Complementarity of Different SDS-PAGE Gel Staining Methods for the Identification of Short Open Reading Frame-Encoded Peptides

Philipp T. Kaulich, Liam Cassidy, Katrin Weidenbach, Ruth A. Schmitz, and Andreas Tholey*

Short open reading frame-encoded peptides (SEP) have been identified across all domains of life and are predicted to be involved in many biochemical processes, however, for the vast majority of SEP their biological function is still unknown. Optimized methodologies have to be used for the mass spectrometric analysis of SEP, because traditional methods of bottom-up proteomics show a bias against small proteins. Here, different staining methods for SDS-PAGE gels prior in-gel digestion following LC-MS/MS analysis for the identification of SEP in the archaeon Methanosarcina mazei are investigated. In total, 45 SEP with at least one high confidence (FDR < 1%) unique peptide and five consecutive b- or y-ions in the MS2 spectrum are identified. The staining methods provide complementary data. The highest number of SEP are identified in the samples stained with Coomassie brilliant blue. However, the highest quality of the identified SEP is achieved in the samples without staining. These comprehensive data sets demonstrate that in-gel digestion is well suited for the identification of SEP.

1. Introduction

Traditional predictions of protein-coding genomic regions utilize algorithms containing a cut-off value for the minimum ORF length to reduce false assignment of non-coding to coding sequences. Therefore, short or small open reading frames (sORF), coding for small proteins (SEP) with lengths between 5 to around 100 amino acids, have for a long time been neglected in genome annotation processes. Moreover, sORF can be located within non-coding RNA sequences or out-of-frame within large ORFs which adds a new class of ORFs, the so-called altORF to the genomic repertoire of organisms. The improvement of the algorithms provided evidences for a high number of sORFs in almost all domains of life, ranging from bacteria to mammals and plants. For example, in the genome of the bacteria in the human microbiome thousands of highly conserved sORF were discovered across the different bacterial species.

Gene-products from sORF, named short/small open reading frame-encoded peptides (SEP) or microproteins, are predicted to be involved in many essential processes. However, for the vast majority of SEP the biological function is still unknown, while for a small number of SEP functions could be identified. For example, small proteins were described to be involved in cell division, sporulation, transport systems or in signaling in bacteria cells. The first SEP from a eukaryote was detected in Drosophila, which fulfills an important function during the embryonal development. In human cells, hundreds of SEP were identified and some of them have been functionally characterized.

The identification of SEP at protein level is therefore an essential task for the comprehensive understanding of the molecular processes in the cell. However, the mass spectrometry-based identification of SEP is challenging, because traditional methodologies of bottom-up proteomics show a bias against small proteins. Therefore, special approaches have to be used, for example, enrichment of small or depletion of larger proteins. A number of methodologies have been described, for example the use of molecular weight cut-off filters, the enrichment of small proteins by the GELFrEE-system, the application of size exclusion chromatography or of solid phase extraction and the depletion of high molecular weight proteins by acetonitrile-precipitation or acetic acid precipitation. However, these methodologies favor and disfavor particular proteins and therefore the development of other potentially complementary methodologies for the analysis of small proteins is necessary.

Here, we investigated classical SDS-PAGE based proteome separation followed by in-gel digestion with trypsin prior to LC-MS/MS (GeLC-MS) analysis for the identification of small proteins, and in particular SEP. We assessed the influence of
different SDS-PAGE gel staining methods prior to in-gel digestion: i) only fixing the gel, ii) zinc-imidazole negative staining,\textsuperscript{[18]} and iii) Coomassie-staining, on the identification of SEP. As a model organism, we used the archaeon \textit{Methanosarcina mazei}, grown under nitrogen starvation conditions; the same biological situation was used earlier to establish alternative strategies for SEP analysis.\textsuperscript{[11,14]} The genome of this organism encodes for $\approx$3400 predicted proteins and additional about 1400 predicted SEP.\textsuperscript{[19–21]}

2. Results and Discussion

2.1. Proteins Identified by In-Gel Digestion/GeLC

We investigated the influence of different staining methods for SDS-PAGE separated proteins on protein identification by means of in-gel digestion and LC-MS/MS analysis (GeLC). For this, 200 $\mu$g \textit{M. mazei} total proteome were separated by SDS-PAGE and three different staining methods were performed, each in triplicates: i) negative-staining (zink-imidazol), ii) no-staining (only fixing the gel), and iii) Coomassie-staining. Eight bands below 30 kDa were cut out and an in-gel digestion with trypsin was performed (Figure S1, Supporting Information). Band 1 contains the proteins with the highest molecular mass ($\approx$30 kDa), band 8 those with the lowest molecular mass.

