AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Bacteriology®

MEETING REVIEW

# A Practical Guide to Small Protein Discovery and Characterization Using Mass Spectrometry

Christian H. Ahrens,[a] Joseph T. Wade,[b,c] Matthew M. Champion,[d] Julian D. Langer[e,f]

[a]Agroscope, Method Development and Analytics & SIB Swiss Institute of Bioinformatics, Wädenswil, Switzerland
[b]Wadsworth Center, New York State Department of Health, Albany, New York, USA
[c]Department of Biomedical Sciences, School of Public Health, University at Albany, Albany, New York, USA
[d]Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana, USA
[e]Mass Spectrometry and Proteomics, Max Planck Institute of Biophysics, Frankfurt am Main, Germany
[f]Proteomics, Max Planck Institute for Brain Research, Frankfurt am Main, Germany

**ABSTRACT** Small proteins of up to ~50 amino acids are an abundant class of biomolecules across all domains of life. Yet due to the challenges inherent in their size, they are often missed in genome annotations, and are difficult to identify and characterize using standard experimental approaches. Consequently, we still know few small proteins even in well-studied prokaryotic model organisms. Mass spectrometry (MS) has great potential for the discovery, validation, and functional characterization of small proteins. However, standard MS approaches are poorly suited to the identification of both known and novel small proteins due to limitations at each step of a typical proteomics workflow, i.e., sample preparation, protease digestion, liquid chromatography, MS data acquisition, and data analysis. Here, we outline the major MS-based workflows and bioinformatic pipelines used for small protein discovery and validation. Special emphasis is placed on highlighting the adjustments required to improve detection and data quality for small proteins. We discuss both the unbiased detection of small proteins and the targeted analysis of small proteins of interest. Finally, we provide guidelines to prioritize novel small proteins, and an outlook on methods with particular potential to further improve comprehensive discovery and characterization of small proteins.

**KEYWORDS** proteomics, small protein, sproteins, SEP, microprotein, genome annotation, LC-MS/MS, shotgun proteomics, top-down proteomics, sample preparation

Large-scale discovery of small proteins (<~50 amino acids; also referred to as "sproteins," "short ORF-encoded proteins" (SEPs), or "microproteins") has relied primarily on computational analysis of genome sequences (1), and genome-scale experimental measurements of transcription and translation such as transcriptome sequencing (RNA-seq) (2) and ribosome profiling (Ribo-seq) (3, 4). Computational analyses can leverage the thousands of available genome sequences, but still require experimental support to serve as conclusive evidence. The major experimental approaches used to date are genome-scale analyses of transcription and translation. These methods share the advantage of amplified RNA-based detection (allowing sensitivity down to single molecules), high data acquisition speeds, and—with appropriate modification as differential RNA-seq (dRNA-seq) (5) or Ribo-RET (6)—the ability to identify transcription and translation initiation sites and potentially even very low-abundance proteins. However, all RNA-based approaches only provide indirect evidence of small protein expression, and some of the translation products detected by Ribo-seq may be unstable. Moreover, RNA-based approaches cannot provide data on post-translational modifications, other processing steps like signal peptide cleavage, or maturation that proteins can undergo.

Mass spectrometry-based proteomics (MS) provides a direct method to detect and quantify small proteins on a global or targeted scale (for general proteomics reviews, see references

**Editor** Tina M. Henkin, The Ohio State University



Address correspondence to Christian H. Ahrens, christian.ahrens@agroscope.admin.ch, Joseph T. Wade, joseph.wade@health.ny.gov Matthew M. Champion, Matthew.M.Champion.8@nd.edu, or Julian D. Langer, julian.langer@brain.mpg.de.

The authors declare no conflict of interest.

**Accepted manuscript posted online** 8 November 2021

**Published** 18 January 2022

7 and 8). In addition, MS can provide information on the different proteoforms, i.e., different forms for a protein product derived from a single gene (9), including genetic variations, splice variants (for eukaryotes), processed forms, and different combinations of posttranslational modifications, all of which can affect protein function (for selected key MS terms, please see Box 1). Many small proteins have been detected as adventitious spectra within traditionally acquired proteomes, mainly using exploratory "shotgun" bottom-up proteomics (10). In this approach, all proteins in a sample (e.g., a cellular lysate or a purified protein complex) are digested using a protease, and subsequently subjected to liquid chromatography-coupled tandem mass spectrometry (LC-MS/MS) (Box 1). Sequences of the proteolytic peptides are inferred from their MS/MS spectra by matching the fragmentation patterns to theoretical spectra of a reference proteome database that contains the sequences of all annotated proteins of the target organism (Fig. 1). However, there are two caveats: First, MS-based detection is "amplification-free" and intrinsically limited by the sensitivity of the instrument and its dynamic range. This leads to preferential detection of more abundant proteins, and proteins that have peptides with high ionization efficiencies. Second, "conventional" bottom-up proteomics approaches are biased toward studying proteins with molecular weights above 10 kDa, which represent roughly >90% of annotated proteomes (as determined from an analysis of 21,400 complete prokaryotic genomes from NCBI's RefSeq). Conventional studies typically report a statistically significant under-representation of experimentally identified small proteins and are thus ill-suited for their comprehensive analysis (11, 12). In fact, limitations for small protein detection exist across the entire MS workflow: experimental sample preparation and digestion, MS detection and acquisition, database search, peptide spectrum matching (PSM), and protein identification. Protease digestion is confounded by the intrinsic scarcity of necessary cleavage sites in small proteins (see "Protease digestion" below). If no MS-detectable peptides with a length of approximately 7 to 40 amino acids (aa) are generated (this range is based on MaxQuant's lower length and upper molecular weight limit of 4,600 Da in the first-pass search [13]), a small protein will not be identifiable via bottom-up proteomics at all, unless alternative proteases are used. Furthermore, database searches to identify proteins require accurately and comprehensively annotated genomes. This is often not the case for small proteins, as gene prediction algorithms apply varying length thresholds for genes encoding proteins below 50 to 100 aa to minimize the number of spurious, false short ORFs (sORFs) (14) (see "Data Analysis" below). In addition, the traditionally applied "two peptide rule" has required more than one unique peptide for confident protein identification (15); this, however, is ill-suited for small proteins, which, due to their small size, can often only be identified by a single peptide, resulting in a higher false discovery rate (FDR) for small proteins that needs to be tightly controlled (see "Stringent FDR control" below). Moreover, quantitation based on single-peptide-hit proteins is often neither accurate nor precise (16, 17). Standard workflows for protein identification, FDR estimation, and quantification are thus less useful for small protein discovery, and have to be modified to allow their efficient and reliable detection and analysis. Targeted proteomics approaches can fill this gap and are frequently used to validate and quantify small proteins (see "Validation of Novel Small Protein Candidates" below).

## BOX 1: MASS SPECTROMETRY AND PROTEOMICS CONCEPTS AND TERMS

**General approaches.**

**(i) Bottom up.** Methods in which protein samples are enzymatically or chemically cleaved into peptides prior to MS and MS/MS analysis. Typically uses liquid chromatography (LC) to separate digested peptides prior to MS and MS/MS analysis. Typically involves fragmentation in the mass spectrometer (MS/MS) to sequence peptides.

**(ii) Top down.** Methods in which protein samples are analyzed directly in the mass spectrometer without digestion. Can involve direct infusion or LC to separate proteins prior to MS and MS/MS analysis. Typically involves fragmentation in the mass

spectrometer (MS/MS) to sequence and identify the proteins. Can require specialized instrumentation or custom setup of existing instruments.

**Mass spectrometry instrumentation and acquisition.**

**(i) LC-MS.** Liquid chromatography-coupled mass spectrometry and tandem mass spectrometry (LC-MS/MS).

**(ii) MS/MS.** Typical arrangement for most MS analyses, also termed "tandem mass spectrometry" Separated proteins and/or peptides are measured in the instrument and then isolated and fragmented to record tandem mass spectra. Can be used for "bottom-up" and "top-down" approaches. Typically uses electrospray ionization (ESI) to generate ions.

**(iii) MALDI/ESI.** Ionization methods for MS. Matrix-assisted laser desorption ionization (MALDI) is rarely attached to LC systems. ESI is typically coupled to LC systems and is the dominant ionization mode for proteomics.

**(iv) DDA.** Data-dependent acquisition mode. Most commonly-used approach for proteomics studies. Candidate ions are selected based on peak intensity and resolution (= data dependent), isolated, and fragmented to provide MS/MS spectra.

**(v) DIA.** Data-independent acquisition mode. A hybrid approach where LC-MS/MS data are acquired without isolating specific ions for tandem analysis. Yields complex MS/MS spectra from multiple precursors. Since all ions are fragmented for MS/MS, no information is lost in the process, and typically data quality for quantitation is higher. Requires libraries of empirical or derived data from samples to extract identification and abundance from acquired data. Existing files can be reanalyzed with new information, as no ions are discarded.

**Data analysis terms.**

**(i) PRM.** Parallel reaction monitoring. Typical targeted acquisition mode for specific detection and robust quantification. Can involve specialized instrumentation.

**(ii) *De novo*.** Direct sequencing of polypeptides from fragment spectra via comparison of mass differences to amino acid residues.

**(iii) Database search.** Peptide and protein identification from fragment spectra by comparison of spectral patterns and fragment ions to databases of theoretical spectra/fragment masses. Redundant peptides measured from this approach are called a peptide spectrum match (PSM).

Bottom-up approaches also face the so-called "protein inference issue": protein identifications are inferred based on the detected peptides, which can be ambiguous and match to more than one protein or isoform (18). Proteins can carry different combinations of posttranslational modifications and exist in multiple proteoforms. In bottom-up proteomics, any information on posttranslational modifications or truncations is lost if the respective segment is not covered by one proteolytic peptide. In addition, there is very limited information on the combinations of these modifications. The only way to address this issue is "top-down" proteomics (19) (Fig. 1; Box 1). This digest-free approach analyzes full-length proteins and directly provides information on the different proteoforms present in a sample. The size of small proteins makes them ideally suited for a top-down proteomics approach. This is, however, a very experimental strategy with limitations in dynamic range and sequence analysis depth; only a limited number of labs conduct top-down proteomics on a routine basis. Hence, we provide guidelines and examples for top-down approaches, but more limited in scope than those for bottom-up proteomics.

Here, we describe a selection of MS-based approaches to identify and characterize small proteins in both exploratory and targeted studies. We focus on (i) approaches for sample preparation and MS data collection for small proteins, (ii) analysis of MS data sets for small protein discovery, (iii) validation of putative small proteins, and (iv) bioinformatic and MS-based prioritization of putative small proteins. While we cannot cover all approaches due to space constraints, we provide a repertoire of approaches that can be modified for a specific research question and that optimize the output for small proteins. These methods were developed by many different groups and now help to level the playing field for small proteins.

