DOI 10.1002/pmic.201400168 *Proteomics* 2014, *14*, 2660–2675

REVIEW

Proteogenomics in microbiology: Taking the right turn at the junction of genomics and proteomics

Veronika Kucharova and Harald G. Wiker

Department of Clinical Science, The Gade Research Group for Infection and Immunity, University of Bergen, Norway

Received: April 28, 2014
Revised: August 18, 2014
Accepted: September 23, 2014

High-accuracy and high-throughput proteomic methods have completely changed the way we can identify and characterize proteins. MS-based proteomics can now provide a unique supplement to genomic data and add a new level of information to the interpretation of genomic sequences. Proteomics-driven genome annotation has become especially relevant in microbiology where genomes are sequenced on a daily basis and limitations of an *in silico* driven annotation process are well recognized. In this review paper, we outline different strategies on how one can design a proteogenomic experiment, for example on genome-sequenced (synonymous proteogenomics) versus unsequenced organisms (ortho-proteogenomics) or with the aid of other "omic" data such as RNA-seq. We touch upon many challenges that are encountered during a typical proteogenomic study, mostly concerning bioinformatics methods and downstream data analysis, but also related to creation and use of sequence databases. A large list of proteogenomic case studies of different microorganisms is provided to illustrate the mapping of MS/MS-derived peptide spectra to genomic DNA sequences. These investigations have led to accurate determination of translational initiation sites, pointed out eventual read-throughs or programmed frameshifts, detected signal peptide processing or other protein maturation events, removed questionable annotation assignments, and provided evidence for predicted hypothetical proteins.

Keywords:
Genome annotation / Microbiology / MS/MS / Synonymous proteogenomics

1 From top-down to bottom-up analysis

Before MS-based methods were developed for large-scale protein analysis, protein characterization was dependent on purification of single protein species from complex samples. This could require substantial quantities of starting material and a method to monitor the protein amount and/or its activity through various purification steps. Once a pure protein was obtained, one could accurately determine the protein sequence using the traditional, but nevertheless laborious, N-terminal Edman degradation method. Two-dimensional PAGE represented a major improvement for proteomic investigations. It is a technique with efficient separation of proteins based on differences in isoelectric points in the first dimension and molecular mass in the second dimension [1]. Individual protein spots from the gel can be picked for digestion by a protease and subsequent measurement of peptide masses by MS. MALDI-TOF instruments were initially used for the mass measurements. Over the years it has been recognized that 2D-PAGE performs quite poorly in the analysis of hydrophobic proteins and the reproducibility of experiments is often compromised. On the other hand, the technique is relatively robust and can also be used to analyze intact proteins at high resolution.

The use of two mass spectrometers in tandem (MS/MS) introduced a new dimension to the field by employing two stages of mass analysis [2]. From the mass spectrum that is produced by using the first MS, a single (precursor) mass of a given compound can be selected. These mass-selected

Correspondence: Professor Harald G. Wiker, Department of Clinical Science, The Gade Research Group for Infection and Immunity, University of Bergen, Laboratory Building, 5th Floor, Jonas Lies Vei 87, 5021 Bergen, Norway
E-mail: harald.wiker@k2.uib.no
Fax: +47-55974689

Abbreviations: **ABPP**, activity-based protein profiling; **COFRADIC**, combined fractional diagonal chromatography; **NCBI**, National Center for Biotechnology Information; **TSS**, translational start site

ions are next fragmented and the resulting fragment (product) ions are analyzed in the second MS. Such setup allows for obtaining more accurate chemical structure related information and for more selective quantitation of target compounds in complex mixtures. The mass of a peptide can then be determined with very high accuracy, down to 1 part per million (ppm) or even better. The identification of a peptide in MS/MS is based on subjecting the peptide to stress-induced fragmentation and then using the measured mass of fragments as a fingerprint for the peptide. In a bottom-up proteomic approach, trypsin is usually used to digest a cell lysate. The resulting peptides are subsequently separated on an LC column coupled to the MS/MS instrument. The combined information of the tryptic peptide (from the first MS) and the fragments masses (from the second MS) gives reliable sequence information when matched against a sequence database. Although comprehensive annotated protein databases (e.g. National Center for Biotechnology Information (NCBI) protein resources, UniProtKB/Swiss-Prot, PROSITE) are reliable and frequently used sources, a common approach in microbiology is to use custom-made databases containing translated nucleotide sequences from the genome of the investigated organism. The concept of searching uninterpreted mass spectra against a translated nucleotide database was first introduced in 1995 by Yates et al. [3] and was quickly followed up by development of various search engines to match the experimental data against tryptic peptides generated in silico from entries in a database [4–8].

1.1 Proteogenomics versus de novo proteomics

Proteogenomics is often defined as using MS to identify proteins predicted from genomic sequences and to use this information to improve the genome annotation [9–11]. In principle, most proteomic investigations in microbiology rely heavily on genomic sequence information that acts as the blueprint for protein production. Programs such as MASCOT [5] and MaxQuant [8] used to interpret MS-derived experimental data do so in the context of translated genomic sequences. As soon as a protein has been observed by one or more peptides, this is of value for genomic annotations. Proteomic analysis is therefore much more than just confirmatory and can be used to both correct and add missing information to the genomic annotations (see, e.g., case studies listed in Table 1). From the above reasoning, it is inferred that bottom-up proteomic investigations, which in some or another way utilize and interact with genomic sequence data, are essentially proteogenomic. Searching MS/MS data against a sequence database is the dominant method for peptide sequencing by MS; however, de novo sequencing of peptides by MS can also be done [12, 13]. In the latter case, the sequencing is performed without any prior knowledge of the amino acid sequence and the observed mass spectra are used for direct reconstruction of the protein sequence without guidance from information in protein sequence databases. Peptide de novo sequencing is, for example, valuable for the identification of proteins without any existing homolog in the database (e.g. characterization of different protein isoforms [14]) and for metaproteome analyses of microbial communities [15, 16]. For such top-down proteomics approaches, proper assembly of de novo sequence data has been met with major challenges, such as low protein sequence coverage. Nevertheless, the top-down approach may be more widely applied in the future, once the technical issues are resolved. Although this review focuses primarily on the bottom-up MS approach, which is the most common, there is an example of a relevant top-down proteomic study [17].

1.2 Evaluation and validation of genomic annotations

The progress in determination of DNA sequences since the Sanger sequencing method was published in 1977 has been breathtaking. Because of the technical advances in next-generation sequencing technologies [18], the cost of sequencing for a bacterial genome is no longer a limiting factor. It can be foreseen that within a few years, a semiautomatic sequencer will be used on a routine basis in many microbiological laboratories [19]. Still, once the genome of an organism is sequenced and assembled, the critical step lies in identification of gene-coding regions and in establishing their annotations [20]. The functional content of the genome is inferred through computational analysis that usually involves recognizing sequence similarity between an anonymous query and characterized matching sequence [21]. Gene annotators are typically based on Basic Local Alignment Search Tool (BLAST) [22] (e.g. IMG [23] and RAST [24]) or probabilistic models such as hidden Markov models (e.g. Glimmer [25], GeneMark [26], and HMMer [27]). Recent evaluation of 54 methods for gene annotation identified a considerable need for improvement of currently available tools [28]. The main conclusion of the study was that second-generation annotation tools that combine a variety of biological and computational concepts outperform the first-generation alignment-based methods. There is also a common opinion that the use of subsequent manual curation of the genome sequence can provide the best annotations [29]. On the other hand, such curation is often slower and more costly, and it has also been considered as insufficient for genome annotation by some [30].

According to the NCBI, the number of sequenced bacterial genomes is currently over 27 thousand (August 2014). From this large number, almost 3400 genomes are listed as complete and the rest are in the draft stage (Fig. 1). This disproportion indicates that the processing of large amounts of sequencing data in order to yield useful results is a rather complex and laborious task. Several in-depth genomic and proteomic studies have evaluated the accuracy of in silico annotation methods and pointed out their limitations [29, 31–36]. For example, a comparison of gene annotations

