Bottom-up and top-down proteomic approaches for the identification, characterization, and quantification of the low molecular weight proteome with focus on short open reading frame-encoded peptides

Liam Cassidy1 | Philipp T. Kaulich1 | Sandra Maaß2 | Jürgen Bartel2 | Dörte Becher2 | Andreas Tholey1

1 Systematic Proteome Research & Bioanalytics, Institute for Experimental Medicine, Christian-Albrechts-Universität zu Kiel, Kiel, Germany
2 Department of Microbial Proteomics, Institute of Microbiology, University of Greifswald, Greifswald, Germany

Correspondence
Andreas Tholey, Systematic Proteome Research & Bioanalytics, Institute for Experimental Medicine, Christian-Albrechts-Universität zu Kiel, 24105 Kiel, Germany.
Email: a.tholey@iem.uni-kiel.de

Abstract
The recent discovery of alternative open reading frames creates a need for suitable analytical approaches to verify their translation and to characterize the corresponding gene products at the molecular level. As the analysis of small proteins within a background proteome by means of classical bottom-up proteomics is challenging, method development for the analysis of small open reading frame encoded peptides (SEPs) have become a focal point for research. Here, we highlight bottom-up and top-down proteomics approaches established for the analysis of SEPs in both pro- and eukaryotes. Major steps of analysis, including sample preparation and (small) proteome isolation, separation and mass spectrometry, data interpretation and quality control, quantification, the analysis of post-translational modifications, and exploration of functional aspects of the SEPs by means of proteomics technologies are described. These methods do not exclusively cover the analytics of SEPs but simultaneously include the low molecular weight proteome, and moreover, can also be used for the proteome-wide analysis of proteolytic processing events.

KEYWORDS
mass spectrometry, LC-MS/MS, peptidomics, post-translational modification analysis, prefractination, sample preparation, SEP, top-down proteomics

1 | INTRODUCTION

The manner in which coding sequences (CDSs) within genomes are identified, has typically been based on constraints given by statistically derived cut offs determining the minimum length of CDS to be assigned by algorithms [1]. A second assumption was the need for clearly defined start codons for a transcription and translation of the particular gene. This traditional approach provided a statistically sound way for analyzing genomes for new genes encoding proteins, building a compromise between false positives and true positives, and finally lead to a lower limit of gene lengths at approximately 100 codons [2]. With the advent of improved sequencing technologies, [3] ribosomal profiling [4] and of novel algorithms [2] in recent years, the assignment of potentially protein encoding genes below this threshold, the so called short short open
reading frames (sORFs) or small open reading frames (smORFs), was accelerated tremendously both in pro- and eukaryotes.

Further, shifts in the paradigm of protein translation occurred by the discovery of alternative coding start and stop sites [5] which introduced additional, more specific terms within the group of short or short ORFs. For example, CDSs reaching into the so-called 5′ or 3′ untranslated regions or even completely contained in these regions were discovered (Figure 1) challenging the classical view of gene translation [6]. These non-classical ORFs have been termed alternative ORFs (altORFs), in order to distinguish them from classical ORFs (reference ORFs). These altORFs can be further divided into different subclasses, [7] such as upstream and downstream ORFs (uORFs and dORFs), small or short ORFs (smORFs or sORFs), ORFs from nontemplate encoded peptides (SEPs). To our knowledge, no clear rules do not meet this criterion.

With the knowledge of this expansion of the potential coding repertoire of genomes the question about its transcription, and in particular, its translation into peptides or proteins arose. Terms such as “hidden proteome,” the “dark proteome,” or “ghost proteome” have come up in recent years to summarize this part of the peptide and protein repertoire of organisms [7,9,10]. This review will further focus on the sORF based gene products unless otherwise stated.

A compounding issue in emerging new fields is the development of a consensus on nomenclature [11]. As the underlying genes are denoted either as small or short ORFs, with both terms being used interchangeably, their gene products are hence named either small open reading frame encoded peptides (SEPs). To our knowledge, no clear rules are defining differences between short or small ORFs. As typically the length of biomolecules, thus the number (and sequence) of monomers building them determines their function, the term “short” seems more logic to us, and is further on used in this review.

Other terms used in the literature for describing SEPs are mini- or microproteins [11,12]. Microproteins are generally defined as small proteins, encoded by sORFs that contain a single protein domain and regulate the activity of multi-protein complexes via protein-protein interactions [13]. As there is a longstanding debate about the boundary between peptides and proteins, the terms mini- and microproteins are used but are a potential matter of discussion. In addition, the terms “ghost proteins” and “novel proteins” are used for altORF gene products, of which SEPs form a subclass [8,14]. Therefore, we recommend to use all these different search terms when a literature search is performed.

A further issue under discussion is the definition of the term short/small in terms of the amino acid length defining SEPs. This definition is not consistent, and for example, 50, 70, or 100 amino acids are considered as upper limit for length, but longer SEPs were also reported [15–17].

Another major problem to be tackled with the emerging field of SEP-analysis is the distinction from other "classical" small proteins (sProteins) [18]. Many small proteins are for a long time known in almost all organisms, often fulfilling important biological functions, for example, ribosomal proteins, insulin, or small neuroproteins, which exhibits protective effects across many cell types as well as in the brain. While some small proteins are formed through proteolytic preprocessing from precursor proteins, others are biosynthesized in a direct manner. SEPs form a subset of small proteins of the proteomes, and are loosely defined as proteins directly translated from sORFs that were not previously annotated in traditional protein annotation workflows [16,19]. The analytical approaches for both SEPs and small proteins are identical.

2 FINDING SEPs IN PROTEOMES—MAJOR CHALLENGES AND GENERAL CONSIDERATIONS

With increasing knowledge about the existence of sORFs, the identification, quantification, and molecular characterization of the SEPs becomes an indispensable requirement towards the study of their biological function(s). While methods such as riboprofiling provide indirect proof of the transcription and translation of these molecules, [20] the direct proof and characterization at peptide or protein level can only be achieved by means of proteomics [16]. Here, the rich methodological repertoire of proteomics and peptidomics, mainly based on the combination of separation technologies (e.g., liquid chromatography (LC)) and various techniques of mass spectrometry (MS) come to the forefront. While numerous approaches for deep coverage of proteomes even in highly complex organisms have been developed, the analysis of smaller proteins and peptides within these proteomes is still characterized by some inherent challenges.

