Recent advances in mass spectrometry–based peptidomics workflows to identify short-open-reading-frame-encoded peptides and explore their functions

Bertrand Fabre, Jean-Philippe Combier, Serge Plaza

Recent advances in mass spectrometry-based peptidomics workflows to identify short open reading frame encoded peptides and explore their functions

Bertrand Fabre[1], Jean-Philippe Combier[1] and Serge Plaza[1]

[1] Laboratoire de Recherche en Sciences Végétales, UMR5546, Université de Toulouse, UPS, CNRS, 31320 Auzeville-Tolosane, France

Corresponding author: Fabre, Bertrand (bertrand.fabre@lrsv.ups-tlse.fr)

Abstract

Short open reading frame (sORF)-encoded polypeptides (SEPs) have recently emerged as key regulators of major cellular processes. Computational methods for the annotation of sORFs combined with transcriptomics and ribosome profiling approaches predicted the existence of tens of thousands of SEPs across the kingdom of life. Though, we still lack unambiguous evidence for most of them. The method of choice to validate the expression of SEPs is mass spectrometry (MS)-based peptidomics. Peptides are less abundant than proteins which tends to hinder their detection. Therefore, optimization and enrichment methods are necessary to validate the existence of SEPs. In the present review, we discuss the challenges for the detection of SEPs by MS and recent developments of biochemical approaches applied to the study of these peptides. We detail the advances made in the different key steps of a typical peptidomics workflow and highlight possible alternatives that have not been explored yet.

Keywords: Short open reading frame-encoded polypeptide (SEP), Microprotein, Peptidomics, Mass spectrometry.

Introduction

Genome sequencing has revolutionized our understanding of physiological and pathological states of living organisms [1]. Two decades later, it becomes obvious that many genes predicted to be non-coding (open reading frames with less than 100 codons) actually encode small polypeptide sequences. These peptides, called short open reading frame (sORF)-encoded-peptides (SEPs) or microproteins, are produced from a variety of RNAs [2]. SEPs are biologically active molecules which seem to be involved in a wide range of biological functions. They were identified as major regulators of several key cellular pathways across the kingdom of life. Examples of active SEPs have been discussed in details in previous review articles [2–5]. Due to their large spectrum of functions, SEPs represent attractive new tools for drug development and agronomical applications [6,7].

Thanks to computational methods based on prediction and conservation for the annotation of sORFs, transcriptomics and ribosome profiling approaches, several thousands of SEPs have been predicted among several species [4]. However, we still lack evidence for the existence of most them, even though hundreds of SEPs have been validated [4]. The roles of fewer SEPs, less than 50 across all species, have been characterized [4]. SEPs still represent a largely unexplored repertoire of active biomolecules with possible important roles in the animal and plant kingdoms.

Mass spectrometry (MS) is particularly adapted to validate the expression of peptides and proteins [8]. In addition to the validation of existence of SEPs, MS can also be used to get insights into the function of these peptides through the identification of their interacting proteins (interactome) [9]. Most of the SEPs described in the literature carry out their functions by affecting proteins activity via direct binding [5]. Thus, MS-based peptidomics has the potential to explore the full spectrum of SEPs in different organisms as well as unravelling their functions. Nonetheless, despite a growing efforts of the scientific community, the number of SEPs validated is still limited and the functions of most of them remain unexplored so far [2]. This is probably explained by technical challenges, and unlike proteomics, peptidomics is still a very recent area of research. The analytical workflow to analyze peptides and proteins differs, and peptidomics based workflows require additional optimization [10,11].

These last few years, several groups around the world have dedicated their efforts to the identification of SEPs across several model organisms. In the present review, we discuss the recent development of MS and biochemical approaches applied to the study of SEPs. We detail the advances made in the different key steps of peptidomics workflow and highlight possible alternatives that have not been explored yet. A table summarizing the different parameters used in recent successful attempts to identify SEPs by MS is provided here and will be useful to compare the different approaches implemented in these studies (Table 1).

Methods to extract and enrich SEPs

One of the key steps for the identification of SEPs using MS based workflows is the extraction of these peptides from biological material (Figure 1). SEPs are generally low abundant compared to proteins, the later hindering the detection of the former in MS analysis if no prior SEPs specific enrichment is used. In addition, the release of peptidases during cell lysis is believed to degrade free peptides. Finally, peptides resulting from proteins degradation, either endogenous, produced by proteases such as the proteasome, or artificial, due to proteases during the lysis step, might also hinder the detection of SEPs by MS. To circumvent these problems, several approaches have been used. First, to limit peptide degradation due to intracellular peptidases, samples are often boiled in water/lysis buffer to denatured protein structures, thus abolishing peptidase and protease activity [10–12] (Table 1 and Figure 1). An alternative method to preserve the integrity of SEPs is to perform protein precipitation using TCA or methanol/chloroform which also inactivates peptidase and protease [11,13] (Table 1 and Figure 1). A combination of both methods was used in several studies [10,13].

