexplored regions is an important goal in the efforts to shed light on the molecular mechanisms underlying biological processes.

A hybrid method which combines antibody-based protein or PTM detection with a mass spectrometric readout to enhance the multiplexing capability of flow cytometry analysis is called mass cytometry. This experimental approach is explained in Box 1, and is a promising way to screen large numbers of samples for the abundance of sets of proteins or protein modifications for which good antibodies are available. Another analytical method which combines antibody-based analyte enrichment with mass spectrometry is called Stable Isotope Standards and Capture by Anti-Peptide Antibodies

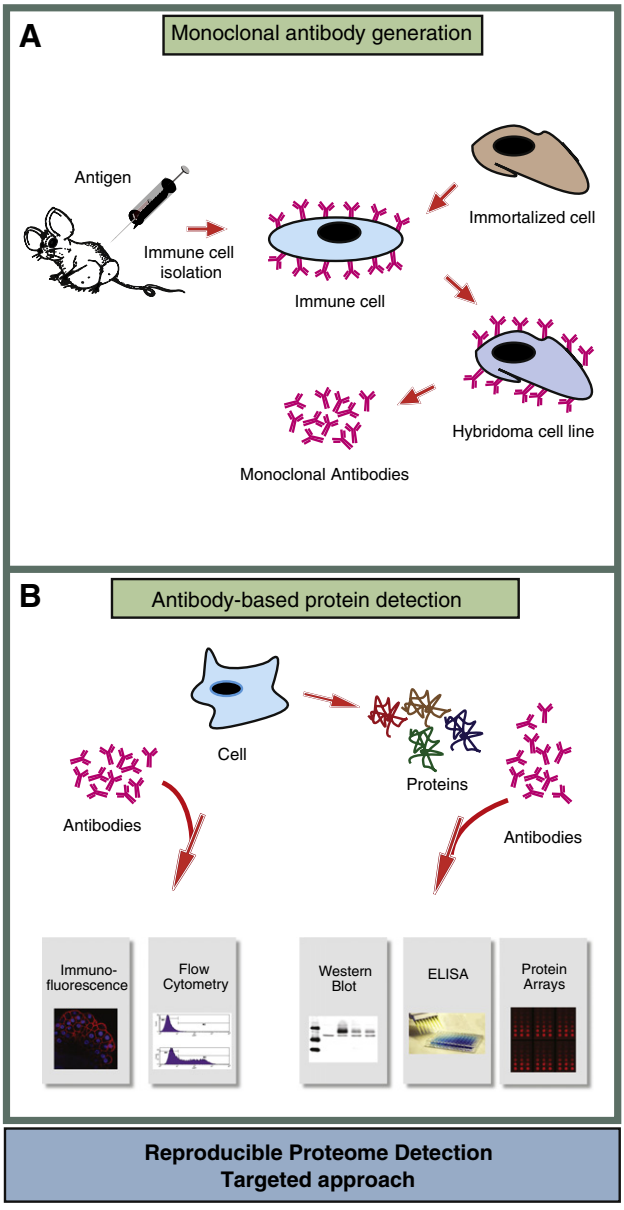

Fig. 3 – Targeted detection of proteins using antibodies. Detection experiments employing antibodies are consist two phases: reagent generation and antibody-based detection of proteins (A-B). Monoclonal antibody generation requires the immunization of host animals such as mice or rabbits by antigen injection. The spleen cells of the animals are isolated and fused with immortalized cells. The generated hybridoma cell lines can be cultured and they represent an permanent resource for antibody production (A). The antibodies are then purified and the detection performances are tested for specific applications. Immunofluorescence and flow cytometry experiments are performed with intact cells and thus allow protein detection in single cells. Similar to mass spectrometry approaches, analyses like western blotting, ELISA and protein array experiments require the extraction of proteins from larger populations of cells, therefore signals detected for proteins are averaged over the entire cellular population (B).

(SISCAPA) [101]. In contrast to mass cytometry, it does not perform single cell measurements but is rather a specialized quantitative peptide-immunoaffinity purification strategy, in which selected peptides derived from proteins of interest are physically enriched on an affinity column containing rabbit polyclonal antibodies raised against those peptides. For quantification purposes, stable isotope-labeled versions of the peptides peptides of interest are spiked into the samples prior to analysis and are therefore co-purified on the affinity column, which is coupled on-line to an analytical LC-MS/MS setup for peptide quantification by SRM. The method has been used primarily in the context of protein biomarker studies to quantify peptides from proteins present in blood plasma or other body fluids in very low concentrations.

3.1.1. *Error estimation for antibody-based assays*

Although technological advances have reduced the time required for the generation of antibodies on a large scale, accurate evaluation of their performance for proteome detection still represent a challenge [35–38]. The validation of an antibody's performance usually consists of controlling its specificity and sensitivity. In small scale experiments, these parameters are usually assessed using dilution series of both positive and negative controls that in the case of Western Blot analysis for example, can be cell extracts in which the specific protein targets are present or absent, respectively. Already in this simple case, such controls, especially for antibodies targeting post-translationally modified residues on proteins, are often not available. This complicates the required thorough evaluation of antibody quality. With the development of initiatives for large-scale antibody production, it became possible to test the performance of thousands of affinity reagents resulting in a systematic estimation of antibody performance. In recent work from the Uhlen group [102], the authors reported that from a collection of 11,000 antibodies, only 531 detected a single protein band in western blots of human plasma, suggesting that antibody monospecificity represents the exception rather than the rule. Although the introduction of antibodies in life science research is anterior to the one of mass spectrometry, the development of strategies for the evaluation of false positive detection remains a serious problem. This is mainly due to the impossibility both to predict and measure antibody specificity on a whole proteome level. Until this issue is addressed and solved, the risk of off-target reactivity compromising measurements performed with affinity reagents such as antibodies cannot be excluded.

3.2. *Targeted mass spectrometry*

There are two different approaches to direct mass spectrometric measurements towards specific peptides of interest. The first one relies on mass inclusion lists and uses the same instrumentation as in discovery-mode experimentation [103]. The second one relies on SRM [44] and is the subject of this review. Targeted proteomics by SRM has matured to a degree that it is now a viable alternative to antibody-based detection of target proteins. SRM experiments focus on the detection of a predefined set of peptide candidates and are typically characterized by two phases: the first one consists of the

Box 1

Mass cytometry — a hybrid approach utilizing elemental detection by mass spectrometry for antibody-based protein analysis in single cells.

