The Small Open Reading Frame-encoded Peptides: Advances in Methodologies and Functional Studies

Lei Chen,[ab] Ying Yang,[a] Yuanliang Zhang,[a] Kecheng Li,[a] Hongmin Cai,[c] Hongwei Wang,[d] and Qian Zhao*[a]

[a] Dr. Q Zhao, Dr. L Chen, Y Yang, YL Zhang, KC Li
State Key Laboratory of Chemical Biology and Drug Discovery, Department of Applied Biology and Chemical Technology
Hong Kong Polytechnic University
Hung Hom, Hong Kong SAR 999077, China
E-mail: q.zhao@polyu.edu.hk

[b] Dr. L Chen
Laboratory for Synthetic Chemistry and Chemical Biology Limited
Hong Kong Science and Technology Park, New Territories, Hong Kong SAR 999077, China

[c] Prof. HM Cai
School of Computer Science and Engineering, South China University of Technology, Guangzhou 510623, China

[d] Prof. HW Wang
State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University; Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangzhou 510623, China

Abstract: Small open reading frames are an important class of genes with less than 100 codons. They were historically annotated as noncoding or even junk sequences. In recent years, accumulating evidence suggested that sORFs could encode a considerable number of polypeptides, many of which played important roles in both physiology and disease pathology. However, it has been technically challenging to directly detect the sORF-encoded peptides (SEPs). Here, we discuss the latest advance in methodologies for identifying SEPs with mass spectrometry, as well as the progress on functional studies of SEPs.

1. Introduction

As is known to all, DNAs are transcribed into RNAs and then translated into proteins in the central dogma. However, less than 2% of human genome have been known to code proteins while a large majority of the detectable transcripts are not fully annotated. They were once believed to be useless or "junk".[1] Only until this decade, it has come to light that many of these noncoding RNAs (ncRNAs) have coding potential to produce polypeptides.[2] Small open reading frames (sORFs) that encode these polypeptides are historically excluded from genome annotation for multiple reasons. For example, the initiation codons of sORFs are not limited to AUG.[3] The length of sORFs is less than 300 nucleotide, which is arbitrarily defined as the minimum length of open reading frames (ORFs).[3a-c, 4] Unlike peptide hormones and neuropeptides that are produced through proteolysis of large precursor proteins,[5] sORF-encoded peptides (SEPs) are translated directly from sORF. Although the ubiquitous existence of sORFs in the genome has been reported in various species,[6] these sORFs are regarded as non-functional and have been neglected for a long time. With the significant development of various technologies, such as ribosome profiling (Ribo-seq) and mass spectrometry (MS), the existence of sORFs and SEPs have been gradually evidenced in recent years.

SEPs have been demonstrated to play an important role in a variety of processes and cellular pathways.[7] The biological functions of a handful of mammalian SEPs involving DNA repair,[8] mitochondrial function,[9] stress signaling,[10] and muscle development,[11] have been characterized in human and other vertebrates. In other organisms like bacteria, SEPs originating from small RNA (sRNA) or other ncRNA are being both discovered and performing indispensable biological functions.[12] Most known SEPs perform their functions by interacting with other proteins to regulate their functions. For example, Venkat et al.[12a] reported VcdRP from *Vibrio cholerae*, which was shown to interact with and regulate the enzymatic activity of citrate synthase. Considering that sORFs comprise at least 5–10% of genomes,[13] a large number of functional SEPs remain to be discovered and characterized. Kubatova et al.[14] investigated the secondary structures and conformation of 27 SEPs from 9 different bacterial and archaeal utilizing nuclear magnetic resonance (NMR) spectroscopy. The discovery of novel SEPs will complete the essential composition of the genome and proteome; moreover, the functional characterization of SEPs will provide us with new insight into fundamental biology, leading to translational applications. Several recently published reviews provide a good summary of the SEPs that have been identified in the past few years.[15]

The classification of sORFs have been beautifully reviewed in prior work.[15-16] sORFs can be classified into long noncoding RNAs (lncRNAs), microRNAs (miRNAs), spliced RNAs, circular RNAs (cirRNAs) and ribosomal RNAs (rRNAs), according to their origins from genome and molecular structures. lncRNAs are transcripts having low expression levels without coding sequence (CDS), and evolve rapidly.[17] miRNAs contain 5' untranslated region (UTR) miRNAs, 3' UTR miRNAs, as well as miRNAs across part of the CDS region and the adjacent 5' untranslated region (UTR) or 3' UTR. These miRNAs used to be known to regulate translation efficiency of the upstream and downstream canonical proteins.[18] Spliced mRNAs that frequently occur in conjunction with tumorigenesis and progression,[19] are completely from CDS regions and undergo alternative splicing. In recent decades, it has become apparent that these sORFs tend to have intriguing functions rather than being pointless sequences.[20]

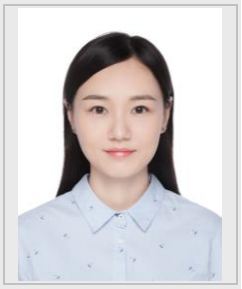
SEPs are recognized as important elements in biology, but discovering novel SEPs and understanding their functions are still

very challenging due to their small size, low abundance and high biology context-dependent specificity.[21] In recent years, scientists around the world have successfully identified a few SEPs from a plethora of organisms using different methods. The reported methods for discovering novel SEPs mainly fall into two categories, namely the sequencing result-based computational prediction and MS-based identification. Although computational prediction could indicate the general existence of sORFs and their coding potential, MS is the only method to provide direct evidence of SEPs.[22] Actually, sequencing results and computational prediction have enabled the identification of SEPs by MS-based methods when integrated approaches are implemented. In this review, we will focus on the current pipelines for SEP discovery using MS-based methods. Particularly, we will detail the progress made in each key step in identifying SEPs and highlight possible alternatives. In the end, the latest advances in studies towards the biological functions of SEPs will also be introduced, followed by the future prospect of this field. There are multiple terms in literature to indicate the translation products of sORFs. Some of them have alternative definitions, and thus using these terms may lead to misinterpretation. Therefore, we will use the terms SEP and polypeptide to indicate the translation production of sORFs all through this review.

Dr. Qian Zhao received her PhD in Chemistry from the University of Hong Kong in 2012 under the supervision of Prof. Dan Yang. She joined the University of California, San Francisco (UCSF) as a postdoctoral fellow working with both Prof. Alma Burlingame and Prof. Jack Taunton. She is currently an assistant professor in Hong Kong Polytechnic University. Her research group investigates protein-small molecule interactions. She is also interested in studying the “dark proteome” with mass spectrometry, such as polypeptides encoded by non-canonical open reading frames and non-canonical immunopeptides.

Dr. Hongwei Wang obtained his PhD degree from the Harbin Medical University under the supervision of Prof. Zheng Guo in the field of Biophysics. He is currently an Associate Professor in Bioinformatics at Zhongshan Ophthalmic Center, Sun Yat-sen University. His group works to decipher noncanonical translational events and translational control of gene expression.

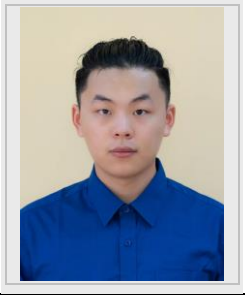
Dr. Hongmin Cai is a Professor at the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He received the B.S. and M.S. degrees in mathematics from the Harbin Institute of Technology, Harbin, China, in 2001 and 2003, respectively, and the Ph.D. degree in mathematics from the University of Hong Kong in 2007. From 2005 to 2006, he was a visiting Professor at Institute of Chemical Research, Kyoto University. His areas of research interests include biomedical image processing and bio-omics data integration.



Dr. Lei Chen received her PhD degree in chemistry from the University of Hong Kong in 2018 under Prof. Dan Yang’s supervision. She is currently a postdoctoral research fellow working in Dr. Qian Zhao’s lab. Her research is focusing on functional SEPs and immunopeptidomics in cancer immunotherapy.

Ying Yang received her BSc degree in 2016 from the China Pharmaceutical University. She obtained MSc in 2019 from University of Macau. Currently, she is a PhD student in Dr. Qian Zhao’s group. Her main research interests are discovery of SEPs with biological functions through MS-based workflow.

Yuanliang Zhang is a PhD student in Dr. Zhao’s group. He received his Bachelor of Science degree from Qingdao University in 2017 and joined Dr. Qian Zhao’s group in 2020. His research mainly focuses on proteomics, mass spectrometry, bioinformatics.



Kecheng Li obtained his bachelor’s degree in 2017 and master’s degree in 2020 from Sun Yat-sen University (under the supervision of Prof. Feng Liu). He joined Dr. Zhao’s group as a PhD student in 2021. He is focusing on proteogenomics and research of SEPs.