Once, SEPs have been stabilized in the lysate following inactivation of peptidases/proteases, it is essential to perform a separation of the peptides from proteins to maximize SEPs identification. The separation methods are based on the chemical properties of the peptides. An example of separation is the use of ultrafiltration to segregate proteins and peptides based on molecular weight cut-off, typically 10 or 30 kDa [10,13–15] (Table 1 and Figure 1). Proteins are then retained by the filter while SEPs, being smaller, are found in the flow-through. An alternative physical separation method is SDS-PAGE (Sodium Dodecyl Sulfate-PolyAcrylamide Gel Electrophoresis) in which the gel is cut after protein separation to analyze only the low molecular weight fractions [11,15] (Table 1 and Figure 1). Another type of SEPs enrichment can be performed based on hydrophobicity using C8 reverse phase which was shown to recover a large number of SEPs in human cells (Table 1 and Figure 1) [10]. However, each extraction and enrichment method has its own set of limitations (e.g. loss of peptides by adsorption on ultrafiltration device filters

or loss of membrane-associated peptides if a lysis buffer without detergents is used for SEPs extraction) and must be taken into account when designing the analytical workflow to identify SEPs.

Another parameter to consider is whether or not to use SEPs enzymatic digestion (Figure 1). Trypsin was used as the enzyme of choice in most studies focusing on SEPs and in which enzymatic digestion was performed [10,11] (Table 1). The size of SEPs being very variable (from few to one hundred amino acids), peptides resulting from the digestion might fit into the observable mass-to-charge ratio (m/z) range from the MS analysis or be too short and missed. Using different enzymes (e.g. Lys-C) to digest peptides was shown to increase the fraction of SEPs identified in a sample compared to trypsin [16].

Overall, the combination of the SEPs extraction method, the enrichment approach and the implementation, or not, of enzyme digestion will allow the identification of a certain set of SEPs. Indeed, depending on their chemical properties and/or their amino acids lengths particular sets of SEPs were identified depending on the workflow used to isolate SEPs [10,11]. Interestingly, recent studies identified a surprisingly high number of SEPs (more than 200 non canonical peptides) in preparation of HLA-I bound antigenic peptides, suggesting that SEPs are often presented by HLA molecules and that this kind of strategy can be used to identify new SEPs in mammalian cells [9,17,18] (Table 1).

In sum, as a general rule of thumb, to maximize the number of SEPs identified, it is preferable to combine several of the previously described approaches [10,11] (Table 1). Though, if one is interested in the identification of a particular SEP, it is probable that a specific combination of extraction and enrichment methods and the use of protein digestion or not, will result in an optimal detection for that peptide. The workflow would also have to be adjusted based on the subcellular localization of the SEPs investigated (e.g. mitochondrial, membrane associated or secreted SEPs) [19,20].

Mass spectrometry analysis applied to SEPs identification

Regarding MS analysis, it is possible to tweak several acquisition parameters to optimize the identification of SEPs. First, using state of the art mass-spectrometers for the MS analysis with increased sensitivity and scan speed favors deeper coverage of the samples analyzed and thus identification of more candidates. Typically, data dependent acquisition (DDA) is used as the method of choice (Table 1). By combining DDA MS analysis with specific extraction/enrichment methods, Ma et al. identified more than one hundred of SEPs in human cell lines [10] (Table 1). Leaning towards sensitive MS settings and using HCD (Higher-energy Collisional Dissociation) as the fragmentation method proved to result in the recovery of more fragment ions and slightly increased the number of SEPs identified [10]. Overall, optimizing MS parameters results in better quality MS2 spectra for SEPs thus increasing the confidence in their identification. Nonetheless, it would be interesting to test the potential of complementary techniques such as ion mobility [21,22] to identify SEPs. This method, based on the separation of ionized molecules in a gas phase through application of an electric field, would provide another layer of separation of peptides [21,22], and thus theoretically increase the number of SEPs identified per MS analysis.

Although DDA has been used as the primary acquisition mode to identify SEPs (Table 1), alternative MS methods would provide some advantages. Targeted peptidomics approaches, such as selected reaction monitoring (SRM) or parallel reaction monitoring (PRM), were used in some studies to validate SEPs identified by DDA [15,23] (Figure 1). Due to the sensitivity and specificity of targeted peptidomics, these

methods have a high potential for SEPs identification. However, it should be noted that designing SRM/PRM assay requires optimization of the MS method for each peptide, which is time consuming (especially as the number of peptides followed increases) [24]. So far, studies on SEPs have mainly focused on the identification of these peptides and very few articles provide quantitative data on changes of SEPs expression in different biological conditions [25,26]. Approaches such as data independent acquisition (DIA) would be particularly interesting to follow changes of expression of SEPs in different biological contexts (Figure 1). Indeed, DIA was shown to provide more complete and reproducible quantitative data compared to DDA, still providing a good depth in the peptidome/proteome [27,28].

Bioinformatics analysis of mass spectrometry data

A critical step to identify SEPs by MS is the bioinformatics analysis of the raw data generated from the MS analysis. In a typical MS analysis, the experimental precursors and fragments m/z values, measured from MS and MS2 spectra respectively, are compared to *in silico* generated databases containing theoretical precursors and fragments m/z values for the proteins/peptides of interest [21]. Using this type of approach, it is possible to identify only peptides/proteins present in the database. There are different possibilities for the creation of databases which include predicted SEPs. It can be generated from RNAseq data from the same organism/cell line of interest [12,29] (Table 1 and Figure 1). Alternatively, it can be designed from ribosome profiling data [23,30] (Table 1 and Figure 1). Finally, another possibility is to perform a six frames *in silico* translation of the entire genome of the organism of interest from stop codon to stop codon [13] (Table 1 and Figure 1). Using this latest strategy, Wang *et al.* recently identified 1,900 and 1,800 non-conventional peptides from Maize and *A. thaliana* samples, respectively [13] (Table 1). Interestingly, most of these peptides were mapped to intergenic regions, illustrating the fact that there are still many genes to discover. The main limitations of database searching in this context is that the genome of the organism of interest must be well sequenced.