An inherent problem in bottom up proteomics is the inference of the proteins from the peptides generated by digestion. This prevents a clear discrimination of protein fragments formed in vivo by endogenous proteases, and is an important consideration to be taken into account when evaluating the data. Protein (group) identifications reported below were based on identifications of at least two peptides, where one is proteotypic (FDR $<1\%$) (strict criteria).

In total, we identified per replicate ($n = 3$, average and standard derivation stated) 1705$\pm$26 protein groups with 160172$\pm$11,306 peptide-spectral matches (PSM) in the negative-stained samples; 1687$\pm$6 protein groups with 135720$\pm$13,940 PSM in the non-stained samples; and 1708$\pm$79 protein groups with 141991$\pm$11,472 PSM in the Coomassie-stained samples.

Some characteristics of the proteins identified in the different gel bands are exemplarily shown for the negative-stained gels in Figure S2, Supporting Information; no significant differences between the staining methods regarding these general aspects were observed.

The median size of the identified proteins decreased from band 1 (340 aa) to band 8 (160 aa) which correlates with the expected sizes according to the location in the gel. In band 7 and 8, however, potentially proteolytic breakdown products of bigger proteins were identified.\textsuperscript{[22]} For example, the 97.1 kDa protein pyruvate phosphate dikinase (Q8PW31) was identified in band 6 ($<10$ kDa region) with a high number of unique peptides (36) and a high sequence coverage (49%); the identified peptides are distributed over the entire protein sequence (Figure S3, Supporting Information).

In all gel-bands, proteins spanning the entire MW-range were identified. This is in agreement with earlier studies showing that 40% of proteins were not detected in the expected MW region of a SDS-PA gel.\textsuperscript{[23]}

Significance Statement

In recent years, formerly widely overlooked short open reading frames (sORF) came in the focus of biology. This creates a demand for the development of suitable analytical strategies to confirm the presence of the corresponding small proteins (sORF-encoded proteins, SEP) at the protein level. We evaluated the usability of one of the most commonly used proteomics approaches—the combination of SDS-PAGE with subsequent proteolytic digestion and LC-MS (GeLC-MS)—for the analysis of SEP. By application of three different staining methods we were able to obtain complementary information with improved reproducibilities and quality of identifications. To our knowledge, the present study is the first which compares different protocols for the use of this well-established analytical approach to the emerging field of SEP analysis and thus for the study of yet widely overlooked parts of proteomes.
Figure 2. Overlap of the identified SEP in GeLC-MS analysis after different staining methods using A) ion-series based, B) strict criteria.

Coomassie- (166±13) and negative-staining (163±3) identified slightly lowered numbers. The reduced numbers of small proteins after negative-staining is potentially due to the lack of a fixing step after SDS-PAGE.

The influence of the Coomassie-staining on the identification of proteins in the mass range from 10 kDa to 600 kDa by GeLC-MS was investigated earlier.[24] It was shown that significantly more proteins below 40 kDa were identified after staining compared to only fixing the gel. This observation was explained by the binding of Coomassie to the proteins, resulting in a lower diffusion due to the larger volume of the Coomassie-protein-complex. In our study, however, there was no significant difference of the number of identifications depending on protein size between the Coomassie- and the non-stained samples.

After in-gel digestion, independent of the staining method, compared to the 1D-LC analysis more than double the number of proteins with a size below 100 aa (≈75 versus ≈170 proteins) were identified. This benefit of multidimensional separations in respect of identifications and data quality is well known and has also been described for GeLC approaches.[25] Several studies described a pronounced loss of small proteins during washing and equilibration steps in in-gel digestion.[26,27] However, our data demonstrate that the benefit of the enrichment of small proteins is much higher than the potential loss.