2. Important Characteristics of SEPs

SEPs of various lengths and with different start codons have been discovered using bioinformatic prediction and MS-based proteomic approach. Using Ribo-seq data and RibORF for scoring translation potential, Martinez et al.[21] found a total of 7,664 sORFs that may be translated into SEPs from HEK293T, HeLa-S3 and K562 cells. Using *de novo* method, Wang et al.[23] found 1,682 peptides from 2,544 human sORFs in Hep3B cells, whereas Zhang et al.,[24] using a conventional database search, found 355 human SEPs from eight human cell lines. According to their discoveries, the length range of the identified SEPs is similar, but the distributions are slightly different (Figure 1A). *De novo* method is likely to identify short SEPs, whereas conventional database search identifies longer ones. Such preferences could be due to the different scoring algorithms of these two methods. In database search, longer peptides, from which more spectra are likely to be collected, tend to obtain higher scores.

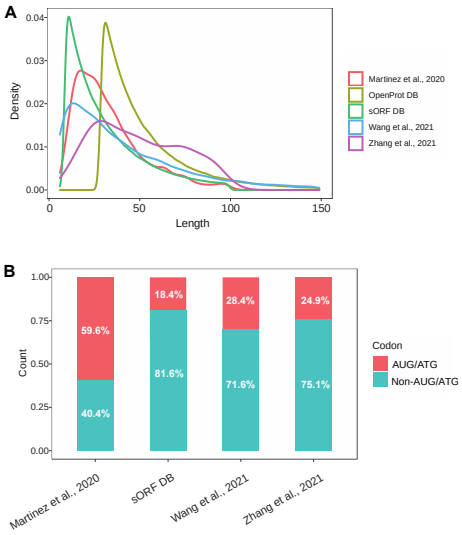


Figure 1. The length distribution (A) and the AUG start codon percentage (B) of SEPs in public databases and several representative studies.

In terms of translation initiation sites, it has been demonstrated that not only AUG, but also GUG and UUG may be employed as alternative translation codon in *Escherichia coli*.[25] Indeed, a substantial percentage of SEPs are translated with non-AUG start codons (Figure 1B). Nevertheless, the SEPs discovered using different approaches are all shorter than 100 amino acids in length and worthy to be further investigated.

3. Proteomics-based Methodology for Identification of SEPs

Like all conventional MS-based bottom-up proteomic studies, identification of SEPs is performed through the workflow (Figure 2), including sample extraction and enrichment, digestion and fractionation, MS data acquisition, and data analysis.

3.1. Sample extraction and enrichment

The first key step to identify SEPs with proteomics approaches is to extract SEPs from complex biological matrices (Figure 2A). Compared with proteins, extraction of SEPs is more challenging, because SEPs are easily hydrolysed by peptidase, or masked by undesired protein degradation products. Several methods have been applied to preserve the integrity of SEPs. In some studies, samples were heated up in water or lysis buffer, or stabilized by adding protease inhibitors to suppress peptidase and protease activity.[24, 26] However, the addition of peptidase and protease inhibitors may interfere with subsequent analysis of SEPs, as some of them are polypeptides. Meanwhile, the inhibition was not complete. An alternative method to circumvent the degradation of SEPs is to induce protein precipitation with hydrochloric acid or acetic acid, which simultaneously inactivates peptidases and proteases. The combination of these two methods has been widely used in the extraction of SEPs. A recent study by Cardon et al.[26b] identified new SEPs that were extracted using boiling water in combination with RIPA lysis buffer, and subsequently enriched by acetic acid precipitation. In this study, a blood marker (AltEDARADD) was found to be related to the diagnosis and prognosis of ovarian cancer. In summary, as a key step prior to the enrichment of SEPs, an appropriate extraction method should be selected according to the purpose of the study and the stability of the biological sample in question.

Once extracted, SEPs need to be separated from other proteins in the same sample. The separation is usually achieved based on various physical properties (such as size, hydrophobicity, and charge) of SEPs using one of the below methods (Figure 2B).

3.1.1. Selective precipitation

Selective precipitation of relatively large proteins with organic solvents, such as methanol,[26b] acetonitrile,[23, 26b, 27] trichloroacetic acid, acetic acid[26a] and chloroform,[23] was shown to effectively retain low-molecular-weight proteins including SEPs in the supernatant. In particular, acetic acid precipitation significantly induced aggregation of large proteins and left proteins of molecular weight lower than 30 kDa in the supernatant.[26a] In another endeavour, Cassidy et al.[27] used acetonitrile to precipitate proteins in the absence of detergent, which reduced the complexity of small protein samples, and successfully detected 11 SEPs smaller than 15 kDa from *Methanosarcina mazei*.

3.1.2. Size selection

Ultrafiltration with 10 or 30 kDa molecular weight cut-off (MWCO) membranes is widely used to achieve selective separation based on protein size. SEPs with low molecular weight pass through the membrane, while higher molecular weight proteins remain on the filter. However, ultrafiltration suffers from several shortcomings. First, the concentrated macromolecules may block the membrane pores, resulting in poor filtration efficiency. Second, non-specific absorption of small proteins by the membrane seems to be inevitable. Third, it is time-consuming to process samples of large volume. Another alternative separation method according to molecular weight is SDS-PAGE by which gel bands corresponding to desired molecular weight can be cut out for subsequent MS analysis. Ma et al.[26c] used 30 kDa MWCO Amicon filters and Tricine gel to discover 90 and 94 SEPs, respectively. He et al.[28] integrated four specific enrichment strategies (Urea-Tricine, HCl-Tricine, Urea-MWCO, and HCl-MWCO) for enhanced sequence coverage and SEP identification.

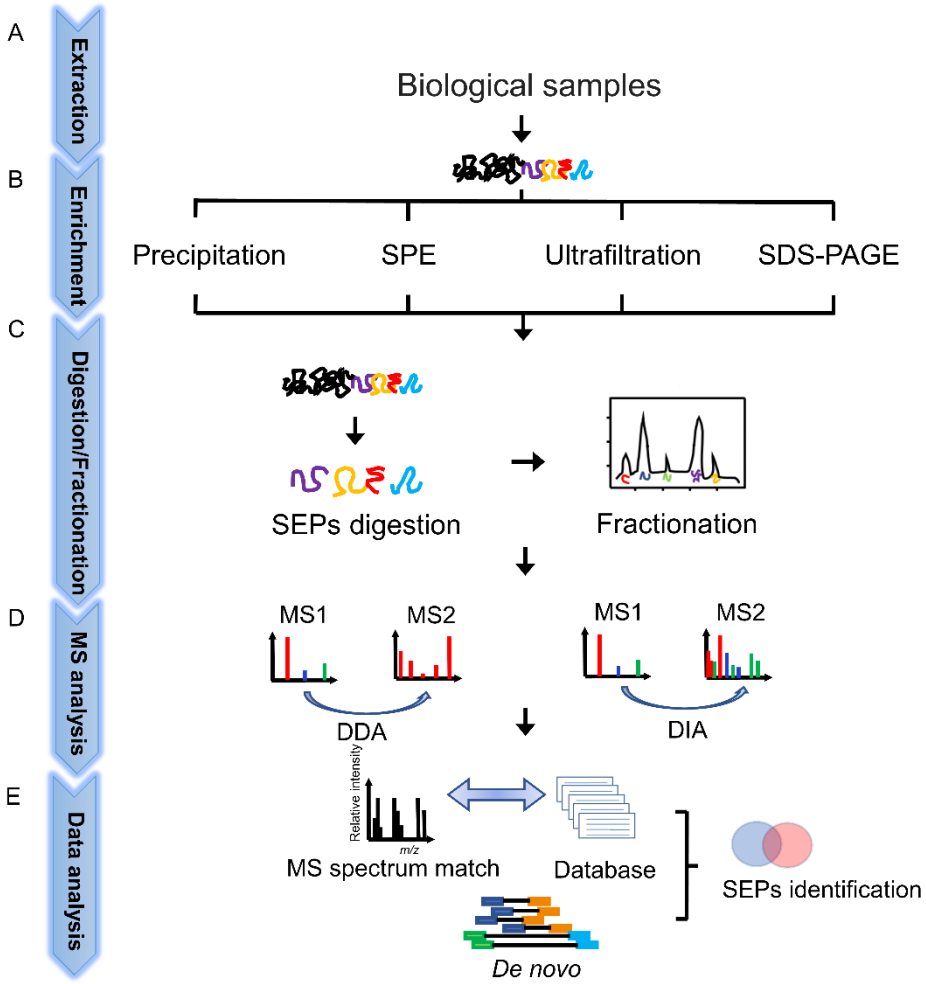


Figure 2 . The MS-based workflow for identification of SEPs. SEPs are extracted from complex biological samples, enriched from the total proteome, and digested with trypsin (or multiple enzymes). The tryptic peptides are subjected to fractionation, MS data acquisition, and data analysis to identify SEPs.