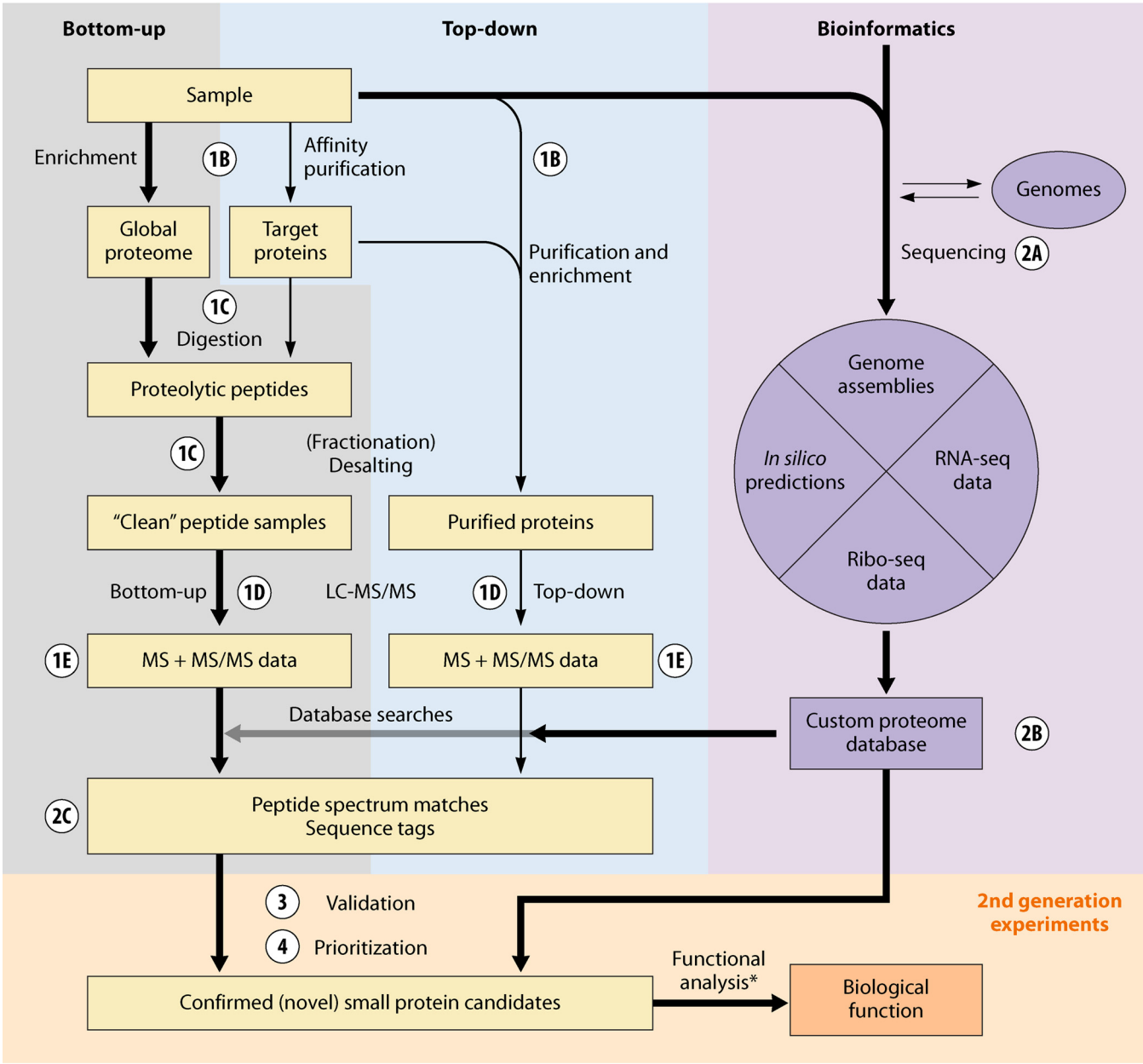


**FIG 1** Overview of the main mass spectrometry-based workflows for small protein discovery, analysis, and characterization. The large majority of studies have relied on a shotgun proteomics discovery approach (bottom-up) to identify small proteins. Top-down approaches are slowly gaining momentum but are not yet widely accessible from core facilities. Bioinformatics is important to assemble complete genomes *de novo* (at times using genomic DNA extracted from the same sample), to integrate small protein predictions with experimental RNA-seq and Ribo-seq data to create custom databases that allow the identification of novel small proteins by MS-based proteomics. Validation and prioritization facilitate focusing on the elucidation of function(s) of the most promising novel small proteins (yellow shading; see asterisk), an aspect that is described in more detail in the accompanying article "Small Proteins; Big Questions" (124). Shading matches that in Fig. 2. Corresponding text sections are indicated by white circles, as follows: 1B, "Sample Preparation and Data Collection—Preparation and enrichment for small proteins"; 1C, "Sample Preparation and Data Collection—Protease Digestion"; 1D, "Sample Preparation and Data Collection—Liquid chromatography"; 1E, "Sample Preparation and Data Collection—Ionization and data acquisition" ; 2A, "Data Analysis—Overview: the relevance of genome sequences for proteogenomics"; 2B, "Data Analysis—Creation of custom search databases"; 2C, "Data Analysis—Stringent FDR control"; 3, "Validation of Novel Small Protein Candidates"; 4, "Prioritization/Selection of Novel Small Proteins."

## SAMPLE PREPARATION AND DATA COLLECTION

**Overview: general considerations.** Small protein detection by MS is most frequently applied to discover small proteins from complex samples such as whole-cell lysates (Fig. 1; Table 1). Below, we outline the different methodological approaches for small protein discovery, highlighting differences to standard methods that are tailored to small proteins specifically (Fig. 2). We then describe parallel methods for detection of small proteins from

**TABLE 1** A selection of small protein discovery studies for bacteria and archaea using shotgun (bottom-up) proteomics and top-down approaches without proteolytic digest[a]

| Organism(s) | Taxonomy | Approach | Notes | Sample | Reference(s) |
|---|---|---|---|---|---|
| *Mycoplasma pneumoniae* | Bacteria | Shotgun | Term proteogenomics introduced; search six-frame translated genome | Whole cell lysate | 56 |
| *Mycoplasma mobile* | Bacteria | Shotgun | Used proteomics data in initial genome annotation of an organism | Whole cell lysate | 126 |
| *Shewanella oneidensis* | Bacteria | Shotgun | MS-based proteomics to improve genome annotation, used PTM data, studied several conditions | Whole cell lysate; subcellular fractionation | 127 |
| *Methanosarcina acetivorans* | Archaea | Top down | Top-down approach identified five unannotated small proteins (40–76 aa) | Whole cell lysate | 19 |
| *Staphylococcus aureus* | Bacteria | Shotgun | Effort to analyze the entire expressed proteome, combining different conditions and proteomics approaches | Subcellular fractionation | 128 |
| *Yersinia pestis* | Bacteria | Shotgun | Required 2 peptides to identify a novel protein | Subcellular fractionation | 57 |
| *Mycobacterium tuberculosis* complex | Bacteria | Shotgun | Custom database approach that merges information from different strains | Various | 129 |
| 46 species (bacteria and archaea) | Bacteria, Archaea | Shotgun | Large proteogenomic study; use of stringent PSM level FDR advocated | Various | 30 |
| *Escherichia coli* | Bacteria | Shotgun | High FDR among peptides implying novel proteins; trypsin + Lys-C | Whole cell lysate | 31 |
| *Helicobacter pylori* | Bacteria | Shotgun | Required 2 peptides to identify a novel protein; also used size exclusion chromatography | Whole cell lysate | 130 |
| 57 bacterial species | Bacteria | Shotgun | Large proteogenomics study on N-terminal methionine excision and PTM (N-terminal acetylation) | Various | 131 |
| *Saccharopolyspora erythrea* | Bacteria | Shotgun | Reannotated genome of organism with high GC content (transcriptomics, shotgun proteomics) | Whole cell lysate | 58 |
| *Bradyrhizobium japonicum* | Bacteria | Shotgun | Custom databases to find longer (84)/shorter proteoforms (132) (integration with dRNA-seq data) | Whole cell lysate | 84, 132 |
| *Synechococcus sp* | Archaea | Shotgun | Proteogenomic study of model cyanobacterium (8 conditions); global profiling for PTMs (1% PSM FDR) | Whole cell lysate | 133 |
| *Roseobacter denitrificans* | Bacteria | Shotgun | N terminomic combined with six-frame translation database to validate/correct N termini (+ alternative proteases; Glu-C, chymotrypsin) | Whole cell lysate | 27 |
| *Mycoplasma pneumoniae* | Bacteria | Shotgun | Integrated anaylsis to re-annotate genome (>300 transcriptome, >70 proteome datasets); evidence for internal starts, new small proteins | Various | 103 |
| *Pseudomonas stutzeri* | Bacteria | Top down (MALDI) | Small transmembrane subunit of cbb3 oxidase | Purified protein complex | 23 |
| Metagenomic study of grassland soil | Bacteria, archaea | Shotgun | Metagenome-assembled genomes as basis for meta-proteomics (custom database); integrate metabolomics; beyond culturable strains | Soil extract | 61 |
| *Listeria monocytogenes* | Bacteria | Shotgun | N-terminal enrichment (COFRADIC approach) (26); 2nd study (92) used spectral libraries and combined DDA and DIA | Whole cell lysate | 26, 29, 92 |
| *Xanthomonas euvesicatoria* | Bacteria | Shotgun | Reannotation of a plant pathogen with Shotgun data; confirmed expression of 5 novel proteins with Immunoblot (c-Myc tag) | Whole cell lysate | 134 |
| *Bartonella henselae* | Bacteria | Shotgun | Broadly applicable proteogenomic approach, custom databases validated 107/138 peptides with PRM (97) | Subcellular fractionation | 81, 97, 141 |
| *Methylobacterium extorquens* | Bacteria | Shotgun | N terminome study of a strain used for microbial rehabilitation and degradation of industrial pollutants | Whole cell lysate | 28 |
| *Escherichia coli* | Bacteria | Top down (MALDI) | Small transmembrane subunit of bd oxidase | Purified protein complex | 43 |
| *Bacillus subtilis* | Bacteria | Shotgun | Explored small protein enrichment strategies, different proteases, database searches; validation by PRM and spectral matching | Small protein enrichment | 12 |
| | Bacteria | Shotgun | | Whole cell lysates | 88 |

(Continued on next page)

**TABLE 1** (Continued)

| Organism(s) | Taxonomy | Approach | Notes | Sample | Reference(s) |
|---|---|---|---|---|---|
| *Salmonella* Typhimurium, *Deinococcus radiodurans* | | | Broadly applicable custom peptide DB; integrated Ribo-seq data and peptide fragmentation prediction | | |
| *Salmonella* Typhimurium | Bacteria | Shotgun | Integrated small protein prediction with Ribo-seq, shotgun, and other OMICS data; elucidated function of novel small proteins | Various | 104 |
| Prokaryotes from human microbiota, *Bacteroides thetaiotaomicron* | Bacteria, archaea | Shotgun | Prediction of ~4500 small protein families <50 aa; experimental evidence for selected examples (meta-transcriptomics/proteomics) Identified several small proteins in *Bacteroides thetaiotaomicron* | Various | 1 |
| Intestinal microbiota model system | Bacteria | Shotgun | Extended custom iPtgxDB to multispecies model (8 strains); meta-transcriptomics + meta-proteomics; spectral matching | Small protein enrichment | 60 |
| *Thermosynechococcus elongatus* | Bacteria | Top down (MALDI) | Small transmembrane subunit of photosystem II | Purified protein complex | 44 |
| *Staphylococcus aureus* | Bacteria | Shotgun | Broadly applicable approach; integrated shotgun and Ribo-seq data; automated evaluation of MS/MS spectra quality | Cytoplasmic extract | 87 |
| *Methanosarcina mazei* | Archaea | Shotgun/top down | Characterization of 36 proteoforms mapping to 12 small proteins with top down (2D-LC-MS) | Small protein enrichment | 89 |
| *Methanosarcina mazei* | Archaea | Shotgun | Multiprotease approach (SDS-PAGE) | Small protein enrichment | 35 |

[a]We apologize to authors of the many important studies we could not reference due to space restrictions. Preference was given to more recent studies, many of which have used higher accuracy MS instruments and small protein enrichment strategies, carried out validation of novel small proteins (e.g., by PRM or spectral matching), and put a larger emphasis on integration of RNA-seq, Ribo-seq or computational small protein prediction algorithms. Proteogenomic studies prior to 2014 are more thoroughly listed by Kucharova and Wiker (125).