It is worth noticing that several online repositories have been developed to gather SEPs prediction and identification in several organisms. SmProt [31], ARA-PEPs [32], PsORF [33], sORFs.org [34] and Openprot [35] provide sequences for predicted and/or validated SEPs (by re-analyzing previously published MS and/or ribosome profiling datasets) that can be downloaded and used as the database for the searching step to analyze MS data data (Table 1). The databases provided by sORF and Openprot are frequently updated when new MS and/or ribosome profiling data become available.

When looking for SEPs in MS data using database searching, it is important to remember that the number of sequences in the database is generally larger than the ones used for typical proteomics data analysis. For example, a database based on six frames *in silico* translation of the entire human genome can be around 70 times bigger than the corresponding Ensembl reference protein database [36]. Besides the increase in computational time, the size of such databases can results in a larger number of false identifications [36]. It is also necessary to include canonical proteins and common contaminants (e.g. human keratins, trypsin) in the database to discard any hit matching both a SEP and one of the former categories of proteins. Thus, special care must be taken when searching MS data using customized databases to identify SEPs and false discovery rates have to be adapted accordingly [35,36].

An alternative strategy to database searching to analyze MS raw data would consist in a peptide *de novo* sequencing approach (Figure 1). *De novo* sequencing is based on the assignment of fragment ions from a

mass spectrum [37]. This kind of analysis is dependent on a good coverage of the peptide sequence from the MS2 spectra. Although, search of MS raw data with the help of a database was successfully used in most studies focusing on the identification of SEPs, it would be interesting to see how *de novo* search would perform in comparison to database searching. Search algorithms have been designed to integrate database into de novo sequencing which would represent a very promising approach to identify new SEPs [37].

Given the high number of proteomics datasets that have been generated over recent years [38], reanalysis of these data can represent an important source of identification of SEPs. Data repository such as ProteomeXchange [38] provides a great opportunity to mine for SEPs and has been used in that sense in several recent studies [9,19,39]. Projects such as sORFs.org [34] and Openprot [35] take full advantage of data repositories and continuously search for new SEPs in several species from ProteomeXchange uploaded datasets.

Validation of mass spectrometry data

When SEPs are identified, several methods can be used to validate the results from the MS analysis (Figure 1). One possibility is the use of synthetic peptides (either isotopically labelled or not) to compare their retention times and fragmentation profiles in the DDA MS2 spectra to the corresponding ones measured for SEPs identification [12,13] (Table 1). The order of the fragment in the MS2 spectra and the measured retention time observed for the SEPs should be similar for the endogenous and synthetic versions of the peptides. The validation using synthetic peptides can be coupled with SRM/PRM approaches to increase the signal to noise ratio as these MS methods are more sensitive than DDA mode [21].

Another option is to produce an antibody directed against the SEPs identified and perform western blot or immunofluorescence analysis [40] (Table 1). However, depending on the number of SEPs to validate, this strategy might be expensive and time consuming. Alternatively, addition of a tag to an endogenous SEP can be performed using knock-in techniques (e.g. CRISPR) and followed by immune-detection to confirm the expression of the peptide [41,42]. Knock-out/knock-down methods (e.g. CRISPR or RNA interference), coupled with western blot analysis were also successfully used to validate the existence of SEPs [19,41,43]. An advantage of these different genetic approaches compared to antibodies directed against endogenous SEPs or synthetic peptides is that they provide clear evidence that a given peptide is produced from a putative sORF.

Affinity purification coupled to mass spectrometry to identify SEPs associated proteins

Once SEPs have been identified, the next step would be to understand their role in the organism/cell line of interest. In order to explore the function of SEPs at the molecular level, one way is to determine what molecules these SEPs interact with. So far, for all the SEPs for which the functions were unraveled were shown to interact with proteins [5]. Alternatively, it was proposed that some SEPs could be produced but remained not functional [44]. Determining what are the interacting proteins would bring crucial information regarding the molecular functions underlined by SEPs. Affinity purification coupled to mass spectrometry (AP-MS) is the method of choice to identify the proteins interacting with a peptide/protein of interest in an unbiased manner [45]. In this kind of approach, SEPs are tagged, either through CRISPR

or overexpression in the organism or cell line of interest, or by producing synthetic versions of the peptides [9,19]. In the case of the purification of overexpressed SEPs, the organism/cells are lysed and the associated proteins are co-purified with the tagged SEP. When synthetic versions are used, the tagged SEPs are added directly into the cell lysate so that they can bind their partners.

A recent study by Chen *et al*. identified new SEPs by MS and further characterized the function of 16 of them [9]. They found interacting partners for 12 of them. Further immunofluorescence-based experiments confirmed that the SEPs co-localized with the proteins they interact with. Surprisingly, when they investigated the interactome of SEPs produced from sORF, they observed, that for 5 out of 10 peptides tested, these SEPs interact with their downstream canonical coding proteins. This suggests that SEPs have a wide range of functions and that, as previously suggested by Roucou and co-workers [46], some peptides produced from sORFs (or alternative ORF) might regulate their canonical associated protein.

In another recent article, Zhang *et al.* explored human mitochondrial SEPs functions [19]. Reanalyzing previously published MS data, they identified 22 SEPs and confirmed that 20 of them, including 16 newly discovered, were localized to mitochondria (Table 1). They further explored the function of a SEP they named BRAWNIN and performed AP-MS to identify its associated partners. They showed that BRAWNIN interacts with the electron transport chain complex III and knocking-out this SEP results in a down-regulation of several subunits of this protein complex. This suggests that the electron transport chain complex III stability and/or assembly is compromised in the absence of BRAWNIN. Finally, they showed that knocking-out of BRAWNIN in zebrafish causes lethal mitochondrial deficiency, highlighting the essential role of this SEP.