The most commonly used proteomics approach is bottom-up proteomics (BUP). This is based on the digestion of all proteins contained in a proteome into peptides. The resulting highly complex peptide mixture is then separated via single- or multidimensional separation techniques before the peptides are analyzed by MS and MS/MS. In the last step, the spectral information is used to infer the identity of the peptides and conclude the original proteins, which can be performed by a plethora of bioinformatics methods. BUP is well established;
FIGURE 1  Nomenclature of ORFs in mRNA (A) or non-coding RNA (B); (C) overlap of the nomenclature for different ORF types. All ORFs that are not reference ORFs (RefORF, classical ORF of the coding sequence) are referred as altORFs, which can be further subdivided into upstream/downstream ORFs (uORFs/dORFs), small ORFs (sORFs), ORFs from non-coding RNA (altORFsnc), and novel ORFs (nORFs). CDS, coding sequence; ncRNA, non-coding RNA; untranslated region (UTR), untranslated region.

However, also bears some inherent drawbacks. In particular, not all peptides formed by digestion are detectable. This can lead to a loss of information about proteolytic processing events or features in the missing sequences, for example, post-translation modifications (PTMs). Secondly, the distinction of proteoforms [21] sharing peptides is not always possible, leading to the identification of protein groups rather than of particular proteins. These shared peptides also have consequences for the quantification of proteins via proteolytic peptides. For the analysis of small proteins, disregarding whether it is an SEP or classical smaller protein, the number of peptides, which can be formed, are usually much smaller than for bigger proteins. In this case, the loss of information about single peptides can easily lead to the complete loss of information about the entire small protein. It has to be noted that some of the predicted SEPs can have sizes less than 30 aa, which makes these analytes close classical bottom-up peptides.

In many ways the methodologies employed in the detection of SEPs represent the next iteration of peptidomics [22].

Two general principles can be used to circumvent, or at least minimize, these problems. First, one can try to reduce the complexity of the proteomic sample, for example, by depletion of the bigger proteins or the enrichment of the smaller ones. A second way is to avoid the bottom-up strategy by direct analysis of the intact proteins. The later approach, called top-down proteomics (TDP), inherently prevents problems caused by shared peptides or missing peptide stretches. However, the separation of intact proteins and their MS analysis are more challenging than peptide-based LC-MS analysis, for example, due to lowered sensitivity, a more challenging data interpretation, or problems with protein solubility [23]. In this review, we will describe both bottom-up and top-down approaches for SEP analysis.

Since the hypothesis of the existence of SEPs arose, a number of techniques and methodologies have been established across a range of organism spanning the kingdoms [15,16,24,25]. Here, we highlight those techniques established both in pro- and eukaryotes and identify space for further improvement in future analyses of these enigmatic protein species. We will address major steps of analysis, including sample preparation and (small) proteome isolation, separation and MS, data interpretation and quality control, the analysis of PTMs, and the study of functional aspects of the SEPs by means of proteomics technologies.

3  |  SAMPLE PREPARATION AND ISOLATION OF LOW MOLECULAR WEIGHT PROTEINS

3.1  |  Cell lysis

Principally all methods used for cell lysis can be applied for the analysis of SEPs. However, the particular physicochemical properties resulting from their short length require special attention to avoid loss during extraction and proteolytic damage. Loss of SEPs might be caused by incomplete solvation of small hydrophobic proteins in aqueous buffering systems or by the formation of tight complexes with larger cellular complexes, such as membrane proteins or DNA, which could sediment in centrifugation steps typically following the cell lysis.

A main goal of sample preparation and enrichment is to deliver a sample that is both amenable to analysis and in a state that most closely represents the in vivo conditions. Therefore, the inhibition of activities of enzymes suitable to modify proteins, for example, by (ir)reversible post-translational processing, is essential. Major players in this respect are proteases, which can exhibit their activities still after cell lysis, leading to protein degradation, resulting in artificial, not biological relevant proteoforms.

Various methodologies have been described to prevent proteolytic protein activity, for example, the addition of protease inhibitors [24,26–28] or using elevated temperatures [16,29–31].

Furthermore, detergents or high concentrations of chaotrophic salts can be used to both disrupt cells and denature all proteins. If subsequent analytical steps are compatible with detergents (e.g.,
Enrichment of Small Proteins

Reduction of sample complexity prior to subsequent separation is one of the main tasks to decrease chromatographic co-elution. In BUP, there is an inherent bias against small proteins, for example, because the digestion of large proteins leads to higher numbers of generated peptides compared with small proteins as discussed above. In TDP, larger proteins hamper both, the chromatographic separation and MS analysis of small proteins [36]. Therefore, prior to LC-MS analysis of small proteins, it is mandatory to increase the relative abundance of small proteins by either enriching small or depleting large proteins.

A recently published review provides an excellent summary of the number of identified SEPs for different protocols reported in the last years [37]. In this review, we focus on a more general overview of the main steps for small protein analysis, with emphasis on the different methodologies.

A number of approaches have been developed to enrich small proteins or peptides based on their physicochemical properties (Figure 2). Technologies using the protein size as discriminator, such as size-exclusion chromatography (SEC) or molecular weight cut-off (MWCO) filters are common methodologies for peptidome analysis [38]. For example, an SEC-based enrichment strategy was developed for the identification of more than 100 low abundant small proteins in human plasma [39]. Another methodology widely used in peptidome analysis is the application of restricted access materials (RAMs). These consist of functionalized size-exclusion material with the outer surface of the particles being coated with a hydrophilic packing that protect from the entry of larger proteins into the pores. The surface of the pores can be functionalized with different affinity matrices. RAMs possess a size
exclusion effect for both enrichment and chromatographic separation of small proteins [38].

Additional one-step enrichment methods include depletion of larger proteins by precipitation with organic solvents, such as in reversed acetone precipitation [40]. Other approaches successfully utilized acetonitrile in a volume ratio of 1:3–1:4 (water:acetonitrile) to precipitate large proteins [17,24]. Interestingly, the salt concentration in the extraction solution is the most important factor for small protein enrichment [17,41]. Moreover, it is possible to extract specifically acidic or basic small proteins adapting the pH of the extraction solution [17]. Besides organic precipitation, an acidic precipitation was applied for the identification of ghost proteins in human glioma cells [29,30]. Based on the same principle, in differential solubilization, first all proteins are precipitated and in a second step, small proteins were resolubilized, for example, with 70% acetonitrile [17,42]. Further, liquid-liquid extraction was used for the enrichment of small proteins; for example, after methanol-chloroform-water precipitation, three phases were obtained (chloroform-rich and water-rich phase and the interlayer where most protein precipitate), with hydrophobic small proteins enriched in the chloroform-rich phase [43].