Using the Urea-Tricine could identify more SEPs than the other three strategies, and the four strategies show complementarity.

3.1.3. Solid phase extraction (SPE)

SEPs can be isolated according to their hydrophilicity and hydrophobicity. Although C8 SPE may lead to hydrophilic protein loss during sample enrichment,[24] the identification number of SEPs that were enriched using C8 SPE and acetic acid precipitation could be higher than using ultrafiltration with 30 kDa Amicon filters.[26a] It is noteworthy that a comparable number of SEPs were identified by Zhang et al.[24] Using C8 SPE, 30 kDa Amicon filters, and a combination of these two methods could lead to the identification of more SEPs in total. Overall, distinct extraction and enrichment methods have their special strengths and limitations, there is so far no individual method that can overperform the others. Therefore, it is necessary to design a systematic sample preparation method to identify SEPs.

3.2. Enzyme digestion and fractionation

The use of a single enzyme or a combination of enzymes to digest samples is also an important factor affecting SEPs identification (Figure 2C). Most studies adopt trypsin alone or Lys-C/trypsin to digest SEPs. When dealing with proteins containing many or no lysine/arginine residues, digestion with trypsin may impair sequence coverage. In addition, the sophisticated structure of SEPs may hinder the accessibility of trypsin and lead to miss cleavage. Due to the hydrolysis of peptide bonds by trypsin that always occurs at the C-terminus of lysine/arginine, the resulting N-terminal tryptic peptides are predominantly double-charged, whereas the C-terminal peptides are single-charged and hard to be detected by mass spectrometers.[29] Therefore, digestion with trypsin alone is insufficient for SEP identification. Multi-protease digestion using trypsin, Lys-C, chymotrypsin, and Glu-C has been shown to benefit the identification of SEPs, particularly in terms of the identification number, spectrum count and sequence coverage of SEPs.[30]

As the resultant peptide mixture can be highly complex after digestion, various off-line fractionation was introduced to enhance the sequencing depth of SEPs before MS analysis (Figure 2D). Electrostatic repulsion-hydrophilic interaction chromatography (ERLIC),[26c, 31] strong cation exchange (SCX),[32] high pH reverse phase fractionation,[9, 23, 30b, 33] and OFFGEL fractionation,[34] have been reported to significantly improve the identification of SEPs in multiple studies.

3.3. Data acquisition with mass spectrometry

Mass spectrometry remains the only method for direct detection and quantification of SEPs to date. Among the various MS methods, there are several that have been used in studying SEPs: data-dependent acquisition (DDA), data-independent acquisition (DIA), and parallel reaction monitoring (PRM) (Figure 2E). DDA is one of the most classic methods and it is based on the shotgun principle where top N abundant precursor ions are fragmented and analysed.[35] Depending on whether the sample is chemically incorporated to stable isotopic labels, DDA can also be applied for labelled quantitation or label-free quantitation. In the labelled quantitative method, the abundance of target proteins is derived from the intensity of the reporter ion in the spectra. DIA, also known as SWATH,[36] was developed by the Aebersold's group. It is gaining popularity in mass spectrometry-based proteomics, where it is one of the prominent label-free quantitative methods. DIA not only enables the simultaneous fragmentation of all precursor ions, which contributes to lower missing value and higher identification rate, but also preserves data that can be reanalyzed multiple times *in silico* using different spectral libraries. Under the PRM data acquisition mode, targeted precursor ions are selectively fragmented regardless of their relative abundance in the sample to generate distinctive spectrum patterns of these targets.[37]

Among the various label-free MS methods, DDA has remained the mostly widely applied proteomics method for years. In the past five years, thousands of SEPs have been discovered by using DDA from different species including human,[24, 38] *Escherichia coli*,[39] and plants.[6c] Although DDA remains relatively low complexity in data and easy to process using several developed workflows, it is intrinsically stochastic because only the top N intensive peaks can be selected for fragmentation and MS/MS analysis. While DDA methods suffer from the overlooking of the precursor ions of low MS signals, DIA methods should offer a broader dynamic range of detection, better sensitivity and reproducibility. With DIA methods, all peptides are theoretically fragmented and detected in parallel regardless of intensity. DIA methods hold great promise for the identification and quantification of SEPs. Pak et al.[40] reported a DIA-based workflow to detect both canonical and noncanonical immunopeptides. By virtue of the workflow and a collection of spectral libraries, identification of immunopeptides could be achieved with up to 3-fold increase.[40] Apart from the large-scale proteomics research that is based on the label-free quantitative method, labelled quantitation also contributes to the detection of SEPs. Zhu et al.[41] and Zhang et al.[9] used the Tandem Mass Tags (TMT) to identify hundreds of SEPs in their studies. Among these SEPs, TATDN2P1 and BRAWNIN were shown to have potential biological functions.[9, 41]

PRM, with improved sensitivity and accuracy for monitoring low abundant analytes in complex samples, is an excellent choice to validate the existence of newly discovered SEPs.[23-24, 42] The information of SEPs, including m/z and retention time of precursor ions, has been collected in a prior shotgun proteomics experiment and therefore can be used to establish the specific PRM methods to compare the spectra of the targets with their corresponding synthetic peptides. Zhu et al.[41] used PRM MS to confirm that 110 out of the 117 SEPs from their preliminary study were true. Moreover, PRM can also be used to measure the abundance of SEPs. For example, Delcourt et al.[43] employed PRM and isotope labelling to quantify the two translation products, reference MiD51 protein and alternative MiD51 protein (AltMiD51) of the same *MIEFI* gene based on its hits detected in Ribo-seq data; and AltMiD51 was identified as the major player in the regulatory function of this gene.

3.4. Database search strategy and database construction

The database search strategy by *in silico* matching against the theoretical spectra of peptides in an appropriate reference database is essential for protein identification.[44] Similar to the database-dependent proteomic studies on reference proteins, the accurate identification and investigation of SEPs heavily depends on the reliability of the database used. An ideal reference database for spectral matching should consist of all bona fide sample specific SEPs, with few irrelevant sequences to reduce false discovery and searching time.[31a, 45] As most SEPs are absent from databases that are commonly used, such as Ensembl, RefSeq, and UniProtKB,[31a, 45] it is particularly important to generate customized reference databases for mining novel SEPs.

3.4.1. Database construction from genomic sequencing data and RNA-seq data

To generate a reference database that covers putative SEPs, one of the most straightforward methods is to use the *in silico* six frame translation of the whole genome sequences. In the 1990s, this particular approach was used in conjunction with mass spectrometry to improve protein identification in *Escherichia coli.*[46] Until 2009, the six-frame genome translation approach was applied to identify SEPs in *Methanosarcina acetivoranswere* by Ferguson et al.[47] Several other groups also utilized the six-frame translation to construct reference databases and subsequently identified hundreds of novel SEPs from multiple species, including *Saccharomyces cerevisiae*,[28] *Bacillus subtilis*,[30b] *Zea mays* and *Arabidopsis thaliana*.[48] However, databases derived from the six-frame translation are restricted in their application to organisms with relatively small and intronless genome only with an obvious reason: a database that is derived from six-frame translation of whole human genome can be 70 times inflated than its corresponding Ensembl database.[15a] On the one hand, such databases contain a large number of spurious sequences and thus are so inflated that they are likely to lead to false identifications and an exponential increase in searching time.[15a] On the other hand, the false peptide sequences present in such databases make it unreliable to assess the confidence of peptide spectral matches (PSMs), further aggravating the difficulty in the discovery of low abundant SEPs.[49]

To compress the size of reference database and elevate the proportion of *bona fide* SEPs, different sources of genomic annotation and transcriptomic data are introduced to construct the database through six- or three-frame (forward only) *in silico* translation. PeptideClassifier and iPtgxDB integrate six-frame genome translation and annotations from different sources to unambiguously identify novel SEPs from prokaryote,[50] such as *B. henselae*, *Bradyrhizobium diazoefficiens,* and *Escherichia coli*. In 2013, Slavoff et al.[31a] successfully uncovered 86 novel SEPs with custom reference databases that were derived from the six-frame translation of RefSeq transcripts and the three-frame translation of RNA-seq data of K562 cells. After their pioneering work, plenty of novel SEPs have been identified from different species, such as human, mouse, zebrafish, fruit fly, and the nematode *C. elegans*.[26a, 26c, 27, 42, 51] The combination of annotated protein database and *in silico* translation of RNA-seq data excludes ORFs that cannot be transcribed, and thus reduces the chance of false discovery. Nevertheless, such method is still haunted by a large proportion of spurious SEPs since it only relies on AUG as the

start codon and the stop codon to recognize ORFs. Hence in subsequent studies, RNA-seq data was used in combination with other evidence of sORFs, such as bioinformatic prediction and Ribo-seq data, to refine the customized SEP reference database for improved reliability.[6c, 51a, 52]