An interesting approach to identify SEPs and their associated proteins has been proposed by Gao *et al*. [47]. They implemented a bioinformatics workflow to search SEPs (or any peptide sequence) in datasets from large-scale interactome from human cell lines. They were able to identify 120 SEPs in these datasets with a searching speed increased 200 times compared to usual bioinformatics workflows. Because the analysis is done on interactome data, in addition to the identification of the SEPs, the workflow designed by Gao *et al.* provides a list of the potential interactors of the SEPs. They validated the interactors of two SEPs by generating tagged versions of these peptides, which were overexpressed in human cells, and affinity purified followed by western blot analysis.

Conclusion

In recent years, an important step forward has been made in the discovery of SEPs. Through the development of ribosome profiling and peptidomics, combined with more sophisticated bioinformatics analytical workflow, it is now clear that small proteins encoded by short ORFs exist in a wide range of species across life kingdom. Moreover, several hundreds of SEPs have been identified and the exploration of their functions seems to point toward important roles for some of them in key cellular pathways. This hidden part of the proteome can no longer be ignored and has thus attracted the attention of the scientific community. As demonstrated in this review, there is a strong effort of several groups to develop specific biochemical and bioinformatics tools to study these peptides. Although the analytical workflow differs from one study to another, some elements are often found in common (such as the use of DDA as the preferred MS acquisition mode or the use of enzymatic digestion with trypsin) (Table 1). Nevertheless, as

discussed above, different SEPs extraction and enrichment methods will result in the identification of different sets of peptides according to their biochemical properties and/or subcellular localizations. Thus, the more diverse the analytical workflows, the higher the number of SEPs identified. The multiplication of new MS-based peptidomics studies focusing on the identification SEPs will then certainly help raising up the number of peptides validated. Following to the validation of the existence of microproteins, further functional studies are necessary to better understand the function of each of these peptides to help access this repertoire of active biomolecules for agronomical or medicinal applications.

Acknowledgements

This work has been supported by the Fondation ARC pour la recherche sur le cancer.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest

* * of outstanding interest

1. Levy SE, Myers RM: **Advancements in Next-Generation Sequencing**. *Annu Rev Genomics Hum Genet* 2016, **17**:95–115.

2. Plaza S, Menschaert G, Payre F: **In search of lost small peptides**. *Annu Rev Cell Dev Biol* 2017, **33**:391–416.

3. Yeasmin F, Yada T, Akimitsu N: **Micropeptides encoded in transcripts previously identified as long noncoding RNAs: A new chapter in transcriptomics and proteomics**. *Front Genet* 2018, **9**:1–10.

4. Pueyo JI, Magny EG, Couso JP: **New Peptides Under the s(ORF)ace of the Genome**. *Trends Biochem Sci* 2016, **41**:665–678.

5. Saghatelian A, Couso JP: **Discovery and characterization of smORF-encoded bioactive polypeptides**. *Nat Chem Biol* 2015, **11**:909–916.

6. Crook ZR, Nairn NW, Olson JM: **Miniproteins as a Powerful Modality in Drug Development**. *Trends Biochem Sci* 2020, **45**:332–346.

7. Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, Nakaminami K, Nishi R, Ohashi C, Iida K, Tanaka M, et al.: **Small open reading frames associated with morphogenesis are hidden in plant genomes**. *Proc Natl Acad Sci U S A* 2013, **110**:2395–2400.

8. Chu Q, Ma J, Saghatelian A: **Identification and characterization of sORF-encoded polypeptides**. *Crit Rev Biochem Mol Biol* 2015, **50**:134–141.

9. Chen J, Brunner AD, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, et al.: **Pervasive functional translation of noncanonical human open reading frames**. *Science (80-)* 2020, **367**:140–146.

* * First article reporting the identification of a large number of SEPs from HLA peptidomics data. Genome-scale CRISPR screens and interactome data reveal that bicistronic mRNAs encode uORF peptides that function in trans of the canonical protein

10. Ma J, Diedrich JK, Jungreis I, Donaldson C, Vaughan J, Kellis M, Yates JR, Saghatelian A: **Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides**. *Anal Chem* 2016, **88**:3967–3975.

* Deep comparison of different MS-based peptidomics workflows for the identification of SEPs

11. Cardon T, Hervé F, Delcourt V, Roucou X, Salzet M, Franck J, Fournier I: **Optimized Sample Preparation Workflow for Improved Identification of Ghost Proteins**. *Anal Chem* 2020, **92**:1122–1129.

* Deep comparison of different MS-based peptidomics workflows for the identification of SEPs

12. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A: **Peptidomic discovery of short open reading frame–encoded peptides in human cells**. *Nat Chem Biol* 2013, **9**:59–64.

13. Wang S, Tian L, Liu H, Li X, Zhang J, Chen X, Jia X, Zheng X, Wu S, Chen Y, et al.: **Large-Scale Discovery of Non-conventional Peptides in Maize and Arabidopsis through an Integrated Peptidogenomic Pipeline**. *Mol Plant* 2020, **13**:1078–1093.