Table 1. List of microbiological proteogenomic studies

Organism	Genome size (Mbp)	Proteins identified	Proteome coverage (%)	MS Instrument[a), h)] (total acquired spectra)	Protein FDR[g), h)] (%)	Search algorithm	Study
Acholeplasma laidlawii	1.50	803	58	1100 Series LC/MSD Trap XCT Ultra[b)]	NA	MASCOT	[54]
Aspergillus niger	33.90–37.10	NA	NA	QTOF I[e)] (19 628)	2.00	MASCOT	[40]
Bartonella henselae	1.93	1250	85	LTQ Orbitrap XL[c)] or LTQ FT Ultra[c)]	1.00	MASCOT,MS-GF+	[70]
Bradyrhizobium japonicum[i)]	9.10	2654	32	(621 176) [i)]	1.00	OMSSA, X!Tandem, InsPecT, MassWiz	[143]
Brucella abortus	3.28	621	20	NA	NA	MASCOT	[144]
Candida glabrata	12.34	4421	83	LTQ Orbitrap Velos[c)]	1.00	SEQUEST, X!Tandem	[41]
Campylobacter concisus	1.80	1369	73	Orbitrap Velos[c)]	NA	MASCOT	[53]
Cryptococcus neoformans	18.92	3674	52	LTQ Orbitrap Velos[c)]	1.00	Proteome Discoverer, SEQUEST, MASCOT	[43]
Deinococcus deserti	3.86	1348	40	LTQ-Orbitrap XL[c)] (264 2519)	0.20	MASCOT	[52, 67]
Escherichia coli K12	4.60	2600	32	Orbitrap Velos[c)] (1 941 724)	1.00	MaxQuant, MASCOT	[48]
Helicobacter pylori	1.67	1115	71	LTQ Orbitrap XL ETD[c)]	1.00	MaxQuant, MASCOT, X!Tandem	[122]
Leishmania donovani	32.40	3999	50	LTQ-Orbitrap Velos ETD[c)]	1.00	SEQUEST, MASCOT	[145]
Leishmania major	32.80	3613	43	LTQ-Orbitrap Velos ETD[c)]	1.00	SEQUEST, MASCOT	[42]
Mycobacterium leprae	3.27	1046	65	LTQ Orbitrap[c)]	0.25	MASCOT	[146]
Mycobacterium smegmatis	6.98	901	13	LCQ DecaXP+(~825 000)	5.00	BioWorks, DTASelect	[147]
		946	14	HCT Ultra ion trap	1.00	MASCOT	[66]
Mycobacterium tuberculosis	4.41	3176	79	LTQ Orbitrap Velos[c)] (~1.8 million)	1.00	MASCOT, SEQUEST, MassWiz	[148]
		2561	65	LTQ Orbitrap[c)]	1.00	MASCOT, MaxQuant	[58]
Mycoplasma mobile	0.78	557	88	LCQ DecaXP Plus[c)]	NA	SEQUEST	[51]
Mycoplasma pneumoniae	0.82	557	81	LCQ Classic ion trap[c)]	NA	SEQUEST	[50]
Plasmodium falciparum	22.9	1289	23	QSTAR[d)] (200 000)	NA	MASCOT	[149]
Rhodopseudomonas palustris	5.47	2814	58	LTQ[c)]	5.00	DBDigger, MASPIC	[150]
Ruegeria pomeroyi	4.60	2006	47	LTQ-Orbitrap XL[c)] (~1.1 million)	0.20	MASCOT	[65]
Saccharopolyspora erythraea	8.21	1139	16	QSTAR-Elite[d)]	5.00	Protein Pilot, Paragon Algorithm	[73]

Table 1. Continued

Organism	Genome size (Mbp)	Proteins identified	Proteome coverage (%)	MS Instrument[a), h)] (total acquired spectra)	Protein FDR[g), h)] (%)	Search algorithm	Study
Salmonella enterica	4.87	2118	44	LTQ Orbitrap[c)]	1.00	SEQUEST	[17]
Shewanella frigidimarina	4.85	1744	43	LCQ[c)] (~0.955 million)	1.00	InsPecT	[60]
Shewanella oneidensis	5.13	1992	47	FT-ICR, LCQ[c)] (~14.5 million)	1.00 5.00% (PTMs)	InsPecT	[60, 99]
Shewanella putrefaciens	4.65	1625	41	LCQ[c)] (~0.768 million)	1.00	InsPecT	[60]
Shigella flexneri	4.83	823	19	Ultraflex III MALDI-TOF/TOF[f)]	<1.00	Biotools, MASCOT	[151]
Streptococcus pyogenes	1.88	567	31	LTQ Orbitrap XL[c)] (~7000)	8.00	MASCOT	[152]
Streptococcus suis	2.10	28 surface proteins	~1	LTQ[c)]	NA	MASCOT	[153]
Synechocystis sp. PCC6803	3.95	1462	41	QTOF[b)]	2.70	MASCOT	[154]
Thermotoga maritima	1.87	1077	57	LTQ Orbitrap Velos[c)]	NA	SEQUEST	[56]
Yersiniae pestis CO92	4.83	1682	41	LTQ Orbitrap Velos[c)] (~1.6 million)	0.80	SEQUEST	[71]
Yersiniae pestis	4.73	1773	44				
Yersiniae pseudotuberculosis	4.77	1603	38				
Yersiniae pestis KIM	4.70	1302	31	LTQ[c)] (~15 million)	5.00	Inspect, PepNovo	[64]

Abbreviations: ETD: electron transfer dissociation; FDR: false discovery rate; LTQ: linear trap quadrupole; MSD: mass selective detector; QTOF: quadrupole TOF.
[a)] MS instruments manufacturers: b) Agilent, c) Thermo Fischer Scientific, d) ABSciex, e) Waters, f) Bruker Daltonics.
[g)] FDR—proportion of incorrect identifications among all identifications judged correctly.
[h)] NA: not available.
[i)] MS data obtained from the PRIDE repository.

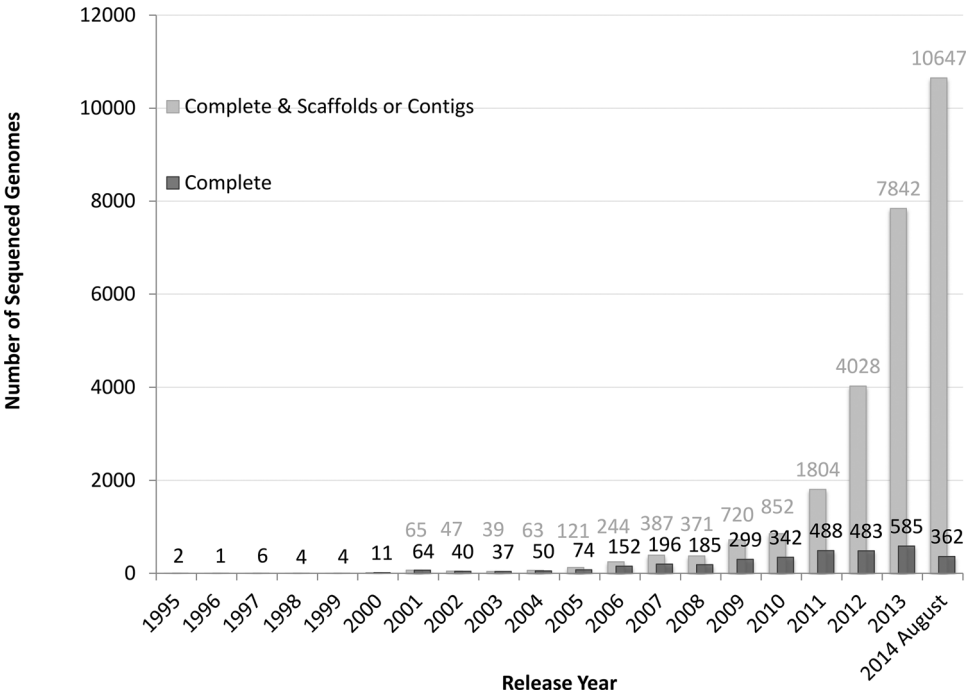

Figure 1. Number of prokaryotic genomes deposited at Entrez Genome database at NCBI. Only assembled genomes designed as complete or draft (scaffolds or contigs) were considered. Data labels above bars show the number of genomes sequenced in the respective year. Data were collected in August 2014.

obtained either from the Sanger Institute or J. Craig Venter Institute for the genomic sequence of *Mycobacterium tuberculosis* H37Rv showed 15 and 50% difference in the gene annotations and in the start codon assignments, respectively [29]. Similarly, a comparison of three different gene-calling platforms found that about half of almost 3000 predicted protein-coding regions of *Halorhabdus utahensis* were inconsistent across the automatic annotations [36].

1.3 Proteogenomics of eukaryotes

Proteogenomic mapping has mainly been used for the validation (and improvement) of the structural annotation of small prokaryotic genomes; nevertheless, the integration of large-scale proteomics data is gaining popularity also in eukaryotic genome annotation projects [37–43]. The larger number of conducted proteogenomic studies of prokaryotes, when compared to eukaryotes, is probably due to the differences in genome complexity between eukaryotes and prokaryotes. Genomic organization of a prokaryote genome is much more economical than that of a eukaryote. Prokaryotic genes are tightly packed on a single chromosome leaving very little space in between genes. Noncoding sequences account for an average of 12% of a prokaryotic genome, while in a eukaryotic genome up to 98% of the genetic material might not code for functional proteins [44]. The tight genome organization in prokaryotes is reflected in the arrangement of most genes into polycistronic operons or clusters of genes that are governed by a single promoter. The situation is more complex for eukaryotes where the DNA sequence for a given gene is organized into coding exons and noncoding introns. Eukaryotic nascent pre-mRNA transcripts therefore must undergo splicing where the introns are removed and exons join. To correctly define gene boundaries and the respective protein products usually requires a great deal of effort in both genome annotation and proteogenomic studies and an account of commonly encountered difficulties has been previously provided [45]. Events such as alternative splicing, exon skipping, and truncation or extension at the introns 5′ or 3′ ends cannot be accurately predicted by bioinformatics methods and proteomics therefore has become invaluable for validation and refinement of eukaryotic genome annotations. However, in this review, we provide systematic account of proteogenomic studies primarily in prokaryotes, with a few exceptions of medically relevant single-cell eukaryotic model organisms.