| | Experiment type | 1A/B Preparation | 1C Digest | 1D LC | 1E MS | 2A-C Data analysis | 3-4 Validation/Prioritization | Notes |
|---|---|---|---|---|---|---|---|---|
| Discovery/Identification | Bottom-up | In-solution preparation, detergent-based lysis (7, 8) | Tryptic (7, 8) | Conventional C18 | ESI-DDA | Integrated data analysis pipelines (13, 16, 72, 73), FDR estimation (67), scoring algorithms (68, 70, 101) | Software-based; statistical tests | Most widely-used approach |
| | Alternatives | Detergent-free lysis, protein precipitation, organic extraction, size exclusion chromatography (SEC); enrichment (25) | Other proteases: Lys-C, Chymotrypsin, Proteinase K, Pepsin, Asp-N, Glu-C (12, 35) | Peptide pre-fractionation (SCX/WAX, high pH) (37, 38), C4/C8; CE | ESI-DDA Ion mobility (PASEF) (46) | Custom database search (81, 83–87); utilize multiple search algorithms; stringent FDR; inspect single peptide IDs; manual data curation | Experimental: synthetic peptides, other proteomics data (PRM (97), data integration (RiboSeq) and transcriptomics (support. evidence) *In silico*: predictions of fragment spectra (101, 102) | Most relevant adaptations for small protein discovery listed |
| | Avoid | Polymer-containing detergents (e.g., Triton, NP-40), FASP (21) | Conventional in gel digests (ext. washing), FASP (21) | Too long or too short gradients | Free MS/MS cycle time | Strict application of "Two peptide" rule (15) | "One-hit wonders" ( i.e., critically assess single PSM identifications) | Manual validation required |
| | Top-down | Detergent-free lysis or stringent cleanup; organic extraction; reversed SEC | N/A | C18/C8/C4/PLRP-S | ESI/MALDI MS/DDA | Deconvolution, manual assignment, Prosit (101) | 2nd generation experiments, e.g., genetic knock-out | Great potential for small protein discovery (20), limited availability |
| Targeted/Quantification | Bottom-up | In-solution perparation, detergent-based lysis | Tryptic | Conventional C18 (135, 136) | ESI/MALDI MS/DDA | Integrated data analysis pipelines (13, 16, 72, 73) | Synthetic peptides, internal standards (110, 112) | Most sensitive approach for quantitation |
| | Alternatives | Detergent-free lysis; organic extraction (23); reversed SEC (20) | *In silico* digest to choose suitable proteases (108) | Adjust chromatography for optimal separation and peak height | Inclusion lists for targets, test collision energies, PRM, alt. fragmentation (ETD, PTR, UVPD); DIA (47) | Publicly and commercially available software packages | PRM for validation and quantitation (12, 81, 111) | DIA most promising strategy for broad quantitative studies |
| | Top-down | Desalting (e.g., ZipTip); organic extractions, SPE (23) | N/A | No LC separation | ES/MALDI (23, 42–44) | Publicly and commercially available software packages | Synthetic peptides, internal standards (110, 112) | Can yield full sequence and proteoform information |
| | Alternatives | Test organic extraction solvents/SPE; reversed SEC | N/A | Attempt LC separation to reduce sample complexity | Modified instruments for high m/z values (23, 42–44) | Deconvolution, manual Open searches (100), manual de novo sequencing (90) | PRM (97, 110) | IM-assisted separation promising alternative |
| Elucidate Function | Forward Genetics WT vs. knockout | Bottom-up proteomics | Tryptic | C18 | ESI-DDA | Differential protein expression analysis | Complementation (104) | Identify function (137), mechanism of action |
| | Interactomics | Immobilized bait, bottom-up proteomics | Tryptic | C18 | ESI-DDA | Data analysis pipelines including statistical evaluation | Negative controls (beads, scrambled bait, reverse IPs...) | Identify interactors (23, 121, 122) |
| | Binding Studies Structural analysis | HDX-MS (138)/ Crosslinking-MS (139)/ Native MS (140) | Pepsin/Trypsin or other proteases/digest-free | Direct infusion, C18 or N/A | ESI-DDA, ESI-DIA | Method-specific data analysis pipelines | Biochemical control experiments (e.g., SDM or alanine-scanning) | Identification and mapping of binding sites (123) |
| | Subcellular Localization | Molecular (fluorescent tag); Enrich subcellular fractions | Tryptic | C18 | ESI-DDA, ESI-DIA | Relative quantification over subcellular localization (e.g., secreted, membrane, cytoplasmic); often combined with machine learning (e.g., (141)) | Microscopy (tagged proteins); PRM on additional biological samples to confirm subcellular localization (81) | For eukaryotes: LOPIT (142) (high resolutiion separation of organelles and subcellular compartments) |

**FIG 2** Overview of the major steps of the most common MS-based workflows for discovery/identification of small proteins, their targeted analysis (for quantification), and for the functional characterization of novel and known small proteins. The numbering of the steps is aligned with Fig. 1, with corresponding text sections indicated by white circles, as follows: 1B, "Sample Preparation and Data Collection—Preparation and enrichment for small proteins"; 1C, "Sample Preparation and Data Collection—Protease digestion"; 1D, "Sample Preparation and Data Collection—Liquid chromatography"; 1E, "Sample Preparation and Data Collection—Ionization and data acquisition"; 2A, "Data Analysis—Overview: the relevance of genome sequences for proteogenomics"; 2B, "Data Analysis—Creation of custom search databases"; 2C, "Data Analysis—Stringent FDR control"; 3, "Validation of Novel Small Protein Candidates"; 4, "Prioritization/Selection of Novel Small Proteins." Alternative approaches are listed and selected references provided.

low-complexity samples, and a separate application that robustly detects small proteins from protein complexes where a small protein has been detected, but its specific identity is unknown.

An important consideration in method choice is the use of bottom-up or top-down proteomics (Box 1). The vast majority of proteomics studies make use of bottom-up methods, although top-down approaches are ideally suited for small proteins. However, practical considerations in method choice are often limited by the expertise and instrumentation available in a core facility. Bottom-up methods are widely available at high quality in most laboratories and core facilities, and sample introduction from bottom-up small protein preparations can be reasonably approached by routine-use setups. In general, we recommend bottom-up analyses for complex samples and top-down approaches for low-complexity samples, due to the ease of data acquisition of the two approaches. Low-complexity samples can also be analyzed directly using matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) mass spectrometry, providing direct information on the different small proteins present in a sample and their molecular weights.

Below, we outline general considerations for each step of an MS-based analysis of a sample for small proteins. The experimental workflow can be divided up in four parts (Fig. 2), described in turn below.

**Preparation and enrichment for small proteins.** Depending on sample complexity and the type of MS analysis, different preparation steps are required. In general, bottom-up analyses focus on efficient denaturation and digestion, while top-down approaches mainly utilize size exclusion steps to enrich small proteins (20). We outline the advantages and limitations below, and suggest modifications to existing protocols as well as alternative strategies (Fig. 2).

**(i) Sample preparation for bottom-up proteomics.** Sample preparation for bottom-up proteomics typically includes denaturation, reduction, and alkylation steps to ensure optimal access of the proteases to unfolded proteins, and to prevent thiol oxidation and disulfide bridge formation. *Per se*, these steps do not impede small protein detection; however, most protocols were designed for protein molecular weights $>$10 kDa, and include efficient removal steps for small molecular weight contaminants. In particular, filter-based methods such as filter-aided sample prep (FASP) (21) and suspension traps (S-Traps) (22) efficiently remove detergents and contaminants prior to digestion, leading to improved data quality and signal-to-noise ratios for studies of large proteins and proteomes. However, these approaches, if used conventionally, also actively de-enrich small proteins (Fig. 2). We therefore suggest using either in-solution digests without any cut-off filters, or modifying solid-phase-assisted digest protocols to avoid removal of small proteins by reducing stringency in the wash steps, described below in "Protease digestion." However, if sample composition requires extensive washing, small protein detection remains challenging using this approach.

**(ii) Sample preparation strategies using protein precipitation steps.** The same considerations apply for sample preparation strategies using protein precipitation steps that are often applied to concentrate and purify target proteins, e.g., organic solvent precipitation using acetone, methanol, chloroform or other solvents (Fig. 2). Small proteins may remain in the soluble fraction due to their high solubility in organic solvents, and hence may be discarded. We recommend avoiding precipitation steps to ensure small proteins are not lost. Despite the challenges associated with organic extraction, membrane proteins benefit from organic extraction due to their high hydrophobicity. Moreover, top-down approaches benefit from organic phase extractions, as detrimental salt adduct formation is minimized. Our and other labs found that a combination of low organic solvent concentration with organic solvent-resistant filter membranes identified peptides that were not accessible using conventional bottom-up analysis, thereby significantly expanding the identified small proteome in different organisms (23, 24).

**(iii) Enrichment using typically discarded fractions.** While the above modifications reduce the likelihood of de-enriching small proteins, it is also possible to actively enrich small proteins by using only the fractions that are typically discarded in a conventional

bottom-up proteomics preparation. This can be achieved by either (i) using molecular weight cut-off membranes, with large proteins being trapped using a filter-membrane with a low molecular weight cutoff, and the flow-through containing small proteins being retained for subsequent analysis, or (ii) subjecting the sample to a standard organic precipitation; the precipitated proteome is discarded, and the supernatant is digested and submitted to LC-MS/MS analysis. Both of these techniques can also be applied to top-down approaches. The small protein fractions isolated from these approaches typically coenrich for other interfering small molecules from the sample and common preparative contaminants. Hence, special care must be taken to remove these prior to MS analysis. In addition to reverse-phase desalting prior to analysis, we recommend desalting using strong cation exchange by solid-phase extraction (SCX-SPE).

**(iv) MS approaches.** MS approaches commonly involve a prefractionation step to reduce sample complexity, with the most widely used methods being separation of the sample by (i) centrifugation, into soluble and membrane fractions; (ii) sucrose density gradients, based on hydrodynamic radius; (iii) size exclusion or ion-exchange chromatography; and (iv) SDS-PAGE. These methods can be applied to study small proteins, but separation by SDS-PAGE requires modification to prevent loss of small proteins from the sample; small proteins often stain poorly with Coomassie, or pass through the gel with minimal retention. Isolating the dye front (typically bromophenol blue) and a narrow range of low MWs will enrich small proteins. If no bands are detected using Coomassie staining, we suggest silver staining due to its higher sensitivity. After excision of the respective gel areas, we recommend only minimal wash steps, as small proteins may diffuse out of the gel into the wash buffer. When small proteins are not efficiently recovered with these conventional gel-based approaches, distinct MW regions can also be selected by electroeluting specific regions or by gel-eluted liquid fraction entrapment electrophoresis (GELFrEE) and downstream processing in the liquid phase (25).

Another useful biochemical purification strategy is a combined fractional diagonal chromatography (COFRADIC) approach (26), where chemical or enzymatic modification prior to and/or after the digestion step allows for specific enrichment for a certain peptide fraction, e.g., N-terminal peptides (Table 1), which greatly reduces the sample complexity. This method requires a suitable database containing alternative start sites (see "Creation of custom search databases" below) but then can provide direct information on translational start sites (27–29).

In general, we recommend fewer preparation and manipulation steps for less complex sample mixtures prior to MS analysis. For purified protein complexes, we suggest minimal desalting and purification using, for example, reversed-phase tips or cartridges, with samples then analyzed directly by MALDI-MS or -MS/MS. For complex protein mixtures and lysates, we recommend modifying S-trap protocols with minimal washing, organic extractions, or reciprocal purification using molecular weight cut-off filters.

**Protease digestion.** Protease digestion is the defining step of bottom-up proteomics. Digestion is often performed in-solution, in-gel, or during filter-based sample preparations. These digestion protocols use extensive washing of the sample that inherently de-enriches small proteins; due to their size, small proteins can pass through filter membranes, may not adsorb sufficiently (or may stick too tightly) to solid phases, and may be flushed during wash steps. Hence, the digest should be conducted in a way that specifically ensures inclusion of, and access to, small proteins, which can range from in-gel and in-solution digests to solid-phase supported protocols. The advantages and limitations of different digestion strategies for small proteins are detailed below.

Digestion of small proteins is challenging due to the intrinsic scarcity of necessary protease cleavage sites, and the often short length of the few peptides that are generated. In general, Trypsin, mostly alone or in combination with LysC, has been the workhorse for bottom-up proteomics due to cleavage specificity (cleaves after K and R) and activity in typical preparation conditions. This excellent enzyme/combination of enzymes, in particular in consecutive denaturing conditions, has also been widely used to identify small proteins in different prokaryotes (30, 31) (Table 1). Alternatively, proteases targeting other amino acids,

such as Asp-N and Glu-C, are used for studying posttranslational modifications such as K acetylation (32), or for specific target protein groups (33). However, the high specificity of these proteases limits their activity against small proteins, and specificity for basic amino acids limits activity against small membrane proteins in particular, which are deficient in charged residues in membrane-spanning regions. Proteases with specificity for other amino acids, or lower specificity, such as chymotrypsin (mainly F, Y, and W > M and L), proteinase K (hydrophobic amino acids), pepsin (nonspecific at pH <2), elastase (A, V, S, G, L, and I), or subtilisin A (nonspecific) (34), are promising alternatives for identifying small membrane proteins (35) (Fig. 2). However, using a protease with lower specificity comes with several challenges: (i) proteases with limited specificity require careful testing and optimization of protease:protein ratio, digestion time, digestion buffer, and digestion temperature; (ii) the abundance of each proteolytic peptide species will be lower due to the statistical nature of the cleavage, i.e., each protein is potentially cleaved into multiple, different, overlapping peptides, each lower in abundance compared to a specific digest where only a single peptide is produced from each segment; (iii) the lack of a charged amino acid at the C terminus will lead to MS/MS spectra with a lower number of usable fragment ions; (iv) the search space for database matching is increased substantially, with consequences for the FDR of protein identification (see protease: protein "Stringent FDR control" below); and (v) nonspecific cleavage compromises stoichiometric recovery, limiting quantitation of peptides and proteins. We therefore recommend to initially use trypsin/LysC as proteases for bottom-up experiments, except if membrane proteins are targeted specifically. The choice of enzyme largely depends on the sample and the small proteins of interest: some hydrophobic small proteins still yield proteolytic peptides with high-quality, information-rich fragment spectra (36), while other small membrane proteins are completely inaccessible to bottom-up approaches and require top-down analysis (23). If trypsin/LysC proves ineffective, in particular for samples with low complexity, alternative proteases can be tried, as described above (12, 27, 35). After successful digestion, regardless of the protease used, proteolytic peptides in very complex mixtures can be further fractionated prior to LC separation for more sequencing and data depth (37, 38).