2.3. Identification of SEP: Staining Methods Provide Complementary Information

The SEP were named with arbitrary accession numbers for database searches. In Table S1, Supporting Information the accession numbers are assigned to the gene identifiers.

A major challenge for the identification of small proteins, including SEP, in bottom-up proteomics is the number of peptides being generated during digestion. Many of these may be undetectable via LC-MS or shorter than 6 aa which is a critical minimal length for confident automated interpretation of the MS/MS spectra. For example, the SEP A00959 consists of 30 residues, and following in silico tryptic digestion only a single peptide 15 residues in length is generated, while the other peptides are shorter than 6 amino acids. To compensate for these limitations, we applied two different criteria to identify SEP: i) strict (at least two high confident peptides and at least one proteotypic peptide); ii) ion series-based, in which an SEP is identified with one high confident proteotypic peptide if in its MS2 spectrum a series of at least five consecutive b- or y-ions is detected.[28] In the following, we report SEP identified applying both criteria (ion series-based/strict); the MS/MS spectra were manually evaluated in addition to the automated assignment for the validation of ion series-based identifications.

Summing up the results achieved with the three staining methods, we identified 45 (ion series-based) or 21 (strict) SEP (Table 1; Table S2, Supporting Information). From these, 30/19 were identified in the negative-stained, 30/18 in the non-stained and 36/20 in the Coomassie-stained samples (Figure 2).

From the 45 SEP identified with ion series-based criteria, 24 (53%) were identified in all three stainings and 17 (81%) were observed with strict criteria (Figure 2). On the other hand, 18 (40%, ion series-based) and 2 (10%, strict) SEP were identified only in one of the three staining methods.

Aside from the numbers (summed up from three technical replicates) and the complementarity of identifications the three staining methods showed different levels of reproducibility in technical replicates. (Figure S4, Supporting Information). The reproducibility in the Coomassie-stained gels was lower than both the negative- or non-stained gels. Furthermore, fewer SEP were identified in all three replicates (47%/85%) compared to the negative- (70%/84%) or the non-stained gels (73%/95%).

The quality of the identified SEP can be evaluated by factors such as the number of associated PSM, the sequence coverage or the Xcorr, which is a quality criterion for MS2-spectra (Figure 3). While the number of identified SEP using ion series-based criteria was comparable between the different staining methods, the number of associated PSM was significantly reduced in the Coomassie-stained gels (783±102 PSM) compared to the non-stained gels (1593±112 PSM) and the negative-stained gels (1392±233 PSM). Moreover, the sequence coverage of the SEP
identified in the negative- and the non-stained gels was slightly higher compared to the Coomassie-stained gels.

An interesting feature from this analysis was the SEP A00184, which was identified with a significantly higher number of PSM in the non-stained (940 PSM per replicate) and negative-stained (820 PSM) compared to Coomassie-stained samples (220 PSM). The Coomassie dye binds to positively charged amino acids. Due to the high number of positively charged amino acids in SEP A00184 (12 of 67 residues), the interaction with Coomassie is potentially very strong and may prevent complete removal of the dye molecules during destaining. This potentially affects the digestion and the following LC-MS/MS analysis. However, for almost all SEP the number of associated PSM in the Coomassie-stained samples was lower compared to the negative- and non-stained gels (Figure S5, Supporting Information). The sequence coverage showed the largest variation and was slightly lower for almost all SEP identified in the Coomassie-gels, again demonstrating a reduced reproducibility achievable with this method. The averaged Xcorr was slightly increased in the non-compared to the negative-stained gels, however, it was the highest in the Coomassie-stained gels (Figure 3). Further, no obvious bias in the physico-chemical properties of the identified peptides in the different staining methods was observed.

In summary, despite any shortcomings within the three staining approaches tested, together they show complementary results allowing the identification of a range of different SEP.