Fractionation using gel-based methods, such as SDS-PAGE or GELFrEE, [44] are widely used for small protein and SEP analysis because proteins can be separated over a wide molecular weight range and sample clean-up occurs simultaneously during electrophoresis. Moreover, for the separation of small proteins and peptides specialized methodologies have been developed, for example, Tricine-PAGE, [45] which have been applied for SEP analysis [24]. Furthermore, a GELFrEE approach was successfully employed in combination with proteolytic digestion of the proteins and subsequent LC-MS analysis to identify SEPs in the archaeon Methanosarcina mazei [28].

MWCO filters are based on the principle that larger proteins than the characteristic pore size, for example, 3, 10, or 30 kDa, are retained, while smaller proteins and peptides pass through the membrane. It was shown that the filter material used by different vendors has a significant impact on the success of isolation of small proteins in human plasma and batch-to-batch variance was observed even for filters of the same material [46,47]. Moreover, the pores of MWCO filters tend to clog at high protein concentrations resulting in a reduced protein recovery [48]. Besides these limitations, MWCO filters were applied in multiple studies for the identification of SEPs [16,29,31,49,50].

Another possibility for the enrichment of SEPs is solid-phase extraction (SPE), which is based on the combination of reversed phase binding (due to the hydrophobicity of the proteins) and size-based separation (due to limited surface-accessibility). Both, C18 and C8 reversed phase stationary phases were successfully used for the enrichment of SEPs [29,33,40,51]. Using SPE, 210 small proteins could be identified in Bacillus subtilis [51].

Due to the heterogeneity of small proteins in terms of their physicochemical properties, it is unlikely that a single approach is suitable for enrichment of all small proteins in a given organism. For example, very hydrophobic and membrane-associated small proteins can be easily lost during cell lysis or bind strongly to devices used to enrich small proteins, such as MWCO filters. Using adapted methods for hydrophobic proteins, however, causes potential loss of hydrophilic proteins. In addition, small proteins that bind tightly to other proteins may be missed when using size-separated approaches [28,52]. For an intensive analysis of SEPs, it is therefore mandatory to use multiple complementary methodologies to improve both the number and the confidence of the identifications [24,29,30,40].

### 3.3 | Digestion

In the case of BUP any of the above methods would be followed by the (enzymatic) digestion to generate peptides. A variety of protocols has been evolved in classical proteomics and the advantages and disadvantages of each protocol apply to the analysis of SEPs. However, additional considerations are required. For gel-based approaches, the influence of the staining method was investigated [27]. It was shown that by using different staining methods (Coomassie, negative staining, and no staining), complementary SEPs could be identified, resulting in a large overall improvement in the number and quality of SEP identifications. Other authors supposed that diffusion of small proteins during the equilibration following protein fixation in PAGE could be employed to enrich this fraction prior digestion [53].

Although the mainly used protease in proteomics is trypsin, the usage of multiple proteases can dramatically increase the sequence coverage and confidence of the identified small proteins [32,51,54]. The advantage of using alternative proteases than trypsin is primarily due to identification of critical sequence regions that cannot be identified with trypsin, for example, sequences with many or lacking any lysine or arginine residues [32].

### 4 | SEPARATION AND MS

After sample preparation, which can also encompass pre-fractionation and separation steps as outlined above, all proteomic analyses include steps centered around analyte separation and MS analysis. Here, with the bottom-up and the top-down approach, two fundamental different strategies can be utilized, which of course also demand for suitable sample preparation as discussed above.

#### 4.1 | Bottom-up and semi-top-down approaches

The most commonly employed approach for the analysis of SEPs is the bottom-up approach as the underlying technologies are well established and elaborated in classical proteomics [37]. Indeed, the analysis of SEPs by BUP does not significantly differ from the classical analyses using reversed-phase LC, mostly employing C18 stationary phases for the separation of peptides. The continued refinement of LC systems, the downsizing of column internal diameters, and the
improvement in consistency of the column packing materials have resulted in highly reproducible columns that offer excellent resolving power and separation for peptides.

In a number of studies aiming to identify and characterize SEPs, after an enrichment step, for example, using MWCO filters, one-dimensional LC separation coupled online to MS was applied [50]. On the other hand, several multidimensional separation schemes were employed, which can be classified in semi-top-down approaches and full bottom-up like strategies.

Semi-top-down approaches employ a separation at intact protein level, followed by proteolytic digestion of collected fractions and subsequent peptide-level LC-MS. Approaches such as the combination of SDS-PAGE/in-gel digestion/LC-MS are typical examples and have been employed successfully for the identification SEPs in bacteria, [55] archaea [27] as well as higher organisms, such as Drosophila and mammalian cells [30].

Ma et al. presented a three-dimensional (3D) separation scheme [56]. Here, after a MWCO pre-fractionation, an SDS-PAGE followed by an electrostatic repulsion hydrophilic interaction chromatography and a RP-LC separation was performed. Application of this approach enables the identification of 94 SEPs from human K562 cells. In the same study, the effect of various combinations of the aforementioned separation dimensions were also evaluated, with the authors clearly demonstrating the benefit of the 3D scheme [56].

A classical peptide based two-dimensional LC-separation, encompassing a semi-orthogonal combination of high pH (pH 10) reversed-phase and low pH (pH 2) ion-pairing reversed-phase LC, [57] was used in the same study mentioned above employing the GELFrEE separation of M. mazei proteins [28]. Both approaches delivered a high overlap of identified SEPs, but each approach also identified SEPs not detect by the other. Overall, using the GELFrEE semi-top-down approach the total number of protein identifications was lower, but sequence coverages for low molecular weight (LMW) proteins were higher, in particular for longer peptides (above about 29 amino acids) [28].

4.2 | Top-down proteomics

Direct analysis at the intact protein level by TDP provides an alternative in which big advances in analytics have been achieved in the past few years [58,59]. Despite these huge advances, the use of TDP for the analytics of SEPs is still not widespread. SEPs are, with their typical sizes between 600 Da and 10 kDa, at the lower edge of the ideal range for top-down analytics that now ranges up to ca. 70 kDa (however, the vast majority of analyses focus on proteins less than 30 kDa) [60]. Indeed, SEPs represent a class of analytes that shares some needs, for example, for LC conditions and MS analysis, with peptides generated in middle-down proteomics. In the latter, enzymatic or chemical cleavage of peptides is performed aiming to produce peptides that are longer than those typically formed by tryptic digestion [61]. Parameters for the LC-MS based analysis of middle-down peptides have been systematically optimized [62]. Improved peak width and area under the curves for peptides > 4 kDa was achievable when using a C18 material with a 300 Å pore size, as compared to the more typical 100 Å pore size commonly utilized in tryptic digests, which was shown to negatively influence both the retention and separation of peptides in the middle down size range.