Protein analysis on a single cell level can be performed by flow cytometry, a technique which employs fluorescently labeled antibodies and is therefore difficult to multiplex beyond 8 to 10 spectral channels. In contrast, mass spectrometry-based protein detection can be easily multiplexed, but requires larger amounts of material i.e. large numbers of cells to measure the average protein concentration. Because most of the current biochemical knowledge is based on averaged measurements from pools of cells from cell culture or tissues, many new questions that were out of reach before could be addressed with a technology that is capable of performing high throughput and multiplexed protein measurements in single cells. Examples include the immuno-phenotyping of distinct cell subsets from the immune system, and the simultaneous monitoring of signal transduction pathways in cell populations subjected to a defined perturbation or stimulus. An emerging technology which features these qualities is mass cytometry, a hybrid technique which combines multiplexed antibody-based protein detection with a quantitative mass spectrometry readout. The experimental setup is according to the principle of flow cytometry and therefore allows the analysis of single cells in high throughput. In conventional flow cytometry analyses, cells are labeled with fluorescent antibodies or dyes, and therefore the number of available channels that can be detected simultaneously is limited by the rather broad excitation and emission spectra and resulting spectral overlap of the fluorophores used. The consequence is that parallel measurement of more than 10 parameters is not readily achievable by fluorescence-based flow cytometry. Newly developed fluorophores called quantum dots have narrower emission bands and improve multiplexing capacity, but the conceptual limitation remains the same. Mass cytometry overcomes this problem by using mass spectrometry instead of fluorescence readout, thereby providing a much higher resolution and multiplexing capacity. Antibodies against the proteins or protein forms (e.g. bearing a specific posttranslational modification) of interest are tagged with different transition element isotopes, which do not occur in living systems at detectable levels. After binding of the antibody conjugates to the cells and introduction of the cell suspension into the instrument, droplets containing single labeled cells are sprayed into inductively coupled argon plasma, which atomizes and ionizes the cellular constituents. A time-of-flight (TOF) analyzer subsequently acquires mass spectra containing quantitative information about the reporter isotopes present in each cell. Parallel measurement of 34 parameters in single cells has been reported, and interrogation of up to 100 parameters per cell is expected to be achievable [136]. Because the targeted protein analysis is based on antibodies in this method, in contrast to the pure mass spectrometry approaches, the choice of proteins of interest and the specificity of detection is fundamentally limited by the availability, specificity and binding properties of the respective antibodies, which can vary considerably.

development of assays, the second one entails the employment of the assays for the peptide quantification in samples of interest (Fig. 4).

3.2.1. *SRM assay development*

The knowledge of mass spectrometry coordinates of the peptides of interest is fundamental to their reliable detection in the digested proteomes by SRM. As for the generation of antibodies, the first step of assay development consists of the selection of the target proteins. The *a priori* selection of these can include sets of protein species related to the same functional networks such as signal transduction pathways, interacting proteins that constitute functional complexes, enzymes involved in metabolic pathways or belonging to the same functional class such as kinases or phosphates, or e.g. groups of proteins encoded by the same genomic region as may be the case in studies focusing on certain genetically transmitted diseases. For each selected protein, typically 2–5 peptides are chosen and chemically synthesized. Mallick et al. observed that few peptides for each protein are repetitively and consistently identified in large shotgun approaches. The authors defined physiochemical properties for more than 16,000 peptides identified in a large-scale investigation of the yeast proteome and used these to generate computational tools for the prediction of peptides with favorable mass spectrometry detection properties [97]. The synthesis of those peptides should be preferred over the others belonging to the same protein.

The introduction of innovative chemical synthesis strategies such as ON-Spot synthesis lead to a significant reduction of costs and increase in scale with a reasonable financial investment (about 10 EUR per peptide) [104,105]. These peptides are the basis for the development of SRM assays. For this assay development, pools of about 100 synthetic peptides are separated by reverse phase chromatography, directly ionized by ESI and typically analyzed on a hybrid QTRAP mass spectrometer in SRM-triggered MS2 mode [40,106], where the fragment ion spectra for the different peptides are acquired. Automatic database search tools identical to the ones used to analyze shotgun data confirm the presence of the synthetic peptides in the samples and identify fragment ion spectra representing the targeted peptides [107] (Fig. 4A). The acquired spectra are then used as a reference for the generation of the assays from which several pieces of information are extracted including the m/z of the precursor and its fragments, the preferred ionization charge state, the relative intensity of the fragment ion signals and the chromatographic retention time. This combined information is commonly referred to as a peptide SRM assay. SRM assays for each peptide can be further improved to increase the detection sensitivity by optimizing ionization