3.4.2. Database construction from RNA-seq and Ribo-seq data: various pipelines to predict sORFs

Since the development of Ribo-seq technology in 2009, scientists have been able to detect the actively translated region of mRNAs in a genome-wide manner by capturing and sequencing the ribosome protected fragments (RPFs).[53]The utilization of translation inhibitors and the 3-nucleotide periodicity of RPFs offers information on active translated ORFs with single-codon resolution.[54] Ribo-seq has proven to be a highly powerful technique for exploring the peptide-coding potential of sORFs.[55] Different from *in silico* methods, Ribo-seq relies on the distribution patterns of RPFs, rather than the canonical initiation codon or transcript annotations, to predict coding sORFs.[56] Thus putative peptides encoded by both AUG- and non-AUG-initiated sORFs are incorporated in the reference database.[16b]

Many tools are available for the construction of reference databases based on Ribo-seq data, and they assist in the discovery of thousands of novel SEPs. RiboTaper, presented by Calviello et al.[57] in 2016 was a relatively early and statistically rigorous method based on 3-nucleotide periodicity of RPFs to identify actively translated regions. RiboTaper was then integrated into a proteogenomic pipeline and identified 218 novel proteins in Chinese hamster tissue and CHO cell lines.[58] Following the debut of RiboTaper, many tools have been developed for analysing Ribo-seq data (Table 1).[54, 59] Although most of them were not originally designed for mining sORFs from Ribo-seq data, they have great potential in the studying of sORFs. Using reference databases constructed with these tools, thousands of SEPs have been identified recently. The probabilistic inference of codon activities by an EM algorithm (PRICE), developed by Erhard et al., looks for the set of codons that is most likely to generate the observed reads, and predict the potential start codons using a machine learning (ML) model with high accuracy. The presence of PRICE-predicted SEPs was validated using two previously published MHC-I immunopeptidome datasets.[16b] Very recently, the application of PRICE has contributed to the identification of 525 noncanonical immunopeptides that are derived from sORFs.[60] Apart from PRICE, RibORF and RiboCode have also been applied to reference database construction for mining SEPs from MS data.[21, 61]

However, it is almost unlikely to eliminate the translation-irrelevant binding between ribosome and transcripts in Ribo-seq, and thus the probability of false discovery is increased.[62] In addition, the difference of algorithm design may lead to fairly different sORF prediction results from the same Ribo-seq data.

Table 2. Software tools available for mining sORFs from Ribo-seq data

Tool	Year	Algorithm	Function
ORF-RATER[63]	2015	Linear regression algorithm	Identifying and quantifying translation of ORFs
RibORF[64]	2015	Support vector machine	Genome-wide translated ORF identification
riboHMM[6b]	2016	Hidden Markov model	ORF prediction
RiboTaper[57]	2016	Multitaper method	ORF prediction
RIBO-TISH[54]	2017	The non-parametric Wilcoxon rank-sum test	Translation initiations analysis and ORF prediction
RP-BP[59a]	2017	Bayesian approach	ORF prediction
PRICE[16b]	2018	Expectation–maximization algorithm and machine-learning model	ORF prediction and resolving overlapping sORFs
Ribowave[59b]	2018	Wavelet transform	ORF prediction, protein abundance estimation, TE calculation, ribosomal frameshift identification
RiboCode[65]	2018	Wilcoxon signed-rank test	*De novo* annotation of the translatome
ORFquant[66]	2020	Multitaper method	ORF prediction and quantification
Ribotricer[59c]	2020	3D to 2D projection	ORF detection across multiple species

To enhance the reliability of database derived from Ribo-seq data, the orchestration of multifaceted methods, including bioinformatic analysis,[52] RNA-seq,[55b] and genome-scale CRISPR screens is imperative.[67] In complement to customized reference databases for discovering SEPs from specific samples, public databases of sORFs predicted using Ribo-seq data such as OpenProt,[68] sORFs.org,[69] ARA-PEPs,[70] PsORF,[71] and MetamORF,[72] are easily accessible. These resources can be readily adopted in database construction for identification of SEPs when sample-specific Ribo-seq data are unavailable.

3.5. *De novo* sequencing

When lacking a flawless database, *de novo* sequencing is a valuable supplement to database search to discover unannotated proteins like SEPs. *De novo* sequencing of MS data is a library-independent approach for deciphering protein or peptide sequences only from the spectrum. The sequence of SEPs that is absent from the database used for searching will never be identified by any of the MS database search algorithms. Since *de novo* sequencing does not require any reference databases to deduce peptide sequences from MS raw data, the spectrum identification rate could be improved.[23] The peptides-spectrum match (PSM) can be assigned by *de novo* sequencing engines, including pNovo3,[73] Novor,[74] PepNovo,[75] and PEAKS.[76] These engines employ divergent algorithms, such as spectrum graph,[75] TagGraph,[77] and deep learning,[78] to evaluate the PSM quality

by generating confident scores. To date, *de novo* sequencing has been implemented to discover SEPs in addition to database search in a few studies. Chen et al.[79] and Wang et al.[23] identified over hundreds of SEPs using PEAKs and pNovo3, respectively. In these studies, stringent filtering criteria, such as confident score cut-offs and sequence similarity to the reference database by BLASTp,[80] are adopted to eliminate false discovery. Notably, in spite of the improved spectrum identification rate, *de novo* sequencing was reported to offer identifications of which only 35% were also identified using database search.[81] The incorporation of *de novo* sequencing in MS-based proteomic studies is still disputable.[82] Very recently, Erhard et al.[83] developed Peptide-PRISM, an FDR-based method to filter noncanonical immunopeptides that were discovered using *de novo* sequencing. The successful identification of 6,636 noncanonical immunopeptides has demonstrated the practical feasibility to use *de novo* sequencing for discovering SEPs.

4. Methods to Predict SEPs Other Than Proteomics Approaches

Bioinformatic and computational analysis of genome sequence are intuitive operations to predict SEPs, predominantly through detecting purifying selection, similarity comparison with any known protein domains, and machine learning algorithms (Table 3).[84] PhyloCSF is one of the most widely used tools to detect evolutionarily conserved coding ORFs. It performs alignment of transcripts from multi-species and takes phylogenetic models into account additionally.[85] Other prediction tools with the same working principle include RNAcode,[86] uPEPeroni,[87] and micPDP.[88] Alternative selection-based tools predict the coding ORFs through evaluation of the nucleotide composition, as exemplified by CRITICA,[89] PhastCons,[90] and sORF finder.[91] An advancement of these tools in predicting coding ORFs is that they take into account the influence from the ORF context in addition to the conservation of sequences. The second most utilized principle for SEP prediction is to test whether the sORFs in question are similarity to any known proteins or protein domains. Tools based on this working principle include BLAST,[92] HMMR,[93] and PFAM.[94] There are also emerging tools that are specially built on machine learning (ML) algorithms for SEP prediction, such as DeepCPP and miPepid.[95] In particular, miPepid is alignment-free and designed specifically to predict the coding potential of sORFs.[95b] Using the cryptic features of coding sORFs that were concealed in the training datasets, miPepid was able to predict whether a given sORF encodes a polypeptides with 96% accuracy.[95b] Nevertheless, all these bioinformatic tools have their own systematic drawbacks, which calls for careful selection of an appropriate one according to the research content. Tools that are based on intrinsic features of exons may omit SEPs with non-AUG initiation codon,[96] whereas those comparing phylogenetic conservation may suffer from the bad quality of the multi-species alignment. Other tools that are restricted to functional polypeptides that assemble known proteins might be counterproductive for newly emerging and species- or tissue-specific SEPs.[84] Therefore, a combination of tools based on different working principles will be beneficial to achieve accurate and comprehensive identification of SEPs.