The distribution of SEP over the eight MW regions was also analyzed (Figure S6, Supporting Information). The vast majority of SEP and associated PSM were identified in bands 5, 6, and 7. For example in band 6 of the negative-stained gels 14±2 SEP with 495±130 PSM were identified. Interestingly, in this band the protein with the highest number of PSM was the SEP A00184. This SEP was also found to be highly abundant in our previous study using a 2D-LC approach. In accordance with the finding for other proteins, SEP were identified in all bands, although based on their molecular size, they were only expected in the LMW bands. Possible reasons for this are discussed below.

2.4. Identified SEP

Overall, we identified 45 SEP via GeLC-MS across the different staining methods under ion series-based criteria (at least one proteotypic peptide (FDR <1%) and five consecutive b- or γ-ions), with 21 SEP identified under strict (two peptide) criteria (Table 1; Table S2, Supporting Information).

The 1D-LC approach identified only 12/6 SEP (ion series-based/strict) (Table S7, Supporting Information). Moreover, the number of identified PSM and the sequence coverage of the SEP were much higher in all GeLC-approaches. The application of the multidimensional GeLC separation had an observable influence on MS quality.

Direct comparison of the MS1 and corresponding MS2 spectra of a proteotypic peptide from the SEP A00075 demonstrates the benefit of the GeLC approach (Figure 4), with reduced isolation interference and a “cleaner” less crowded MS2 spectra that allows complete γ-ion series annotation.

In GeLC analysis a number of SEP were unexpectedly identified either additionally (18 SEP) or exclusively (11 SEP) in the gel...
bands representing higher MW regions (band 1–3, 30–20 kDa). For example, the SEP A00954 (12 amino acids) was identified with 12 PSM and a sequence coverage of 58% in the 30 kDa band. Interestingly, it was primarily small SEP that were identified in the 20–30 kDa bands (Tables S3–S6, Supporting Information).

We hypothesized that a potential overloading of the gel caused the occurrence of SEP in higher MW regions. However, analysis of a reduced amount (20 µg) of *M. mazei* proteome via SDS-PAGE gel with fixing and in-gel digestion of the proteins in the eight bands below 30 kDa showed a distribution of SEP that was approximately the same as observed in gels run with 200 µg protein (Figure S7, Supporting Information).

Interestingly, in our previously employed GELFrEE separation strategy, also based on intact protein separation in first dimension, we also identified SEP in MW fractions not corresponding to the particular molecular masses of the analytes.\(^{[11]}\)
There are different explanations for the unexpected migration behavior of proteins in SDS-PAGE gels,\[23,30\] for example disordered structures and extreme amino acid composition. However, we did not observe unusual amino acid composition of these SEP.

In our dataset, two SEP (A00380 and A00937) were identified by single PSM as intact proteins (21 and 36 amino acids, respectively) using ion series-based criteria. The MS2 spectra, however, were very crowded and contained a high number of unassigned peaks (Figure S8, Supporting Information). For 19 SEP, N-terminal peptides were identified of which 8 (42%) had a truncated start methionine.

A particular challenge for the identification of SEP by means of LC-MS-based proteomics is highlighted by the SEP A00213. This small protein shares 100% sequence homology with the N-terminal region of the highly abundant (based on PSM count) methyl-coenzyme M reductase, gamma subunit (Q8PXH7) (residues 4 through 47). MS data cannot be used for unambiguous discrimination between these two proteins, because A00213 could also be formed through proteolytic processing of Q8PXH7. Interestingly, only the methionine cleaved N-terminus from A00213, but not from Q8PXH7 was identified. As the protein Q8PXH7 was identified with more than 100 unique peptides, but only the N-terminal sequence MHEM was not identified, one can hypothesize that the annotated sequence for the start methionine is incorrect. To confirm this hypothesis, a pBLAST search was performed. Among others, the Methyl-coenzyme M reductase, gamma subunit (P07964) from *Methanosarcina barkeri* was found with a high sequence identity, which starts with the methionine that conforms to the second methionine in the sequence of Q8PXH7. If there is a 100% sequence identity beginning on the first amino acid, then only the identification of the C-terminal peptide allows an unambiguous identification of the SEP A00213. In GeLC analysis, there is an additional information about the molecular mass of the identified proteins due to the location of the analyzed band in the SDS-PAGE gel. We visualized the number of identified peptides assigned to A00213 and Q8PXH7 and the sequence coverage in the eight gel bands (Figure S9, Supporting Information). We discovered for A00213 two maxima for the number of PSM (only from the N-terminal peptide) in band 1 and band 7. For Q8PXH7 the highest number of PSM were found in band 1 and the sequence coverage decreases from band 1 to band 8. The proteins were expected in band 1 (Q8PXH7, 28.1 kDa) and band 7 (A00213, 4.8 kDa). Thus, it is likely, that the identified N-terminal peptide in the first band derives from (the intact protein) Q8PXH7 and in band 7 from A00213. However, the C-terminus from A00213 was identified only with a single PSM (in band 1).