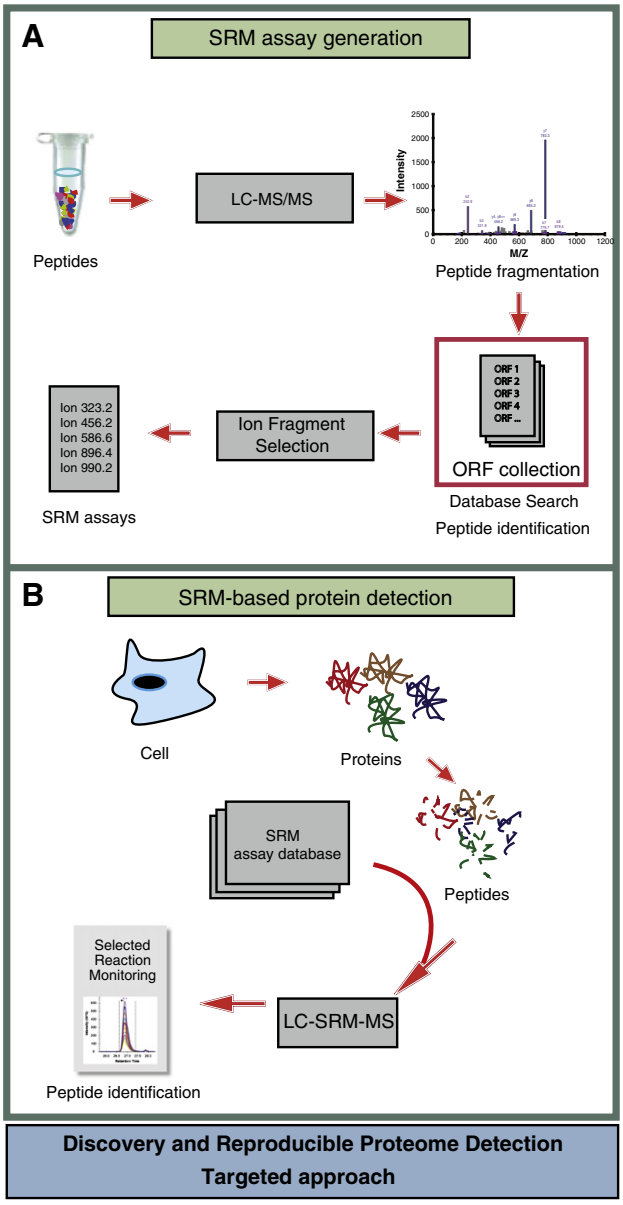

Fig. 4 – SRM mass spectrometry approach for targeted peptide identification. Similar to antibody-based approaches SRM experiments are divided in two phases: SRM assays development and employment (A-B). SRM assay development: pools of synthetic peptides are analyzed by LC-MS/MS on a triple quadruple mass spectrometer. Fragment ion spectra are searched using same panel of tools available for shotgun MS-based peptide identification. Several pieces of information are extracted for the generation of the SRM assay: the preferred precursor charge, the relative intensities of the fragment ions and the retention time of the peptide. The entirety of this information is often referred as the SRM assay for the specific peptide. Assays for about 100 peptides can be developed in a single 60 min mass spectrometry analysis (A). SRM assay employment: the developed SRM assays are deposited in databases from which they can be downloaded and used for targeted detection of peptide by SRM. The proteins extracted from cells are digested into peptides and those are analyzed by LC-SRM-MS. Differently from the shotgun approach, the detection of the peptide does not require any database search tool at this stage, and is only based on the identification of a series of coeluting peaks (B).

parameters such as the declustering potential (DP) or cone voltage (CV) and/or fragmentation parameters such as the collision energy (CE) [106,108,109]. Furthermore, accurate extraction of peptide elution times is necessary for their consecutive scheduled detection by SRM [110]. The elution time of each peptide should be related to standard retention time peptides to facilitate the exchange of detection methods between different mass spectrometry platforms. So far, proteome-scale SRM assay libraries have been developed for the complete *Saccharomyces cerevisiae* [111], *Mus musculus and Homo sapiens* proteomes. These libraries are deposited in the publicly accessible SRMAtlas website [112] and are continuously expanded to reach full proteome coverage. These public repositories are of great value for the scientific community and are part of a number of free computational tools which allow the generation or retrieval of suitable SRM assays from either own experimental measurements or data present in large databases such as PeptideAtlas or the Global Proteome Machine [114].

3.2.2. *Use of SRM assays*

Once SRM assays have been developed, they can be readily used for peptide detection in complex proteomes similar to the use of antibodies for protein detection. The SRM assays generated for the peptides of interest represent the set of information necessary to train the mass spectrometer for the specific detection of a peptide in protease digested samples. Each peptide is detected as a series of co-eluting fragment ion signal traces (transition traces) (Fig. 4B). Each transition represents a pair of mass-charge ratios both of the precursor and one of its fragments. To reach the specificity to conclusively detect a targeted peptide in a complex background, several transitions (usually 3–5) need to be measured for each peptide. The SRM cycle time is the sum of the individual time windows the instrument spends to monitor each transition, and this is often referred to as the dwell time of the transition. The lower the number of transitions per experiment, the higher is the sensitivity for each specific signal. This is because the instrument dedicates more time for the detection of each individual transition. On the other hand, having a too long cycle time will diminish the number of data points measured for each peptide during its chromatographic elution, and this can compromise both the detection and the accurate quantification of peptides. The choice of an optimal cycle time for the experiment and a sufficient dwell time for the detection of each transition mainly depends on the performance of the liquid chromatography and the mass spectrometric detectability of each peptide. For example, assuming that a peptide elutes with a peak width of 30 s, the cycle time of the instrument should not be higher than 2.5 s in order to have at least 12 measured points over the entire

elution of the peak. The prediction of the optimal dwell time for the detection of each transition is more difficult, as this mainly depends on the abundance and detectability of the specific peptide present in the sample under investigation. In scheduled SRM measurements, the transitions for the detection of a particular peptide are only acquired around the time of the expected elution of the peptide. The dwell time for each transition depends on the number of transitions targeted in a specific time window and therefore changes over the gradient. The detection of relatively abundant peptides normally can be achieved by setting a dwell time of 10 ms, while for less abundant species this could be increased up to 20–30 times. Based on the sensitivity one wants to reach in the SRM experiment, the number of transitions that can be monitored per analysis can range from 300 to more than 1000 transitions.

3.2.3. *Protein and peptide quantification by SRM*

SRM-based quantification is supported by a range of quantification techniques for both relative and absolute quantification. Although label-free approaches have recently been developed [115], more often quantification experiments use stable isotope labeling-based methods where a labeled reference peptide is always present in the sample and represents the standard for the quantification of the peptide [116,117]. Stable isotope labeling can be performed either *in vivo* or by chemical means. Very frequently heavy peptides are used as a reference for absolute peptide and protein quantification, but the very high costs of such peptides prevent their routine use for large-scale analyses. Several computational tools including SkyLine [118] and TIQAM [119] assist users via a visualization interface during all stages of SRM-based proteomic workflows, including target selection, transition optimization and data evaluation for quantification experiments. However, most of the mentioned tools do not support statistical analysis of the identified and quantified proteins. SRMstats is, to the best of our knowledge, the only software package available for significance analysis of large quantitative SRM experiments [120]. It is applicable both to label-based and label-free strategies and supports a variety of experimental set-ups ranging from two-condition treatments to more complex time course experiments in which many technical and biological experimental replicates are acquired.

3.2.4. *Error estimation for SRM measurements*

For both SRM assay development and use, an accurate evaluation of error rates in the peptide identification is required. During assay development, fragment ion spectra obtained for each of the synthetic peptides analyzed on triple quadruple instruments are searched and the FDR is estimated using the same strategies and tools as for shotgun mass spectrometry approaches [51]. However, peptide identifications are more straightforward in this case compared to pure discovery experiments for two reasons. First, false assignments are easily identified since the sequences of the peptides present in the samples are known and second, because the fragmentation of relatively high abundant peptides generally produces higher quality fragmentation spectra, which simplifies peptide identifications [106]. At this stage, tools like AuDIT [121] can filter out interfering transitions, or the SRM collider tool [122,123] can identify redundant transitions for a given peptide considering the specific proteome background under investigation. Estimation of the FDR for the use in SRM assays is critical for the objective analysis of large SRM datasets. SRM transition signals corresponding to the targeted peptide detected as a series of co-eluting peaks are combined into peak groups. Error analysis involves the identification of the correct peak group (the one that truly represents the targeted peptide) among the (potentially many) false peak groups that represent other analytes present in the sample.

Recently, Reiter and colleagues described an automated system called mProphet [124] that computes error rates in peptide identification experiments performed with SRM. Similarly to some decoy strategies frequently used for the analysis of shotgun datasets [55], mProphet aims at discriminating false positive peptide detections from the true positive identifications to calculate the error rate associated with the SRM experiment performed. To achieve this, false positive SRM traces are generated by measuring unspecific transitions called “decoy transitions”. These should not lead to the detection of any real peptides in the samples under investigation, and therefore reflect the level of unspecific background which has to be taken into account for an estimation of the FDR. In contrast to target-decoy strategies in the analysis of shotgun measurements where the false positive identifications are generated during the database search after the measurement, in SRM these control signals are recorded during data acquisition in the mass spectrometer and therefore need to be pre-computed prior to data acquisition.