TDP was employed for proteome analysis in the archaeon Methanosarcina acetivorans utilizing a 8.5 Tesla quadrupole-Fourier transform MS [63]. While not specifically targeting SEPs, 99 new proteins were identified, of which a number would be classified as SEPs under today’s nomenclature.

A workflow encompassing a pre-fractionation and multidimensional separation strategy at intact protein level was developed recently for the analysis and molecular characterization of SEPs in another archaeon, M. mazei [33]. Key steps, including the application of a 5% formic acid dilution step followed by SPE, were shown to support enrichment of proteins below ca. 20 kDa. This sub-proteome was further pre-fractionated using strong cation exchange chromatography, followed by RP-LC MS analysis. In particular, this approach allowed the removal of a number of unidentified analytes with low charge states (primarily singly charged) which hampered deeper coverage of the LMW proteome. Using both HCD and EThCD ion activation for MS/MS, 36 proteoforms deriving from 12 SEPs could be identified. Most importantly, a number of PTMs of SEPs could be assigned, including proteolytic processing and disulfide formations (see chapter Post-translational Modifications) [33]. This study also highlighted the value of applying different MS/MS ion activation mechanisms such as HCD and EThCD, and moreover of the optimization of MS/MS parameters for intact protein analysis [33,54,62,64].

5 | DATA ANALYSIS AND QUALITY CONTROL

5.1 | Databases for identification of SEPs from MS-spectra

Acquired MS spectra are usually identified by comparing the detected masses to those calculated after an in silico digestion of sequence database. In order to support the discovery of unannotated SEPs by MS, reference genome sequences may be refined with proteogenomic tools to broaden the explorable proteome. These methods combine genomic methods and proteomics to detect expression of yet non-annotated proteins or peptides with the aim to enhance or refine genome annotations [65].

One proteogenomic approach is the application of six-frame translation-based protein databases [66,67]. A more specific approach would be to only add sequence information of expressed genes to reference databases with the help of mRNA sequencing (RNA-Seq). After collecting high-quality RNA-Seq data, the information is assembled into mRNA transcripts. Three frame translation of these mRNAs generate a custom proteomic database that should contain all polypeptide sequences theoretically produced in the cell under the conditions used for the RNA-Seq experiment. With such a database, Slavoff and colleagues were able to identify 90 SEPs in a human cell line out
of which 86 were novel [16]. Another proteogenomic solution is the integrated proteogenomics search database (iPtgxDB) concept that integrates and consolidates annotations and predictions from different sources and simultaneously captures information on their overlap and differences [68]. As alternative start sites are also considered, the entire protein-coding potential of a prokaryotic genome is captured. The hierarchical integration of available annotations and a peptide classification scheme ensures that the vast majority of peptides in the iPtgxDB uniquely identify a single protein [69]. Database search with an iPtgxDB allowed the identification of 22 novel ORFs in Bartonella henselae with a median length of 48 amino acids [68].

However, with expanded databases, the search space becomes much larger, which is a major limitation as the identified peptides often map to multiple possible proteins, thus complicating protein inference [70]. As the vast majority of entries in a six-frame or proteogenomic database belong to novel and very small proteins, the likelihood for a random hit in this subset is larger [71].

In proteogenomic applications, the FDR of novel peptides can differ from the FDR of annotated peptides by several orders of magnitude and the degree of this effect depends mainly on the genome annotation completeness [72]. Hence, additional filters, [16,67] as well as strict downstream validation steps, have to be considered to avoid reporting of spurious novel protein identifications (see below). In order to address the trade-off between sensitivity and rate of false positive identifications from custom made databases derived from six-frame translations, RNA-Seq data, or consolidated annotations, it has been shown to be beneficial to use the number of identified target spectra weighted for the respective source of a database entry (less PSMs required for more credible sources) as a filter for valid results [65,68,73].

After novel SEPs have been reliably detected, the use of spectral libraries that consider all major qualitative and quantitative features of fragment spectra during searches [74,75] can offer a number of benefits in terms of peptide validation and robust identification in large datasets.

### 5.2 Database search parameters and filter criteria for SEP discovery

Independent of the final repository used for spectrum identification at least one database search is included in a proteomic workflow. For the discovery of SEPs, the parallel application of multiple search engines has shown to be beneficial [51,54,76]. This is most likely due to differences in spectrum preprocessing and scoring functions programmed into the different search algorithms that results in slightly different sets of reported peptides [77]. In order to maintain a constant FDR across peptides identified by multiple search algorithms, a subsequent merging step is required. However, a proper merging of search results led to a decreased number of SEP candidates [51]. Notably, this effect was more pronounced for SEPs than for RefSeq annotated small proteins. In order to retain as many SEP candidates as possible during discovery-based analyses, it is possible to append results from the different database searches. However, in order to preserve confidence in the candidate protein identifications validation via external methods is required [51].

In addition to the search algorithms, parameters applied during database searches can greatly affect the identification of small proteins. While strict enzymatic cleavage specificity is usually assumed in proteomics, it seems to be beneficial for SEP-enriched samples to allow peptides with only one enzyme-specific terminus. A decrease in specificity during the database search is unfortunately accompanied by a vastly inflated search space resulting in increased search times, and more strikingly, a larger number of false-positive peptide hits. Still, with a constant FDR of 0.1%, an increased number of identified spectra, peptide sequences, and proteins was observed for semi-specific cleavage compared to full enzyme specificity [51].

Another parameter to be considered for SEP discovery is the minimal number of unique peptides required for a successful protein identification. In a classical proteome experiment, only proteins are accepted for which the search results contain at least two unique peptides. However, only a limited number of peptides can be observed in SEPs. This issue was overcome by the application of filter criteria that are independent of the number of unique peptides and are, for example, based on the MS/MS-spectra quality [16,17,67]. Ion series-based criteria were developed where an SEP is considered as identified if one high confident peptide has a sequence tag of at least five consecutive b- or y-ions in its MS2 spectrum [16]. Other filter criteria for single peptide identification of SEPs were used, for example, higher peptide identification scores, exclusion of peptides with more than one missed cleavage, less than eight amino acids in length, atypical charge state, or variable modifications [16,67,78,79].