Table 3. Representative bioinformatic tools for SEP prediction

Tool	Year	Working Principle
BLAST[92]	1990	Sequence similarity to known proteins
HMMER[93]	1995	Sequence similarity to known proteins
CRITICA[89]	1999	Nucleotide composition
PhastCons[90]	2005	Nucleotide composition
sORF finder[91]	2009	Nucleotide composition
PhyloCSF[85]	2011	Evolutionary conservation, codon substitution, multispecies transcript alignment
RNAcode[86]	2011	Codon substitution
micPDP[88]	2014	Codon substitution
uPEPperoni[87]	2014	Codon substitution
ELM[97]	2018	Similarity to linear proteins
miPepid[95b]	2019	Machine learning algorithms
PFAM[94]	2019	Similarity to linear proteins
DeepCPP[95a]	2020	Machine learning algorithms
RNAsamba[98]	2020	Similarity to known proteins

5. Molecular Functions of SEPs

Quite a few SEPs that are translated from lncRNA, circRNA, and miRNA have been reported to function in multiple biological processes, such as DNA repair,[8] tumorigenesis,[99] ion signalling and muscle development,[11, 100] metabolism,[9] and transcriptional regulation.[101]

5.1. lncRNA-encoded polypeptides

lncRNAs that are conventionally more than 200-nt in length make up the most of known noncoding RNAs (ncRNAs).[102] Multiple biological and molecular functions of lncRNA-encoded polypeptides have been found in i) regulation of DNA repair,[8] cancer proliferation,[99a] invasion, metastasis and prognosis,[26b, 99a, 103] ii) modulation of metabolism and muscle cell growth,[9, 11, 100] and iii) cellular response to stress.[10, 21, 99b, 104] SEPs and their corresponding sORFs can independently regulate biological processes through different mechanisms, like the case for non-annotated P-body dissociating polypeptide (NoBody). It is encoded by a lncRNA *LINC01420*,[38] and the lncRNA is negatively correlated to the overall survival rate of patients with nasopharyngeal carcinoma.[105] Increase in NoBody abundance leads to decrease in the amount of aberrant transcripts that are substrates of the nonsense-mediated mRNA decay (NMD).[38] NoBody binds to the enhancer of decapping 4 (EDC4) as a novel component of the mRNA decapping complex to accelerate mRNA turnover through NMD.[38, 106] In addition to acting as modulators,

SEPs have also been found to be a biomarker of many cancers, showing their relationship with to diagnosis and prognosis.[26b, 103] In colorectal cancer (CRC), a 53-aa peptide encoded by lncRNA *HOXB-AS3* was found to have inhibitory effect on cancer growth, and it could suppress glucose metabolism. HOXB-AS3 is down-regulated in highly metastatic and primary CRC tissues, and patients with low HOXB-AS3 peptide level had poorer prognoses.[103a] Other studies have shown that many lncRNAs bind to RPS6 in cancer cells, and SEP SMIM30, which is encoded by *LINC00998,* could promote tumorigenesis by modulating cell proliferation and migration.[99a] Furthermore, the SMIM30 level was correlated with poor survival in patients with hepatocellular carcinoma (HCC).[99a]

Polypeptides encoded by lncRNAs also function as important regulators of metabolism and cell growth.[100a] In muscle cells, Sarco/endoplasmic reticulum Ca^{2+} ATPase (SERCA) is a central pump mediating the reuptake of Ca^{2+} into the sarcoplasmic reticulum (SR).[100a] The Ca^{2+} uptaking system is directly impeded by three muscle-specific SEPs, myoregulin (MLN),[11c] phospholamban (PLN),[11b] and sarcolipin (SLN).[11a, 100b] In contrast, the only known endogenous SEP, DWORF, can enhance SERCA activity by interfering with MLN, SLN, and PLN.[100c] Multiple independent studies have shown that SEPs are able to regulate general biological pathways and processes as well. Small regulatory polypeptide of amino acid response (SPAR), a polypeptide encoded by the lncRNA *LINC00961*, interacts with the lysosomal v-ATPase to deactivate mTORC1.[100d] Myomixer, encoded by *Gm7325*, promotes fibroblast-fibroblast and fibroblast-myoblast fusions in association with Myomaker, which is the critical step in myofiber formation during muscle development.[100e] BRAWNIN, a mitochondrial-localized SEP, encoded by *C12orf73* gene, was identified by using Ribo-seq combined with proteomic prediction pipelines.[9] BRAWNIN is an essential regulator of oxidative metabolism for respiratory chain complex III assembly, and is induced by AMPK pathway. Down-regulation or loss of BRAWNIN impairs mitochondrial ATP production.[9] In another study, several lncRNAs were detected to have coding capability using Ribo-seq data, and 11 polypeptides encoded by these lncRNAs were validated.[107] Three of the SEPs were subsequently identified to regulate cardiomyocyte hypertrophy by being involved in modulating oxidative phosphorylation, calcium signalling pathway, and the MAPK pathway.[107] Very recently, a short ORF-encoded histone binding protein (SEBHP) was identified as a transcriptional regulator that interacted with chromatin-bounded proteins. Strikingly, SEBHP is capable of modulating more than 15% of the active transcriptome.[101]

The early studies on aberrant translation of ncRNAs in response to cellular stress started with prokaryotes, and the functions of several bacterial SEPs have been fully investigated at both the phenotypic and molecular levels.[108] It was not until the 2010s that there were reports about the regulated expression of SEPs in human cells in response to cellular stress. One fundamental work is to identify essential polypeptides that is encoded by lncRNA *Aw112010* for modulating mucosal immunity in bacterial infection and colitis.[104a] Additionally, a mitochondrial SEP termed PIGBOS which is encoded by the opposite strand of *PIGB* gene is identified to regulate the unfolded protein response (UPR).[104b] This SEP is critical for many cellular activities, and depletion of PIGBOS would lead to severe UPR and increased cell death upon endoplasmic reticulum (ER) stress.[104b] The FOXA1-regulated conserved small protein (FORCP), which is encoded by a putative gastrointestinal specific lncRNA *LINC00675*, is overexpressed to regulate apoptosis and tumorigenesis upon ER stress in well-differentiated CRC cells.[99b] In addition to cancer regulation, recent evidence has demonstrated that lncRNAs are also involved in the proliferation of pulmonary artery smooth muscle cells (PASMCs) in response to hypoxia in pulmonary hypertension.[10] The lnc-Rps4l-encoded peptide 40S ribosomal protein S4 X isoform-like (RPS4XL) is found to inhibit the proliferation of PASMCs and the phosphorylation of RPS6 through its interaction with RPS6 under hypoxic conditions. [10]

5.2. circRNA-encoded polypeptides

circRNAs are covalently closed loops of ncRNAs containing no 5'cap or polyA tail.[109] They are produced during alternative splicing,[110] and driven by the internal ribosome entry site (IRES)- or N6-methyladenosine (m6A)-mediated initiation.[111] To date, several circRNA-encoded polypeptides have been identified in different cancers, such as glioma, HCC, and CRC, with their functions in physiology and cancer development have also been validated extensively.[109] The circular form of the SNF2 histone linker PHD RING helicase gene (*circ-SHPRH*) is translated into a novel tumor suppressor termed SHPRH-146aa.[99c] This SEP protects the full-length SHPRH encoded by *SHPRH* from ubiquitination-mediated protein degradation, and the sORF is known for the inhibitory effect in regulating cell proliferation and tumorigenicity.[99d] Other circRNA-encoded peptides, including PINT-87aa encoded by the circular form of the long intergenic non-protein-coding RNA p53-induced transcript (LINC-PINT), and FBXW7-185aa encoded by *circ-FBXW7,* were also reported.[99e, 99f] Particularly, PINT-87aa interacts with polymerase associated factor complex (PAF1c) directly, and thereby inhibits the transcriptional elongation of many oncogenes.[99e] In parallel to SHPRH-146aa and PINT-87aa, FBXW7-185aa shows an agonistic effect in cell proliferation and cell cycle acceleration as a cancer suppressor through destabilization of c-Myc.[99f] Particularly in glioblastoma, these three SEPs are observed downregulated, and their corresponding circRNAs are also positively correlated to the overall survival rate of patients.[99c, 99e, 99f] Very recently, a secretory SEP E-cadherin variant (C-E-Cad), which is encoded by a circular E-cadherin RNA, has been identified to stimulate EGFR signalling and promote tumorigenicity in glioblastoma.[112] All these studies have demonstrated the essential regulatory functions of circRNA-encoded peptides in tumorigenesis.