2 Proteomics-driven annotation in microbiology

Ideally one may think that genome annotations should be completely proteomics driven. Indeed, large-scale proteomics data were early recognized as a potentially rich source for validation and reevaluation of genome annotations [46]. Proteomics technologies have now reached a level where they can provide a platform for annotation of genomes that is mainly proteomics driven combined with gap filling using theoretical interpretation. Using six-frame translational data, the entire coding repertoire in a genome should in principle be represented. However, the augmented nature of a six-frame database brings certain difficulties for correct statistical interpretation of the data. In such a database, for every target

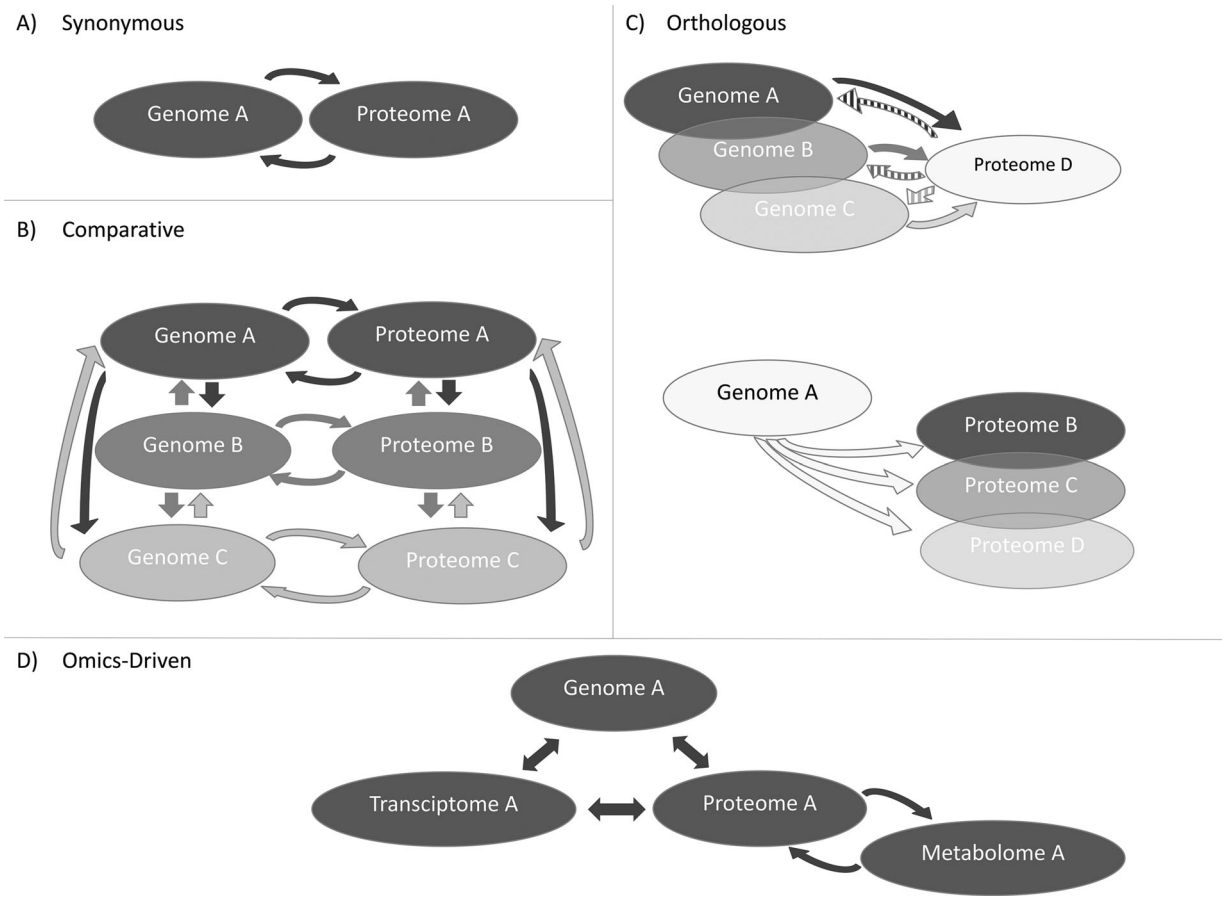

Figure 2. Strategies for a proteogenomic experiment. (A) In a synonymous proteogenomic study, genome sequencing and MS-based proteomics is performed on the same isolate of a species. (B) The joint power of comparative genomics and proteomics can be employed for annotation validation and refinement of several members of the same species. (C) Ortho-proteogenomics also utilize homology-based searches against protein-coding genes from closely related organisms. This strategy allows for proteome characterization of an unsequenced clinical or environmental isolate and investigation of multiple proteomes at once. (D) A comprehensive approach for genome annotation refinement takes advantage of large–scale data from several omic fields.

protein sequence there exist five false targets, which in turn leads to low S/N and reduced sensitivity under the search criteria needed to maintain a low false discovery rate [47, 48]. In order to avoid too many entries, one can apply a size cut off, for example excluding proteins of less than 20 amino acids. In addition, redundancy removal guidelines have been proposed in order to choose the most likely reading frame [49]. The idea of mapping peptides back to their source genome in order to validate the existing annotation was introduced 10 years ago [50]. Since then different proteogenomic strategies have been described and these are discussed below (Fig. 2). A list of different microorganisms for which the mapping of peptides onto the corresponding genome has been applied is given in Table 1.

2.1 Synonymous proteogenomics

Synonymous proteogenomics may be defined as proteomics performed on exactly the same isolate or strain that has been sequenced (Fig. 2A). High-throughput sequencing of a bacterial genome has become quite efficient and the possibilities to combine proteomic and genomic experiments are feasible [17, 51–56]. This will be the ideal approach for future proteogenomics, as the likelihood of inconsistent characteristics will be kept to a minimum. However, it is not always the case that the genomic sequence matches exactly the sequence present in the model organism chosen for a proteomics study. Some bacteria are more prone to genomic changes than other bacteria, specifically, *Neisseria meningitidis* and *Helicobacter pylori* are examples of bacteria that tend to change frequently through rearrangements of genetic elements and mutations [57]. Moreover, bacterial strains tend to change when subcultured in the laboratory, and strain collections have rigorous routines to keep such changes at a minimal level. It basically implies using the seed-lot system where a batch is frozen down in numerous aliquots, and there are rules as to how many passages one may apply until it is necessary to go back to the seed-lot for a new sample. When working with

proteogenomics, it is in principal necessary to know the history of the sequenced strain as well as the sample one is working with to create the proteomic data. Historical records of bacterial strains are therefore invaluable, but are not always available. For example, when working on the proteome of *M. tuberculosis* H37Rv, it became clear that the original genomic sequence was performed on a sample of this strain, suggesting that the strain was not recently acquired from a repository. When searching our proteomic data obtained using the American Type Culture Collection (ATCC) strain of *M. tuberculosis* H37Rv against other genomic sequences of *M. tuberculosis*, protein products that were not encoded in the H37Rv genome but present in the other genomes were discovered [58]. The most likely explanation would be that a genetic element had been lost from the sequenced sample of H37Rv. This shows that one needs to have a focus on the origin of the genomic sequence as well as the sample being studied for proteomics.

2.2 Comparative proteogenomics

The ever-growing collection of sequenced prokaryotic genomes is giving the opportunity for comparative genomic and proteomic analysis of related species (i.e. sharing high level of sequence similarity) (Fig. 2B). As demonstrated in a study of exoproteomes of the *Roseobacter* clade (marine bacteria) and proteogenomic analysis of three *Shewanella* bacteria [59, 60], such comparative approaches allow for annotation of multiple genomes at once. Moreover, the simultaneous analysis of genomes containing orthologous genes has been used for more reliable interpretation of protein identifications based only on a single peptide hit. Such "one-hit-wonders" usually need manual validations otherwise they are discarded by the publication guidelines of proteomics journals. In a typical proteomic analysis, the percentage of one-hit-wonders can be as high as 30% [60] and using the "two peptide per protein" inference rule for protein identification might result in loss of a large number of protein identifications [61].

2.3 Nonsynonymous (ortho-) proteogenomics

In this case, there is no genomic sequence that matches the proteomic data exactly. This approach is relevant for bottom-up proteomic investigations performed on clinical and environmental isolates (Fig. 2C). Initially one needs to assess which genomic sequence(s) is/are the most relevant for data interpretation. The first step would be to identify the bacterial isolate at species level or better, and then to identify the peptides using the most closely matching genome. The disadvantage of such approach is that sequence differences between the available genome and the proteins being studied result in lost identifications. There may be more than one genome available for one species and in that case it will be of value to use a database that includes all relevant genomes. Building and the use of a species-specific custom-made database has been shown as an effective way to improve annotation of members of the same species [58, 62]. Collated genome sequences can also be used to identify variants in positions with single nucleotide/amino acid polymorphism. In this case, it is useful to apply a database with tagged sequences for more convenient identification of such differentiating peptides [58].

Nonsynonymous proteogenomic strategies have also been used for mapping the proteomes of unsequenced pathogens [58, 63]. Similarly, a combination of proteomics and comparative genomics has used only one member of a group of organisms for the proteomic analysis and extrapolated the findings on orthologous genes to other members of the clade [64–67].

2.4 Integration of multiple omics datasets

An emerging trend in the functional annotation of genome-scale data is the integration of various omics datasets (Fig. 2D). Transciptome profiling by the means of high-density tilling arrays or RNA-seq technology (high-throughput RNA sequencing using next-generation sequencing technologies) enables identification and also quantification of both rare and common transcripts with over five orders of magnitude of dynamic range [68]. Although transcriptional information does not confirm the protein expression, it can provide necessary supporting evidence in cases of protein identifications based on a single peptide hit. Studies integrating transcriptomic and proteomic analyses can be found, for example, for *Escherichia coli* and pathogenic bacteria *Bartonella henselae* and *Yersiniae* spp. [69–71]. In the case of the intracellular pathogen *B. henselae*, an endpoint estimate of the number of actively transcribed protein coding genes based on mRNA-seq data was shown to better represent the expressed protein catalog than considering all annotated protein-coding genes [70]. RNA-seq analysis was also complemented with proteomics for the radiation-tolerant bacterium *Deinococcus deserti* [72]. An interesting finding of that study was a very high number and proportion of leaderless mRNA in *D. deserti* (60%), an exceptionally high number for a bacterial species. An illustrative multi-omics approach was used to functionally reannotate the genome of a model actinomycetes and an important antibiotic producer *Saccharopolyspora erythraea* [73]. The authors integrated data from proteomics, RNA sequencing, and previously determined genome-scale metabolic reconstruction [74] to experimentally validate and improve the annotation of this model G+C rich genome. An outlook to the future of genome annotation studies can be a very thorough functional genome description of a hyperthermophilic bacterium, *Thermotoga maritime* [56]. Proteomic profiling was one part of an all-inclusive combination of whole-genome resequencing, transcriptome profiling, and various bioinformatics tools, which resulted in a more accurate genome annotation.