Based also on the presence of the decoy transitions in the dataset, mProphet aims at maximizing the specificity and sensitivity of analysis by combining relevant features of SRM data, such as the chromatographic co-elution of transitions belonging to the same peptide, and the conservation of ion fragment intensity in cases where spectral libraries are available. Often several transitions are acquired for the same peptide and each peak group detected in the SRM analysis is scored.

The presence of the “decoy peak groups” allows for an estimation of the error rate in a SRM datasets and this could be expressed as the fraction of false positive peptide detections over the true positive identifications for a specific score cut-off value.

4. Applications of SRM in biological research

Tools for the accurate quantification of protein networks are essential for the understanding of cellular function, and the reproducible measurement of protein sets in clinical samples is a key task in translational research. Antibody-based methods are characterized by an exquisite reproducibility to monitor protein abundances or the occurrence of specific PTMs in complex samples over the course of several points in time, in pharmacological dosage series, or in different biological or clinical samples. For the reasons discussed above, SRM-based targeted proteomics is a particularly valuable technique for the study of cellular protein network dynamics where a defined number of proteins or PTMs can be monitored in many different cellular conditions.

4.1. SRM in proteomic studies focusing on protein levels

Enzymatic digestion of the cellular proteome produces millions of peptides both due to the intrinsic complexity of the proteomes and modification introduced by sample preparation protocols. However, only a fraction of the peptides are both indicative of the presence of specific proteins and have a good chance to be detected and identified in mass spectrometry experiments. Such peptides are normally referred to as proteotypic peptides (PTPs) [97]. Indeed, targeted mass spectrometry approaches able to detect specific PTPs have two advantages: they maximize the chances for peptide detection in a complex background and speed up analysis time by eliminating the problem of redundant peptide sampling. Recently, Picotti et al. demonstrated the great potential and the exquisite sensitivity of SRM for peptide detection [125]. By targeting PTPs, the authors were able to detect and quantify proteins over 4–5 orders of magnitude in a total yeast cell extract, resulting in the detection of proteins expressed as low as 50 copies per cell. The authors targeted 45 proteins by SRM that were never identified before by mass spectrometry, and they were able to unambiguously detect 37 of these, opening up a previously inaccessible abundance range. This work demonstrates that targeted analysis can assist classical approaches in the ambitious effort of covering parts of proteomes that are still escaping detection. Using these new protein assays, the abundance profile of a network of 45 enzymes involved in central carbohydrate metabolism was quantified over the course of a metabolic adaptation called the diauxic shift. The speed of the SRM analysis enabled the quantification of the central metabolism network in a single run, therefore the protein abundance profiles could be acquired in triplicate biological samples over 10 points in time. In a different context, the reproducibility of SRM measurements was recently exploited by the Pawson group for the characterization of protein network dynamics of the GRB2 complex [126]. The authors used a combination of affinity purification and SRM to monitor abundance changes of 90 proteins in a time course experiment after EGF stimulation. The entire network of proteins was successively quantified after stimulation of cells with six different growth factors. The speed of this approach enabled the measurement of protein fluctuations over several points in time in response to different growth factors, providing novel insights into dynamic signaling processes. Targeted mass spectrometry analysis also proved to be a useful tool to experimentally validate software predictions of biological events. Jovanovic and colleagues applied targeted protein quantification by SRM to validate computationally predicted miRNA targets in *C. elegans* [127]. The authors reproducibly quantified protein abundances for 161 proteins in biological replicates between a wild-type and a *let-7* mutant *C.elegans* strains. From the predicted protein targets, only 29 were significantly regulated by the absence of let-7 miRNA. The authors conducted further genetic experiments confirming that the list of 29 regulated proteins was enriched in let-7 miRNA targets, confirming that SRM is a valuable tool to experimentally investigate software prediction in a targeted manner.

4.2. SRM in proteomic studies focusing on the detection of protein isoforms and alternative splice variants

The utility of targeted mass spectrometry became apparent while probing the presence of protein isoforms or a SNPs in specific samples.

Recently, the defense response in conifers was investigated by SRM [128]. It is mediated by a family of enzymes called terpene synthases, which are implicated in the biosynthesis of a number of different molecules called terpenoids used by the plants to react against both physical and chemical attacks of pathogens, insects or herbivores. Terpene synthases (TPS) belong to a highly conserved family of plant enzymes and their study is complicated for different reasons. First, the biochemical function of individual TPS family members cannot be predicted based on the amino acid sequence of the protein, as few amino acid substitutions can lead to changes in the structure of the specific terpene synthesized. Second, the high sequence homology among members of this protein family makes it difficult to use available antibodies as they cross-react with the different isoforms of enzymes. Third, the conifers express terpene synthase isoforms at very low levels, therefore very sensitive detection assays need to be developed. The authors used SRM to detect proteotypic peptides specific for 19 isoforms. Five proteins were both identified and quantified by SRM in digested proteome samples. Changes in protein amount were monitored over time in biological replicate measurements, starting from the induction of the defense response in plants.

Recently, Coestenoble et al. employed SRM to measure the abundances of more than 200 proteins implicated in the amino acid and central carbon metabolism in *S. cerevisiae*, including a family of isoenzymes with highly similar amino acid sequences. The authors *a priori* selected peptides unique to each protein isoform, and subsequently quantified these by SRM in 5 different nutritional conditions to investigate how this protein network responds to the availability of different nutrients. Interestingly, the data support a functional divergence for most of the isoenzyme families [129].

The Vogelstein lab used a targeted mass spectrometry approach to identify and quantify missense mutations in the small GTPase Ras. This kind of mutation is very frequent in human cancers, but difficult to detect in proteins as they alter their sequence by changing only one amino acid. Although it is theoretically possible to detect these mutant proteins by antibodies directed against mutant epitopes, this has been very difficult to achieve in practice. The detection of splice variants or SNPs by targeted analysis represents a relatively new field in proteomics, and we expect that in the future an increasing number of studies like the one just described will populate the literature, since SRM approaches have adequate specificity and sensitivity to facilitate SNP or splice variant detection in complex protein samples.