5.3. miRNA-encoded polypeptides

Primary miRNAs are the long precursors of miRNAs that function mainly in mRNA silencing and post-transcriptional regulations.[113] During the maturation of miRNAs, they are processed and spliced into miRNAs as short single-stranded ncRNAs.[114] There are a few pieces of evidence that demonstrate the coding potential of miRNAs in plant biology, such as the discovery of primary miR171b and miR165a in the root development in *alfalfa* and *Arabidopsis thaliana*, respectively.[16a] In human cells, miRNAs prefer to interact with mRNA through base pairing to regulate the expression of most mRNAs.[16a, 105, 115] In 2017, miPEP-200a and miPEP-200b are the first two polypeptides encoded by primary miRNAs miR-200a and miR-200b, respectively.[116] These two SEPs have been demonstrated to regulate epithelial-mesenchymal transition of prostate cancer cells by inhibiting the vimentin-mediated pathway.[116]

6. SEPs in HLA-I complexes

In addition to performing biological functions as polypeptides, SEPs are also substrates of the antigen processing and presenting machinery, leading to the discovery of SEP-derived noncanonical antigens.[16d, 117] In human cells, the major histocompatibility (MHC) complexes present peptide fragments or immunopeptides from cellular proteins that are digested by proteosomes. While most of the immunopeptides are recognized as “self-antigen”, a few tumors associated/specific antigens (TAAs/TSAs) are promising targets for immunotherapy. Pioneering works have demonstrated that SEPs offer a new source of immunopeptides supplementary to canonical proteins.[16d, 117] Cancer antigens derived from SEPs that were specifically upregulated in tumor could trigger immune responses.[117e, 117f] Since 2018, several independent research groups including our group have reported the prevalence of noncanonical immunopeptides in cancer. By utilizing a proteogenomic approach, 40 TSAs have been identified from human primary tumors, 90% of which are from noncoding regions.[118] By virtue of Ribo-seq, 320 noncanonical immunopeptides derived from SEPs were identified from a previous study.[21] In another work, thousands of noncanonical immunopeptides were identified by using Ribo-seq and MS, and their source proteins were tumor-specific SEPs expressed in multiple cancers.[119] With the incorporation of bulk and single cell sequencing, Ribo-seq and MS, hundreds of shared and tumor-specific noncanonical immunopeptides derived from SEPs have been discovered.[42] Alternatively, 240 noncanonical peptides were identified from human induced pluripotent stem cells (iPSCs) using MS, Ribo-seq and CRISPR-based screening methods.[67] Indeed, immunopeptidome is an ideal enrichment for SEPs with 2,503 out of 14,498 proteins are SEPs.[60] All these studies demonstrate the existence of SEPs and their involvement in antigen presentation pathway and cancer immunology.

7. Conclusions

As hidden gems for decades, SEPs have attracted much attention and research on them has made significant progress in recent years. Many methods for improving the identification of SEPs have emerged. Addition to the MS-based bottom-up proteomic methods we reviewed in the paper, there are a few recent works using top-down strategies to identify SEPs. Zhang et al.[23] applied top-down methods and found 241 SEPs in Hep3B cell line. Cassidy et al.[32a] identified 12 SEPs and characterized corresponding 36 proteoforms by top-down strategy from *Methanosarcina mazei*. These findings suggest that the top-down strategy is also a feasible alternative method to study SEPs. It is noticed that almost every step in the mass spectrometry-based proteomics workflow have been optimized substantially, including SEPs extraction in sample preparation, sample separation, data acquisition and analysis, etc. However, we must admit that the current identification number is still not reaching the predicted numbers of SEPs. With the rapid development of mass spectrometry methodologies, it is foreseeable that there will be a dramatic increase in number of SEPs. Meanwhile, the quantification of SEPs will become more accurate, making it feasible to study the biological functions of SEPs in large scale.

Acknowledgements

Our sORFs work is funded by Research Grants Council (RGC) ECS 25301518, GRF 15305821 and RGC-CRF Equipment C5033-19E. We also acknowledge the funding support from Laboratory for Synthetic Chemistry and Chemical Biology under the Health@InnoHK Program launched by Innovation and Technology Commission, The Government of Hong Kong Special Administrative Region of the People's Republic of China.

Keywords: small open reading frames • sORF-encoded peptides • mass spectrometry • functions • proteomics

[1] J. T. Kung, D. Colognori, J. T. Lee, *Genetics* **2013**, *193*, 651-669.

[2] a) I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B.-K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, I. Dunham, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kellis, J. Khatun, P. Kheradpour, A. Kundaje, T. Lassmann, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. J. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, M. Snyder, M. J. Pazin, R. F. Lowdon, L. A. L. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, E. D. Green, P. J. Good, E. A. Feingold, B. E. Bernstein, E. Birney, G. E. Crawford, J. Dekker, L. Elnitski, P. J. Farnham, M. Gerstein, M. C. Giddings, T. R. Gingeras, E. D. Green, R. Guigó, R. C. Hardison, T. J. Hubbard, M. Kellis, W. J. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Stamatoyannopoulos, S. A. Tenenbaum, et al., *Nature* **2012**, *489*, 57-74; b) T. R. Cech, J. A. Steitz, *Cell* **2014**, *157*, 77-94.

[3] a) X. Yang, T. J. Tschaplinski, G. B. Hurst, S. Jawdy, P. E. Abraham, P. K. Lankford, R. M. Adams, M. B. Shah, R. L. Hettich, E. Lindquist, U. C. Kalluri, L. E. Gunter, C. Pennacchio, G. A. Tuskan, *Genome Res* **2011**, *21*, 634-641; b) M. Su, Y. Ling, J. Yu, J. Wu, J. Xiao, *Front Genet* **2013**, *4*, 286; c) P. Sieber, M. Platzer, S. Schuster, *Trends in genetics : TIG* **2018**, *34*, 167-170; d) M. Kozak, *Annual Review of Cell Biology* **1992**, *8*, 197-225; e) M. Kozak, *Proc Natl Acad Sci U S A* **1995**, *92*, 2662-2666.

[4] M. Furuno, T. Kasukawa, R. Saito, J. Adachi, H. Suzuki, R. Baldarelli, Y. Hayashizaki, Y. Okazaki, *Genome Res* **2003**, *13*, 1478-1487.

[5] V. Hook, L. Funkelstein, D. Lu, S. Bark, J. Wegrzyn, S.-R. Hwang, *Annu Rev Pharmacol Toxicol* **2008**, *48*, 393-423.

[6] a) E. Ladoukakis, V. Pereira, E. G. Magny, A. Eyre-Walker, J. P. Couso, *Genome Biol* **2011**, *12*, R118; b) A. Raj, S. H. Wang, H. Shim, A. Harpak, Y. I. Li, B. Engelmann, M. Stephens, Y. Gilad, J. K. Pritchard, *Elife* **2016**, *5*; c) I. Fesenko, I. Kirov, A. Kniazev, R. Khazigaleeva, V. Lazarev, D. Kharlampieva, E. Grafskaia, V. Zgoda, I. Butenko, G. Arapidi, A. Mamaeva, V. Ivanov, V. Govorun, *Genome Res* **2019**, *29*, 1464-1477.

[7] a) J.-P. Couso, P. Patraquim, *Nature Reviews Molecular Cell Biology* **2017**, *18*, 575-589; b) Q. Chu, J. Ma, A. Saghatelian, *Crit Rev Biochem Mol Biol* **2015**, *50*, 134-141.

[8] S. A. Slavoff, J. Heo, B. A. Budnik, L. A. Hanakahi, A. Saghatelian, *Journal of Biological Chemistry* **2014**, *289*, 10950-10957.

[9] S. Zhang, B. Reljić, C. Liang, B. Kerouanton, J. C. Francisco, J. H. Peh, C. Mary, N. S. Jagannathan, V. Olexiouk, C. Tang, G. Fidelito, S. Nama, R.-K. Cheng, C. L. Wee, L. C. Wang, P. Duek Roggli, P. Sampath, L. Lane, E. Petretto, R. M. Sobota, S. Jesuthasan, L. Tucker-Kellogg, B. Reversade, G. Menschaert, L. Sun, D. A. Stroud, L. Ho, *Nature Communications* **2020**, *11*, 1312.

[10] Y. Li, J. Zhang, H. Sun, Y. Chen, W. Li, X. Yu, X. Zhao, L. Zhang, J. Yang, W. Xin, Y. Jiang, G. Wang, W. Shi, D. Zhu, *Mol Ther* **2021**, *29*, 1411-1424.

[11] a) N. C. Bal, S. K. Maurya, D. H. Sopariwala, S. K. Sahoo, S. C. Gupta, S. A. Shaikh, M. Pant, L. A. Rowland, E. Bombardier, S. A. Goonasekera, A. R. Tupling, J. D. Molkentin, M. Periasamy, *Nat Med* **2012**, *18*, 1575-

1579; b) D. H. MacLennan, E. G. Kranias, *Nature Reviews Molecular Cell Biology* **2003**, *4*, 566-577; c) D. M. Anderson, K. M. Anderson, C. L. Chang, C. A. Makarewich, B. R. Nelson, J. R. McAnally, P. Kasaragod, J. M. Shelton, J. Liou, R. Bassel-Duby, E. N. Olson, *Cell* **2015**, *160*, 595-606.

[12] a) K. Venkat, M. Hoyos, J. R. Haycocks, L. Cassidy, B. Engelmann, U. Rolle-Kampczyk, M. von Bergen, A. Tholey, D. C. Grainger, K. Papenfort, *The EMBO Journal*, *n/a*, e108542; b) S. Cao, X. Liu, Y. Huang, Y. Yan, C. Zhou, C. Shao, R. Yang, W. Zhu, Z. Du, C. Jia, *Communications Biology* **2021**, *4*, 1248.