14. Tharakan R, Kreimer S, Ubaida-Mohien C, Lavoie J, Olexiouk V, Menschaert G, Ingolia NT, Cole RN, Ishizuka K, Sawa A, et al.: **A methodology for discovering novel brain-relevant peptides: Combination of ribosome profiling and peptidomics**. *Neurosci Res* 2020, **151**:31–37.

15. Ma J, Ward CC, Jungreis I, Slavoff SA, Schwaid AG, Neveu J, Budnik BA, Kellis M, Saghatelian A: **Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue**. *J Proteome Res* 2014, **13**:1757–1765.

16. Bartel J, Varadarajan AR, Sura T, Ahrens CH, Maaß S, Becher D: **Optimized Proteomics Workflow for the Detection of Small Proteins**. *J Proteome Res* 2020, doi:10.1021/acs.jproteome.0c00286.

17. Martinez TF, Chu Q, Donaldson C, Tan D, Shokhirev MN, Saghatelian A: **Accurate annotation of human protein-coding small open reading frames**. *Nat Chem Biol* 2020, **16**:458–468.

* Accurate annotation of sORFs from ribosome profiling data and identification of a large number of the corresponding SEPs from HLA peptidomics data

18. Chong C, Müller M, Pak HS, Harnett D, Huber F, Grun D, Leleu M, Auger A, Arnaud M, Stevenson BJ, et al.: **Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes**. *Nat Commun* 2020, **11**.

* Identification of a large number of SEPs predicted from ribosome profiling in HLA peptidomics data. This study shows that human tumors present specific SEPs on their HLA molecules

19. Zhang S, Reljić B, Liang C, Kerouanton B, Francisco JC, Peh JH, Mary C, Jagannathan NS, Olexiouk V, Tang C, et al.: **Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly**. *Nat Commun* 2020, **11**.

* Identification of mitochondrial SEPs. One of them, called BRAWNIN, is required for respiratory complex III assembly and activity

20. Makarewich CA: **The hidden world of membrane microproteins**. *Exp Cell Res* 2020, **388**:111853.

21. Dupree E, Jayathirtha M, Yorkey H, Mihasan M, Petre BA, Darie C: **A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field**. *Proteomes* 2020, **8**:14.

22. Meier F, Brunner AD, Koch S, Koch H, Lubeck M, Krause M, Goedecke N, Decker J, Kosinski T, Park MA, et al.: **Online parallel accumulation–serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer**. *Mol Cell Proteomics* 2018, **17**:2534–2545.

23. van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, Kirchner M, Maatz H, Blachut S, Sandmann CL, et al.: **The Translational Landscape of the Human Heart**. *Cell* 2019, **178**:242-260.e29.

* * Detection of SEPs produced from lncRNAs and circRNAs in human hearts *in vivo*. Some of these SEPs are localized in mitochondria

24. Lange V, Picotti P, Domon B, Aebersold R: **Selected reaction monitoring for quantitative proteomics: A tutorial**. *Mol Syst Biol* 2008, **4**.

25. Cao X, Khitun A, Na Z, Dumitrescu DG, Kubica M, Olatunji E, Slavoff SA: **Comparative Proteomic Profiling of Unannotated Microproteins and Alternative Proteins in Human Cell Lines**. *J Proteome Res* 2020, **19**:3418–3426.

26. Casas-Vila N, Bluhm A, Sayols S, Dinges N, Dejung M, Altenhein T, Kappei D, Altenhein B, Roignant J-Y, Butter F: **The developmental proteome ofDrosophila melanogaster.** *Genome Res* 2017, **27**:1273–1285.

27. Bruderer R, Bernhardt OM, Gandhi T, Xuan Y, Sondermann J, Schmidt M, Gomez-Varela D, Reiter L: **Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results**. *Mol Cell Proteomics* 2017, **16**:2296–2309.

28. Fabre B, Korona D, Mata CI, Parsons HT, Deery MJ, Hertog MLATM, Nicolaï BM, Russell S, Lilley KS: **Spectral Libraries for SWATH-MS Assays for Drosophila melanogaster and Solanum lycopersicum**. *Proteomics* 2017, **17**.

29. Ma J, Saghatelian A, Shokhirev MN: **The influence of transcript assembly on the proteogenomics discovery of microproteins**. *PLoS One* 2018, **13**:1–19.

30. Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M, Couso JP: **Extensive translation of small open reading frames revealed by poly-ribo-seq**. *Elife* 2014, **3**:1–19.

31. Hao Y, Zhang L, Niu Y, Cai T, Luo J, He S, Zhang B, Zhang D, Qin Y, Yang F, et al.: **SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci**. *Brief Bioinform* 2018, **19**:636–643.

32. Hazarika RR, De Coninck B, Yamamoto LR, Martin LR, Cammue BPA, Van Noort V: **ARA-PEPs: A repository of putative SORF-encoded peptides in Arabidopsis thaliana**. *BMC Bioinformatics* 2017, **18**:1–9.

33. Chen Y, Li D, Fan W, Zheng X, Zhou Y, Ye H, Liang X, Du W, Zhou Y, Wang K: **PsORF: a database of small ORFs in plants**. *Plant Biotechnol J* 2020, doi:10.1111/pbi.13389.

34. Olexiouk V, Van Criekinge W, Menschaert G: **An update on sORFs.org: A repository of small**

ORFs identified by ribosome profiling. *Nucleic Acids Res* 2018, **46**:D497–D502.