4.3. SRM in proteomic studies focusing on post-translational protein modifications

Differential quantification of protein PTMs in two or more cellular states is not trivial. In contrast to the quantification of protein levels, where multiple peptides are normally used to

calculate the abundance, the quantification of PTMs is based only on the detection of single modified peptides. For this reason, their quantification is likely more error prone compared to the quantification of proteins, and therefore the acquisition of measurements from biological replicates is even more critical. White and colleagues monitored the tyrosine phosphorylation cascade downstream of the epidermal growth factor (EGF) receptor by SRM [130]. The authors measured the phosphorylation of 222 tyrosine residues over 8 points in time after EGF stimulation. Using an elegant end efficient hybrid approach, the authors first identified nodes in the network by shotgun analysis and subsequently used SRM to quantify phosphorylation dynamics with a reproducibility of about 90% across four biological replicates [130]. Since then, different posttranslational modifications were studied by SRM such as oxidation of cysteines, glycosylation, acetylation, methylation or ubiquitination. Furthermore, several groups exploited the specificity of SRM to determine the stoichiometry of post-translational modifications. This requires the quantification of both the modified and corresponding non-modified peptides. Jin and colleagues measured the phosphorylation stoichiometry on two different tyrosine residues including phosphorylation of the activation loop of the Lyn kinase [131]. Deregulation of this kinase has been reported to correlate with B-cell related malignancies including multiple myeloma. The sensitivity of SRM enables the quantification of the phosphorylation sites both in cell lines and tumor samples [131]. An interesting field to exploit both the reproducibility and multiplexing capacity of SRM is the study of histone modification crosstalk mechanisms. Darwanto et al. quantified about 20 different histone modifications by SRM, including acetylation, propionylation, methylation and ubiquitination in four different cell lines and replicate experiments. The analysis revealed that H2B ubiquitination is inversely correlated with methylation of Histone H3. The authors proposed the existences of a novel inhibitory loop mechanism to describe the crosstalk between these two modifications [132]. This work demonstrated the possibility to systematically and reproducibly quantify different PTMs on proteins by SRM, and we expect that in the future similar approaches will be more frequently used for the systematic characterization of signaling pathway crosstalk in the context of biological processes.

5. Conclusion and perspectives

Due to the extraordinary technological advances of the last 10–20 years, we are now able to routinely use mass spectrometry for the large-scale analysis of cellular proteomes. The cataloging phase of both proteins and post-translational modification is drawing to an end, with the rate of discovery of new proteins and their modifications reaching a level of saturation. A new and probably more exciting period for proteomics is beginning in which scientists will increasingly focus on the consistent and systematic measurement of proteins in a variety of different cellular conditions. The SRM approaches we described here are evolving into this new scenario as complementary to the shotgun analyses in the ambitious project to accurately measure proteomes with high confidence. At this time, the number of peptides that can be monitored by SRM is limited compared to current shotgun techniques. However, SRM-based approaches offer the possibility to perform experiments that would be otherwise difficult to conduct mainly because methods with adequate sensitivity and reproducibility have been lacking [98]. Further, it can also be expected that technical developments such as the targeted analysis of DIA datasets, exemplified by the SWATH-MS technique [46] will greatly increase the number of proteins targeted in a single measurement. One important point that is critical for the development of SRM assays is the synthesis of peptides. With technologies available today, the amount of peptide synthesized per array element is still orders of magnitude too large for the mass spectrometric measurements. Typically 60 nmol of each peptide are produced by the SPOT synthesis [104,105] while e.g. SRM assay development only requires less than 15 pmol per peptide, about 4000 times less. Developments of innovative approaches which are able to drastically reduce the amount of peptide synthesized could also significantly reduce the costs of peptide synthesis. The lowered costs will facilitate the creation of proteome-wide SRM assay collections for an increasing number of organisms, and this aspect becomes even more important when considering the synthesis of modified peptides such as phosphopeptides. In addition to the technical limitation that not all peptides, especially modified ones, can be synthesized efficiently, the financial investments required for their synthesis are still hindering the conception of projects such as large-scale SRM assay generation for the targeted measurements of the thousands of phosphopeptides that have already been identified in shotgun experiments and deposited in public databases. The same holds true for the synthesis of heavy reference peptides used for quantitative studies. Miniaturization of the synthesis would enable large quantitative experiments, which would facilitate the analysis of the proteome of cells that cannot routinely be metabolically labeled, such as primary cells and tissue samples. In addition to its apparent importance for SRM workflows, the synthesis of proteotypic peptides and the acquisition of their fragmentation spectra both with targeted and data-dependent workflows also opens up new perspectives for the refinement of database searching approaches employed in the analysis of proteomic data. Information like the chromatographic retention time of a peptide and the relative intensity of its fragment ions could be integrated in software tools to increase the accuracy of peptide identifications, and therefore reduce the risk of incorrect assignments. Initiatives are in progress that aim at formulating the retention time of peptides as a unitless value (independent retention time, iRT) expressed as a function of the retention time of standard peptides that are spiked into every sample before analysis [133,134]. An iRT peptide kit has been commercialized by the company Biognosys [135]. If generally adopted, this strategy will have the immediate effect of standardizing the peptide retention times in experiments performed in different laboratories worldwide allowing a publicly accessible database of measured retention times to be created. The knowledge of the elution time of peptides is beneficial both for targeted and shotgun approaches. Indeed it is possible to tune the mass spectrometer to their detection

only in the time window in which they are expected to elute, increasing both the throughput and confidence of the measurements.

Acknowledgments

We would like to thank the whole Aebersold group and Dr. Paola Picotti and Dr. Christopher A. Barnes for fruitful discussions. A.M. is the recipient of EMBO Long-Term Fellowship ALTF 386–2010. I.E. is the recipient of EMBO Short Term Fellowship ASTF 186–2011. R.A. is supported by the European Research Council (grant #ERC-2008-AdG 233226), SystemsX.ch, the Swiss initiative for systems biology (project PhosphonetX), by the European Union Seventh Framework Program PROSPECTS (Proteomics Specification in Space and Time, Grant HEALTH-F4-2008) and by Swiss National Science Foundation (Grant 3100A0-130530).