[13] A. Saghatelian, J. P. Couso, *Nature chemical biology* **2015**, *11*, 909-916.

[14] N. Kubatova, D. J. Pyper, H. R. Jonker, K. Saxena, L. Remmel, C. Richter, S. Brantl, E. Evguenieva‐Hackenberg, W. R. Hess, G. Klug, *Chembiochem* **2020**, *21*, 1178.

[15] a) B. Fabre, J.-P. Combier, S. Plaza, *Current Opinion in Chemical Biology* **2021**, *60*, 122-130; b) L. Cassidy, P. T. Kaulich, S. Maaß, J. Bartel, D. Becher, A. Tholey, *Proteomics* **2021**, 2100008; c) P. V. Sergiev, M. P. Rubtsova, *Biochemistry (Moscow)* **2021**, *86*, 1139-1150.

[16] a) D. Lauressergues, J.-M. Couzigou, H. S. Clemente, Y. Martinez, C. Dunand, G. Bécard, J.-P. Combier, *Nature* **2015**, *520*, 90-93; b) F. Erhard, A. Halenius, C. Zimmermann, A. L'Hernault, D. J. Kowalewski, M. P. Weekes, S. Stevanovic, R. Zimmer, L. Dölken, *Nat Methods* **2018**, *15*, 363-366; c) S. van Heesch, F. Witte, V. Schneider-Lunitz, J. F. Schulz, E. Adami, A. B. Faber, M. Kirchner, H. Maatz, S. Blachut, C. L. Sandmann, M. Kanda, C. L. Worth, S. Schafer, L. Calviello, R. Merriott, G. Patone, O. Hummel, E. Wyler, B. Obermayer, M. B. Mücke, E. L. Lindberg, F. Trnka, S. Memczak, M. Schilling, L. E. Felkin, P. J. R. Barton, N. M. Quaife, K. Vanezis, S. Diecke, M. Mukai, N. Mah, S. J. Oh, A. Kurtz, C. Schramm, D. Schwinge, M. Sebode, M. Harakalova, F. W. Asselbergs, A. Vink, R. A. de Weger, S. Viswanathan, A. A. Widjaja, A. Gärtner-Rommel, H. Milting, C. Dos Remedios, C. Knosalla, P. Mertins, M. Landthaler, M. Vingron, W. A. Linke, J. G. Seidman, C. E. Seidman, N. Rajewsky, U. Ohler, S. A. Cook, N. Hubner, *Cell* **2019**, *178*, 242-260.e229; d) D. Schlesinger, S. J. Elsässer, *The FEBS Journal* **2021**, *n/a*; e) D. Guerra-Almeida, D. A. Tschoeke, R. Nunes-da-Fonseca, *DNA Research* **2021**, *28*.

[17] H. Hezroni, D. Koppstein, Matthew G. Schwartz, A. Avrutin, David P. Bartel, I. Ulitsky, *Cell Reports* **2015**, *11*, 1110-1122.

[18] a) M. A. Valencia-Sanchez, J. Liu, G. J. Hannon, R. Parker, *Genes Dev* **2006**, *20*, 515-524; b) P. McGillivray, R. Ault, M. Pawashe, R. Kitchen, S. Balasubramanian, M. Gerstein, *Nucleic Acids Res* **2018**, *46*, 3326-3338.

[19] Y. Zhang, J. Qian, C. Gu, Y. Yang, *Signal Transduction and Targeted Therapy* **2021**, *6*, 78.

[20] S. J. Andrews, J. A. Rothnagel, *Nat Rev Genet* **2014**, *15*, 193-204.

[21] T. F. Martinez, Q. Chu, C. Donaldson, D. Tan, M. N. Shokhirev, A. Saghatelian, *Nat Chem Biol* **2020**, *16*, 458-468.

[22] M. K. R. Peeters, G. Menschaert, *Experimental Cell Research* **2020**, *391*, 111923.

[23] B. Wang, Z. Wang, N. Pan, J. Huang, C. Wan, *Int J Mol Sci* **2021**, *22*.

[24] Q. Zhang, E. Wu, Y. Tang, L. Zhang, J. Wang, Y. Hao, B. Zhang, Y. Zhou, X. Guo, J. Luo, T. Cai, R. Chen, F. Yang, *Molecular & Cellular Proteomics* **2021**, 100109.

[25] A. Hecht, J. Glasgow, P. R. Jaschke, L. A. Bawazer, M. S. Munson, J. R. Cochran, D. Endy, M. Salit, *Nucleic Acids Res* **2017**, *45*, 3615-3626.

[26] a) J. Ma, J. K. Diedrich, I. Jungreis, C. Donaldson, J. Vaughan, M. Kellis, J. R. Yates, A. Saghatelian, *Analytical Chemistry* **2016**, *88*, 3967-3975; b) T. Cardon, F. Hervé, V. Delcourt, X. Roucou, M. Salzet, J. Franck, I. Fournier, *Analytical Chemistry* **2020**, *92*, 1122-1129; c) J. Ma, C. C. Ward, I. Jungreis, S. A. Slavoff, A. G. Schwaid, J. Neveu, B. A. Budnik, M. Kellis, A. Saghatelian, *Journal of proteome research* **2014**, *13*, 1757-1765.

[27] L. Cassidy, P. T. Kaulich, A. Tholey, *J Proteome Res* **2019**, *18*, 1725-1734.

[28] C. He, C. Jia, Y. Zhang, P. Xu, *J Proteome Res* **2018**, *17*, 2335-2344.

[29] P. F. Huesgen, P. F. Lange, L. D. Rogers, N. Solis, U. Eckhard, O. Kleifeld, T. Goulas, F. X. Gomis-Rüth, C. M. Overall, *Nat Methods* **2015**, *12*, 55-58.

[30] a) P. T. Kaulich, L. Cassidy, J. Bartel, R. A. Schmitz, A. Tholey, *J Proteome Res* **2021**, *20*, 2895-2903; b) J. Bartel, A. R. Varadarajan, T. Sura, C. H. Ahrens, S. Maaß, D. Becher, *J Proteome Res* **2020**, *19*, 4004-4018.

[31] a) 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, *Nature Chemical Biology* **2013**, *9*, 59-64; b) B. Wang, J. Hao, N. Pan, Z. Wang, Y. Chen, C. Wan, *J Proteomics* **2021**, *230*, 103965.

[32] a) L. Cassidy, A. O. Helbig, P. T. Kaulich, K. Weidenbach, R. A. Schmitz, A. Tholey, *J Proteomics* **2021**, *230*, 103988; b) B. Huraiova, J. Kanovits, S. B. Polakova, L. Cipak, Z. Benko, A. Sevcovicova, D. Anrather, G. Ammerer, C. D. S. Duncan, J. Mata, J. Gregan, *Cell Cycle* **2020**, *19*, 1777-1785; c) C. H. Na, N. Sharma, A. K. Madugundu, R. Chen, M. A. Aksit, G. D. Rosson, G. R. Cutting, A. Pandey, *Molecular & cellular proteomics : MCP* **2019**, *18*, 1382-1395; d) Y. G. Kim, A. M. Lone, A. Saghatelian, *Nat Protoc* **2013**, *8*, 1730-1742.

[33] M. Yang, X. Shang, Y. Zhou, C. Wang, G. Wei, J. Tang, M. Zhang, Y. Liu, J. Cao, Q. Zhang, *Frontiers in Cellular and Infection Microbiology* **2021**, *11*.

[34] S. Prabakaran, M. Hemberg, R. Chauhan, D. Winter, R. Y. Tweedie-Cullen, C. Dittrich, E. Hong, J. Gunawardena, H. Steen, G. Kreiman, J. A. Steen, *Nat Commun* **2014**, *5*, 5429.

[35] W. H. McDonald, J. R. Yates, 3rd, *Curr Opin Mol Ther* **2003**, *5*, 302-309.

[36] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, R. Aebersold, *Mol Cell Proteomics* **2012**, *11*, O111.016717.

[37] A. C. Peterson, J. D. Russell, D. J. Bailey, M. S. Westphall, J. J. Coon, *Molecular & cellular proteomics : MCP* **2012**, *11*, 1475-1488.

[38] N. G. D'Lima, J. Ma, L. Winkler, Q. Chu, K. H. Loh, E. O. Corpuz, B. A. Budnik, J. Lykke-Andersen, A. Saghatelian, S. A. Slavoff, *Nat Chem Biol* **2017**, *13*, 174-180.

[39] M. R. Hemm, J. Weaver, G. Storz, *EcoSal Plus* **2020**, *9*.