35. Brunet MA, Brunelle M, Lucier JF, Delcourt V, Levesque M, Grenier F, Samandi S, Leblanc S, Aguilar JD, Dufour P, et al.: **OpenProt: A more comprehensive guide to explore eukaryotic coding potential and proteomes**. *Nucleic Acids Res* 2019, **47**:D403–D410.

36. Nesvizhskii AI: **Proteogenomics: concepts, applications and computational strategies**. *Nat Methods* 2014, **11**:1114–1125.

37. Allmer J: **Algorithms for the de novo sequencing of peptides from tandem mass spectra**. *Expert Rev Proteomics* 2011, **8**:645–657.

38. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M, et al.: **The PRIDE database and related tools and resources in 2019: Improving support for quantification data**. *Nucleic Acids Res* 2019, **47**:D442–D450.

39. Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, Mastrobuoni G, Rajewsky N, Kempa S, Selbach M, et al.: **Extensive identification and analysis of conserved small ORFs in animals**. *Genome Biol* 2015, **16**:1–21.

40. Lauressergues D, Couzigou JM, San Clemente H, Martinez Y, Dunand C, Bécard G, Combier JP: **Primary transcripts of microRNAs encode regulatory peptides**. *Nature* 2015, **520**:90–93.

41. Matsumoto A, Pasut A, Matsumoto M, Yamashita R, Fung J, Monteleone E, Saghatelian A, Nakayama KI, Clohessy JG, Pandolfi PP: **MTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide**. *Nature* 2017, **541**:228–232.

42. Jackson R, Kroehling L, Khitun A, Bailis W, Jarret A, York AG, Khan OM, Brewer JR, Skadow MH, Duizer C, et al.: **The translation of non-canonical open reading frames controls mucosal immunity**. *Nature* 2018, **564**:434–438.

43. Na Z, Luo Y, Schofield JA, Smelyansky S, Khitun A, Muthukumar S, Valkov E, Simon MD, Slavoff SA: **The NBDY Microprotein Regulates Cellular RNA Decapping**. *Biochemistry* 2020, **59**:4131–4142.

44. Patraquim P, Mumtaz MAS, Pueyo JI, Aspden JL, Couso JP: **Developmental regulation of canonical and small ORF translation from mRNAs**. *Genome Biol* 2020, **21**:1–26.

45. Brunet MA, Leblanc S, Roucou X: **Reconsidering proteomic diversity with functional investigation of small ORFs and alternative ORFs**. *Exp Cell Res* 2020, **393**:112057.

46. Samandi S, Roy A V., Delcourt V, Lucier JF, Gagnon J, Beaudoin MC, Vanderperre B, Breton MA, Motard J, Jacques JF, et al.: **Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins**. *Elife* 2017, **6**:1–32.

47. Gao Y, Ma J, Saghatelian A, Yates J: **Targeted Searches for Novel Peptides in Big Mass Spectrometry Data Sets**. 2017, doi:10.1101/239863.

48. Cassidy L, Prasse D, Linke D, Schmitz RA, Tholey A: **Combination of Bottom-up 2D-LC-MS and Semi-top-down GelFree-LC-MS Enhances Coverage of Proteome and Low Molecular Weight Short Open Reading Frame Encoded Peptides of the Archaeon Methanosarcina mazei**. *J Proteome Res* 2016, **15**:3773–3783.

49. Cassidy L, Kaulich PT, Tholey A: **Depletion of High-Molecular-Mass Proteins for the Identification of Small Proteins and Short Open Reading Frame Encoded Peptides in Cellular Proteomes**. *J*

Proteome Res 2019, **18**:1725–1734.

50. Kaulich PT, Cassidy L, Weidenbach K, Schmitz RA, Tholey A: **Complementarity of Different SDS-PAGE Gel Staining Methods for the Identification of Short Open Reading Frame-Encoded Peptides**. *Proteomics* 2020, doi:10.1002/pmic.202000084.

51. He C, Jia C, Zhang Y, Xu P: **Enrichment-Based Proteogenomics Identifies Microproteins, Missing Proteins, and Novel smORFs in Saccharomyces cerevisiae**. *J Proteome Res* 2018, **17**:2335–2344.

52. Cardon T, Franck J, Coyaud E, Laurent EMN, Damato M, Maffia M, Vergara D, Fournier I, Salzet M: **Alternative proteins are functional regulators in cell reprogramming by PKA activation**. *Nucleic Acids Res* 2020, **48**:7864–7882.

53. Na CH, Sharma N, MadugUndu AK, Chen R, Aksit MA, Rosson GD, Cutting GR, Pandey A: **Integrated transcriptomic and proteomic analysis of human eccrine sweat glands identifies missing and novel proteins**. *Mol Cell Proteomics* 2019, **18**:1382–1395.

54. Li N, Zhou Y, Wang J, Niu L, Zhang Q, Sun L, Ding X, Guo X, Xie Z, Zhu N, et al.: **Sequential Precipitation and Delipidation Enables Efficient Enrichment of Low-Molecular Weight Proteins and Peptides from Human Plasma**. *J Proteome Res* 2020, **19**:3340–3351.

55. Wang B, Hao J, Pan N, Wang Z, Chen Y, Wan C: **Identification and analysis of small proteins and short open reading frame encoded peptides in Hep3B cell**. *J Proteomics* 2021, **230**:103965.

56. Fesenko I, Kirov I, Kniazev A, Khazigaleeva R, Lazarev V, Kharlampieva D, Grafskaia E, Zgoda V, Butenko I, Arapidi G, et al.: **Distinct types of short open reading frames are translated in plant cells**. *Genome Res* 2019, **29**:1464–1477.