REFERENCES

[1] Wasinger VC, Cordwell SJ, Cerpa-Poljak A, Yan JX, Gooley AA, Wilkins MR, et al. Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. Electrophoresis 1995;16:1090–4.
[2] Brent MR. Genome annotation past, present, and future: how to define an ORF at each locus. Genome Res 2005;15:1777–86.
[3] Wang J, Li S, Zhang Y, Zheng H, Xu Z, Ye J, et al. Vertebrate gene predictions and the problem of large genes. Nat Rev Genet 2003;4:741–9.
[4] Zhang MQ. Computational prediction of eukaryotic protein-coding genes. Nat Rev Genet 2002;3:698–709.
[5] Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 2008;40:1413–5.
[6] Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature 2008;456:470–6.
[7] Smith CW, Valcarcel J. Alternative pre-mRNA splicing: the logic of combinatorial control. Trends Biochem Sci 2000;25:381–8.
[8] Xing Y, Lee C. Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. Nat Rev Genet 2006;7:499–509.
[9] Talavera D, Vogel C, Orozco M, Teichmann SA, de la Cruz X. The (in)dependence of alternative splicing and gene duplication. PLoS Comput Biol 2007;3:e33.
[10] Tress ML, Bodenmiller B, Aebersold R, Valencia A. Proteomics studies confirm the presence of alternative protein isoforms on a large scale. Genome Biol 2008;9:R162.
[11] Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 2003;21:577–81.
[12] Ryu GM, Song P, Kim KW, Oh KS, Park KJ, Kim JH. Genome-wide analysis to predict protein sequence variations that change phosphorylation sites or their corresponding kinases. Nucleic Acids Res 2009;37:1297–307.
[13] Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L. Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. Nature 2009;458:342–5.
[14] Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, et al. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science 2007;317:338–42.
[15] Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, et al. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature 2007;448:1050–3.
[16] Consortium GP. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061–73.
[17] Mann M, Jensen ON. Proteomic analysis of post-translational modifications. Nat Biotechnol 2003;21:255–61.
[18] Cox J, Mann M. Quantitative, high-resolution proteomics for data-driven systems biology. Annu Rev Biochem 2011;80:273–99.
[19] Lemeer S, Heck AJ. The phosphoproteomics data explosion. Curr Opin Chem Biol 2009;13:414–20.
[20] Ahrens CH, Brunner E, Qeli E, Basler K, Aebersold R. Generating and navigating proteome maps using mass spectrometry. Nat Rev Mol Cell Biol 2010;11:789–801.
[21] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001;291:1304–51.
[22] Roest Crollius H, Jaillon O, Bernot A, Dasilva C, Bouneau L, Fischer C, et al. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. Nat Genet 2000;25:235–8.
[23] Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, et al. WormBase: a comprehensive resource for nematode research. Nucleic Acids Res 2010;38:D463–7.
[24] Celniker SE, Rubin GM. The Drosophila melanogaster genome. Annu Rev Genomics Hum Genet 2003;4:89–117.
[25] Claverie JM. Gene number. What if there are only 30,000 human genes? Science (New York, NY) 2001;291:1255–7.
[26] Huang L, Guan RJ, Pardee AB. Evolution of transcriptional control from prokaryotic beginnings to eukaryotic complexities. Crit Rev Eukaryot Gene Expr 1999;9:175–82.
[27] Fickett JW, Wasserman WW. Discovery and modeling of transcriptional regulatory regions. Curr Opin Biotechnol 2000;11:19–24.
[28] Godovac-Zimmermann J, Brown LR. Perspectives for mass spectrometry and functional proteomics. Mass Spectrom Rev 2001;20:1–57.
[29] Alberts B. The cell as a collection of protein machines: Preparing the next generation of molecular biologists. Cell 1998;92:291–4.
[30] Beck M, Malmstrom JA, Lange V, Schmidt A, Deutsch EW, Aebersold R. Visual proteomics of the human pathogen Leptospira interrogans. Nat Methods 2009;6:817–23.
[31] Gstaiger M, Aebersold R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. Nat Rev Genet 2009;10:617–27.
[32] Choudhary C, Mann M. Decoding signalling networks by mass spectrometry-based proteomics. Nat Rev Mol Cell Biol 2010;11:427–39.
[33] ProteinAtlas. http://www.proteinatlas.org/ Accessed 2012 Apr 26.
[34] Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol 2010;28:1248–50.
[35] Uhlen M, Hober S. Generation and validation of affinity reagents on a proteome-wide level. J Mol Recognit 2009;22:57–64.
[36] Holm A, Wu W, Lund-Johansen F. Antibody array analysis of labelled proteomes: how should we control specificity? N Biotechnol 2011:1–8.
[37] Stoevesandt O, Taussig MJ. Affinity reagent resources for human proteome detection: initiatives and perspectives. Proteomics 2007;7:2738–50.
[38] Colwill K, Graslund S. A roadmap to generate renewable protein binders to the human proteome. Nat Methods 2011;8:551–8.
[39] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198–207.

[40] Domon B, Aebersold R. Options and considerations when selecting a quantitative proteomics strategy. Nat Biotechnol 2010;28:710–21.
[41] Yamashita M, Fenn JB. Electrospray ion source. Another variation on the free-jet theme — The Journal of Physical Chemistry (ACS Publications). J Phys Chem 1984;88(20):4451–9.
[42] Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science 1989;246:64–71.
[43] Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem Oct. 15 1988;60(20):2299–301.
[44] Lange V, Picotti P, Domon B, Aebersold R. Selected reaction monitoring for quantitative proteomics: a tutorial. Mol Syst Biol 2008;4:222.
[45] Geiger T, Cox J, Mann M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. Mol Cell Proteomics 2010;9:2252–61.
[46] Gillet LC, Navarro P, Tate S, Roest H, Selevsek N, Reiter L, et al. Targeted data extraction of the MS/MS spectra generated by data independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics in press.
[47] Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Mol Cell Proteomics 2006;5:144–56.
[48] Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Methods 2004;1:39–45.
[49] Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res 2003;2:43–50.
[50] Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 2007;4:207–14.
[51] Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 2007;4:787–97.
[52] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74:5383–92.
[53] Claassen M, Aebersold R, Buhmann JM. Proteome coverage prediction for integrated proteomics datasets. J Comput Biol 2011;18:283–93.
[54] Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, et al. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol Cell Proteomics 2009;8:2405–17.
[55] Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 2003;75:4646–58.
[56] Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics 2011;10:M111 (007690).
[57] Picotti P, Aebersold R, Domon B. The implications of proteolytic background for shotgun proteomics. Mol Cell Proteomics 2007;6:1589–98.
[58] Beck M, Schmidt A, Malmstroem J, Claassen M, Ori A, Szymborska A, et al. The quantitative proteome of a human cell line. Mol Syst Biol 2011;7:549.
[59] Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, et al. Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol 2011;7:548.
[60] Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 2008;26:1367–72.
[61] Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 2011;10:1794–805.
[62] Abu-Farha M, Elisma F, Zhou H, Tian R, Zhou H, Asmer MS, et al. Proteomics: from technology developments to biological applications. Anal Chem 2009;81:4585–99.
[63] Gatlin CL, Eng JK, Cross ST, Detter JC, Yates III JR. Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. Anal Chem 2000;72:757–63.
[64] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein–protein interaction network. Nature 2005;437:1173–8.
[65] Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, et al. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol Cell Proteomics 2006;5:652–70.
[66] Bunger MK, Cargile BJ, Sevinsky JR, Deyanova E, Yates NA, Hendrickson RC, et al. Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data. J Proteome Res 2007;6:2331–40.
[67] Brosch M, Saunders GI, Frankish A, Collins MO, Yu L, Wright J, et al. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. Genome Res 2011;21:756–67.
[68] Ezkurdia I, Del Pozo A, Frankish A, Rodriguez JM, Harrow J, Ashman K, et al. Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. Mol Biol Evol in press.
[69] Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP. Discovery and revision of Arabidopsis genes by proteogenomics. Proc Natl Acad Sci U S A 2008;105:21034–8.
[70] Severing EI, van Dijk AD, van Ham RC. Assessing the contribution of alternative splicing to proteome diversity in Arabidopsis thaliana using proteomics data. BMC Plant Biol 2011;11:82.
[71] Chang KY, Muddiman DC. Identification of alternative splice variants in *Aspergillus flavus* through comparison of multiple tandem MS search algorithms. BMC Genomics 2011;12:358.
[72] Ning K, Nesvizhskii AI. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. BMC Bioinformatics 2010;11(Suppl. 11):S14.
[73] Adamidi C, Wang Y, Gruen D, Mastrobuoni G, You X, Tolle D, et al. De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. Genome Res 2011;21:1193–200.
[74] Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, et al. Defining the transcriptome and proteome in three functionally different human cell lines. Mol Syst Biol 2010;6:450.
[75] Wang X, Slebos RJ, Wang D, Halvey PJ, Tabb DL, Liebler DC, et al. Protein identification using customized protein sequence databases derived from RNA-Seq data. J Proteome Res 2012;11:1009–17.
[76] Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, Potthast F, et al. A high-quality catalog of the *Drosophila melanogaster* proteome. Nat Biotechnol 2007;25:576–83.