**Liquid chromatography.** In bottom-up methods, proteolytic peptides are typically separated by liquid chromatography prior to acquisition of MS (ionized peptide precursors) and MS/MS data (fragmentation pattern of peptides) (135, 136). The most widely used solid phases for separation are different derivatives of $C_{18}$ bed materials (Fig. 2) that separate the peptides based on hydrophobic interactions with the column bed, and a gradient elution of aqueous into hydrophobic solvent. A typical LC-MS setup might include a trap column to facilitate removal of small molecular weight contaminants prior to separation on the main chromatographic column. Current LC-MS setups for bottom-up proteomics are optimized for peptide lengths between 8 and 25 aa. They will thus perform equally well for proteolytic peptides as well as for short small proteins, and will also work well for top-down experiments for many small proteins. Depending on their amino acid composition, however, small proteins >40 aa may not elute efficiently from a conventional $C_{18}$ column, even at high concentrations of organic solvent. Membrane proteins also often bind strongly to $C_{18}$ phases due to their high hydrophobicity, and even small membrane protein-derived peptides may stick to the $C_{18}$ phase irreversibly. In these cases, other bed materials with lower hydrophobicity ($C_4$/$C_8$) and larger pore sizes (>300 Å) may offer higher recovery in bottom-up and top-down experiments. This often comes at the cost of decreased retention and recovery of hydrophilic peptides and small proteins.

For the analysis of complex bottom-up or top-down samples, extending the length of the chromatographic gradient may offer significant improvements in data depth, as more acquisition time is available (39). However, gradient length should be adjusted to fit the complexity of the sample: chromatographic peaks broaden with increasing overall gradient length, and the associated decrease in intensity per spectrum impairs detection and analysis of low-intensity candidate ions. In general, we recommend short gradients (15 to 60 min) for samples with low complexity and long gradients for

samples with high complexity (90 to 180 min). The chromatographic separation of peptides and small proteins may well be further improved in the future by different bed materials, narrow column diameters, or micropillar array columns (40, 41).

**Ionization and data acquisition.** Most proteomics labs operate electrospray ionization (ESI)-based mass spectrometers due to their direct interface with LC and high ionization efficiency. This combination allows fast acquisition of high quality data, and is equally applicable to conventional bottom-up proteomics approaches and top-down experiments (Fig. 2). ESI-based ionization only falls off for membrane proteins that are highly hydrophobic. For dedicated small membrane protein studies, MALDI ionization can be a promising alternative, although MALDI data acquisition to LC is time and cost intensive and rarely used.

For low-complexity samples, we recommend performing MALDI-MS measurements to get an initial overview of small proteins present. In MALDI-MS, samples are spotted in a solid matrix on a conductive plate and ionization is achieved using a laser that transfers energy via the matrix molecules to the proteins of interest. MALDI usually imparts lower charge states on the analytes and thus makes MS and MS/MS spectra much easier to interpret than ESI spectra, where most peptides typically carry 3 to 5 charges. MALDI spectra can also provide initial information not only on the molecular weight of any candidate small proteins, but also whether multiple candidates are present that cannot be separated efficiently using SDS-PAGE. Sample preparation could either use organic extraction or a desalting step using either $C_{18}$ (soluble small protein) or $C_4$/$C_8$ (membrane small protein) tips or SCX-SPE-based desalting as outlined above. After solvent removal, these purified complexes can then be directly spotted and submitted to MALDI-MS analysis. If a MALDI-MS/MS instrument with sufficient isolation power and resolution is available, small proteins can also be identified directly from these spectra (23, 42–44). If protein identities cannot be directly determined, MALDI-MS survey spectra can already yield valuable information on the number of small proteins and their proteoforms in a sample.

For LC-MS/MS analysis of low-complexity samples, more measurement time in an LC-MS experiment can be spent on each candidate ion. We therefore recommend testing multiple collision energies and other "slow" fragmentation methods (ETD/ETHCD/UVPD/PTR) to generate complementary data, and increase the fragment coverage of the small proteins of interest.

We anticipate that ion mobility-coupled separation of ions will improve the scope and depth of conventional Data Dependent Analysis (DDA)-based bottom-up proteomics analyses also for small proteins (45, 46). In addition, data-independent analysis (DIA) based methods may well further improve bottom-up and top-down analyses of small proteins in the future (47) (Box 1). DIA is an acquisition approach where large *m/z* ranges are sequentially fragmented independent of precursor abundance, providing information-rich MS and MS/MS data sets that can be mined. Since no precursors are "discarded," presumably numerous undiscovered small proteins reside in the data awaiting detection. This approach, facilitated in recent years by faster and more sensitive instruments, is expected to benefit substantially from further improvements in analytical pipelines.

## DATA ANALYSIS

**Overview: the relevance of genome sequences for proteogenomics.** Unbiased identification of small proteins from MS data requires a reference genome sequence. Owing to advances in DNA sequencing technologies and assembly algorithms, complete genome sequences can be readily and cost efficiently assembled *de novo* (48, 49). Complete genomes provide an optimal basis for functional genomics (50) and for small protein discovery. They can be used to create both reference or custom search databases that link to genomic coordinates (e.g., NCBI's RefSeq or Ensembl), an advantage over the widely used universal protein (UniProt) database that provides additional functional information and annotations but lacks the link to the genome sequence (51). Notably, advances in genome annotation have lagged behind the sequencing

revolution (52), and several unresolved issues remain. These include discrepancies in the number of protein coding sequences (CDSs) predicted by different reference annotation centers for the same genome sequence, discrepancies with respect to the precise protein start sites (53, 54), and underrepresentation of genes encoding stably expressed and functional small proteins, i.e., false negatives (55). Furthermore, prediction of some spurious ORFs (false positives) (11, 56, 57) is another issue, especially for genomes with a high GC content, e.g., many actinomycetes (58). As current gene prediction algorithms cannot separate truly coding sORFs from the overwhelming majority of spurious, random sORFs (14), varying size thresholds between 50 and 100 amino acids have been used to minimize the number of false small protein predictions in genome annotations. As a standard database search of MS data will only identify proteins contained in the list of annotated genes, custom databases are needed for small protein discovery (Fig. 1).

Initiatives like the Genomic Encyclopedia of Bacteria and Archaea will significantly boost small protein studies in taxonomically diverse organisms (59). We expect more studies to explore novel small proteins from moderately complex samples such as synthetic consortia (60) and from complex metagenomes (1, 61) (Table 1), based on as-complete-as-possible metagenome-assembled genomes (62, 63).

Proteogenomics, a term first coined in 2004 (56), refers to the use of MS data to provide expression evidence for CDSs missed in genome sequences (for reviews, see references 64 to 66). Searching tandem MS/MS data against one of several different flavors of custom databases allows identification of small proteins plus novel start sites and evidence for expressed pseudogenes, which would be missed if using standard reference databases; for a concise selection of small protein discovery studies using this approach in bacteria and archaea, see Table 1. Several data analysis solutions have been developed that represented important advances for the proteomics field in general and improved overall data quality: (i) search strategies that allow estimation of the number of false positive PSMs and the FDR, including so-called "target-decoy searches," where a peptide spectrum that matches a decoy sequence such as a reversed or randomized protein sequence is considered a false positive identification (67); (ii) statistical models that employ different scoring functions or machine learning approaches to improve the accuracy and robustness of the automated PSM step, i.e., that maximize the number of confident peptide matches with minimal false positive rates (68–70); and (iii) protein inference algorithms that deduce which proteins (or protein groups) were initially present in a sample based on the experimentally observed peptides, many of which (especially in eukaryotes, but also in custom databases [see below]) are ambiguous and imply several proteins (71). Integration of these tools into publicly available data analysis pipelines for shotgun MS data has enabled researchers to carry out large parts of the data analysis themselves (13, 16, 72, 73) (Fig. 2). However, even when using highly accurate MS instruments and optimized small protein sample preparation strategies (12, 25), MS-based novel small protein discovery is still challenging and requires custom databases and a stringent control of the FDR.

**Creation of custom search databases.** All customized databases add short ORFs that potentially encode true small proteins. Most often, custom databases are based on a six-frame translation of the genome sequence. While this approach guarantees inclusion of all possible gene products, it creates a substantially larger search space compared to the standard reference search database: typically, 20 to 50 times more proteins and ~4 to 8 times more distinct tryptic peptides (74). Hence, it requires substantial analysis time to identify the respective reading frame and the precise novel small protein boundaries, and to ensure that the peptides are unique (i.e., unambiguously identify one protein). Consequently, various approaches have been developed that aim to reduce the search space. To limit the database size and thereby improve PSM statistics (75), transcriptomic or Ribo-seq data from the conditions of interest can be used to create a smaller database that only considers the genomic regions transcribed or translated. Tailored transcriptome- or Ribo-seq based databases were

successfully used to identify small proteins in different eukaryotes (76–79) and prokaryotes (80). For more complex eukaryotes, where larger numbers of protein families and splicing lead to a substantial percentage of peptides being ambiguous, this is a promising approach that simplifies protein inference. However, for prokaryotes that lack splicing and have smaller genomes, creating a single database applicable to all conditions is more versatile, and allows identification of small proteins that are conditionally expressed and hence could be missing from RNA-seq or Ribo-seq data. We recently developed an approach to address the dilemma of largely differing genome annotations, missing sORFs, and the higher fraction of ambiguous peptides in large custom databases, by constructing "integrated proteogenomics databases for the protease of choice" (iPtgxDBs) (81). In a pre-processing step, all annotation sources (reference annotations, *ab initio* gene predictions, and a modified form of a six-frame translation that considers the most common alternative start codons GTG, TTG, and CTG [82]), plus all proteins down to a user-specified length threshold, are integrated and consolidated, capturing both overlap and differences. By adding peptides that imply potential new start sites, an iPtgxDB can be kept minimally redundant; it contains ~1/3 fewer peptides than a 6-frame translation. Notably, up to 95% of the peptides unambiguously match one protein. For more information about preprocessing and a public web server to create such iPtgxDBs, see https://iptgxdb.expasy.org. Several other public proteogenomics software solutions or custom search database approaches have been developed to support researchers with small protein identification (83–87), some of which also provide the useful option for integrative data visualization in a genome viewer, or include an option to consider predicted peptide ion intensities (88). The more recent proteogenomics studies emphasize efforts for data validation and integration of orthogonal data sets (see Table 1 for some examples). Top-down studies where novel small proteins were identified via a custom database search are slowly gaining momentum (Table 1) (19, 23, 89). In addition, one can also rely on top-down data in combination with *de novo* sequencing, a completely database-independent approach that infers small protein sequences by searching for amino acid-specific mass increments between adjacent fragment peaks (90). We expect that improvements in the accurate and comprehensive prediction of small proteins from DNA sequence alone will facilitate the creation of more focused databases, thereby also improving the PSM statistics.

**Stringent FDR control.** A second key consideration for small protein discovery concerns false discovery rate (FDR) control (Fig. 1). Historically, more confident protein identification from MS/MS data has relied upon the "two-peptide rule" (see the introduction and reference 15). However, this rule is ill suited for small proteins, which, due to their small size, are often only identified by a single peptide (91, 92). To minimize the identification of so-called "one-hit wonders," several PSMs (independent observations of the same peptide sequence) should be required. As small proteins are often of lower abundance (81, 93) and hence are expected to produce few MS-detectable peptides, researchers have analyzed multiple biological conditions to increase their odds of identification, and used proteases other than trypsin (Table 1) that can add more spectral evidence to support novel small protein identification (12, 35, 60). It is important to note that the number of false positive identifications will also increase as the size of the MS data set increases (94, 95). We suggest performing the steps outlined below before assessing any potential novel hits, as it can be rather discouraging to see your "top novel small protein candidates" fail to withstand rigorous evaluation.