It is important to note that with this approach, it is not possible to clarify the origin of the identified peptides and therefore other methods such as riboprofiling and genome tagging would be necessary to confirm the presence of this SEP.

### 3. Concluding Remarks

The combination of SDS-PAGE, in-gel digestion and LC-MS analysis is one of the most successful and widely used approaches for bottom-up proteomics. However, this methodology has seldom been used for analyses directed toward SEP.\[16,17,28\] Here,
we show that the application of either non-staining or a negative-
staining can increase both the number and confidence of small protein identifications. In particular, the number of PSM and the reproductibility of these approaches leads to improved identification of this novel class of biomolecules. Thus, we suggest the parallel application of the three staining procedures in order to achieve a high coverage of SEP.

Some general caveats inherent for bottom-up proteomics have to be taken into consideration. Here, the distinction of proteolytically processed forms of the proteins and also of SEP and the loss of information due to the missed detection of complete sequence stretches can lead to the loss of information, for example about potential posttranslational modifications.

Due to the reduced number of detectable peptides, a major issue in the analysis of SEP in bottom-up proteomics is the validation of the identifications. Hence the application of strict rules for identification in order to reduce the number of false positives is mandatory. In this study, we applied ion series-based criteria where a peptide is identified if in its MS2 spectrum a series of at least five consecutive b- or y-ions is detected.[28] Recently, Cao et al. suggested a peptide-centric approach[31] to validate SEP identified with only one PSM.[17] We performed this additional filtering step and in total 39 out of 45 SEP identified in or study passed this criterion, while six peptides assigned to SEP identified with only one PSM were negative (Table 1). We further manually interpreted the six spectra of the peptides not passed the criterion and observed extensive precursor co-isolation of other peptides. It is not the aim of this study to compare the different identification criteria; however, to improve confidence of identifications will be a major issue for future studies.

Further improvements in the field of SEP research, which are in part related or applicable to our approach, are possible. A general issue is the increase of MS sensitivity for the detection of low abundant peptides what is also influenced by reducing the complexity of the samples by suitable separation technologies as presented here. Further improvements in detection of SEP can be potentially achieved by the usage of multiple proteases to increase the number of detectable peptides to improve the sequence coverage. Matching LC-MS data with riboprofiling data will potentially deliver important complementary data about SEP expression, which will, together with improved criteria for acceptance of single peptide identifications enhance the confidence of identifications.

Overall, the well-established GeLC-MS approach using different staining methods presented herein provides an easily accessible workflow that we believe should be added to the toolbox available to delve deeper into unknown regions of the small protein proteome.

4. Experimental Section

Chemicals: Trypsin was from Promega (Madison, USA), cComplete EDTA-free protease inhibitor cocktail from Roche (Penzberg, Germany). All other chemicals were from Sigma-Aldrich (Steinheim, Germany). Deionized water (18.2 MΩ) was prepared by an arium611 VF system (Sartorius, Göttingen, Germany).