[77] Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, et al. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol 2005;6:R9.
[78] Jaffe JD, Berg HC, Church GM. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 2004;4:59–77.
[79] Kuster B, Mortensen P, Andersen JS, Mann M. Mass spectrometry allows direct identification of proteins in large genomes. Proteomics 2001;1:641–50.
[80] Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, Pain A, Sauerwein RW, et al. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. Nature 2002;419:537–42.
[81] Merrihew GE, Davis C, Ewing B, Williams G, Kall L, Frewen BE, et al. Use of shotgun proteomics for the identification, confirmation, and correction of *C. elegans* gene annotations. Genome Res 2008;18:1660–9.
[82] Pandey A, Lewitter F. Nucleotide sequence databases: a gold mine for biologists. Trends Biochem Sci 1999;24:276–80.
[83] Loevenich SN, Brunner E, King NL, Deutsch EW, Stein SE, Aebersold R, et al. The *Drosophila melanogaster* PeptideAtlas facilitates the use of peptide data for improved fly proteomics and genome annotation. BMC Bioinformatics 2009;10:59.
[84] de Souza GA, Malen H, Softeland T, Saelensminde G, Prasad S, Jonassen I, et al. High accuracy mass spectrometry analysis as a tool to verify and improve gene annotation using *Mycobacterium tuberculosis* as an example. BMC Genomics 2008;9:316.
[85] Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol 2006;7(Suppl. 1):S4.1–9.
[86] Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B. PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics 2004;4:1551–61.
[87] Gnad F, Gunawardena J, Mann M. PHOSIDA 2011: the posttranslational modification database. Nucleic Acids Res 2011;39:D253–60.
[88] Bodenmiller B, Malmstrom J, Gerrits B, Campbell D, Lam H, Schmidt A, et al. PhosphoPep — a phosphoproteome resource for systems biology research in Drosophila Kc167 cells. Mol Syst Biol 2007;3:139.
[89] UniProt. http://www.uniprot.org/ Accessed 2012 Apr 26.
[90] Eisenhaber B, Eisenhaber F. Prediction of posttranslational modification of proteins from their amino acid sequence. Methods Mol Biol 2010;609:365–84.
[91] Obenauer JC, Cantley LC, Yaffe MB. Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 2003;31:3635–41.
[92] Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, Heyen JW, et al. Identification, analysis, and prediction of protein ubiquitination sites. Proteins 2010;78:365–80.
[93] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997;268:78–94.
[94] Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 2004;20:2878–9.
[95] Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 2005;33:W451–4.
[96] Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, et al. Widespread RNA and DNA sequence differences in the human transcriptome. Science 2011;333:53–8.
[97] Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, et al. Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol 2007;25:125–31.
[98] Edwards A, Isserlin R, Bader G, Frye S. Too many roads not taken. Nature Feb. 10 2011;470(7333):163–5.
[99] Nygren PA. Alternative binding proteins: affibody binding proteins developed from a small three-helix bundle scaffold. FEBS J 2008;275:2668–76.
[100] Du J, Bernasconi P, Clauser KR, Mani DR, Finn SP, Beroukhim R, et al. Bead-based profiling of tyrosine kinase phosphorylation identifies SRC as a potential target for glioblastoma therapy. Nat Biotechnol 2009;27:77–83.
[101] Anderson NL, Anderson NG, Haines LR, Hardie DB, Olafson RW, Pearson TW. Mass spectrometric quantitation of peptides and proteins using stable isotope standards and capture by anti-peptide antibodies (SISCAPA). J Proteome Res 2004;3:235–44.
[102] Schwenk JM, Igel U, Neiman M, Langen H, Becker C, Bjartell A, et al. Toward next generation plasma profiling via heat-induced epitope retrieval and array-based assays. Mol Cell Proteomics 2010;9:2497–507.
[103] Schmidt A, Gehlenborg N, Bodenmiller B, Mueller LN, Campbell D, Mueller M, et al. An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. Mol Cell Proteomics 2008;7:2138–50.
[104] Wenschuh H, Volkmer-Engert R, Schmidt M, Schulz M, Schneider-Mergener J, Reineke U. Coherent membrane supports for parallel microsynthesis and screening of bioactive peptides. Biopolymers 2000;55:188–206.
[105] Frank R, Overwin H. SPOT synthesis. Epitope analysis with arrays of synthetic peptides prepared on cellulose membranes. Methods Mol Biol 1996;66:149–69.
[106] Picotti P, Rinner O, Stallmach R, Dautel F, Farrah T, Domon B, et al. High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. Nat Methods 2010;7:43–6.
[107] Nesvizhskii AI. Protein identification by tandem mass spectrometry and sequence database searching. Methods Mol Biol 2007;367:87–119.
[108] Maclean B, Tomazela DM, Abbatiello SE, Zhang S, Whiteaker JR, Paulovich AG, et al. Effect of collision energy optimization on the measurement of peptides by selected reaction monitoring (SRM) mass spectrometry. Anal Chem 2010;82:10116–24.
[109] Holstein Sherwood CA, Gafken PR, Martin DB. Collision energy optimization of b- and y-ions for multiple reaction monitoring mass spectrometry. J Proteome Res 2011;10:231–40.
[110] Stahl-Zeng J, Lange V, Ossola R, Eckhardt K, Krek W, Aebersold R, et al. High sensitivity detection of plasma proteins by multiple reaction monitoring of N-glycosites. Mol Cell Proteomics 2007;6:1809–17.
[111] Picotti P, Lam H, Campbell D, Deutsch EW, Mirzaei H, Ranish J, et al. A database of mass spectrometric assays for the yeast proteome. Nat Methods 2008;5:913–4.
[112] SRMAtlas. http://www.srmatlas.org/ Accessed 2012 Apr 26.
[113] Deutsch EW. The PeptideAtlas Project. Methods Mol Biol 2010;604:285–96.
[114] Cham Mead JA, Bianco L, Bessant C. Free computational resources for designing selected reaction monitoring transitions. Proteomics 2010;10:1106–26.
[115] Eissler CL, Bremmer SC, Martinez JS, Parker LL, Charbonneau H, Hall MC. A general strategy for studying multisite protein phosphorylation using label-free selected reaction monitoring mass spectrometry. Anal Biochem 2011;418:267–75.
[116] Gevaert K, Impens F, Ghesquiere B, Van Damme P, Lambrechts A, Vandekerckhove J. Stable isotopic labeling in proteomics. Proteomics 2008;8:4873–85.
[117] Pan S, Aebersold R, Chen R, Rush J, Goodlett DR, McIntosh MW, et al. Mass spectrometry based targeted protein