Another critical consideration when interpreting FDR estimates from MS data analysis is that the false positive protein identifications are distributed unevenly between large (mostly annotated) and small (mostly unannotated) proteins; custom databases will contain many more predicted small proteins than annotated proteins, such that the likelihood of a PSM matching a predicted small protein by chance is much higher than the likelihood of matching an annotated protein by chance. Moreover, there will be proportionally more PSM matches to annotated proteins than to small proteins,

causing the overall FDR estimate to be skewed toward the value for annotated proteins. Consequently, reported protein-level FDRs substantially underestimate the true FDR for novel small proteins (71, 74, 94, 95). We and others have advised to strictly control the PSM FDR to 0.1% or (for very large data sets) even lower (30, 81), in order to achieve an estimated protein-level FDR for small proteins of around 1% (Table 1). Alternatively, if a more relaxed FDR is applied, the potential for false positives can be addressed with more extensive validation experiments (see below).

We also recommend applying a resource-dependent filter on top of the global FDR filtering step, as proposed earlier (64), which essentially requires more spectral evidence from less credible prediction sources for novel small proteins, e.g., from *ab initio* gene predictions and a six-frame translation (74); we have used thresholds of 2 PSMs for peptides implying RefSeq-predicted proteins, 3 PSMs for peptides of candidates from the excellent *ab initio* gene predictor Prodigal (96), and 4 PSMs for *in silico* ORFs solely predicted based on potential start codons. These values were in part motivated by a study on *Bartonella henselae* (81), where we successfully validated a novel small protein identified by one peptide and 3 PSMs in a very large data set with parallel reaction monitoring (Box 1) (97). Moreover, we recommend considering aspects like repeated identification in biological replicates, high spectral quality score (e.g., using the MS-GF+ search engine scores [98]), *q* values (a statistical confidence measure provided along with the posterior error probability by Percolator [99]), and second peptide searches, which can identify less abundant peptides in chimeric spectra (overlapping fragmentation patterns of co-eluting peptides), potentially including novel small proteins (88). As proteomics workflows rely on database searches, in some cases, modified peptides of abundant proteins may represent a better match than the top-scoring PSM implying a novel small protein. They were not considered because the modification was not specified in the database search; typically, only few modifications are specified in closed searches to keep the search time manageable. Fast open search software solutions like MSFragger (100) represent a valuable option to identify and eliminate such false positives.

Recently, software tools were developed that accurately predict the retention time of peptides and peptide fragmentation intensity purely *in silico* with very encouraging results for large proteomics studies and improving the number of correct PSMs (101, 102). In the near future, we foresee that such tools will also allow users to increasingly leverage data from data independent analysis (DIA) workflows.

## VALIDATION OF NOVEL SMALL PROTEIN CANDIDATES

**Overview: options for validation.** Newly identified small protein candidates from MS or non-MS data, such as RNA-seq, Ribo-seq, computational prediction, or genetic inference, need to be validated before they can subsequently be prioritized for further study (Fig. 1). As genome/proteome-scale methods identify false positives as described above, it is critical to provide independent lines of evidence to support the existence of a novel small protein identified by these methods. Transcriptomic data (ideally obtained from the same samples used for proteomic analysis [11, 81, 103, 104]) can be considered supporting evidence for MS-identified small proteins, with products of strongly expressed genes more likely to be detected at the protein level; however, not all RNAs are translated into a stable or MS-detectable protein. (105). Ribo-seq data can also complement MS-based evidence (104), and are particularly well-suited for identification of protein start sites (6, 93, 106), which is more challenging with MS-based approaches. Fifteen out of 16 novel small proteins jointly implied by Ribo-seq data and predicted by sPepFinder (107) in *Salmonella enterica* serovar Typhimurium were validated by MS, highlighting the complementarity of MS and Ribo-seq as methods to identify small proteins (104). Even more valuable than Ribo-seq data is direct evidence for the expression of a small protein, which can be obtained using classical immunochemistry such as Western blot analysis. However, either antibodies would have to be raised against each candidate, or a chromosomal tagging approach that introduces

small sequence tags would need to be carried out (106). A simpler alternative is to use additional MS-based approaches that can be applied on a larger scale.

**MS-based validation of small proteins identified using non-MS methods.** For MS validation of putative small proteins identified by non-MS methods, we recommend using the approaches described above for sample preparation and MS data collection. Since the sequence of the putative novel small protein is known, protease cleavage prediction tools such as "Peptidecutter" (UniProt) or "ProteaseGuru" (108) serve as valuable tools to choose a suitable protease that produces detectable proteolytic peptides. Additionally, *in silico* prediction of the proteolytic peptides, based on their sequence, can be used to reanalyze previously acquired data to look for any signals in the respective retention time and *m/z* windows as a primary indication if the candidate peptides were present in the first place (101, 109). However, to date, *in silico* predictions only work well for tryptic peptides (101), so this approach cannot be used reliably for targets requiring alternative proteases or for top-down approaches. If a signal is detected, targeted methods like parallel reaction monitoring can be used to selectively accumulate and fragment ions in the specific retention time and *m/z* windows. In parallel reaction monitoring, specific RT and *m/z* values are generated for synthetic peptides that uniquely identify novel small proteins (110) to ensure that the recorded transitions do not stem from a coeluting peptide from another protein (111).

**Validation of MS-identified small proteins using synthetic peptide mimics.** For small proteins identified from complex samples by MS, a broadly applicable and economic approach for validation is to purchase synthetic peptides that match the putative small protein(s), determine their retention time, capture their fragmentation patterns, and compare these spectra to the experimentally observed small protein spectra. High-scoring matches add confidence to the initial MS identifications (60). Due to the relatively low cost of peptide synthesis, this approach can be applied to confirm a large number of proteins, thereby supporting the option of using a less stringent FDR for initial data analysis, and investing more effort into the validation. In a large study on *B. henselae*, overall almost 80% of 136 peptides implying novel sORFs, novel start sites, or expressed pseudogenes, were successfully validated by parallel reaction monitoring, with lower success rates for N-terminal extensions and *in silico* ORFs (81). Alternatively, databases of spectra associated with specific peptides (spectral libraries) can be generated from experimental data (typically DDA data, which results in cleaner, higher quality reference spectra) and used to more robustly identify and quantify candidates with low signal intensities or incomplete fragment spectra. However, in a study of *Bacillus subtilis*, parallel reaction monitoring-based validation proved more discriminatory than a spectral library approach (12).

A modification of this approach that is more sensitive but also more expensive, is based on synthetic, "heavy" isotope-labeled "AQUA" peptides (112). These heavy peptides can be used both for validating small protein candidates as well as for quantitation of small protein candidates with low signal intensities or insufficient fragment spectrum data for an unambiguous identification, but have to be designed and synthesized for each small protein of interest. The heavy peptides are added to samples containing the putative small protein(s) of interest, and provide a direct reference for retention time, signal intensity, and fragmentation pattern, facilitating unambiguous verification of novel small proteins.

## PRIORITIZATION/SELECTION OF NOVEL SMALL PROTEINS

There is increasing evidence that some, perhaps many, small proteins that can be detected by transcriptomic or MS-based approaches are nonfunctional, as defined by the lack of an effect on cell fitness (6, 93, 113, 114). Hence, prioritization of novel small proteins for further study is a critical step (Fig. 1). Classical annotation approaches that use similarity to functionally characterized proteins are generally ineffective for small proteins due to their size. For example, the overall percentage of CDSs annotated as hypothetical (i.e., lacking any functional annotation) is around 12% (based on a meta-analysis of ~21,400 completely sequenced prokaryotes and 80 million encoded proteins). For the

subsets of CDSs with a length below 100 aa or below 50 aa, this percentage rises to 41% for each subset, respectively. Similarly, in a landmark study, 4,500 conserved small protein families were identified using comparative genomics of metagenomic data sets, but 90% of the families lack a predicted protein domain, and 50% were not even annotated in reference databases (1). Important caveats are that the ability to predict a domain is somewhat dependent on protein size, and the lack of a predicted protein domain does not necessarily indicate the lack of a biological function. Small protein functional annotation is expected to improve as more small proteins are characterized, but alternative approaches for prioritization are recommended.

As an initial prioritization step, we recommend searching databases of protein annotations (e.g., eggNOG, whose predictions can be retrieved for a particular taxonomic level [115]), domains/motifs (e.g., InterProScan or LipoP [116, 117]), and predicted protein-protein interactions (118). The large majority of small proteins are not expected to generate significant hits with these tools due to the small protein size, but this is an extremely easy approach to try, and it has proven to be effective in a few cases. For example, two novel small proteins identified in *B. henselae* were predicted by LipoP (117) to contain signal peptides for the Sec/SRP secretion machinery, and both proteins were experimentally detected in membrane fractions and later validated with parallel reaction monitoring (81).

Analysis of phylogenetic sequence conservation, including determining the ratio of synonymous to nonsynonymous SNPs between homologues, is highly recommended for prioritization of novel small proteins, since evidence of purifying selection supports a functional role (1, 93, 113). The genomic context of the candidate ORFs is yet another important consideration, since short ORFs overlapping larger genes may be conserved due to selective pressure on the overlapping gene. Genomic context can also be informative for predicting the specific functions of small proteins or short ORFs, although providing evidence for function *per se* (rather than a specific function) is an important prerequisite. Short ORFs immediately upstream of genes/operons may function to regulate downstream transcription or translation (119). Short ORFs can also encode proteins that have related functions to the proteins encoded by the neighboring genes, such as a novel ribosome-related small protein family identified in the human microbiome study, where the novel small protein was encoded just downstream of two annotated ribosomal proteins (1). We therefore recommend using model organism resources and databases that contain operon assignments and predictions.

An additional level of prioritization can be achieved by integrating differential gene and protein expression data, where available (104). Regulated expression provides evidence for function, but can also suggest specific functions, based on the conditions and strains tested. Lastly, codon usage can provide evidence for function, since codon sequences are non-random with respect to the nucleotide content of the genome (93, 120). This is particularly true for genomes with extreme A/T or G/C content. As for analysis of sequence conservation, care must be taken to analyze only the codon usage of ORF regions that do not overlap other genes.

Each of the approaches described above can provide evidence for small protein functions that can be used for prioritization. However, we recommend relying on multiple, independent lines of evidence, since any individual method is likely to generate a substantial number of false positives. It is also the case that the statistical power of analyses of sequence conservation and codon usage is strongly length-dependent. Hence, failure to identify evidence for small protein function using these methods should not be interpreted as a lack of function.

## CONCLUDING REMARKS

The approaches described above are intended to provide a set of best practices for the detection and validation of small proteins using MS. We also discuss further bioinformatic analysis and prioritization strategies, which are critical for choosing the most promising candidates for subsequent physiological studies. MS approaches can also be

used for these functional studies, and provide valuable information on the roles and functions of the novel small proteins. (See Fig. 2 for details [137–142].) For example, MS analyses can be combined with subcellular fractionation to associate small proteins with specific cellular compartments or complexes (81, 104, 141, 142). MS-based proteomics also represent a powerful tool for identification of protein interaction partners for small proteins of interest, when combined with pulldown approaches (for general reviews, see references 121 and 122). A unique advantage of small proteins for these approaches is the ability to chemically synthesize the proteins on a resin, or with modified amino acids that facilitate cross-linking to resin. This obviates the need for epitope tags, which can disrupt protein function. The use of nonnative amino acids can also be incorporated into pulldown methods. A recent study relied on incorporating a nonnative amino acid into small proteins to facilitate cross-linking to interaction partners; while a tag was required to precipitate the small protein, the cross-linking step allowed for the detection of weaker or more transient interactions, as shown for the interaction partners of 24 human small proteins (123). This also highlights the potential of cross-linking mass spectrometry to gain further insights into small protein function. A more detailed description of the options available to elucidate the functions of novel small proteins is provided in the accompanying article of this special issue (124).

In summary, mass spectrometry-based proteomics approaches represent a powerful tool to identify and characterize small proteins. However, the standard workflows available in the field need to be adjusted depending on sample type and complexity, as outlined above. We anticipate that these strategies will provide an excellent starting point for exploratory studies of prokaryotic small proteins.