2.5 Functional protein annotation

In any newly sequenced bacterial genome annotated through the computational methods of functional and comparative genomics, proteins with unknown functions account for 30–40% of all encoded proteins [75–77]. There are two major bioinformatic approaches for predicting protein function of uncharacterized genes: (1) structural analysis based on homologous proteins [78] and (2) comparative analysis aiming to identify conserved coexpressed genes [79, 80]. Activity-based protein profiling (ABPP) is a functional proteomic technique to label and enrich various enzymatic activities [81, 82]. The contribution of ABPP to functional annotation of genomes lies in its ability to specifically detect active enzymes in a sample through activity-based probes. Several comprehensive reviews have addressed various designs and applications of the activity-based probes [83–85]. Practically any enzyme with an active catalytic site is amenable to study through ABPP, however, most of the research has focused on the diverse class of enzymes able to catalyze the hydrolysis of biomolecules (e.g. proteins, fats, oils, and carbohydrates) [86–88]. Proteolytic enzymes play a crucial part in the course of bacterial infection and the overall pathogenic process [89, 90], and the ABPP method can be of a value to scientists investigating, for example, proteins with unknown function but with a link to pathogenesis (e.g. confirmed expression in the presence of an antibiotic).

An ABPP approach that targets enzymes facilitating the hydrolysis of the β-lactam ring of penicillin antibiotics has been successfully applied to several pathogenic microbes. A library composed of antibiotic-inspired synthetic β-lactam probes was first utilized in a screen against proteomes of the *Pseudomonas putida*, *Listeria welshimeri*, and *Bacillus licheniformis* [91]. The study identified a number of β-lactam-binding enzymes under in vivo and in vitro conditions. Interestingly, in addition to expected penicillin binding proteins, several bacterial resistance- and virulence-associated enzymes were also detected and characterized. In a follow-up study, the β-lactam probe library was employed in a comparative analysis of antibiotic-sensitive *Staphylococcus aureus* and methicillin-resistant *S. aureus* to identify resistance-associated enzymes [92]. An example of proteome-wide application of ABPP was recently provided by Deng et al., who performed global profiling of reactive cysteines in *Pseudomonas aeruginosa* and *S. aureus* [93]. Modification of cysteines by reactive species (e.g. superoxide, hydrogen peroxide, and other reactive molecules containing oxygen) is a widespread regulatory PTM and bacteria use this modification as a part of complex responses to oxidative-stress challenges [94]. The proteomic study identified 200 proteins containing hydrogen peroxide sensitive cysteines across diverse classes of proteins, including metabolic enzymes, transcription factors, and uncharacterized proteins. Another recently reported chemical proteomics screen combined ABPP with quantitative MS-based proteomics, and facilitated high-throughput experimental functional annotation of ATP-binding proteins in the *M. tuberculosis* genome [95]. By using an ATP-based activity probe, Ansong and colleagues identified about 600 ATP-binding proteins in the *M. tuberculosis* proteome, including approximately 120 hypothetical proteins with unknown function.

3 Systematic genome annotation

Correct assignment of gene boundaries and various protein PTMs, which are often prerequisite for correct biological function, is generally invaluable for functional studies. PTMs can include chemical modifications of specific residues, processing of precursors into mature proteins by proteolytic cleavage, or signal-peptide removal during translocation across the cytoplasmic membrane. Protein modifications cannot be computationally predicted from genomic data in a straightforward way. Mapping of protein modifications therefore adds important details to the description of a proteome and to the functional annotation of the respective genome.

3.1 Translational start sites

Incorrect predictions of translation initiation codons are common errors introduced during an in silico driven annotation process [96] and in some prokaryotic genomes nearly 60% of genes can have incorrectly assigned start sites [35]. Usually, there are many potential translational start sites (TSSs) for a given gene. ATG coding for methionine is most frequently used as start codon, but GTG or TTG are also possible start codons. In some organisms, TTG is the most frequently used start codon [97]. Moreover, alternative TSSs both downstream and upstream of the originally annotated site are often observed [62, 98]. Identification of the N-terminal peptide by MS may confirm a TSS if it does not coincide with a peptide produced by the protease used for digestion (e.g. containing a nontryptic N-terminus) and if it is located at the protein N-terminus or in its close proximity [99]. This type of information is invaluable for genomic annotation and some of the large-scale proteomic investigations have addressed this issue [66, 67, 100]. However, because of the rather low sequence coverage in a typical bottom-up approach (~30%), methods for specific enrichment of N-terminal peptides are frequently employed [98, 101, 102]. Several reviews have provided expert views on the current state of N- and C-terminal protein analysis by proteomics [11, 103–105].

Profiling of N-terminally acetylated protein termini has recently emerged as a powerful technique for determination of TSS [106]. N-terminal acetylation together with N-terminal methionine excision is the most common protein PTM, which is widespread both in eukaryotes and prokaryotes [107, 108]. Recent proteogenomic analysis of encapsulated yeast *Cryptococcus neoformans*, which is an opportunistic human pathogen capable of causing potentially lethal disease cryptococcosis [109], described experimental

validation of 52% of the predicted proteome [43]. By defining N-terminal acetylation of proteins as variable modification in the MASCOT search engine, the authors of the study identified 392 N-terminally acetylated peptides that subsequently lead to confirmation of TSSs for 329 proteins (63% of all identified TSSs). In addition, the same proteogenomic study determined two novel start sites by mapping N-terminal acetylated peptides, identified from searching six-frame genome translational database against the MS/MS data, onto *C. neoformans* genome. First large-scale proteomic identification of N-terminal peptides from prokaryotes was produced for two archaea *Halobacterium salinarum* and *Natronomonas pharaonis* [100]. By combining MS with two specific enrichment methods, combined fractional diagonal chromatography (COFRADIC) and strong cation exchange chromatography [110], the authors were able to identify 606 N-terminal peptides from *H. salinarum* and 328 from *N. pharaonis* (29 and 12% of the predicted proteome, respectively). N-terminal COFRADIC is a well-established N-terminomics technology based on a negative selection for N-terminal peptides, that is removing non-N-terminal peptides [110, 111]. Helsens et al. analyzed the proteome of *Saccharomyces cerevisiae* also by applying N-terminal COFRADIC [112]. The analysis identified totally 706 TSSs out of which 89 represented potential alternate TSSs.

The second category of current N-terminomics protocols comprises positive selection procedures and includes chemical derivatization methods that allow for specific targeting of the protein N-termini (reviewed in [103]), for example variations of trimethoxyphenyl phosphonium labeling approach (N-TOP) [66, 67, 98, 101, 102]. The different studies showed, for example, the correction of 19% of translation start sites in *M. smegmatis* and 601 start sites in 16 other mycobacterial species [66], the validation of 278 and the correction of 73 translation initiation codons in the *D. deserti* genome, the annotation refinement of 534 proteins of the model marine bacterium *Roseobacter denitrificans* [98,102], and the characterization of 447 proteins (13.6% of the predicted proteome) for arsenite-oxidizing bacterium *Herminiimonas arsenicoxydans* [101]. Moreover, targeting of N-terminal peptides has led to recognition of rare noncanonical start codons. For example, ATC and CTG start codons for translation of DnaA and RpsL, respectively, were described in *D. deserti* [67] and an ultrarare start codon ATT for protein chain initiation factor IF-3 in *Yersinia pestis* [64].

3.2 Protein processing

Modifications of the protein N-terminal, such as N-methionine excision and N-acetylation, are widespread among bacteria, as discussed above [107,113], as well as the removal of the N-terminal signal peptide [114]. Signal peptides are cleaved by signal peptidase I or II after translocation of the protein through the cytoplasmic membrane. Several algorithms exist for prediction of such cleavage sites (e.g. SignalP [115] and Phobius [116]), but more exact experimental result confirming the cleavage site provides essential information [117].

Putative signal peptides and proteolytic events can be deduced from MS raw data by observing spectra matching to the peptides with nontryptic N-termini. In order to identify true signal peptides, filtering based on peptide length, typical structure (e.g. core hydrophobic patch), and signal peptidase cleavage site has usually been applied [118, 119]. MS-based proteomic techniques have accelerated the experimental verification of secretory proteins (and hence signal peptides), for example in *Salmonella enterica* [17], *Shewanella oneidensis* [99], *Y. pestis* [64], *Novosphingobium aromaticivorans* [120], and a microbial biofilm community [121]. In *H. pylori*, 63 previously unknown signal peptide sequences could be annotated by interpreting MS spectra with a search strategy allowing for semispecifically cleaved peptides and revealed the predominant recognition motif LXA for signal peptidases [122]. A recent study evaluated how many signal peptides can generally be identified in a single proteogenomic experiment by using *E. coli* K-12 as an example of a well-annotated model bacterium [119]. The paper by Ivankov et al. showed that approximately one-third of all experimentally known *E. coli* signal peptides could be validated. Moreover, in accordance with predictions from the latest version of the SignalP program, the authors of the study estimated that about 10% of the *E. coli* genes contain signal peptides, half of previous estimates.