quantification: methods and applications. J Proteome Res 2009;8:787–97.
[118] MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 2010;26:966–8.
[119] Lange V, Malmstrom JA, Didion J, King NL, Johansson BP, Schafer J, et al. Targeted quantitative analysis of Streptococcus pyogenes virulence factors by multiple reaction monitoring. Mol Cell Proteomics 2008;7:1489–500.
[120] Chang CY, Picotti P, Huttenhain R, Heinzelmann-Schwarz V, Jovanovic M, Aebersold R, et al. Protein Significance Analysis in Selected Reaction Monitoring (SRM) measurements. Mol Cell Proteomics 2012;11:M111 (014662).
[121] Abbatiello SE, Mani DR, Keshishian H, Carr SA. Automated detection of inaccurate and imprecise transitions in peptide quantification by multiple reaction monitoring mass spectrometry. Clin Chem 2010;56:291–305.
[122] SRMCollider. http://www.srmcollider.org/ Accessed 2012 Apr 26.
[123] Rost HL, Malmstrom L, Aebersold R. A computational tool to detect and avoid redundancy in selected reaction monitoring. Mol Cell Proteomics in press.
[124] Reiter L, Rinner O, Picotti P, Huttenhain R, Beck M, Brusniak MY, et al. mProphet: automated data processing and statistical validation for large-scale SRM experiments. Nat Methods 2011;8:430–5.
[125] Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R. Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. Cell 2009;138:795–806.
[126] Bisson N, James DA, Ivosev G, Tate SA, Bonner R, Taylor L, et al. Selected reaction monitoring mass spectrometry reveals the dynamics of signaling through the GRB2 adaptor. Nat Biotechnol 2011;29:653–8.
[127] Jovanovic M, Reiter L, Picotti P, Lange V, Bogan E, Hurschler BA, et al. A quantitative targeted proteomics approach to validate predicted microRNA targets in *C. elegans*. Nat Methods 2010;7:837–42.
[128] Zulak KG, Lippert DN, Kuzyk MA, Domanski D, Chou T, Borchers CH, et al. Targeted proteomics using selected reaction monitoring reveals the induction of specific terpene synthases in a multi-level study of methyl jasmonate-treated Norway spruce (*Picea abies*). Plant J 2009;60:1015–30.
[129] Costenoble R, Picotti P, Reiter L, Stallmach R, Heinemann M, Sauer U, et al. Comprehensive quantitative analysis of central carbon and amino-acid metabolism in *Saccharomyces cerevisiae* under multiple conditions by targeted proteomics. Mol Syst Biol 2011;7:464.
[130] Wolf-Yadlin A, Hautaniemi S, Lauffenburger DA, White FM. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. Proc Natl Acad Sci U S A 2007;104:5860–5.
[131] Jin LL, Tong J, Prakash A, Peterman SM, St-Germain JR, Taylor P, et al. Measurement of protein phosphorylation stoichiometry by selected reaction monitoring mass spectrometry. J Proteome Res 2010;9:2752–61.
[132] Darwanto A, Curtis MP, Schrag M, Kirsch W, Liu P, Xu G, et al. A modified "cross-talk" between histone H2B Lys-120 ubiquitination and H3 Lys-79 methylation. J Biol Chem 2010;285:21868–76.
[133] Escher C, Reiter L, Maclean B, Ossola R, Herzog F, Chilton J, et al. Using iRT, a normalized retention time for more targeted measurement of peptides. Proteomics Apr. 2012;12(8):1111–21.
[134] Skyline. iRT Tutorial. https://skyline.gs.washington.edu/labkey/wiki/home/software/Skyline/page.view?name=tutorial_irt Accessed 2012 Apr 26.
[135] Biognosys. http://www.biognosys.ch/ Accessed 2012 Apr 26.
[136] Bendall SC, Simonds EF, Qiu P, Amir el AD, Krutzik PO, Finck R, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 2011;332:687–96.
[137] Salamov AA, Solovyev VV. Ab initio gene finding in Drosophila genomic DNA. Genome Res 2000;10:516–22.
[138] Catrein I, Herrmann R. The proteome of Mycoplasma pneumoniae, a supposedly "simple" cell. Proteomics 2011;11:3614–32.
[139] Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O. The comprehensive microbial resource. Nucleic Acids Res 2001;29:123–5.
[140] Sun N, Pan C, Nickell S, Mann M, Baumeister W, Nagy I. Quantitative proteome and transcriptome analysis of the archaeon *Thermoplasma acidophilum* cultured under aerobic and anaerobic conditions. J Proteome Res 2010;9:4839–50.
[141] Becher D, Hempel K, Sievers S, Zuhlke D, Pane-Farre J, Otto A, et al. A proteomic view of an important human pathogen — towards the quantification of the entire *Staphylococcus aureus* proteome. PLoS One 2009;4:e8176.
[142] de Godoy LM, Olsen JV, Cox J, Nielsen ML, Hubner NC, Frohlich F, et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 2008;455:1251–4.
[143] Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C, et al. Genetic and physical maps of *Saccharomyces cerevisiae*. Nature 1997;387:67–73.
[144] Borchert N, Dieterich C, Krug K, Schutz W, Jung S, Nordheim A, et al. Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models. Genome Res 2010;20:837–46.
[145] Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S, et al. Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. Science 2008;320:938–41.