## ACKNOWLEDGMENTS

We thank Benjamin Heiniger (Agroscope) and Jakob Meier-Credo (Max-Planck Institute for Biophysics) for meta-analyses and feedback on the manuscript. We thank Gisela Storz for comments on the manuscript.

C.H.A. acknowledges funding from the Swiss National Science Foundation (grant 197391). J.T.W. and M.M.C. are supported by the National Institutes of Health (R01GM139277). J.D.L. gratefully acknowledges financial support by the Max Planck Society and by Special Priority Program 2002 of the German Research Council (SPP2002).

## REFERENCES

1. Sberro H, Fremin BJ, Zlitni S, Edfors F, Greenfield N, Snyder MP, Pavlopoulos GA, Kyrpides NC, Bhatt AS. 2019. Large-scale analyses of human microbiomes reveal thousands of small, novel genes. Cell 178:1245–1259.e14. https://doi.org/10.1016/j.cell.2019.07.016.
2. Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63. https://doi.org/10.1038/nrg2484.
3. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324:218–223. https://doi.org/10.1126/science.1168978.
4. Vazquez-Laslop N, Sharma C, Mankin A, Buskirk A. 2021. Identifying small ORFs in prokaryotes with ribosome profiling. J Bacteriol https://doi.org/10.1128/JB.00294-21.
5. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, Stadler PF, Vogel J. 2010. The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 464:250–255. https://doi.org/10.1038/nature08756.
6. Meydan S, Marks J, Klepacki D, Sharma V, Baranov PV, Firth AE, Margus T, Kefi A, Vázquez-Laslop N, Mankin AS. 2019. Retapamulin-assisted ribosome profiling reveals the alternative bacterial proteome. Mol Cell 74:481–493. https://doi.org/10.1016/j.molcel.2019.02.017.
7. Aebersold R, Mann M. 2016. Mass-spectrometric exploration of proteome structure and function. Nature 537:347–355. https://doi.org/10.1038/nature19949.
8. Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR. 2013. Protein analysis by shotgun/bottom-up proteomics. Chem Rev 113:2343–2394. https://doi.org/10.1021/cr3003533.
9. Smith LM, Kelleher NL, The Consortium for Top Down Proteomics. 2013. Proteoform: a single term describing protein complexity. Nat Methods 10:186–187. https://doi.org/10.1038/nmeth.2369.
10. Wolters DA, Washburn MP, Yates JR. 2001. An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem 73:5683–5690. https://doi.org/10.1021/ac010617.e.
11. Omasits U, Quebatte M, Stekhoven DJ, Fortes C, Roschitzki B, Robinson MD, Dehio C, Ahrens CH. 2013. Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome. Genome Res 23:1916–1927. https://doi.org/10.1101/gr.151035.112.
12. Bartel J, Varadarajan AR, Sura T, Ahrens CH, Maaß S, Becher D. 2020. Optimized proteomics workflow for the detection of small proteins. J Proteome Res 19:4004–4018. https://doi.org/10.1021/acs.jproteome.0c00286.
13. Tyanova S, Temu T, Cox J. 2016. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc 11:2301–2319. https://doi.org/10.1038/nprot.2016.136.
14. Dinger ME, Pang KC, Mercer TR, Mattick JS. 2008. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol 4:e1000176. https://doi.org/10.1371/journal.pcbi.1000176.
15. Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A, 2004. The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. Mol Cell Proteomics 3:531–533. https://doi.org/10.1074/mcp.T400006-MCP200.
16. Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, Sun Z, Nilsson E, Pratt B, Prazen B, Eng JK, Martin DB, Nesvizhskii AI, Aebersold R. 2010. A guided tour of the Trans-Proteomic Pipeline. Proteomics 10:1150–1159. https://doi.org/10.1002/pmic.200900375.
17. Ludwig C, Claassen M, Schmidt A, Aebersold R. 2012. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry. Mol Cell Proteomics 11:M111.013987. https://doi.org/10.1074/mcp.M111.013987.

18. Nesvizhskii AI, Aebersold R. 2005. Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics 4:1419–1440. https://doi.org/10.1074/mcp.R500012-MCP200.
19. Ferguson JT, Wenger CD, Metcalf WW, Kelleher NL. 2009. Top-down proteomics reveals novel protein forms expressed in *Methanosarcina acetivorans*. J Am Soc Mass Spectrom 20:1743–1750. https://doi.org/10.1016/j.jasms.2009.05.014.
20. Donnelly DP, Rawlins CM, DeHart CJ, Fornelli L, Schachner LF, Lin Z, Lippens JL, Aluri KC, Sarin R, Chen B, Lantz C, Jung W, Johnson KR, Koller A, Wolff JJ, Campuzano IDG, Auclair JR, Ivanov AR, Whitelegge JP, Pasa-Tolic L, Chamot-Rooke J, Danis PO, Smith LM, Tsybin YO, Loo JA, Ge Y, Kelleher NL, Agar JN. 2019. Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. Nat Methods 16:587–594. https://doi.org/10.1038/s41592-019-0457-0.
21. Wisniewski JR. 2017. Filter-aided sample preparation: the versatile and efficient method for proteomic analysis. Methods Enzymol 585:15–27. https://doi.org/10.1016/bs.mie.2016.09.013.
22. Zougman A, Selby PJ, Banks RE. 2014. Suspension trapping (STrap) sample preparation method for bottom-up proteomics analysis. Proteomics 14:1006–1000. https://doi.org/10.1002/pmic.201300553.
23. Kohlstaedt M, Buschmann S, Xie H, Resemann A, Warkentin E, Langer JD, Michel H. 2016. Identification and characterization of the novel subunit CcoM in the cbb3(3)Cytochrome c oxidase from *Pseudomonas stutzeri* ZoBell. mBio 7:e01921-15. https://doi.org/10.1128/mBio.01921-15.
24. Kaulich PT, Cassidy L, Weidenbach K, Schmitz RA, Tholey A. 2020. Complementarity of different SDS-PAGE gel staining methods for the identification of short open reading frame-encoded peptides. Proteomics 20:e2000084. https://doi.org/10.1002/pmic.202000084.
25. Petruschke H, Anders J, Stadler PF, Jehmlich N, von Bergen M. 2020. Enrichment and identification of small proteins in a simplified human gut microbiome. J Proteomics 213:103604. https://doi.org/10.1016/j.jprot.2019.103604.
26. Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, Thomas GR, Vandekerckhove J. 2003. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. Nat Biotechnol 21:566–569. https://doi.org/10.1038/nbt810.
27. Bland C, Hartmann EM, Christie-Oleza JA, Fernandez B, Armengaud J. 2014. N-Terminal-oriented proteogenomics of the marine bacterium *Roseobacter denitrificans* Och114 using N-Succinimidyloxycarbonylmethyl)tris(2,4,6-trimethoxyphenyl)phosphonium bromide (TMPP) labeling and diagonal chromatography. Mol Cell Proteomics 13:1369–1381. https://doi.org/10.1074/mcp.O113.032854.
28. Bibi-Triki S, Husson G, Maucourt B, Vuilleumier S, Carapito C, Bringel F. 2018. N-terminome and proteogenomic analysis of the *Methylobacterium extorquens* DM4 reference strain for dichloromethane utilization. J Proteomics 179:131–139. https://doi.org/10.1016/j.jprot.2018.03.012.
29. Impens F, Rolhion N, Radoshevich L, Becavin C, Duval M, Mellin J, Garcia Del Portillo F, Pucciarelli MG, Williams AH, Cossart P. 2017. N-terminomics identifies Prli42 as a membrane miniprotein conserved in Firmicutes and critical for stressosome activation in *Listeria monocytogenes*. Nat Microbiol 2:17005. https://doi.org/10.1038/nmicrobiol.2017.5.
30. Venter E, Smith RD, Payne SH. 2011. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. PLoS One 6:e27587. https://doi.org/10.1371/journal.pone.0027587.
31. Krug K, Carpy A, Behrends G, Matic K, Soares NC, Macek B. 2013. Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments. Mol Cell Proteomics 12:3420–3430. https://doi.org/10.1074/mcp.M113.029165.
32. Marini F, Carregari VC, Greco V, Ronci M, Iavarone F, Persichilli S, Castagnola M, Urbani A, Pieroni L. 2020. Exploring the HeLa dark mitochondrial proteome. Front Cell Dev Biol 8:137. https://doi.org/10.3389/fcell.2020.00137.
33. Ruprecht B, Roesli C, Lemeer S, Kuster B. 2016. MALDI-TOF and nESI Orbitrap MS/MS identify orthogonal parts of the phosphoproteome. Proteomics 16:1447–1456. https://doi.org/10.1002/pmic.201500523.
34. Gonczarowska-Jorge H, Loroch S, Dell'Aica M, Sickmann A, Roos A, Zahedi RP. 2017. Quantifying missing (phospho)proteome regions with the broad-specificity protease subtilisin. Anal Chem 89:13137–13145. https://doi.org/10.1021/acs.analchem.7b02395.
35. Kaulich PT, Cassidy L, Bartel J, Schmitz RA, Tholey A. 2021. Multi-protease approach for the improved identification and molecular characterization of small proteins and short open reading frame-encoded peptides. J Proteome Res 20:2895–2903. https://doi.org/10.1021/acs.jproteome.1c00115.
36. Bausewein T, Mills DJ, Langer JD, Nitschke B, Nussberger S, Kühlbrandt W. 2017. Cryo-EM structure of the TOM core complex from *Neurospora crassa*. Cell 170:693–700.e7. https://doi.org/10.1016/j.cell.2017.07.012.
37. Kulak NA, Geyer PE, Mann M. 2017. Loss-less nano-fractionator for high sensitivity, high coverage proteomics. Mol Cell Proteomics 16:694–705. https://doi.org/10.1074/mcp.O116.065136.
38. Alpert AJ. 2016. Protein fractionation and enrichment prior to proteomics sample preparation. Adv Exp Med Biol 919:23–41. https://doi.org/10.1007/978-3-319-41448-5_2.
39. Wang H, Yang Y, Li Y, Bai B, Wang X, Tan H, Liu T, Beach TG, Peng J, Wu Z. 2015. Systematic optimization of long gradient chromatography mass spectrometry for deep analysis of brain proteome. J Proteome Res 14:829–838. https://doi.org/10.1021/pr500882h.
40. Cong Y, Liang Y, Motamedchaboki K, Huguet R, Truong T, Zhao R, Shen Y, Lopez-Ferrer D, Zhu Y, Kelly RT. 2020. Improved single-cell proteome coverage using narrow-bore packed NanoLC columns and ultrasensitive mass spectrometry. Anal Chem 92:2665–2671. https://doi.org/10.1021/acs.analchem.9b04631.
41. Toth G, Panic-Jankovic T, Mitulovic G. 2019. Pillar array columns for peptide separations in nanoscale reversed-phase chromatography. J Chromatogr A 1603:426–432. https://doi.org/10.1016/j.chroma.2019.06.067.
42. Safarian S, Rajendran C, Muller H, Preu J, Langer JD, Ovchinnikov S, Hirose T, Kusumoto T, Sakamoto J, Michel H. 2016. Structure of a bd oxidase indicates similar mechanisms for membrane-integrated oxygen reductases. Science 352:583–586. https://doi.org/10.1126/science.aaf2477.
43. Safarian S, Hahn A, Mills DJ, Radloff M, Eisinger ML, Nikolaev A, Meier-Credo J, Melin F, Miyoshi H, Gennis RB, Sakamoto J, Langer JD, Hellwig P, Kuhlbrandt W, Michel H. 2019. Active site rearrangement and structural divergence in prokaryotic respiratory oxidases. Science 366:100–104. https://doi.org/10.1126/science.aay0967.
44. Zabret J, Bohn S, Schuller SK, Arnolds O, Moller M, Meier-Credo J, Liauw P, Chan A, Tajkhorshid E, Langer JD, Stoll R, Krieger-Liszkay A, Engel BD, Rudack T, Schuller JM, Nowaczyk MM. 2021. Structural insights into photosystem II assembly. Nat Plants 7:524–538. https://doi.org/10.1038/s41477-021-00895-0.
45. Meier F, Beck S, Grassl N, Lubeck M, Park MA, Raether O, Mann M. 2015. Parallel accumulation-serial fragmentation (PASEF): multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. J Proteome Res 14:5378–5387. https://doi.org/10.1021/acs.jproteome.5b00932.
46. Meier F, Brunner AD, Koch S, Koch H, Lubeck M, Krause M, Goedecke N, Decker J, Kosinski T, Park MA, Bache N, Hoerning O, Cox J, Rather O, Mann M. 2018. Online parallel accumulation-serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. Mol Cell Proteomics 17:2534–2545. https://doi.org/10.1074/mcp.TIR118.000900.
47. Zhang F, Ge W, Ruan G, Cai X, Guo T. 2020. Data-independent acquisition mass spectrometry-based proteomics and software tools: a glimpse in 2020. Proteomics 20:e1900276. https://doi.org/10.1002/pmic.201900276.
48. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ. 2012. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol 10:599–606. https://doi.org/10.1038/nrmicro2850.
49. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. 2015. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15:141–161. https://doi.org/10.1007/s10142-015-0433-4.
50. Fraser CM, Eisen JA, Nelson KE, Paulsen IT, Salzberg SL. 2002. The value of complete microbial genome sequencing (you get what you pay for). J Bacteriol 184:6403–6405. https://doi.org/10.1128/JB.184.23.6403-6405.2002.
51. Sinitcyn P, Rudolph JD, Cox J. 2018. Computational methods for understanding mass spectrometry-based shotgun proteomics data. Annu Rev Biomed Data Sci 1:207–234. https://doi.org/10.1146/annurev-biodatasci-080917-013516.
52. Salzberg SL. 2019. Next-generation genome annotation: we still struggle to get it right. Genome Biol 20:92. https://doi.org/10.1186/s13059-019-1715-2.
53. Poole FL, Gerwe BA, Hopkins RC, Schut GJ, Weinberg MV, Jenney FE, Adams MW. 2005. Defining genes in the genome of the hyperthermophilic archaeon *Pyrococcus furiosus*: implications for all microbial genomes. J Bacteriol 187:7325–7332. https://doi.org/10.1128/JB.187.21.7325-7332.2005.
54. Bakke P, Carney N, Deloache W, Gearing M, Ingvorsen K, Lotz M, McNair J, Penumetcha P, Simpson S, Voss L, Win M, Heyer LJ, Campbell AM.