### 5.3 Validation of SEP identifications

As the availability of identified high-quality MS/MS spectra alone sometimes does not provide sufficient evidence to validate expression of an sORF into a polypeptide, additional validation might be necessary. As mentioned before, the use of orthogonal ion activation types can be of high value.

Another possibility would be the application of available software such as PepQuery [27,79,80]. This tool is conceptually similar to BLAST and was designed for validating putative novel protein CDSs. PepQuery allows users to query a spectral library with a novel peptide or DNA sequence of interest to look for PSMs. For quality control, PepQuery provides a direct statistical measurement for each PSM, where a 1% p-value cut-off has shown to result in well-controlled FDRs.

The analytical validity of spectra assigned to novel SEPs may also be confirmed by matching the observed spectrum to that of a synthetic peptide with the expected sequence in a spectral library-based approach [73]. Synthetic peptides can also be used with targeted MS methods (MRM or PRM) to confirm the identification of SEPs. In this
case, validation of native SEPs is achieved by adding isotopically labeled synthetic peptides to the sample and subsequently comparing peak profiles, retention times, and fragment ion pattern (considering the mass shift introduced in the synthetic peptide by the isotopic label) [26,51,81].

Validation of SEP-identification can also be achieved by application of non-proteomics methods, such as ribosome profiling [20] or epitope-tagged versions for antibody detection via western blotting [55].

6 | QUANTIFICATION OF SEP

Besides the analytical challenges in SEP discovery and detection, the quantitative comparison of SEP abundance in different samples is an even more difficult task. Usually, quantifications that are based on only a single peptide, are discarded as any statistical analysis within the protein is impossible. Hence, quantitative analyses on SEPs are rarely described and most are limited to selected candidates of a given organism. Classical methods for quantitative peptide and SEP analysis rely mostly on immunoassay approaches that require specific antibodies to recognize the amino acid sequence of interest [82]. Comparative abundance changes of SEPs can also be performed through relative quantification of native forms, with the help of (isotopically labeled) synthetic peptides [39,81]. However, neither immunoassays nor the use of synthetic standards can be easily applied for large-scale quantification of the LMW proteome.

Due to their straightforward implementation in many proteomics workflows, label-free quantification approaches have been applied to detect differently abundant SEPs on a large-scale. However, label-free approaches are the least accurate among the MS quantification techniques because all variations between experiments are reflected in the obtained data. Consequently, the number of experimental steps should be kept to a minimum and every effort should be made to control reproducibility at each step [83]. This is a highly critical factor for the analysis of SEPs, as the majority of workflows employ multiple fractionations and many rely on enrichment or depletion strategies. In 2010, a peptidome-wide quantification based on ion peak integration was compared to spectral counting methods and was shown to be superior [84] which has also been described for large(r) proteins [85].

Compared to label free methods, higher quantitation accuracy can be achieved with stable isotope labeling approaches [86,87]. As the physicochemical properties of labeled and native peptides are almost identical, quantification of a sample of interest can be performed by comparing its MS intensity with that of a labeled peptide standard present in the same sample. However, most of the efforts towards the development of stable isotopic tags for large-scale quantification have been directed at large(r) proteins rather than SEPs or peptides. For quantification of differences in peptide abundance, the isotopic labeling with succinic anhydride to modify N-terminal amines and lysine residues prior to MS has been shown to be a suitable approach [88,89]. Similar to isobaric labels, no proteolytic digest is necessary for successful labeling of peptides. The possibility to omit proteolytic processing of samples makes labeling with isotopically modified succinic anhydride or isobaric tags especially suited for the quantitative description of the LMW proteome.

7 | PTMs

PTM of proteins alters their physicochemical properties, and in consequence, their biological function. As with all classical proteins, knowledge in regard to the presence or absence of modifications on SEPs is essential to understand their biological functions.

The use of MS-based proteomics for the detection of PTMs is well established, [90] and pitfalls and limitations are widely known [91]. A major challenge in BUP is the loss of information when modified peptides are not detected. Therefore, special approaches, for example, the enrichment of modified peptides, are necessary. However, this requires a prior hypothesis in regard to which potential modification may be occurring. In some cases, modified peptides can also be detected without enrichment, for example, due to their high abundance. These issues may explain why only few reports of modified SEPs have been published to date.

A recent publication described the application of in silico modelings for SEPs, showing that they have an increased level of disorder compared to canonical proteins [8]. In addition, the PTM prevalence in SEPs, based on amino acid sequence, was predicted. For most PTM types, the PTM density was predicted to be equivalent or higher in SEPs compared to canonical proteins. In addition, LC-MS data provided evidence on the presence of six phosphorylation events on sORF-encoded peptides and 297 phosphorylations across 259 nORF gene products [8].

In contrast to bottom-up approaches, TDP identifies intact proteoforms, thus offering the opportunity to recognize both proteolytic truncations and other PTMs occurring mainly in the side chains of amino acids. It should be noted that proteolytic truncations can have biological functions going far beyond degradation of proteins but are heavily involved in regulatory processes.

In an early top-down analysis of the archaeon *M. acetivorans*, five sORF-encoded proteins were detected. For one SEP, two proteoforms of the same coding sequence (CDS) were identified with and without the N-terminal methionine residue [63]. TDP was also employed for the analysis of microproteins from mouse brain, and a number of PTMs, including N-terminal acetylation, C-terminal amidation, disulfide bonding, and dehydration could be identified in these proteins [92].

More recently, top-down analysis of the archaeon *M. mazei* identified not only a number of methionine truncations but also the presence of an N-terminal formylation, which is, in contrast to bacteria, less commonly known in archaea [33]. Moreover, TDP allowed the identification of N- and C-terminal regions, which are often missed in bottom-up analyses. This led to the identification of N- or C-terminal truncated proteoforms of several SEPs. Furthermore, through the comparison of reduced and non-reduced samples by combination of EThcD and HCD ion activation, identification and mapping of a number of disulphide linkages in SEPs were possible [33].
Besides the identification of SEPs, the characterization of their biological activity is of particular interest to better understand their role in molecular processes. The vast majority of SEPs have novel sequences with no known functional domains (based on homology to larger proteins) and thus, bioinformatics prediction of protein domains assigning a functional role is difficult. Moreover, the prediction of secondary structure elements is challenging due to the small size of the proteins [52].

Due to the lack of catalytic protein domains, it has been suggested that SEPs elicit an effect through interaction with other proteins, such as increasing or decreasing catalytic activity, stabilization of protein complexes, peptide-assisted degradation, or mediation of other protein-protein interactions. Meanwhile, the functions of some SEPs in, for example, cell division, membrane transport, embryonic development, or cell death have been elucidated. Different biological functions of SEPs have been reviewed [12,15,93,94].