3.3 PTMs of specific residues

Chemical modifications at specific residues such as phosphorylation, oxidation, methylation, etc. (therein referred to as PTMs) are known to play a significant role in many biological functions [123]. PTMs identification by the bottom-up MS approach is based on a change in the peptide fragmentation pattern (e.g. shifts in the masses of fragments containing the modification). There are several challenges in analysis of PTMs by MS (reviewed in detail in [124–127]). First, the detection of modified peptides is far from being straightforward because of the labile nature of many modifications during the peptide fragmentation. The modified peptides are usually present in low amounts in the complex sample and therefore specific enrichment methods have become an essential part of the proteomic workflow. Additionally, when several possible PTMs are included as an optional parameter during the sequence database search, there will be a combinatorial explosion of the search space and subsequently a lower statistical confidence in the search results. Nevertheless, bottom-up proteomic analysis has been successful in identifying some PTM sites, for example in bioremediation-relevant bacteria *Shewanella* [60, 99] and the sulfate-reducing bacterium *Desulfovibrio desulfuricans* [128].

In bottom-up MS approach, the identified PTMs are restricted to individual peptides. Multiple PTMs occurring in a single protein represent multiple combinatorial possibilities and hence such different proteoforms (i.e. specific

molecular form of a protein product arising from a specific gene) cannot be accurately defined. Consequently, by using the bottom-up approach, it is not possible to obtain information on how many protein isoforms there are or what combinations of PTMs are present in a single proteoform. A top-down proteomic approach (analysis of intact proteins) offers advantages in accurately localizing PTMs. However, throughput, sensitivity, and insufficient downstream bioinformatics tools have frequently been considered its major limiting factors [129, 130]. Moreover, if the goal is to analyze protein isoforms by top-down proteomics, there can arise issues concerning their isolation and the purity of individual isoforms. On account of advances in intact protein LC separations and MS instrumentation in recent years, top-down MS has shifted from analyzing single proteins to investigating multiple proteins (<50 kDa) in complex samples [131–133]. Top-down proteomic analysis of *S. enterica* Typhimurium identified 563 unique proteins (40% of the predicted proteome) corresponding to 1665 proteoforms and enabled discovery of the differential utilization of the protein S-thiolation forms, S-glutathionylation, and S-cysteinylation, in response to infection-like conditions [17].

Protein phosphorylation is one of the most extensively studied PTMs, not only because of its general importance in signal transduction in a living cell but also because of its relevance for bacterial virulence and pathogenesis [134–136]. A comprehensive list of phosphoproteomic studies performed on various bacterial species up to year 2013 can be found in [123]. Studies published later that are worth mentioning are, for example, a description of the phosphoproteome of the human pathogen *S. aureus* [137] and quantitative phosphoproteome analysis of *B. subtilis* [138].

4 Software for proteogenomics

Searching large spectral datasets against large sequence databases would not be possible without computational support. Computational proteomics has become a dynamically growing field. Bioinformaticians designing proteomics software tools have to tackle substantial challenges associated with the assignment of peptide sequences to MS/MS spectra and correct protein identifications [139]. Because of that, methods for assessing the quality of the match between an MS/MS spectrum and a theorized peptide sequence have been proposed. A popular approach is to simultaneously search the spectra against the target and decoy databases, the latter being of equal or known size and similar redundancy as the former [140]. Peptides identified using the decoy are then regarded as spurious and can be used to estimate false discovery rate. Altogether, the computational reconstruction of protein identities from proteomic data is nontrivial task and several excellent reviews described in detail the various problems commonly encountered and their current solutions [45, 141, 142]. Several automated software pipelines have been developed for integration of MS-based proteomic evidence into genome databases, as well as a number of visualization and database-building tools (Table 2 provides a list of these approaches with corresponding references).

Table 2. List of various open-source tools for proteogenomic research

Software name[a)]	Description	Reference
customProDB	Software package for generation of customized protein databases from RNA-Seq data.	[155]
GFS	Mapping of protein-derived MS data directly to genomic and/or transcript sequences.	[156]
Genosuite	Integrated proteogenomic pipeline for annotating prokaryotic genomes.	[143]
InsPecT	Identification of posttranslationally modified peptides from MS/MS spectra.	[157]
iPiG	Visualization of peptide identifications in genome browsers.	[158]
MINOMICS	Visualization of prokaryotic transcriptomic and proteomic data in conjunction with genomic data.	[159]
MSMSpdbb	Merging and clustering of protein sequences inferred from multiple genomic sequences.	[62]
PG Nexus with IGV	Covisualization of peptides in the context of genomes, genomic contigs or RNA-seq reads.	[160]
PepLine	Mapping of MS/MS fragmentation spectra of trypsic peptides to genomic DNA sequences.	[161]
Peppy	Integrated software package for proteogenomic analysis.	[162]
PGP	Proteogenomic annotation pipeline for improving existing genomic annotations.	[163]
PMT	Mapping of MS identified peptides to a target genome for structural genome annotation.	[164]
Protter	Web-based application for protein feature visualization and integration with experimental data.	[165]
TopFIND	Knowledgebase for protein termini, terminus modifications and underlying proteolytic processing.	[166]
VESPA	Visual analysis software integrating proteomics and transcriptomics data into a genomic context.	[167]

[a)] Abbreviations not defined by description: GFS: genome fingerprint scanning; IPiG: integrating peptide spectrum matches into genome browser visualizations; IGV: integrated genome viewer; MSMSpdbb: multistrain MS prokaryotic database builder; PMT: Proteogenomic Mapping Tool

5 Concluding remarks

Proteogenomic methods are still facing several key challenges that stand in the way of large-scale application of proteomics to genome annotation. One of the main concerns for nearly all MS-based proteomic studies is low sequence coverage. In addition, there is a notorious need for improved data mining methods and bioinformatics tools. Finally, in order to obtain high proteome coverage, one often needs to apply multiple growth conditions together with several separation and/or fractionation techniques prior to MS/MS analysis. Despite all of its shortcomings, proteogenomics analysis provides the ultimate validation of expressed gene products on a large scale and leads to correct interpretation of genomic sequences. Experimental verification of predicted hypothetical proteins and discovery of novel coding regions can be considered as one of the most important outcomes of proteogenomic studies. Moreover, specific applications designed to characterize various protein-processing events and PTMs are invaluable in deciphering the actual biological function.

This work was supported by the Research Council of Norway (grant 204743).

The authors have declared no conflict of interest.

6 References

[1] Rabilloud, T., Chevallet, M., Luche, S., Lelong, C., Two-dimensional gel electrophoresis in proteomics: past, present and future. *J. Proteomics* 2010, *73*, 2064–2077.

[2] Nesvizhskii, A. I., Vitek, O., Aebersold, R., Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* 2007, *4*, 787–797.

[3] Yates, J. R., Eng, J. K., McCormack, A. L., Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal. Chem.* 1995, *67*, 3202–3210.

[4] Yates, J., Eng, J., Clauser, K., Burlingame, A., Search of sequence databases with uninterpreted high-energy collision-induced dissociation spectra of peptides. *J. Am. Soc. Mass Spectrom.* 1996, *7*, 1089–1098.

[5] Perkins, D. N., Pappin, D. J. C., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, *20*, 3551–3567.

[6] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L. et al., Open mass spectrometry search algorithm. *J. Proteome Res.* 2004, *3*, 958–964.

[7] Craig, R., Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, *20*, 1466–1467.

[8] Cox, J., Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotech.* 2008, *26*, 1367–1372.

[9] Ansong, C., Purvine, S. O., Adkins, J. N., Lipton, M. S., Smith, R. D., Proteogenomics: needs and roles to be filled by proteomics in genome annotation. *Brief. Funct. Genomics Proteomics* 2008, *7*, 50–62.

[10] Renuse, S., Chaerkady, R., Pandey, A., Proteogenomics. *Proteomics* 2011, *11*, 620–630.

[11] Armengaud, J., Marie Hartmann, E., Bland, C., Proteogenomics for environmental microbiology. *Proteomics* 2013, *13*, 2731–2742.

[12] Seidler, J., Zinn, N., Boehm, M. E., Lehmann, W. D., De novo sequencing of peptides by MS/MS. *Proteomics* 2010, *10*, 634–649.

[13] Ma, B., Johnson, R., De novo sequencing and homology searching. *Mol. Cell. Proteomics* 2012, *11*, O111.014902.

[14] Siuti, N., Kelleher, N. L., Decoding protein modifications using top-down mass spectrometry. *Nat. Methods* 2007, *4*, 817–821.

[15] VerBerkmoes, N. C., Denef, V. J., Hettich, R. L., Banfield, J. F., Systems biology: functional analysis of natural microbial consortia using community proteomics. *Nat. Rev. Microbiol.* 2009, *7*, 196–205.

[16] Hettich, R. L., Pan, C., Chourey, K., Giannone, R. J., Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. *Anal. Chem.* 2013, *85*, 4203–4214.

[17] Ansong, C., Tolic, N., Purvine, S., Porwollik, S. et al., Experimental annotation of post-translational features and translated coding regions in the pathogen *Salmonella typhimurium*. *BMC Genomics* 2011, *12*, 433.

[18] Metzker, M. L., Sequencing technologies—the next generation. *Nat. Rev. Genet.* 2010, *11*, 31–46.

[19] Fricke, W. F., Rasko, D. A., Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nat. Rev. Genet.* 2014, *15*, 49–55.

[20] Stein, L., Genome annotation: from sequence to biology. *Nat. Rev. Genet.* 2001, *2*, 493–503.

[21] Siezen, R. J., Van Hijum, S. A. F. T., Genome (re-)annotation and open-source annotation pipelines. *Microbial Biotechnol.* 2010, *3*, 362–369.

[22] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J. et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997, *25*, 3389–3402.

[23] Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K. et al., IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* 2012, *40*, D115–D122.

[24] Aziz, R., Bartels, D., Best, A., DeJongh, M. et al., The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 2008, *9*, 75.