57. Budamgunta H, Olexiouk V, Luyten W, Schildermans K, Maes E, Boonen K, Menschaert G, Baggerman G: **Comprehensive Peptide Analysis of Mouse Brain Striatum Identifies Novel sORF-Encoded Polypeptides**. *Proteomics* 2018, **18**:1–12.

58. Murgoci AN, Cardon T, Aboulouard S, Duhamel M, Fournier I, Cizkova D, Salzet M: **Reference and Ghost Proteins Identification in Rat C6 Glioma Extracellular Vesicles**. *iScience* 2020, **23**:101045.

Table 1: Summary of recent studies aiming at identify SEPs by MS.

Ref	Species	Extraction method	Enrichment method	Enzymatic digestion	MS type	Database	Number of SEPs identified	Validation method
[48]	Archaea	Rapigest and disruption with steal ball	High pH reversed phase and native PAGE (GELFREE)	Trypsin	DDA	RNA seq	28	-
[49]	Archaea	Freeze–thaw and homogenization with glass beads	Protein precipitation	Trypsin	DDA	RNA seq	11	-
[50]	Archaea	SDS and freeze–thaw and ultrasonic homogenization	SDS-PAGE	Trypsin	DDA	RNA seq	45	-
[16]	Bacillus subtilis	Ultrasonication	Reversed-phase	Trypsin, Lys-C, Chymotrypsin, Arg-C	DDA	Genome six-frame translation	210	Comparison of MS2 spectra of synthetic peptides
[51]	Budding yeast	Urea and HCl	SDS-PAGE and 30 kDa MWCO	Trypsin	DDA	Genome six-frame translation	117	Comparison of MS2 spectra of synthetic peptides
[30]	Fruit fly	SDS	SDS-PAGE	Trypsin	DDA	RNA seq	119	IF of tagged SEPs
[26]	Fruit fly	Bead milling in isotonic lysis buffer	SDS-PAGE	Trypsin	DDA	Not specified	268	-
[25]	Human	SDS	SDS-PAGE	Trypsin	DDA	RNA seq	28	IF and WB of tagged SEPs
[11]	Human	SDS, RIPA, Water/Acetic acid/Methanol, boiling water	Protein precipitation and SDS-PAGE	Trypsin	DDA	OpenProt	52	-
[52]	Human	Urea buffer	-	Trypsin/Lys-C	DDA	Ensembl and RefSeq	13	-
[9]	Human	Sodium Deoxycholate	High-pH reversed-phase and HLA type 1 affinity purification	Trypsin/Lys-C	DDA	Ribosome profilling	More than 250	IF and WB of tagged SEPs and CRISPR knock out

[18]	Human	Sodium Deoxycholate	HLA type 1 affinity purification	-	DDA	RNA seq	452	PRM with heavy peptide standard
[15]	Human	Acetic acid	30 kDa MWCO, SDS-PAGE and ERLIC fractionation	Trypsin	DDA	RNA seq	237	SRM
[10]	Human	Acetic acid, Triton X-100, Boiling water	Protein precipitation, 30 kDa MWCO, reverse-phase (C8)	Trypsin	DDA	RNA seq	Around 200	-
[53]	Human	SDS and Guanidine HCl	High pH reversed phase and SAX fractionation	Trypsin/Lys-C	DDA	RNA seq	5	Comparison of MS2 spectra of synthetic peptides
[12]	Human	Boiling water	Acetic acid, 10 and 30 kDa MWCO	Trypsin	DDA	RNA seq	90	Isotopically labelled synthetic peptides
[19]	Human	Urea	Mitochondria isolation	Trypsin	DDA	Ribosome profilling	22	IF of tagged SEPs
[54]	Human	-	Protein precipitation and delipidation	Trypsin/-	DDA	RNA seq	19	PRM and heavy isotope labeled peptides
[55]	Human	Protein precipitation	SDS-PAGE and ERLIC fractionation	Trypsin	DDA	OpenProt	271	-
[39]	Human, mouse, zebrafish, fruit fly and C. elegans	Urea	SAX and IEF fractionation	Trypsin/Lys-C	DDA	RNA seq and conservation	70	-
[13]	Maize and Arabidopsis	Protein precipitation (TCA/acetone)	10 kDa MWCO	-	DDA	Genome six-frame translation	1993 (M) and 1860 (At)	Comparison of MS2 spectra of synthetic peptides
[56]	Moss	Acetic acid	Gel filtration	-	DDA	RNA seq	46	-
[57]	Mouse	Water/Acetic acid/Methanol	Reverse phase	-	DDA	Ribosome profilling	4	-
[58]	Rat	Extracellular vesicles	10 kDa MWCO	Trypsin/Lys-C	DDA	OpenProt	6	-