Samples: Methanotherma mazei was cultivated under anaerobic conditions (37 °C, 80% N2, 20% CO2) on minimal medium (nitrogen starvation) with 150 mM methanol and 40 mM acetate. The cells were grown until a turbidity at 600 nm of 0.6 and then centrifuged 30 min at 3100 x g at 4 °C. The cells were washed with 100 mM triethyl ammonium bicarbonate (TEAB) and disrupted in lysis buffer (10 mM TRIS (pH = 8.8), 1% SDS, cComplete EDTA-free protease inhibitor cocktail) via freeze-thaw cycling and ultrasonic homogenizing (Sonopuls HD 2070, Bandelin, Berlin, Germany). For 1D-LC-MS analysis, cells were disrupted in 100 mM TEAB with cComplete EDTA-free inhibitor. The lysed cells were centrifuged (20 min, 21100 x g, 4 °C) and the protein concentration was determined by the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific, Bremen, Germany).

SDS-PAGE and Staining Methods: Aliquots of M. mazei proteome (200 µg) were separated in triplicates on discontinuous SDS-PAGE gels (4% stacking gel, 16% resolving gel, 0.1 cm thick) using Mini-PROTEAN Tetra Cell (Bio-Rad, Hercules, USA). Prior to loading, the samples were mixed with Laemmli buffer (62 mM TRIS- HCl (pH 6.8), 25% Glycerin (w/v), 5% β-mercaptoethanol (w/v), 2% SDS (w/v), 0.01% bromophenol blue (w/v)) and heated 5 min at 95 °C. Separation was performed with a TRIS-glycine running buffer (0.3% TRIS, 1.44% glycine, 0.1% SDS (w/v)) 15 min at 40 V and >80 min at 120 V until the running front reached the end of the gel.

After separation, for the non-stained condition the gel was incubated for 20 min in fixing solution (30% methanol, 10% acetic acid in MilliQ-water (w/v)). For Coomassie-staining, the gel was first incubated 20 min in fixing solution and then incubated 20 min in staining solution (0.2% (w/v) Coomassie R250, 0.005% (w/v) Coomassie G250, 40% methanol, 10% ethanol, 2% glycerin in MilliQ-water (w/v)). The gel was destained with 10% acetic acid, 5% ethanol in MilliQ-water (w/v) overnight until a clear background was observed. For negative-staining the gel was incubated in an imidazole solution (0.2 mM imidazole in MilliQ-water, 0.3% SDS (w/v)) and washed 1 min in MilliQ-water.[18] The gel was developed 1–2 min in 0.2 mM zinc sulfate until the background was white. In-gel digestion was performed directly after staining.

In-Gel Digestion: Eight bands below 30 kDa with approximately the same size were cut out. All following steps were performed at 20 °C in a thermomixer with 800 rpm unless otherwise stated. Negative-stained samples were washed 10 min with 150 µL of 0.1% acetic acid in 50% (v/v) ACN and the peptides were extracted 20 min with 60% ACN, 1% TFA (v/v) and heated 5 min at 95 °C. Separation was performed by mixing 300 µL washing solution with 300 µL methylene blue solution (0.025% methylene blue (w/v), 5% sodium sulfate (w/v), 1% H2SO4 (w/v) in MilliQ-water) and 1.2 mL chloroform. After vigorous shaking, it was incubated 10 min at room temperature and centrifuged with a table centrifuge. The lower chloroform-rich phase was transferred into a reaction vial containing 50–100 mg sodium sulfate. After mixing and centrifugation the absorption of the supernatant at 651 nm was determined.[12]

When SDS was no longer detectable in the washing solutions, the gel bands were washed 5 min in 400 µL 100% ACN and dried via vacuum evaporation on a Concentrator Plus (Eppendorf, Germany). For digestion, 5 µL of 20 ng µL−1 trypsin in 50 mM ABC (10 min, RT) and 80 µL digestion buffer (50 mM ABC, 3% ACN) was added to the gel bands and digested for 18 h at 37 °C. The supernatant was transferred into a new reaction vial and the peptides were extracted 20 min with 60% ACN, 1% TFA (w/v) and with 100% ACN. The peptides were concentrated to complete dryness, resuspended in 20 µL loading buffer (3% ACN, 0.1% TFA in MilliQ-water) and analyzed without further clean up.