[25] Delcher, A. L., Bratke, K. A., Powers, E. C., Salzberg, S. L., Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 2007, *23*, 673–679.

[26] Besemer, J., Borodovsky, M., GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* 2005, *33*, W451–W454.

[27] Finn, R. D., Clements, J., Eddy, S. R., HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 2011, *39*, W29–W37.

[28] Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M. et al., A large-scale evaluation of computational protein function prediction. *Nat. Methods* 2013, *10*, 221–227.

[29] de Souza, G., Malen, H., Softeland, T., Saelensminde, G. 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.

[30] Baumgartner, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., Hunter, L., Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics* 2007, *23*, i41–i48.

[31] Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L. et al., An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* 2002, *2*, 513–523.

[32] Ueberle, B., Frank, R., Herrmann, R., The proteome of the bacterium *Mycoplasma pneumoniae*: comparing predicted open reading frames to identified gene products. *Proteomics* 2002, *2*, 754–764.

[33] Lipton, M. S., Paša-Tolić, L., Anderson, G. A., Anderson, D. J. et al., Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc. Natl. Acad. Sci.* 2002, *99*, 11049–11054.

[34] Jungblut, P. R., Müller, E.-C., Mattow, J., Kaufmann, S. H. E., Proteomics reveals open reading frames in *Mycobacterium tuberculosis* H37Rv not predicted by genomics. *Infect. Immun.* 2001, *69*, 5905–5907.

[35] Nielsen, P., Krogh, A., Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics* 2005, *21*, 4322–4329.

[36] Bakke, P., Carney, N., DeLoache, W., Gearing, M. et al., Evaluation of three automated genome annotations for *Halorhabdus utahensis*. *PLoS ONE* 2009, *4*, e6291.

[37] Branca, R. M. M., Orre, L. M., Johansson, H. J., Granholm, V. et al., HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods* 2014, *11*, 59–62.

[38] Volkening, J. D., Bailey, D. J., Rose, C. M., Grimsrud, P. A. et al., A proteogenomic survey of the *Medicago truncatula* genome. *Mol. Cell. Proteomics* 2012, *11*, 933–944.

[39] Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M. et al., Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc. Natl. Acad. Sci.* 2008, *105*, 21034–21038.

[40] Wright, J., Sugden, D., Francis-McIntyre, S., Riba-Garcia, I. et al., Exploiting proteomic data for genome annotation and gene model validation in *Aspergillus niger*. *BMC Genomics* 2009, *10*, 61.

[41] Prasad, T. S. K., Harsha, H. C., Keerthikumar, S., Sekhar, N. R. et al., Proteogenomic analysis of *Candida glabrata* using high resolution mass spectrometry. *J. Proteome Res.* 2011, *11*, 247–260.

[42] Pawar, H., Renuse, S., Khobragade, S. N., Chavan, S. et al., Neglected tropical diseases and omics science: proteogenomics analysis of the promastigote stage of *Leishmania major* parasite. *OMICS* 2014, *18*, 499–512.

[43] Nagarajha Selvan, L., Kaviyil, J., Nirujogi, R., Muthusamy, B. et al., Proteogenomic analysis of pathogenic yeast *Cryptococcus neoformans* using high resolution mass spectrometry. *Clin. Proteomics* 2014, *11*, 5.

[44] Ahnert, S. E., Fink, T. M. A., Zinovyev, A., How much non-coding DNA do eukaryotes require? *J. Theor. Biol.* 2008, *252*, 587–592.

[45] Castellana, N., Bafna, V., Proteogenomics to discover the full coding content of genomes: a computational perspective. *J. Proteomics* 2010, *73*, 2124–2135.

[46] Mann, M., Pandey, A., Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases. *Trends Biochem. Sci.* 2001, *26*, 54–61.

[47] Nesvizhskii, A. I., A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* 2010, *73*, 2092–2123.

[48] Krug, K., Carpy, A., Behrends, G., Matic, K. et al., Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Mol. Cell. Proteomics* 2013, *12*, 3420–3430.

[49] Blakeley, P., Overton, I. M., Hubbard, S. J., Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J. Proteome Res.* 2012, *11*, 5221–5234.

[50] Jaffe, J. D., Berg, H. C., Church, G. M., Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* 2004, *4*, 59–77.

[51] Jaffe, J. D., Stange-Thomann, N., Smith, C., DeCaprio, D. et al., The complete genome and proteome of *Mycoplasma mobile*. *Genome Res.* 2004, *14*, 1447–1461.

[52] de Groot, A., Dulermo, R., Ortet, P., Blanchard, L. et al., Alliance of proteomics and genomics to unravel the specificities of sahara bacterium *Deinococcus deserti*. *PLoS Genet.* 2009, *5*, e1000434.

[53] Deshpande, N. P., Kaakoush, N. O., Mitchell, H., Janitz, K. et al., Sequencing and validation of the genome of a *Campylobacter concisus* reveals intra-species diversity. *PLoS ONE* 2011, *6*, e22170.

[54] Lazarev, V. N., Levitskii, S. A., Basovskii, Y. I., Chukin, M. M. et al., Complete genome and proteome of *Acholeplasma laidlawii*. *J. Bacteriol.* 2011, *193*, 4943–4953.

[55] Zivanovic, Y., Armengaud, J., Lagorce, A., Leplat, C. et al., Genome analysis and genome-wide proteomics of *Thermococcus gammatolerans*, the most radioresistant organism known amongst the Archaea. *Genome Biol.* 2009, *10*, R70.

[56] Latif, H., Lerman, J. A., Portnoy, V. A., Tarasova, Y. et al., The genome organization of *Thermotoga maritima* reflects is lifestyle. *PLoS Genet.* 2013, *9*, e1003485.

[57] Darmon, E., Leach, D. R. F., Bacterial genome instability. *Microbiol. Mol. Biol. Rev.* 2014, *78*, 1–39.

[58] de Souza, G. A., Arntzen, M. Ø., Fortuin, S., Schürch, A. C. et al., Proteogenomic analysis of polymorphisms and

gene annotation divergences in prokaryotes using a clustered mass spectrometry-friendly database. *Mol. Cell. Proteomics* 2011, *10*, M110.002527.

[59] Christie-Oleza, J. A., Piña-Villalonga, J. M., Bosch, R., Nogales, B., Armengaud, J., Comparative proteogenomics of twelve *Roseobacter* exoproteomes reveals different adaptive strategies among these marine bacteria. *Mol. Cell. Proteomics* 2012, *11*, M111.013110.

[60] Gupta, N., Benhamida, J., Bhargava, V., Goodman, D. et al., Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Res.* 2008, *18*, 1133–1142.

[61] Gupta, N., Pevzner, P. A., False discovery rates of protein identifications: a strike against the two-peptide rule. *J. Proteome Res.* 2009, *8*, 4173–4181.

[62] de Souza, G. A., Arntzen, M. Ø., Wiker, H. G., MSMSpdbb: providing protein databases of closely related organisms to improve proteomic characterization of prokaryotic microbes. *Bioinformatics* 2010, *26*, 698–699.

[63] Pawar, H., Sahasrabuddhe, N. A., Renuse, S., Keerthikumar, S. et al., A proteogenomic approach to map the proteome of an unsequenced pathogen—*Leishmania donovani*. *Proteomics* 2012, *12*, 832–844.

[64] Payne, S., Huang, S.-T., Pieper, R., A proteogenomic update to Yersinia: enhancing genome annotation. *BMC Genomics* 2010, *11*, 460.

[65] Christie-Oleza, J., Miotello, G., Armengaud, J., High-throughput proteogenomics of *Ruegeria pomeroyi*: seeding a better genomic annotation for the whole marine *Roseobacter* clade. *BMC Genomics* 2012, *13*, 73.

[66] Gallien, S., Perrodou, E., Carapito, C., Deshayes, C. et al., Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. *Genome Res.* 2009, *19*, 128–135.

[67] Baudet, M., Ortet, P., Gaillard, J.-C., Fernandez, B. et al., Proteomics-based refinement of *Deinococcus deserti* genome annotation reveals an unwonted use of non-canonical translation initiation codons. *Mol. Cell. Proteomics* 2010, *9*, 415–426.

[68] Wang, Z., Gerstein, M., Snyder, M., RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 2009, *10*, 57–63.

[69] Cho, B.-K., Zengler, K., Qiu, Y., Park, Y. S. et al., The transcription unit architecture of the *Escherichia coli* genome. *Nat. Biotech.* 2009, *27*, 1043–1049.

[70] Omasits, U., Quebatte, M., Stekhoven, D. J., Fortes, C. et al., Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome. *Genome Res.* 2013, *23*, 1916–1927.

[71] Schrimpe-Rutledge, A. C., Jones, M. B., Chauhan, S., Purvine, S. O. et al., Comparative omics-driven genome annotation refinement: application across *Yersiniae*. *PLoS ONE* 2012, *7*, e33903.

[72] de Groot, A., Roche, D., Fernandez, B., Ludanyi, M. et al., RNA sequencing and proteogenomics reveal the importance of leaderless mRNAs in the radiation-tolerant bacterium *Deinococcus deserti*. *Genome Biol. Evol.* 2014, *6*, 932–948.

[73] Marcellin, E., Licona-Cassani, C., Mercer, T., Palfreyman, R., Nielsen, L., Re-annotation of the *Saccharopolyspora erythraea* genome using a systems biology approach. *BMC Genomics* 2013, *14*, 699.

[74] Licona-Cassani, C., Marcellin, E., Quek, L.-E., Jacob, S., Nielsen, L., Reconstruction of the *Saccharopolyspora erythraea* genome-scale model and its use for enhancing erythromycin production. *A. van Leeuw.* 2012, *102*, 493–502.