Numerous functional genomic approaches for the analysis of SEPs exist, such as gene deletion, and mutation-based analysis; [95] however, in this review, we focus primarily on proteomics-based workflows.

One common approach is based on quantitative MS proteome analysis of cells or organisms grown under different growth conditions [67,96,97]. The evaluation of the change in protein abundance provides potential information about the function of these proteins. In BUP, synthetic peptides can be used for the absolute quantification of targeted proteins by MRM. This approach provided data showing that the protein expression of the SEP spRNA36 is induced in the archaeon M. mazei, cultivated under nitrogen starvation conditions and that the protein might have a regulatory function in nitrogen metabolism [25]. For proteome-wide quantification, extracted ion chromatograms for peptide quantification for the comparative profiling of SEP between different growth conditions was used, for example, for the identification of SEP involved in cold-shock stress response in Escherichia coli [67]. The development of specific antibodies (e.g., anti-NBDY and anti-SEP) enabled the functional characterization of a number of SEP. The development of specific antibodies (e.g., anti-NBDY and anti-SEP) enabled the functional characterization of a number of SEP.

For proteome-wide quantification, extracted ion chromatograms for peptide quantification for the comparative profiling of SEP between different growth conditions was used, for example, for the identification of SEP involved in cold-shock stress response in Escherichia coli [67]. The development of specific antibodies (e.g., anti-NBDY and anti-SEP) enabled the functional characterization of a number of SEP. Despite this, the direct experimental functional characterization of small proteins is more challenging than that of larger proteins, since intrinsic properties (amino acid composition, protein size and abundance) complicate many classical biochemical techniques. However, optimized approaches for the analysis of SEP-protein interaction using co-immunoprecipitation [100] or APEX tagging [101] have been presented. For example, in a split fluorescent protein experiment a small non-fluorescent tag (e.g., 16 aa in length) of a fluorescent protein is fused to the small protein of interest [100]. The small tag reduces the risk of affecting the structure and function of the SEPs. Co-expression with the remainder of the fluorescent protein results in a fluorescent signal that can be used for localization. Moreover, this method provides an easy way for co-immunoprecipitation using an antibody against the fluorescent protein. Using this approach, the cellular localization and interaction partners of several SEPs in human cells were identified, with several SEPs from uORF forming complexes with the reference protein on the same mRNA [100]. A recently published review provides a detailed overview of SEP-protein interactions that were identified using affinity purification coupled to MS [37].

A methodology for the proteome-wide determination of protein-protein interactions is cross-linking mass spectrometry (XL-MS). Before or after cell lysis, a cross-linker is added that reacts with defined amino acid residues located at a maximum distance from each other, depending on the cross-linker used. After digestion, the cross-linked peptides provide information about possible protein-protein interactions [102]. To study proteome-wide interactions between reference and ghost proteins in human HeLa cells, Cardon et al. re-analyzed a XL-MS study using a database containing predicted ghost proteins and identified 292 interactions where ghost proteins were involved [14].

Using bioinformatics prediction algorithms of protein structure and docking models, it was suggested that one ghost protein potentially regulates ribosome activity through interaction with a ribosomal protein [14].

Bioinformatics prediction of altORFs has greatly improved in the past decade, and several thousand altORFs have now been predicted for many organisms. In addition, the number of identified and functional characterized sORFs and SEPs has tremendously increased, demonstrating the importance of this emerging field. SEPs are involved in many different processes in the cell, for example, embryonic development, regulation of activity from multi-protein complexes, or membrane transport. Therefore, the identification and characterization of SEPs is crucial and has the potential to revolutionize the understanding of molecular processes within the cell. As an example, numerous highly conserved sORFs have been identified in the human microbiome [103] and their gene products play a potential role in host-microbiome interaction [26]. Furthermore, SEPs are potentially very interesting targets for medicine and biotechnology [104,105].

Due to the multiple terms used for SEPs, such as novel proteins, micro-, mini-, or ghost proteins, it would be beneficial to the field if a more uniform definition and nomenclature was adopted. We recommend to use the term SEP for all gene products of sORFs, and altORFs of short size that have not previously been annotated [16,19].

Although BUP is a powerful tool for small protein identification, it encounters several inherent limitations. To overcome these limitations, TDP is an emerging approach, for example, for the identification of both truncated versions and PTMs of small proteins and SEPs. Thus, TDP is a valuable tool for the identification and characterization of proteoforms from SEPs, although the technology needs to be greatly improved regarding chromatographic separation, MS fragmentation, and data analysis.

In addition, classical proteomic workflows depend on protein databases and only proteins in this database can be identified. Therefore, either six-frame translations or de novo sequencing, which is completely independent of any protein database, can offer interesting analytical approaches for the identification of novel SEP [8]. It is worth mentioning that this will also require further development.
of bioinformatics methods for quality control and validation of the identifications.

For functional analysis, XL-MS combined with bioinformatics docking predictions offers a promising method for proteome-wide identification of SEP-protein interactions that can help elucidate SEP functions [14].

In summary, future improvements in the identification and characterization of SEPs will enrich the understanding of cellular molecular processes and potentially offers many approaches for new medical and biotechnological applications.

ACKNOWLEDGMENT

Funding was provided by the Deutsche Forschungsgemeinschaft (DFG) within the priority program SPP2002, project Z1 (TH 872/10-1, BE 3869/5-1).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTION

All authors contribute to manuscript design and writing.