2009. Evaluation of three automated genome annotations for *Halorhabdus utahensis*. PLoS One 4:e6291. https://doi.org/10.1371/journal.pone.0006291.
55. Storz G, Wolf YI, Ramamurthi KS. 2014. Small proteins can no longer be ignored. Annu Rev Biochem 83:753–777. https://doi.org/10.1146/annurev-biochem-070611-102400.
56. Jaffe JD, Berg HC, Church GM. 2004. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 4:59–77. https://doi.org/10.1002/pmic.200300511.
57. Payne SH, Huang S-T, Pieper R. 2010. A proteogenomic update to Yersinia: enhancing genome annotation. BMC Genomics 11:460. https://doi.org/10.1186/1471-2164-11-460.
58. Marcellin E, Licona-Cassani C, Mercer TR, Palfreyman RW, Nielsen LK. 2013. Re-annotation of the *Saccharopolyspora erythraea* genome using a systems biology approach. BMC Genomics 14:699. https://doi.org/10.1186/1471-2164-14-699.
59. Mukherjee S, Seshadri R, Varghese NJ, Eloe-Fadrosh EA, Meier-Kolthoff JP, Goker M, Coates RC, Hadjithomas M, Pavlopoulos GA, Paez-Espino D, Yoshikuni Y, Visel A, Whitman WB, Garrity GM, Eisen JA, Hugenholtz P, Pati A, Ivanova NN, Woyke T, Klenk HP, Kyrpides NC. 2017. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. Nat Biotechnol 35:676–683. https://doi.org/10.1038/nbt.3886.
60. Petruschke H, Schori C, Canzler S, Riesbeck S, Poehlein A, Daniel R, Frei D, Segessemann T, Zimmerman J, Marinos G, Kaleta C, Jehmlich N, Ahrens CH, von Bergen M. 2021. Discovery of novel community-relevant small proteins in a simplified human intestinal microbiome. Microbiome 9:55. https://doi.org/10.1186/s40168-020-00981-z.
61. Butterfield CN, Li Z, Andeer PF, Spaulding S, Thomas BC, Singh A, Hettich RL, Suttle KB, Probst AJ, Tringe SG, Northen T, Pan C, Banfield JF. 2016. Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. PeerJ 4:e2687. https://doi.org/10.7717/peerj.2687.
62. Moss EL, Maghini DG, Bhatt AS. 2020. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat Biotechnol 38:701–707. https://doi.org/10.1038/s41587-020-0422-6.
63. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2:1533–1542. https://doi.org/10.1038/s41564-017-0012-7.
64. Nesvizhskii AI. 2014. Proteogenomics: concepts, applications and computational strategies. Nat Methods 11:1114–1125. https://doi.org/10.1038/nmeth.3144.
65. Menschaert G, Fenyo D. 2017. Proteogenomics from a bioinformatics angle: a growing field. Mass Spectrom Rev 36:584–599. https://doi.org/10.1002/mas.21483.
66. Ruggles KV, Krug K, Wang X, Clauser KR, Wang J, Payne SH, Fenyo D, Zhang B, Mani DR. 2017. Methods, tools and current perspectives in proteogenomics. Mol Cell Proteomics 16:959–981. https://doi.org/10.1074/mcp.MR117.000024.
67. Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4:207–214. https://doi.org/10.1038/nmeth1019.
68. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. 2007. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 4:923–925. https://doi.org/10.1038/nmeth1113.
69. Kim S, Gupta N, Pevzner PA. 2008. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. J Proteome Res 7:3354–3363. https://doi.org/10.1021/pr8001244.
70. Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, Mendoza L, Moritz RL, Aebersold R, Nesvizhskii AI. 2011. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics 10:M111.007690. https://doi.org/10.1074/mcp.M111.007690.
71. Nesvizhskii AI. 2010. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics 73:2092–2123. https://doi.org/10.1016/j.jprot.2010.08.009.
72. Rost HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, Andreotti S, Ehrlich HC, Gutenbrunner P, Kenar E, Liang X, Nahnsen S, Nilse L, Pfeuffer J, Rosenberger G, Rurik M, Schmitt U, Veit J, Walzer M, Wojnar D, Wolski WE, Schilling O, Choudhary JS, Malmstrom L, Aebersold R, Reinert K, Kohlbacher O. 2016. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods 13:741–748. https://doi.org/10.1038/nmeth.3959.
73. da Veiga Leprevost F, Haynes SE, Avtonomov DM, Chang H-Y, Shanmugam AK, Mellacheruvu D, Kong AT, Nesvizhskii AI. 2020. Philosopher: a versatile toolkit for shotgun proteomics data analysis. Nat Methods 17:869–870. https://doi.org/10.1038/s41592-020-0912-y.
74. Li H, Joh YS, Kim H, Paek E, Lee SW, Hwang KB. 2016. Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification. BMC Genomics 17:1031. https://doi.org/10.1186/s12864-016-3327-5.
75. Blakeley P, Overton IM, Hubbard SJ. 2012. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. J Proteome Res 11:5221–5234. https://doi.org/10.1021/pr300411q.
76. Wang X, Slebos RJ, Wang D, Halvey PJ, Tabb DL, Liebler DC, Zhang B. 2012. Protein identification using customized protein sequence databases derived from RNA-Seq data. J Proteome Res 11:1009–1017. https://doi.org/10.1021/pr200766z.
77. Woo S, Cha SW, Merrihew G, He Y, Castellana N, Guest C, MacCoss M, Bafna V. 2014. Proteogenomic database construction driven from large scale RNA-seq data. J Proteome Res 13:21–28. https://doi.org/10.1021/pr400294c.
78. Komor MA, Pham TV, Hiemstra AC, Piersma SR, Bolijn AS, Schelfhorst T, Delis-van Diemen PM, Tijssen M, Sebra RP, Ashby M, Meijer GA, Jimenez CR, Fijneman RJA. 2017. Identification of differentially expressed splice variants by the proteogenomic pipeline Splicify. Mol Cell Proteomics 16:1850–1863. https://doi.org/10.1074/mcp.TIR117.000056.
79. Ma J, Saghatelian A, Shokhirev MN. 2018. The influence of transcript assembly on the proteogenomics discovery of microproteins. PLoS One 13:e0194518. https://doi.org/10.1371/journal.pone.0194518.
80. Zickmann F, Renard BY. 2015. MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. Bioinformatics 31:i106–15. https://doi.org/10.1093/bioinformatics/btv236.
81. Omasits U, Varadarajan AR, Schmid M, Goetze S, Melidis D, Bourqui M, Nikolayeva O, Québatte M, Patrignani A, Dehio C, Frey JE, Robinson MD, Wollscheid B, Ahrens CH. 2017. An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. Genome Res 27:2083–2095. https://doi.org/10.1101/gr.218255.116.
82. Hecht A, Glasgow J, Jaschke PR, Bawazer LA, Munson MS, Cochran JR, Endy D, Salit M. 2017. Measurements of translation initiation from all 64 codons in E. coli. Nucleic Acids Res 45:3615–3626. https://doi.org/10.1093/nar/gkx070.
83. Uszkoreit J, Plohnke N, Rexroth S, Marcus K, Eisenacher M. 2014. The bacterial proteogenomic pipeline. BMC Genomics 15(Suppl 9):S19. https://doi.org/10.1186/1471-2164-15-S9-S19.
84. Kumar D, Yadav AK, Kadimi PK, Nagaraj SH, Grimmond SM, Dash D. 2013. Proteogenomic analysis of *Bradyrhizobium japonicum* USDA110 using GenoSuite, an automated multi-algorithmic pipeline. Mol Cell Proteomics 12:3388–3397. https://doi.org/10.1074/mcp.M112.027169.
85. Tovchigrechko A, Venepally P, Payne SH. 2014. PGP: parallel prokaryotic proteogenomics pipeline for MPI clusters, high-throughput batch clusters and multicore workstations. Bioinformatics 30:1469–1470. https://doi.org/10.1093/bioinformatics/btu051.
86. Nagaraj SH, Waddell N, Madugundu AK, Wood S, Jones A, Mandyam RA, Nones K, Pearson JV, Grimmond SM. 2015. PGTools: A software suite for proteogenomic data analysis and visualization. J Proteome Res 14:2255–2266. https://doi.org/10.1021/acs.jproteome.5b00029.
87. Fuchs S, Kucklick M, Lehmann E, Beckmann A, Wilkens M, Kolte B, Mustafayeva A, Ludwig T, Diwo M, Wissing J, Jänsch L, Ahrens CH, Ignatova Z, Engelmann S. 2021. Towards the characterization of the hidden world of small proteins in *Staphylococcus aureus*, a proteogenomics approach. PLoS Genet 17:e1009585. https://doi.org/10.1371/journal.pgen.1009585.
88. Willems P, Fijalkowski I, Van Damme P. 2020. Lost and found: re-searching and re-scoring proteomics data aids genome annotation and improves proteome coverage. mSystems 5:e00833-20. https://doi.org/10.1128/mSystems.00833-20.
89. Cassidy L, Helbig AO, Kaulich PT, Weidenbach K, Schmitz RA, Tholey A. 2021. Multidimensional separation schemes enhance the identification and molecular characterization of low molecular weight proteomes and short open reading frame-encoded peptides in top-down proteomics. J Proteomics 230:103988. https://doi.org/10.1016/j.jprot.2020.103988.
90. Wang B, Wang Z, Pan N, Huang J, Wan C. 2021. Improved identification of small open reading frames encoded peptides by top-down proteomic approaches and de novo sequencing. Int J Mol Sci 22:5476. https://doi.org/10.3390/ijms22115476.