[75] Galperin, M. Y., Kolker, E., New metrics for comparative genomics. *Curr. Opin. Biotechnol.* 2006, *17*, 440–447.

[76] Bork, P., Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome Res.* 2000, *10*, 398–400.

[77] Bernal, A., Ear, U., Kyrpides, N., Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.* 2001, *29*, 126–127.

[78] Binkowski, T. A., Joachimiak, A., Liang, J., Protein surface analysis for function annotation in high-throughput structural genomics pipeline. *Protein Sci.* 2005, *14*, 2972–2981.

[79] van Noort, V., Snel, B., Huynen, M. A., Predicting gene function by conserved co-expression. *Trends Genet.* 2003, *19*, 238–242.

[80] Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., Yeates, T. O., Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* 1999, *96*, 4285–4288.

[81] Li, N., Overkleeft, H. S., Florea, B. I., Activity-based protein profiling: an enabling technology in chemical biology research. *Curr. Opin. Chem. Biol.* 2012, *16*, 227–233.

[82] Barglow, K. T., Cravatt, B. F., Activity-based protein profiling for the functional annotation of enzymes. *Nat. Methods* 2007, *4*, 822–827.

[83] Cravatt, B. F., Wright, A. T., Kozarich, J. W., Activity-based protein profiling: from enzyme chemistry to proteomic chemistry. *Annu. Rev. Biochem.* 2008, *77*, 383–414.

[84] Edgington, L. E., Verdoes, M., Bogyo, M., Functional imaging of proteases: recent advances in the design and application of substrate-based and activity-based probes. *Curr. Opin. Chem. Biol.* 2011, *15*, 798–805.

[85] Sadaghiani, A. M., Verhelst, S. H. L., Bogyo, M., Tagging and detection strategies for activity-based proteomics. *Curr. Opin. Chem. Biol.* 2007, *11*, 20–28.

[86] Puri, A. W., Bogyo, M., Applications of small molecule probes in dissecting mechanisms of bacterial virulence and host responses. *Biochemistry* 2013, *52*, 5985–5996.

[87] Evans, M. J., Cravatt, B. F., Mechanism-based profiling of enzyme families. *Chem. Rev.* 2006, *106*, 3279–3301.

[88] Simon, G. M., Cravatt, B. F., Activity-based proteomics of enzyme superfamilies: serine hydrolases as a case study. *J. Biol. Chem.* 2010, *285*, 11051–11055.

[89] Shen, A., Autoproteolytic activation of bacterial toxins. *Toxins* 2010, *2*, 963–977.

[90] Gur, E., Biran, D., Ron, E. Z., Regulated proteolysis in Gram-negative bacteria—how and when? *Nat. Rev. Microbiol.* 2011, *9*, 839–848.

[91] Staub, I., Sieber, S. A., β-Lactams as selective chemical probes for the in vivo labeling of bacterial enzymes involved in cell wall biosynthesis, antibiotic resistance, and virulence. *J. Am. Chem. Soc.* 2008, *130*, 13400–13409.

[92] Staub, I., Sieber, S. A., β-Lactam probes as selective chemical-proteomic tools for the identification and functional characterization of resistance associated enzymes in MRSA. *J. Am. Chem. Soc.* 2009, *131*, 6271–6276.

[93] Deng, X., Weerapana, E., Ulanovskaya, O., Sun, F. et al., Proteome-wide quantification and characterization of oxidation-sensitive cysteines in pathogenic bacteria. *Cell Host Microbe* 2013, *13*, 358–370.

[94] Wall, S. B., Oh, J.-Y., Diers, A. R., Landar, A., Oxidative modification of proteins: an emerging mechanism of cell signaling. *Front. Physiol.* 2012, *3*, 1–9.

[95] Ansong, C., Ortega, C., Payne, Samuel H., Haft, Daniel H. et al., Identification of widespread adenosine nucleotide binding in *Mycobacterium tuberculosis*. *Chem. Biol.* 2013, *20*, 123–133.

[96] Sato, N., Tajima, N., Statistics of N-terminal alignment as a guide for refining prokaryotic gene annotation. *Genomics* 2012, *99*, 138–143.

[97] Yamazaki, S., Yamazaki, J., Nishijima, K., Otsuka, R. et al., Proteome analysis of an aerobic hyperthermophilic Crenarchaeon, *Aeropyrum pernix* K1. *Mol. Cell. Proteomics* 2006, *5*, 811–823.

[98] Bland, C., Hartmann, E. M., Christie-Oleza, J. A., Fernandez, B., Armengaud, J., N-terminal-oriented proteogenomics of the marine bacterium *Roseobacter denitrificans* OCh114 using TMPP labeling and diagonal chromatography. *Mol. Cell. Proteomics* 2014, *13*, 1369–1381.

[99] Gupta, N., Tanner, S., Jaitly, N., Adkins, J. N. et al., Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res.* 2007, *17*, 1362–1377.

[100] Aivaliotis, M., Gevaert, K., Falb, M., Tebbe, A. et al., Large-scale identification of N-terminal peptides in the halophilic Archaea *Halobacterium salinarum* and *Natronomonas pharaonis*. *J. Proteome Res.* 2007, *6*, 2195–2204.

[101] Bertaccini, D., Vaca, S., Carapito, C., Arsène-Ploetze, F. et al., An improved stable isotope n-terminal labeling approach with light/heavy tmpp to automate proteogenomics data validation: dN-TOP. *J. Proteome Res.* 2013, *12*, 3063–3070.

[102] Bland, C., Bellanger, L., Armengaud, J., Magnetic immunoaffinity enrichment for selective capture and ms/ms analysis of n-terminal-TMPP-labeled peptides. *J. Proteome Res.* 2013, *13*, 668–680.

[103] Armengaud, J., A perfect genome annotation is within reach with the proteomics and genomics alliance. *Curr. Opin. Microbiol.* 2009, *12*, 292–300.

[104] Nakazawa, T., Yamaguchi, M., Okamura, T.-a., Ando, E. et al., Terminal proteomics: N- and C-terminal analyses for high-fidelity identification of proteins using MS. *Proteomics* 2008, *8*, 673–685.

[105] Rogers, L. D., Overall, C. M., Proteolytic post-translational modification of proteins: proteomic tools and methodology. *Mol. Cell. Proteomics* 2013, *12*, 3532–3542.

[106] Van Damme, P., Arnesen, T., Gevaert, K., Protein alpha-N-acetylation studied by N-terminomics. *FEBS J.* 2011, *278*, 3822–3834.

[107] Bonissone, S., Gupta, N., Romine, M., Bradshaw, R. A., Pevzner, P. A., N-terminal protein processing: a comparative proteogenomic analysis. *Mol. Cell. Proteomics* 2013, *12*, 14–28.

[108] Rison, S. C. G., Mattow, J., Jungblut, P. R., Stoker, N. G., Experimental determination of translational starts using peptide mass mapping and tandem mass spectrometry within the proteome of *Mycobacterium tuberculosis*. *Microbiology* 2007, *153*, 521–528.

[109] Idnurm, A., Bahn, Y.-S., Nielsen, K., Lin, X. et al., Deciphering the model pathogenic fungus *Cryptococcus neoformans*. *Nat. Rev. Microbiol.* 2005, *3*, 753–764.

[110] Staes, A., Impens, F., Van Damme, P., Ruttens, B. et al., Selecting protein N-terminal peptides by combined fractional diagonal chromatography. *Nat. Protocols* 2011, *6*, 1130–1141.

[111] Gevaert, K., Goethals, M., Martens, L., Van Damme, J. et al., Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat. Biotech.* 2003, *21*, 566–569.

[112] Helsens, K., Van Damme, P., Degroeve, S., Martens, L. et al., Bioinformatics snalysis of a *Saccharomyces cerevisiae* N-terminal proteome provides evidence of alternative translation initiation and post-translational N-terminal acetylation. *J. Proteome Res.* 2011, *10*, 3578–3589.

[113] Frottin, F., Martinez, A., Peynot, P., Mitra, S. et al., The proteomics of N-terminal methionine cleavage. *Mol. Cell. Proteomics* 2006, *5*, 2336–2349.

[114] Payne, S. H., Bonissone, S., Wu, S., Brown, R. N. et al., Unexpected diversity of signal peptides in prokaryotes. *mBio* 2012, *3*, e00339–12.

[115] Petersen, T. N., Brunak, S., von Heijne, G., Nielsen, H., SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 2011, *8*, 785–786.

[116] Käll, L., Krogh, A., Sonnhammer, E. L. L., Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res.* 2007, *35*, W429–W432.

[117] Leversen, N. A., de Souza, G. A., Målen, H., Prasad, S. et al., Evaluation of signal peptide prediction algorithms for identification of mycobacterial signal peptides using sequence data from proteomic methods. *Microbiology* 2009, *155*, 2375–2383.

[118] Venter, E., Smith, R. D., Payne, S. H., Proteogenomic analysis of bacteria and archaea: a 46 organism case study. *PLoS ONE* 2011, *6*, e27587.

[119] Ivankov, D. N., Payne, S. H., Galperin, M. Y., Bonissone, S. et al., How many signal peptides are there in bacteria? *Environ. Microbiol.* 2013, *15*, 983–990.

[120] Wu, S., Brown, R. N., Payne, S. H., Meng, D. et al., Top-down characterization of the post-translationally modified intact periplasmic proteome from the bacterium *Novosphingobium aromaticivorans*. *Int. J. Proteomics* 2013, *2013*, 10.

[121] Erickson, B. K., Mueller, R. S., VerBerkmoes, N. C., Shah, M. et al., Computational prediction and experimental validation of signal peptide cleavages in the extracellular proteome of a natural microbial community. *J. Proteome Res.* 2010, *9*, 2148–2159.