ORCID

Andreas Tholey https://orcid.org/0000-0002-8687-6817

REFERENCES

1. Fickett, J. W. (1995). ORFs and genes: How strong a connection? Journal of Computational Biology, 2, 117–123.
2. Dinger, M. E., Pang, K. C., Mercer, T. R., & Mattick, J. S. (2008). Differentiating protein-coding and noncoding RNA: Challenges and ambiguities. PLoS Computational Biology, 4, e1000176.
3. Goodwin, S., Mcpherson, J. D., & Mccombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. Nature Reviews Genetics, 17, 333–351.
4. Mumtaz, M. A. S., & Couso, J. P. (2015). Ribosomal profiling adds new coding sequences to the proteome. Biochemical Society Transactions, 43, 1271–1276.
5. Kochevtov, A. V. (2008). Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. Bioessays, 30, 683–691.
6. Renz, P. F., Valdivia-Francia, F., & Sendoel, A. (2020). Some like it hot: Current peptidomics: Applications, purification, identification, quantification, and functional analysis. Proteomics, 15, 1026–1038.
7. Toby, T. K., Fornelli, L., & Kelleher, N. L. (2016). Progress in top-down proteomics and the analysis of proteoforms. Annual Review of Analytical Chemistry, 9, 499–519.
8. Wang, B., Hao, J., Pan, N., Wang, Z., Chen, Y., & Wan, C. (2021). Identification and analysis of small proteins and short open reading frame encoded peptides in Hep3B cell. Journal of Proteomics, 230, 103965.
9. Prasse, D., Thomsen, J., De Santis, R., Muntel, J., Becher, D., & Schmitz, R. A. (2015). First description of small proteins encoded by spRNAs in Methanosarcina mazei strain G01. Microbiology, 113, 138–148.
10. Cardon, T., Fournier, I., & Salzet, M. (2021). Shedding Light on the ghost proteome. Trends in Biochemical Sciences, 46, 239–250.
11. Scheidler, C. M., Kick, L. M., & Schneider, S. (2019). Ribosomal peptides and small proteins on the rise. Chembiochem, 20, 1479–1486.
12. Schlesinger, D., & Elsässer, S. J. (2021). Revisiting sORFs: Overcoming challenges to identify and characterize functional microproteins. The F1000 Research Journal. https://doi.org/10.1180/f1000research.15769.
13. Andrews, S. J., & Rothnagel, J. A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. Nature Reviews Genetics, 15, 193–204.
14. Cardon, T., Salzet, M., Franck, J., & Fournier, I. (2019). Nuclei of HeLa cells interactomes unravel a network of ghost proteins involved in proteins translation. Biochimica et biophysica acta. General subjects, 1863, 1458–1470.
15. Storz, G., Wolf, Y. L., & Ramamurthi, K. S. (2014). Small proteins can no longer be ignored. Annual Review of Biochemistry, 83, 753–777.
16. Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., Karger, A. D., Budnik, B. A., Rinn, J. L., & Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nature Chemical Biology, 9, 59–64.
17. Cassidy, L., Kaulich, P. T., & Tholey, A. (2019). Depletion of high-molecular-mass proteins for the identification of small proteins and short open reading frame encoded peptides in cellular proteomes. Journal of Proteome Research, 18, 1725–1734.
18. Hollis, E. C., Fontaine, F., Yin, X., & Storz, G. (2011). An expanding universe of small proteins. Current Opinion in Microbiology, 14, 167–173.
19. Orr, M. W., Mao, Y., Storz, G., & Qian, S.-B. (2021). Alternative ORFs and small ORFs: Shedding light on the dark proteome. Nucleic Acids Research, 48, 1029–1042.
20. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science, 324, 218–223.
21. Smith, L. M., & Kelleher, N. L. (2013). Proteoform: A single term describing protein complexity. Nature Methods, 10, 186–187.
22. Dallas, D. C., Guerrero, A., Parker, E. A., Robinson, R. C., Gan, J., German, J. B., Barile, D., & Lebrilla, C. B. (2015). Current peptidomics: Applications, purification, identification, quantification, and functional analysis. Proteomics, 15, 1026–1038.
23. Petruschke, H., Schori, C., Anzler, S., Riesbeck, S., Poehlein, A., Daniel, R., Frei, D., Segesmann, T., Zimmerman, J., Marinos, G., Kaleta, C., Jehmlich, N., Ahrens, C. H., & von Bergen, M. (2021). Discovery of novel community-relevant small proteins in a simplified human intestinal microbiome. Microbiome, 9, 55.
30. Cardon, T., Hervé, F., Delcourt, V., Roucou, X., Salzet, M., Franck, J., & Fournier, I. (2020). Optimized sample preparation workflow for improved identification of ghost proteins. Analytical Chemistry, 92, 1122–1129.

31. Wang, S., Tian, L., Liu, H., Li, X., Zhang, J., Chen, X., Jia, X., Zheng, Xu, Wu, S., Chen, Y., Yan, J., & Wu, L. (2020). Large-scale discovery of non-conventional peptides in maize and arabidopsis through an integrated peptidogenomic pipeline. Molecular Plant, 13, 1078–1093.

32. Kaulich, P. T., Cassidy, L., Bartel, J., Schmitz, R. A., & Tholey, A. (2021). Multi-protease approach for the improved identification and molecular characterization of small proteins and short open reading frame-encoded peptides. Journal of Proteome Research, 20, 2895–2903.

33. Cassidy, L., Helbig, A. O., Kaulich, P. T., Weidenbach, K., Schmitz, R. A., & 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. Journal of Proteomics, 230, 103988.

34. Trötschel, C., & Poetsch, A. (2015). Current approaches and challenges in targeted absolute quantification of membrane proteins. Proteomics, 15, 915–929.

35. Crowell, A. M. J., Wall, M. J., & Doucette, A. A. (2013). Maximizing recovery of water-soluble proteins through acetone precipitation. Analytica Chimica Acta, 796, 48–54.

36. Baghalabadi, V., & Doucette, A. A. (2020). Mass spectrometry profiling of low molecular weight proteins and peptides isolated by acetone precipitation. Analytical Chemistry, 91, 11085–11091.

37. Fabre, B., Comibier, J.-P., & Plaza, S. (2021). Recent advances in mass spectrometry–based peptidomics workflows to identify short-open-reading-frame-encoded peptides and explore their functions. Current Opinion in Chemical Biology, 60, 122–130.

38. Finoulst, I., Pinkse, M., van Dongen, W., & Verhaert, P. (2011). Sample preparation techniques for the untargeted LC-MS-based discovery of peptides in complex biological matrices. Journal of Biomedicine and Biotechnology, 2011, 245291.

39. Kaulich, P. T., Cassidy, L., Bartel, J., Schmitz, R. A., & 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. Journal of Proteomics, 230, 103988.

40. Trötschel, C., & Poetsch, A. (2015). Current approaches and challenges in targeted absolute quantification of membrane proteins. Proteomics, 15, 915–929.

41. Kawashima, Y., Fukutomi, T., Tomonaga, T., Takahashi, H., Nomura, F., Maeda, T., & Kodera, Y. (2010). High-yield peptide-extraction method in top-down proteomics identifies microproteins, missing proteins, and novel smORFs in Saccharomyces cerevisiae. Journal of Proteome Research, 17, 2335–2344.

42. He, C., Jia, C., Zhang, Y., & Xu, P. (2018). Enrichment-based proteogenomics identifies microproteins, missing proteins, and novel novel small pro- teins in the human gut microbiome. Proceedings of the National Academy of Sciences of the United States of America, 101, 13417–13422.