In-Solution Digestion: Proteins in 100 µL 100% MEAB were reduced (10 mM DTT, 56 °C, 1 h), alkylated (55 mM IAA, 20 °C, 30 min in the dark) and digested with trypsin 1:50 trypsin-protein ratio (w/w) overnight at 37 °C. The peptides were cleaned using ZipTips (Thermo Scientific, PierceC18, 10 µL). Peptides were dried via vacuum evaporation and resuspended in loading buffer.
LC-MS/MS Analysis: Peptides were separated on a Dionex U3000 UHPLC system (Thermo, Dreieich, Germany) equipped with an Acclaim PepMap 100 column (2 μm, 75 μm × 500 mm) coupled online to a QExactive Plus Orbitrap MS (Thermo, Bremen, Germany). Separation was performed across a 2 h gradient: 4% B for 2 min, followed by a linear gradient from 4% to 20% B over 80 min, and a linear gradient to 40% B over 40 min, an 8 min linear increase to 90% B, and 10 min at 90% B. Inter-run equilibration of the column: 10 min at 4% Eluent B. Eluent A was 0.05% formic acid (FA) and eluent B 80% ACN, 0.04% FA. After in-gel digestion 2 μL of the peptides were injected. For the 1D-LC analysis 1 μg of the in-solution digested sample was injected.

A full-scan MS was acquired (resolution 70000, automatic gain control (AGC) target 3e6, max injection time (IT) 50 ms), with subsequent MS/MS acquisition (resolution 17500, isolation window 1.4 m/z, AGC target 1e5, max IT 100 ms) of the top 15 most intense ions (exclusion of +1 and >+6 ions). Higher-energy collisional dissociation (HCD) was performed with a normalized collision energy (NCE) of 27.5. Dynamic exclusion (30 s duration) and lock mass at 445.12003 m/z were enabled.

MS-data were processed using the Proteome Discoverer software package (V2.2; Thermo, Germany). The files were searched individually using SequestHT algorithm node against a protein database containing genome-derived proteins of M. mazei strain G01 (Uniprot proteome ID UP000000595; 12.09.2019), a list of predicted SEP and the cRAP list of genome-derived proteins of M. mazei strain Gö1 (Uniprot proteome ID M. mazei/1z). The files were searched individually and lock mass at 445.12003 m/z ions). Higher-energy collisional dissociation (HCD) was performed with a normalized collision energy (NCE) of 27.5. Dynamic exclusion (30 s duration) and lock mass at 445.12003 m/z were enabled.

Protein identification criteria i) strict: at least two high confidence (FDR <1%) peptides, where one was proteotypic; ii) ion series-based: at least one high confidence unique peptide with at least five consecutive b- or y-ions in the MS2 spectrum;[19] iii) peptide-centric validation: SEP identified under ion series-based criteria but with only one peptide-spectral match[17] were further analyzed with a peptide-centric approach.[11] For this, the raw data were converted into mgf-files and the same search parameters as for SequestHT database search were used. For p-value calculation, the default number of 1000 random peptides and as scoring algorithm Hyperscore was used.

All LC-MS data was deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXDD18269.[13]

Supporting Information
Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements
The present study was supported by the Deutsche Forschungsgemeinschaft (DFG) within the priority program SPP2002, project Z1 (TH872/10-1) and SCHM1052/20-1.

Open access funding enabled and organized by Projekt DEAL

Conflict of Interest
The authors declare no conflict of interest.

Keywords
alternative open reading frames, GeLC-MS, microproteins, peptidomics, short or small open reading frames