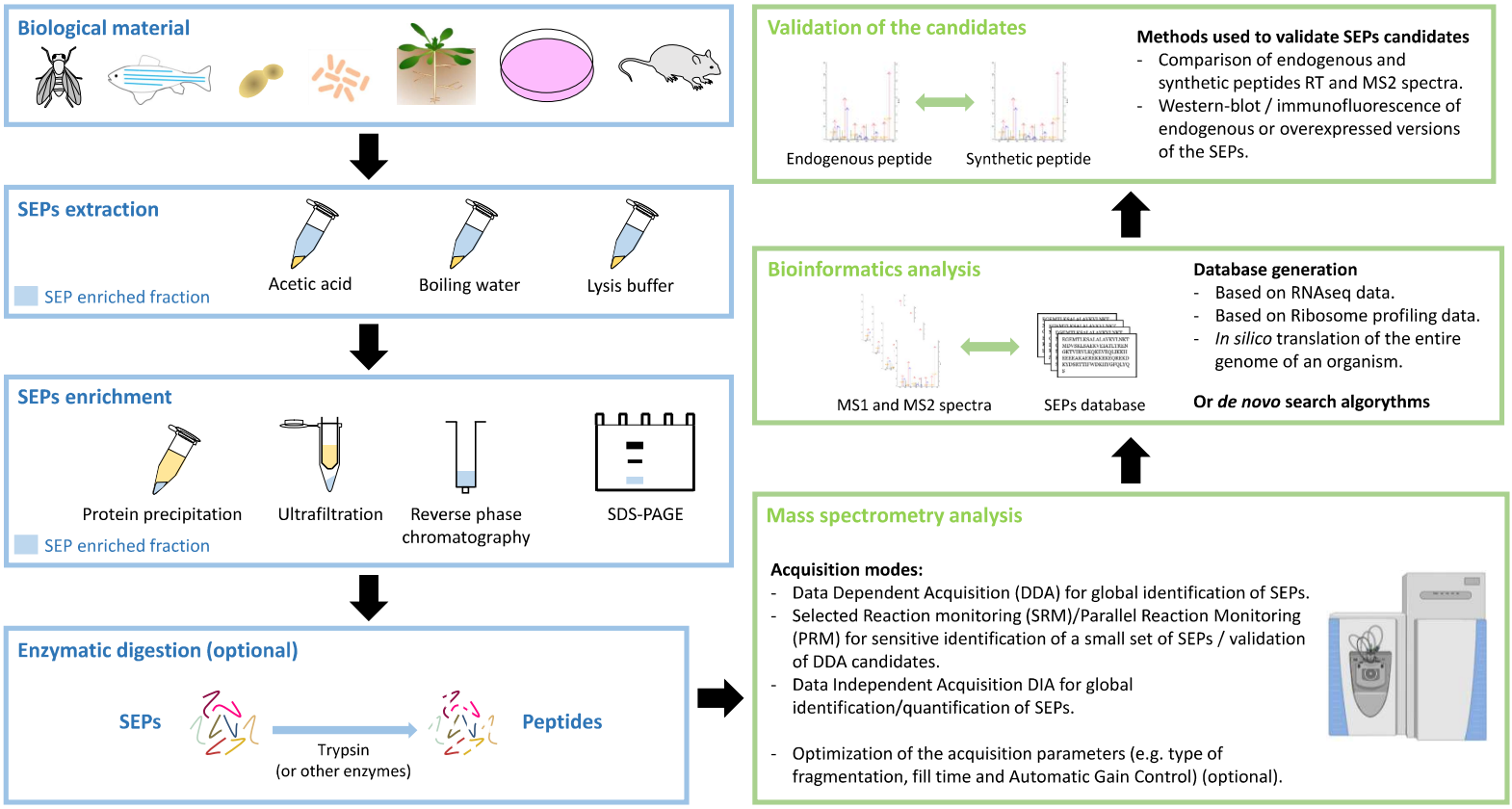
[a] *DDA: Data dependent acquisition; SDS: Sodium dodecyl sulfate; IF: Immunofluorescence; WB: Western blot; MWCO: Molecular weight cut-off; SAX: Strong anion exchange; IEF: Isoelectric focusing; ERLIC: Electrostatic*

repulsion-hydrophilic Interaction chromatography; M: Maize; At; Arabidopsis thaliana; PRM: Parallel reaction monitoring; SRM: Selected reaction monitoring; TCA: Trichloroacetic acid.

Figure legends:

Figure 1: General workflow of a peptidomics approach to identify short open reading frame encoded peptides (SEPs). Typically, SEPs are extracted from the biological material, enriched from the total protein pool and digested (or not) with trypsin (or other enzymes). The resulting peptides are then injected on a mass spectrometer and bioinformatics analysis is performed using custom databases to identify SEPs. Finally, several methods, such as comparison of retention times and MS2 spectra profiles or western blots, can be used to validate the results obtained from the mass spectrometry analysis.

Figure


Biological material
SEPs extraction
SEP enriched fraction
Acetic acid
Boiling water
Lysis buffer
SEPs enrichment
Protein precipitation
Ultrafiltration
Reverse phase chromatography
SDS-PAGE
SEP enriched fraction
Enzymatic digestion (optional)
SEPs
Trypsin
(or other enzymes)
Peptides
Mass spectrometry analysis
Acquisition modes:
- Data Dependent Acquisition (DDA) for global identification of SEPs.
- Selected Reaction monitoring (SRM)/Parallel Reaction Monitoring (PRM) for sensitive identification of a small set of SEPs / validation of DDA candidates.
- Data Independent Acquisition DIA for global identification/quantification of SEPs.
- Optimization of the acquisition parameters (e.g. type of fragmentation, fill time and Automatic Gain Control) (optional).
Bioinformatics analysis
MS1 and MS2 spectra
SEPs database
Database generation
- Based on RNAseq data.
- Based on Ribosome profiling data.
- In silico translation of the entire genome of an organism.
Or de novo search algorythms
Validation of the candidates
Endogenous peptide
Synthetic peptide
Methods used to validate SEPs candidates
- Comparison of endogenous and synthetic peptides RT and MS2 spectra.
- Western-blot / immunofluorescence of endogenous or overexpressed versions of the SEPs.