43. Thomas, S. L., Thacker, J. B., Schug, K. A., & Maráková, K. (2021). Sampl etion and validation of novel small proteins in Pseudomonas putida. Environmental Microbiology Reports, 8, 966–974.

44. Tran, J. C., & Doucette, A. A. (2008). Gel-eluted liquid fraction entrapment electrophoresis: An electrophoretic method for broad molecular weight range proteome separation. Analytical Chemistry, 80, 1568–1573.

45. Schägger, H., & Von Jagow, G. (1987). Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. Analytical Biochemistry, 166, 368–379.

46. Greening, D. W., & Simpson, R. J. (2010). A centrifugal ultrafiltration strategy for isolating the low-molecular weight (< 25K) component of human plasma proteome. Journal of Proteomics, 73, 637–648.
66. Krug, K., Carpy, A., Behrends, G., Matic, K., Soares, N. C., & Mack, B. (2013). Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Molecular & Cellular Proteomics: MCP*, 12, 3420–3430.

67. D’Ilia, N. G., Khitun, A., Rosenbloom, A. D., Yuan, P., Gassaway, B. M., Barber, K. W., Rinehart, J., & Slavoff, S. A. (2017). Comparative proteomics enables identification of nonannotated cold shock proteins in *E. coli*. *Journal of Proteome Research*, 16, 3722–3731.

68. Omasits, U., Varadarajan, A. R., Schmid, M., Goetze, S., Pavlou, M. P., Grosboillot, V., Shen, Y., Burger, T. (2018). Gentle introduction to the statistical foundations of mass spectrometry in proteomics. *Angew. Chem. Int. Ed.*, 57, 1017–1031.

69. Nesvizhskii, A. I. (2010). A survey of computational methods and their roles in proteomics research. *Journal of Proteome Research*, 9, 729–740.

70. Omasits, U., Varadarajan, A. R., Schmid, M., Goetze, S., Pavlou, M. P., Grosboillot, V., Shen, Y., Burger, T. (2018). Gentle introduction to the statistical foundations of mass spectrometry in proteomics. *Angew. Chem. Int. Ed.*, 57, 1017–1031.

71. Burger, T. (2018). Gentle introduction to the statistical foundations of false discovery rate in quantitative proteomics. *Journal of Proteome Research*, 17, 12–22.

72. Zhang, K., Fu, Y., Zeng, W.-F., He, K., Chi, H., Liu, C., Li, Y.-C., Gao, Y., Xu, P., & He, S.-M. (2015). A note on the false discovery rate of novel peptides in proteogenomics. *Bioinformatics*, 31, 3249–3253.

73. Varadarajan, A. R., Goetze, S., Pavlou, M. P., Grosboillot, V., Shen, Y., Loessner, M. J., Matic, K., Soares, N. C., & Ahrens, C. H. (2017). An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. *Genome Research*, 27, 2083–2095.

74. Qeli, E., & Ahrens, C. H. (2010). PeptideClassifier for protein inference and targeted quantitative proteomics. *Nature Biotechnology*, 28, 647–650.

75. Nesvizhskii, A. I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteome Research*, 73, 2092–2123.

76. Wen, B., Du, C., Li, G., Ghali, F., Jones, A. R., Käll, L., Xu, S., Zhou, R., Xu, P., & He, S.-M. (2015). A note on the false discovery rate of novel peptides in proteogenomics. *Bioinformatics*, 31, 3249–3253.

77. Varadarajan, A. R., Goetze, S., Pavlou, M. P., Grosboillot, V., Shen, Y., Loessner, M. J., Matic, K., Soares, N. C., & Ahrens, C. H. (2017). An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. *Genome Research*, 27, 2083–2095.

78. Melior, H., Li, S., Stötzel, M., Maaß, S., Schütz, R., Azarderakhsh, S., Shevkoplias, A., Barth-Weber, S., Baumgardt, K., Ziebuhr, J., Förster, K. U., Chervonska, Z., Becher, D., & Evgeniou-Hackenberg, E. (2021). Reprogramming of sRNA target specificity by the leader peptide pETpRL in response to antibiotic exposure. *Nucleic Acids Research*, 49, 2894–2915.

79. Hobbs, E. C., Yin, X., Paul, B. J., Astarita, J. L., & Storz, G. (2012). Conserved small protein associates with the multidrug efflux pump AcrB and differentially affects antibiotic resistance. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 16696–16701.

80. Bantscheidt, M., Schirle, M., Sweetman, G., Rick, J., & Kuster, B. (2007). Quantitative mass spectrometry in proteomics: A critical review. *Analytical and Bioanalytical Chemistry*, 389, 1017–1031.

81. Chen, J., Brunner, A. D., Cogan, J. Z., Núñez, J. K., Fields, A. P., Adams, B., Itzhak, D. N., Li, Y. J., Mann, M., Leonti, M. D., & Weissman, J. S. (2020). Pervasive functional translation of noncanonical human open reading frames. *Science*, 367, 1140–1146.
101. Chu, Q., Rathore, A., Diedrich, J. K., Donaldson, C. J., Yates, J. R., & Saghatelian, A. (2017). Identification of microprotein-protein interactions via APEX tagging. *Biochemistry, 56*, 3299–3306.

102. Klykov, O., Steigenberger, B., Pektaš, S., Fasci, D., Heck, A. J. R., & Scheltema, R. A. (2018). Efficient and robust proteome-wide approaches for cross-linking mass spectrometry. *Nature Protocols, 13*, 2964–2990.

103. Sberro, H., Fremin, B. J., Zlitni, S., Edfors, F., Greenfield, N., Snyder, M. P., Pavlopoulos, G. A., Kyrpides, N. C., & Bhatt, A. S. (2019). Large-scale analyses of human microbiomes reveal thousands of small. *Novel Genes Cell, 178*, 1245–1259.e14.

104. Couso, J.-P., & Patraquim, P. (2017). Classification and function of small open reading frames. *Nature Reviews Molecular Cell Biology, 18*, 575–589.

105. Brandenburg, F., & Klähn, S. (2020). Small but smart: On the diverse role of small proteins in the regulation of cyanobacterial metabolism. *Life, 10*, 1–15.

How to cite this article: Cassidy, L., Kaulich, P. T., Maaß, S., Bartel, J., Becher, D., & Tholey, A. (2021). Bottom-up and top-down proteomic approaches for the identification, characterization, and quantification of the low molecular weight proteome with focus on short open reading frame-encoded peptides. *Proteomics, e2100008*. https://doi.org/10.1002/pmic.202100008