[1] M. E. Dinger, K. C. Pang, T. R. Mercer, J. S. Mattick, PLoS Comput. Biol. 2008, 4, e1000176.
[2] S. J. Andrews, J. A. Rothnagel, Nat. Rev. Gen. 2014, 15, 193.
[3] L. Cassiday, P. T. Kaulich, A. Tholey, J. Proteome Res. 2016, 15, 3773.
[4] J.-P. Couso, P. Patraquim, Nat. Rev. Mol. Cell Biol. 2017, 18, 575.
[5] M. W. Orr, Y. Mao, G. Storz, S.-B. Qian, Nucleic Acids Res. 2019, 3, 1029.
[6] G. Storz, Y. I. Wolf, K. S. Ramamurthi, Annu. Rev. Biochem. 2014, 83, 753.
[7] M. I. Galindo, J. I. Pueyo, S. Fouix, S. A. Bishop, J. P. Couso, PLoS Biol. 2007, 5, e106.
[8] Q. Chu, J. Ma, A. Saghatelian, Crit. Rev. Biochem. Mol. 2015, 50, 134.
[9] D. W. Greening, R. J. Simpson, J. Proteomics 2010, 73, 637.
[10] L. Cassidy, D. Prasse, D. Linke, R. A. Schmitz, A. Tholey, J. Proteome Res. 2016, 15, 3773.
[11] D. J. Harney, A. T. Hutchison, Z. Su, M. Hatchwell, L. K. Heilbronn, S. Hocking, D. E. James, M. Larance, Mol. Cell Proteomics 2019, 18, 1899.
[12] H. Peteruschke, J. Anders, P. F. Stadler, N. Jehmlich, M. von Bergen, J. Proteomics 2020, 213, 103604.
[13] L. Cassidy, P. T. Kaulich, A. Tholey, J. Proteome Res. 2016, 15, 3773.
[14] T. Cardon, F. Hervé, D. Delcourt, Y. Roucou, M. Salzet, J. Franck, I. Fournier, Anal. Chem. 2020, 92, 1122.
[15] J. Ma, J. K. Dieckrich, J. Jungreis, C. Donaldson, J. Vaughan, M. Kellis, R. Yates, A. Saghatelian, Anal. Chem. 2016, 88, 3967.
[16] X. Gao, A. Khitan, Z. Na, D. C. Dumitrescu, M. Kubica, E. Olatunji, S. A. Slavoff, J. Proteome Res. 2020, https://doi.org/10.1021/acs.jproteome.0c00254.
[17] L. Castellanos-Serra, E. Hardy, Nat. Protoc. 2006, 1, 1544.
[18] U. Deppermeier, A. Johann, T. Hartsch, R. Merki, R. A. Schmitz, R. Martinez-Arias, A. Henne, A. Wieser, S. Bäumer, C. Jacobi, H. Brügge mann, T. Lienard, A. Christmann, M. Bömeke, S. Steckel, A. Bhattacharya, A. Lykidis, R. Overbeek, H.-P. Klenk, R. P. Gunsalus, H.-J. Fritz, G. Gottschalk, J. Mol. Microbiol. Biotechnol. 2002, 4, 453.
[27] C. Klein, M. Aivaliotis, J. V. Olsen, M. Falb, H. Besir, B. Scheffer, B. Bisle, A. Tebbe, K. Konstantinidis, F. Siedler, F. Pfeiffer, M. Mann, D. Oesterhelt, J. Proteome Res. 2007, 6, 1510.

[28] S. A. Slavoff, A. J. Mitchell, A. G. Schwaid, M. N. Cabili, J. Ma, J. Z. Levin, A. D. Karger, B. A. Budnik, J. L. Rinn, A. Saghatelian, Nat. Chem. Biol. 2013, 9, 59.

[29] M. R. de Moreno, J. F. Smith, R. V. Smith, Anal. Biochem. 1985, 151, 466.

[30] L. M. Iakoucheva, A. L. Kimzey, C. D. Masselon, R. D. Smith, A. K. Dunker, E. J. Ackerman, Protein Sci. 2001, 10, 1353.

[31] B. Wen, X. Wang, B. Zhang, Gen. Res. 2019, 29, 485.

[32] M. Arand, T. Friedberg, F. Oesch, Anal. Biochem. 1992, 207, 73.

[33] J. A. Vizcaino, E. W. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Ríos, J. A. Dianes, Z. Sun, T. Farrah, N. Bandeira, P.-A. Binz, I. Xenarios, M. Eisenacher, G. Mayer, L. Gatto, A. Campos, R. J. Chalkley, H.-J. Kraus, J. P. Albar, S. Martinez-Bartolomé, R. Apweiler, C. S. Omenn, L. Martens, A. R. Jones, H. Hermjakob, Nat. Biotechnol. 2014, 32, 223.