91. Gupta N, Pevzner PA. 2009. False discovery rates of protein identifications: a strike against the two-peptide rule. J Proteome Res 8:4173–4181. https://doi.org/10.1021/pr9004794.
92. Varadarajan AR, Goetze S, Pavlou MP, Grosboillot V, Shen Y, Loessner MJ, Ahrens CH, Wollscheid B. 2020. A proteogenomic resource enabling integrated analysis of listeria genotype-proteotype-phenotype relationships. J Proteome Res 19:1647–1662. https://doi.org/10.1021/acs.jproteome.9b00842.
93. Smith C, Canestrari JG, Wang J, Derbyshire KM, Gray TA, Wade JT. 2019. Pervasive translation in *Mycobacterium tuberculosis*. bioRxiv 665208.
94. Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, Hengartner MO, Aebersold R. 2009. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol Cell Proteomics 8:2405–2417. https://doi.org/10.1074/mcp.M900317-MCP200.
95. Savitski MM, Wilhelm M, Hahne H, Kuster B, Bantscheff M. 2015. A scalable approach for protein false discovery rate estimation in large proteomic data sets. Mol Cell Proteomics 14:2394–2404. https://doi.org/10.1074/mcp.M114.046995.
96. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.
97. Peterson AC, Russell JD, Bailey DJ, Westphall MS, Coon JJ. 2012. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. Mol Cell Proteomics 11:1475–1488. https://doi.org/10.1074/mcp.O112.020131.
98. Kim S, Pevzner PA. 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 5:5277. https://doi.org/10.1038/ncomms6277.
99. The M, MacCoss MJ, Noble WS, Kall L. 2016. Fast and accurate protein false discovery rates on large-scale proteomics data sets with Percolator 3.0. J Am Soc Mass Spectrom 27:1719–1727. https://doi.org/10.1007/s13361-016-1460-7.
100. Yu F, Teo GC, Kong AT, Haynes SE, Avtonomov DM, Geiszler DJ, Nesvizhskii AI. 2020. Identification of modified peptides using localization-aware open search. Nat Commun 11:4065. https://doi.org/10.1038/s41467-020-17921-y.
101. Gessulat S, Schmidt T, Zolg DP, Samaras P, Schnatbaum K, Zerweck J, Knaute T, Rechenberger J, Delanghe B, Huhmer A, Reimer U, Ehrlich HC, Aiche S, Kuster B, Wilhelm M. 2019. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods 16:509–518. https://doi.org/10.1038/s41592-019-0426-7.
102. Verbruggen S, Gessulat S, Gabriels R, Matsaroki A, Van de Voorde H, Kuster B, Degroeve S, Martens L, Van Criekinge W, Wilhelm M, Menschaert G. 2021. Spectral prediction features as a solution for the search space size problem in proteogenomics. Mol Cell Proteomics 20:100076. https://doi.org/10.1016/j.mcpro.2021.100076.
103. Chen WH, van Noort V, Lluch-Senar M, Hennrich ML, Wodke JA, Yus E, Alibes A, Roma G, Mende DR, Pesavento C, Typas A, Gavin AC, Serrano L, Bork P. 2016. Integration of multi-omics data of a genome-reduced bacterium: Prevalence of post-transcriptional regulation and its correlation with protein abundances. Nucleic Acids Res 44:1192–1202. https://doi.org/10.1093/nar/gkw004.
104. Venturini E, Svensson SL, Maaß S, Gelhausen R, Eggenhofer F, Li L, Cain AK, Parkhill J, Becher D, Backofen R, Barquist L, Sharma CM, Westermann AJ, Vogel J. 2020. A global data-driven census of *Salmonella* small proteins and their potential functions in bacterial virulence. microLife 1:uqaa002. https://doi.org/10.1093/femsml/uqaa002.
105. Maier T, Schmidt A, Guell M, Kuhner S, Gavin AC, Aebersold R, Serrano L. 2011. Quantification of mRNA and protein and integration with protein turnover in a bacterium. Mol Syst Biol 7:511. https://doi.org/10.1038/msb.2011.38.
106. Weaver J, Mohammad F, Buskirk AR, Storz G. 2019. Identifying small proteins by ribosome profiling with stalled initiation complexes. mBio 10. https://doi.org/10.1128/mBio.02819-18.
107. Li L, Chao Y. 2020. sPepFinder expedites genome-wide identification of small proteins in bacteria. bioRxiv 05.05.079178.
108. Miller RM, Ibrahim K, Smith LM. 2021. ProteaseGuru: a tool for protease selection in bottom-up proteomics. J Proteome Res 20:1936–1942. https://doi.org/10.1021/acs.jproteome.0c00954.
109. Solntsev SK, Shortreed MR, Frey BL, Smith LM. 2018. Enhanced global post-translational modification discovery with MetaMorpheus. J Proteome Res 17:1844–1851. https://doi.org/10.1021/acs.jproteome.7b00873.
110. Picotti P, Aebersold R. 2012. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Methods 9:555–566. https://doi.org/10.1038/nmeth.2015.
111. Erez Z, Steinberger-Levy I, Shamir M, Doron S, Stokar-Avihail A, Peleg Y, Melamed S, Leavitt A, Savidor A, Albeck S, Amitai G, Sorek R. 2017. Communication between viruses guides lysis-lysogeny decisions. Nature 541:488–493. https://doi.org/10.1038/nature21049.
112. Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. 2003. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Natl Acad Sci U S A 100:6940–6945. https://doi.org/10.1073/pnas.0832254100.
113. Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas JL, Messeguer X, Albà MM. 2018. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. Nat Ecol Evol 2:890–896. https://doi.org/10.1038/s41559-018-0506-6.
114. Keeling DM, Garza P, Nartey CM, Carvunis A-R. 2019. The meanings of "function" in biology and the problematic case of de novo gene emergence. Elife 8. https://doi.org/10.7554/eLife.47014.
115. Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47:D309–D314. https://doi.org/10.1093/nar/gky1085.
116. Mulder N, Apweiler R. 2007. InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol Biol 396:59–70. https://doi.org/10.1007/978-1-59745-515-2_5.
117. Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A. 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci 12:1652–1662. https://doi.org/10.1110/ps.0303703.
118. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ, von Mering C. 2021. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 49:D605–D612. https://doi.org/10.1093/nar/gkaa1074.
119. Dever TE, Ivanov IP, Sachs MS. 2020. Conserved upstream open reading frame nascent peptides that control translation. Annu Rev Genet 54:237–264. https://doi.org/10.1146/annurev-genet-112618-043822.
120. Puigbò P, Aragonès L, Garcia-Vallvé S. 2010. RCDI/eRCDI: a web-server to estimate codon usage deoptimization. BMC Res Notes 3:87. https://doi.org/10.1186/1756-0500-3-87.
121. Smits AH, Vermeulen M. 2016. Characterizing protein-protein interactions using mass spectrometry: challenges and opportunities. Trends Biotechnol 34:825–834. https://doi.org/10.1016/j.tibtech.2016.02.014.
122. Richards AL, Eckhardt M, Krogan NJ. 2021. Mass spectrometry-based protein-protein interaction networks for the study of human diseases. Mol Syst Biol 17:e8792. https://doi.org/10.15252/msb.20188792.
123. Koh M, Ahmad I, Ko Y, Zhang Y, Martinez TF, Diedrich JK, Chu Q, Moresco JJ, Erb MA, Saghatelian A, Schultz PG, Bollong MJ. 2021. A short ORF-encoded transcriptional regulator. Proc Natl Acad Sci U S A 118:e2021943118. https://doi.org/10.1073/pnas.2021943118.
124. Gray T, Storz G, Papenfort K. 2021. Small proteins; big questions. J Bacteriol https://doi.org/10.1128/JB.00341-21.
125. Kucharova V, Wiker HG. 2014. Proteogenomics in microbiology: taking the right turn at the junction of genomics and proteomics. Proteomics 14:2360–2675. https://doi.org/10.1002/pmic.201400168.
126. Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, Butler J, Calvo S, Elkins T, FitzGerald MG, Hafez N, Kodira CD, Major J, Wang S, Wilkinson J, Nicol R, Nusbaum C, Birren B, Berg HC, Church GM. 2004. The complete genome and proteome of *Mycoplasma mobile*. Genome Res 14:1447–1461. https://doi.org/10.1101/gr.2674004.
127. Gupta N, Tanner S, Jaitly N, Adkins JN, Lipton M, Edwards R, Romine M, Osterman A, Bafna V, Smith RD, Pevzner PA. 2007. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. Genome Res 17:1362–1377. https://doi.org/10.1101/gr.6427907.
128. Becher D, Hempel K, Sievers S, Zühlke D, Pané-Farré J, Otto A, Fuchs S, Albrecht D, Bernhardt J, Engelmann S, Völker U, van Dijl JM, Hecker M. 2009. A proteomic view of an important human pathogen—towards the quantification of the entire *Staphylococcus aureus* proteome. PLoS One 4:e8176. https://doi.org/10.1371/journal.pone.0008176.
129. de Souza GA, Arntzen MO, Fortuin S, Schurch AC, Malen H, McEvoy CR, van Soolingen D, Thiede B, Warren RM, Wiker HG. 2011. Proteogenomic analysis of polymorphisms and gene annotation divergences in prokaryotes using a clustered mass spectrometry-friendly database. Mol Cell Proteomics 10:M110,002527. https://doi.org/10.1074/mcp.M110.002527.
130. Müller SA, Findeiß S, Pernitzsch SR, Wissenbach DK, Stadler PF, Hofacker IL, von Bergen M, Kalkhof S. 2013. Identification of new protein coding sequences and signal peptidase cleavage sites of *Helicobacter pylori*

strain 26695 by proteogenomics. J Proteomics 86:27–42. https://doi.org/10.1016/j.jprot.2013.04.036.
131. Bonissone S, Gupta N, Romine M, Bradshaw RA, Pevzner PA. 2013. N-terminal protein processing: a comparative proteogenomic analysis. Mol Cell Proteomics 12:14–28. https://doi.org/10.1074/mcp.M112.019075.
132. Čuklina J, Hahn J, Imakaev M, Omasits U, Förstner KU, Ljubimov N, Goebel M, Pessi G, Fischer H-M, Ahrens CH, Gelfand MS, Evguenieva-Hackenberg E. 2016. Genome-wide transcription start site mapping of *Bradyrhizobium japonicum* grown free-living or in symbiosis—a rich resource to identify new transcripts, proteins and to study gene regulation. BMC Genomics 17:302. https://doi.org/10.1186/s12864-016-2602-9.
133. Yang M, Yang Y, Chen Z, Zhang J, Lin Y, Wang Y, Xiong Q, Li T, Ge F, Bryant DA, Zhao J. 2014. Proteogenomic analysis and global discovery of posttranslational modifications in prokaryotes. Proc Natl Acad Sci U S A 111:E5633–5642. https://doi.org/10.1073/pnas.1412722111.
134. Abendroth U, Adlung N, Otto A, Gruneisen B, Becher D, Bonas U. 2017. Identification of new protein-coding genes with a potential role in the virulence of the plant pathogen *Xanthomonas euvesicatoria*. BMC Genomics 18:625. https://doi.org/10.1186/s12864-017-4041-7.
135. Bupp CR, Wirth MJ. 2020. Making sharper peaks for reverse-phase liquid chromatography of proteins. Annu Rev Anal Chem (Palo Alto Calif) 13:363–380. https://doi.org/10.1146/annurev-anchem-061318-115009.
136. Gillet LC, Leitner A, Aebersold R. 2016. Mass spectrometry applied to bottom-up proteomics: entering the high-throughput era for hypothesis testing. Annu Rev Anal Chem (Palo Alto Calif) 9:449–472. https://doi.org/10.1146/annurev-anchem-071015-041535.
137. Schlesinger D, Elsässer SJ. 2021. Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. FEBS J https://doi.org/10.1111/febs.15769.
138. Zheng J, Strutzenberg T, Pascal BD, Griffin PR. 2019. Protein dynamics and conformational changes explored by hydrogen/deuterium exchange mass spectrometry. Curr Opin Struct Biol 58:305–313. https://doi.org/10.1016/j.sbi.2019.06.007.
139. Iacobucci C, Götze M, Sinz A. 2020. Cross-linking/mass spectrometry to get a closer view on protein interaction networks. Curr Opin Biotechnol 63:48–53. https://doi.org/10.1016/j.copbio.2019.12.009.
140. Liko I, Allison TM, Hopper JT, Robinson CV. 2016. Mass spectrometry guided structural biology. Curr Opin Struct Biol 40:136–144. https://doi.org/10.1016/j.sbi.2016.09.008.
141. Stekhoven DJ, Omasits U, Quebatte M, Dehio C, Ahrens CH. 2014. Proteome-wide identification of predominant subcellular protein localizations in a bacterial model organism. J Proteomics 99:123–137. https://doi.org/10.1016/j.jprot.2014.01.015.
142. Geladaki A, Kocevar Britovsek N, Breckels LM, Smith TS, Vennard OL, Mulvey CM, Crook OM, Gatto L, Lilley KS. 2019. Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. Nat Commun 10:331. https://doi.org/10.1038/s41467-018-08191-w.