[122] Müller, S. A., Findeiß, S., Pernitzsch, S. R., Wissenbach, D. K. et al., Identification of new protein coding sequences and signal peptidase cleavage sites of *Helicobacter pylori* strain 26695 by proteogenomics. *J. Proteomics* 2013, *86*, 27–42.

[123] Cain, J. A., Solis, N., Cordwell, S. J., Beyond gene expression: the impact of protein post-translational modifications in bacteria. *J. Proteomics* 2014, *97*, 265–286.

[124] Zhao, Y., Jensen, O. N., Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* 2009, *9*, 4632–4641.

[125] Cox, J., Mann, M., Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* 2011, *80*, 273–299.

[126] Silva, A. M. N., Vitorino, R., Domingues, M. R. M., Spickett, C. M., Domingues, P., Post-translational modifications and mass spectrometry detection. *Free Radical Biol. Med.* 2013, *65*, 925–941.

[127] Olsen, J. V., Mann, M., Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol. Cell. Proteomics* 2013, *12*, 3444–3452.

[128] Gaucher, S. P., Redding, A. M., Mukhopadhyay, A., Keasling, J. D., Singh, A. K., Post-translational modifications of *Desulfovibrio vulgaris* Hildenborough sulfate reduction pathway proteins. *J. Proteome Res.* 2008, *7*, 2320–2331.

[129] Garcia, B., What does the future hold for top down mass spectrometry? *J. Am. Soc. Mass Spectrom.* 2010, *21*, 193–202.

[130] Liu, X., Sirotkin, Y., Shen, Y., Anderson, G. et al., Protein identification using top-down spectra. *Mol. Cell. Proteomics* 2012, *11*, M111.008524.

[131] Sharma, S., Simpson, D. C., Tolić, N., Jaitly, N. et al., Proteomic profiling of intact proteins using WAX-RPLC 2-D separations and FTICR mass spectrometry. *J. Proteome Res.* 2006, *6*, 602–610.

[132] Patrie, S. M., Ferguson, J. T., Robinson, D. E., Whipple, D. et al., Top down mass spectrometry of <60-kda proteins from *Methanosarcina acetivorans* using quadrupole ftms with automated octopole collisionally activated dissociation. *Mol. Cell. Proteomics* 2006, *5*, 14–25.

[133] Vellaichamy, A., Tran, J. C., Catherman, A. D., Lee, J. E. et al., Size-sorting combined with improved nanocapillary liquid chromatography−mass spectrometry for identification of intact proteins up to 80 kDa. *Anal. Chem.* 2010, *82*, 1234–1244.

[134] Mijakovic, I., Macek, B., Impact of phosphoproteomics on studies of bacterial physiology. *FEMS Microbiol. Rev.* 2012, *36*, 877–892.

[135] Hansen, A.-M., Chaerkady, R., Sharma, J., Díaz-Mejía, J. J. et al., The *Escherichia coli* phosphotyrosine proteome relates to core pathways and virulence. *PLoS Pathog.* 2013, *9*, e1003403.

[136] Ge, R., Shan, W., Bacterial phosphoproteomic analysis reveals the correlation between protein phosphorylation and bacterial pathogenicity. *Genomics Proteomics Bioinformatics* 2011, *9*, 119–127.

[137] Bäsell, K., Otto, A., Junker, S., Zühlke, D. et al., The phosphoproteome and its physiological dynamics in *Staphylococcus aureus*. *Int. J. Med. Microbiol.* 2014, *304*, 121–132.

[138] Ravikumar, V., Shi, L., Krug, K., Derouiche, A. et al., Quantitative phosphoproteome analysis of *Bacillus subtilis* reveals novel substrates of the kinase PrkC and phosphatase PrpC. *Mol. Cell. Proteomics* 2014, *13*, 1965–1978.

[139] Kim, S., Gupta, N., Pevzner, P. A., Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* 2008, *7*, 3354–3363.

[140] Elias, J. E., Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 2007, *4*, 207–214.

[141] Colinge, J., Bennett, K. L., Introduction to computational proteomics. *PLoS Comput. Biol.* 2007, *3*, e114.

[142] Claassen, M., Inference and validation of protein identifications. *Mol. Cell. Proteomics* 2012, *11*, 1097–1104.

[143] Kumar, D., Yadav, A. K., Kadimi, P. K., Nagaraj, S. H. et al., Proteogenomic analysis of *Bradyrhizobium japonicum* USDA110 using Genosuite, an automated multi-algorithmic pipeline. *Mol. Cell. Proteomics* 2013, *12*, 3388–3397.

[144] Lamontagne, J., Beland, M., Forest, A., Cote-Martin, A. et al., Proteomics-based confirmation of protein expression and correction of annotation errors in the *Brucella abortus* genome. *BMC Genomics* 2010, *11*, 300.

[145] Nirujogi, R. S., Pawar, H., Renuse, S., Kumar, P. et al., Moving from unsequenced to sequenced genome: reanalysis of the proteome of *Leishmania donovani*. *J. Proteomics* 2014, *97*, 48–61.

[146] de Souza, G. A., Søfteland, T., Koehler, C. J., Thiede, B., Wiker, H. G., Validating divergent ORF annotation of the *Mycobacterium leprae* genome through a full translation data set and peptide identification by tandem mass spectrometry. *Proteomics* 2009, *9*, 3233–3243.

[147] Wang, R., Prince, J. T., Marcotte, E. M., Mass spectrometry of the *M. smegmatis* proteome: protein expression levels correlate with function, operons, and codon bias. *Genome Res.* 2005, *15*, 1118–1126.

[148] Kelkar, D. S., Kumar, D., Kumar, P., Balakrishnan, L. et al., Proteogenomic analysis of *Mycobacterium tuberculosis* by high resolution mass spectrometry. *Mol. Cell. Proteomics* 2011, *10*, M111.011627.

[149] Lasonder, E., Ishihama, Y., Andersen, J. S., Vermunt, A. M. W. et al., Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* 2002, *419*, 537–542.

[150] Savidor, A., Donahoo, R. S., Hurtado-Gonzales, O., VerBerkmoes, N. C. et al., Expressed peptide tags: an additional layer of data for genome annotation. *J. Proteome Res.* 2006, *5*, 3048–3058.

[151] Zhao, L., Liu, L., Leng, W., Wei, C., Jin, Q., A proteogenomic analysis of *Shigella flexneri* using 2D LC-MALDI TOF/TOF. *BMC Genomics* 2011, *12*, 528.

[152] Okamoto, A., Yamada, K., Proteome driven re-evaluation and functional annotation of the *Streptococcus pyogenes* SF370 genome. *BMC Microbiol.* 2011, *11*, 249.

[153] Rodriguez-Ortega, M., Luque, I., Tarradas, C., Barcena, J., Overcoming function annotation errors in the Gram-positive pathogen *Streptococcus suis* by a proteomics-driven approach. *BMC Genomics* 2008, *9*, 588.

[154] Ishino, Y., Okada, H., Ikeuchi, M., Taniguchi, H., Mass spectrometry-based prokaryote gene annotation. *Proteomics* 2007, *7*, 4053–4065.

[155] Wang, X., Zhang, B., customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* 2013, *29*, 3235–3237.

[156] Giddings, M. C., Shah, A. A., Gesteland, R., Moore, B., Genome-based peptide fingerprint scanning. *Proc. Natl. Acad. Sci.* 2003, *100*, 20–25.

[157] Tanner, S., Shu, H., Frank, A., Wang, L.-C. et al., InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* 2005, *77*, 4626–4639.

[158] Kuhring, M., Renard, B. Y., iPiG: integrating peptide spectrum matches into genome browser visualizations. *PLoS ONE* 2012, *7*, e50246.

[159] Brouwer, R. W. W., van Hijum, S. A. F. T., Kuipers, O. P., MINOMICS: visualizing prokaryote transcriptomics and proteomics data in a genomic context. *Bioinformatics* 2009, *25*, 139–140.

[160] Pang, C. N. I., Tay, A. P., Aya, C., Twine, N. A. et al., Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing. *J. Proteome Res.* 2013, *13*, 84–98.

[161] Ferro, M., Tardif, M., Reguer, E., Cahuzac, R. et al., PepLine: a software pipeline for high-throughput direct mapping of tandem mass spectrometry data on genomic sequences. *J. Proteome Res.* 2008, *7*, 1873–1883.

[162] Risk, B. A., Spitzer, W. J., Giddings, M. C., Peppy: proteogenomic search software. *J. Proteome Res.* 2013, *12*, 3019–3025.

[163] Tovchigrechko, A., Venepally, P., Payne, S. H., PGP: parallel prokaryotic proteogenomics pipeline for MPI clusters, high-throughput batch clusters and multicore workstations. *Bioinformatics* 2014, *30*, 1469–1470.

[164] Sanders, W., Wang, N., Bridges, S., Malone, B. et al., The Proteogenomic mapping tool. *BMC Bioinformatics* 2011, *12*, 115.

[165] Omasits, U., Ahrens, C. H., Müller, S., Wollscheid, B., Protter: interactive protein feature visualization and integration with experimental proteomic data. *Bioinformatics* 2013, *30*, 884–886.

[166] Lange, P. F., Huesgen, P. F., Overall, C. M., TopFIND 2.0—linking protein termini with proteolytic processing and modifications altering protein function. *Nucleic Acids Res.* 2012, *40*, D351–D361.

[167] Peterson, E., McCue, L. A., Schrimpe-Rutledge, A., Jensen, J. et al., VESPA: software to facilitate genomic annotation of prokaryotic organisms through integration of proteomic and transcriptomic data. *BMC Genomics* 2012, *13*, 131.