[40] H. Pak, J. Michaux, F. Huber, C. Chong, B. J. Stevenson, M. Müller, G. Coukos, M. Bassani-Sternberg, *Molecular & cellular proteomics : MCP* **2021**, *20*, 100080.

[41] Y. Zhu, L. M. Orre, H. J. Johansson, M. Huss, J. Boekel, M. Vesterlund, A. Fernandez-Woodbridge, R. M. M. Branca, J. Lehtiö, *Nat Commun* **2018**, *9*, 903.

[42] C. Chong, M. Müller, H. Pak, D. Harnett, F. Huber, D. Grun, M. Leleu, A. Auger, M. Arnaud, B. J. Stevenson, J. Michaux, I. Bilic, A. Hirsekorn, L. Calviello, L. Simó-Riudalbas, E. Planet, J. Lubiński, M. Bryśkiewicz, M. Wiznerowicz, I. Xenarios, L. Zhang, D. Trono, A. Harari, U. Ohler, G. Coukos, M. Bassani-Sternberg, *Nat Commun* **2020**, *11*, 1293.

[43] V. Delcourt, M. Brunelle, A. V. Roy, J. F. Jacques, M. Salzet, I. Fournier, X. Roucou, *Mol Cell Proteomics* **2018**, *17*, 2402-2411.

[44] a) C. Chen, J. Hou, J. J. Tanner, J. Cheng, *Int J Mol Sci* **2020**, *21*, 2873; b) R. Matthiesen, O. N. Jensen, *Methods in molecular biology (Clifton, N.J.)* **2008**, *453*, 105-122; c) A. I. Nesvizhskii, *Nature Methods* **2014**, *11*, 1114-1125.

[45] K. V. Ruggles, K. Krug, X. Wang, K. R. Clauser, J. Wang, S. H. Payne, D. Fenyö, B. Zhang, D. R. Mani, *Molecular & cellular proteomics : MCP* **2017**, *16*, 959-981.

[46] P. Dainese, W. Staudenmann, M. Quadroni, C. Korostensky, G. Gonnet, M. Kertesz, P. James, *Electrophoresis* **1997**, *18*, 432-442.

[47] J. T. Ferguson, C. D. Wenger, W. W. Metcalf, N. L. Kelleher, *J Am Soc Mass Spectrom* **2009**, *20*, 1743-1750.

[48] S. Wang, L. Tian, H. Liu, X. Li, J. Zhang, X. Chen, X. Jia, X. Zheng, S. Wu, Y. Chen, J. Yan, L. Wu, *Molecular Plant* **2020**, *13*, 1078-1093.

[49] K. Krug, A. Carpy, G. Behrends, K. Matic, N. C. Soares, B. Macek, *Molecular & cellular proteomics : MCP* **2013**, *12*, 3420-3430.

[50] a) E. Qeli, C. H. Ahrens, *Nature Biotechnology* **2010**, *28*, 647-650; b) U. Omasits, A. R. Varadarajan, M. Schmid, S. Goetze, D. Melidis, M. Bourqui, O. Nikolayeva, M. Québatte, A. Patrignani, C. Dehio, J. E. Frey, M. D. Robinson, B. Wollscheid, C. H. Ahrens, *Genome Res* **2017**, *27*, 2083-2095.

[51] a) S. D. Mackowiak, H. Zauber, C. Bielow, D. Thiel, K. Kutz, L. Calviello, G. Mastrobuoni, N. Rajewsky, S. Kempa, M. Selbach, B. Obermayer, *Genome biology* **2015**, *16*, 179-179; b) L. Cassidy, D. Prasse, D. Linke, R. A. Schmitz, A. Tholey, *Journal of proteome research* **2016**, *15*, 3773-3783; c) X. Cao, A. Khitun, Z. Na, D. G. Dumitrescu, M. Kubica, E. Olatunji, S. A. Slavoff, *Journal of proteome research* **2020**, *19*, 3418-3426.

[52] J. Crappé, W. Van Criekinge, G. Trooskens, E. Hayakawa, W. Luyten, G. Baggerman, G. Menschaert, *BMC genomics* **2013**, *14*, 648-648.

[53] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, J. S. Weissman, *Science* **2009**, *324*, 218-223.

[54] P. Zhang, D. He, Y. Xu, J. Hou, B.-F. Pan, Y. Wang, T. Liu, C. M. Davis, E. A. Ehli, L. Tan, F. Zhou, J. Hu, Y. Yu, X. Chen, T. M. Nguyen, J. M. Rosen, D. H. Hawke, Z. Ji, Y. Chen, *Nature Communications* **2017**, *8*, 1749.

[55] a) N. T. Ingolia, L. F. Lareau, J. S. Weissman, *Cell* **2011**, *147*, 789-802; b) P. Juntawong, T. Girke, J. Bazin, J. Bailey-Serres, *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, E203-E212; c) G. A. Brar, M. Yassour, N. Friedman, A. Regev, N. T. Ingolia, J. S. Weissman, *Science* **2012**, *335*, 552-557; d) Jenna E. Smith, Juan R. Alvarez-Dominguez, N. Kline, Nathan J. Huynh, S. Geisler, W. Hu, J. Coller, Kristian E. Baker, *Cell Reports* **2014**, *7*, 1858-1866.

[56] N. T. Ingolia, G. A. Brar, N. Stern-Ginossar, M. S. Harris, G. J. Talhouarne, S. E. Jackson, M. R. Wills, J. S. Weissman, *Cell Rep* **2014**, *8*, 1365-1379.

[57] L. Calviello, N. Mukherjee, E. Wyler, H. Zauber, A. Hirsekorn, M. Selbach, M. Landthaler, B. Obermayer, U. Ohler, *Nature Methods* **2016**, *13*, 165-170.

[58] S. Li, S. W. Cha, K. Heffner, D. B. Hizal, M. A. Bowen, R. Chaerkady, R. N. Cole, V. Tejwani, P. Kaushik, M. Henry, P. Meleady, S. T. Sharfstein, M. J. Betenbaugh, V. Bafna, N. E. Lewis, *Journal of proteome research* **2019**, *18*, 2433-2445.

[59] a) B. Malone, I. Atanassov, F. Aeschimann, X. Li, H. Großhans, C. Dieterich, *Nucleic Acids Res.* **2017**, *45*, 2960-2972; b) Z. Xu, L. Hu, B. Shi, S. Geng, L. Xu, D. Wang, Z. J. Lu, *Nucleic Acids Res.* **2018**, *46*, e109-e109; c) S. Choudhary, W. Li, D. S. A, *Bioinformatics (Oxford, England)* **2020**, *36*, 2053-2059; d) G. Monteuuis, A. Miścicka, M. Świrski, L. Zenad, O. Niemitalo, L. Wrobel, J. Alam, A. Chacinska, A. J. Kastaniotis, J. Kufel, *Nucleic Acids Res* **2019**, *47*, 5777-5791; e) H. Wang, Y. Wang, J. Yang, Q. Zhao, N. Tang, C. Chen, H. Li, C. Cheng, M. Xie, Y. Yang, Z. Xie, *Nucleic Acids Res.* **2021**, *49*, 6165-6180.

[60] M. V. Ruiz Cuevas, M.-P. Hardy, J. Hollý, É. Bonneil, C. Durette, M. Courcelles, J. Lanoix, C. Côté, L. M. Staudt, S. Lemieux, P. Thibault, C. Perreault, J. W. Yewdell, *Cell Reports* **2021**, *34*, 108815.

[61] Y. Liang, W. Zhu, S. Chen, J. Qian, L. Li, *Front Plant Sci* **2021**, *12*, 695439-695439.

[62] J. L. Aspden, Y. C. Eyre-Walker, R. J. Phillips, U. Amin, M. A. S. Mumtaz, M. Brocard, J.-P. Couso, *eLife* **2014**, *3*, e03528.

[63] A. P. Fields, E. H. Rodriguez, M. Jovanovic, N. Stern-Ginossar, B. J. Haas, P. Mertins, R. Raychowdhury, N. Hacohen, S. A. Carr, N. T. Ingolia, A. Regev, J. S. Weissman, *Mol Cell* **2015**, *60*, 816-827.

[64] Z. Ji, R. Song, A. Regev, K. Struhl, *eLife* **2015**, *4*, e08890.

[65] Z. Xiao, R. Huang, X. Xing, Y. Chen, H. Deng, X. Yang, *Nucleic Acids Res.* **2018**, *46*, e61-e61.

[66] L. Calviello, A. Hirsekorn, U. Ohler, *Nat Struct Mol Biol* **2020**, *27*, 717-725.

[67] J. Chen, A.-D. Brunner, J. Z. Cogan, J. K. Nuñez, A. P. Fields, B. Adamson, D. N. Itzhak, J. Y. Li, M. Mann, M. D. Leonetti, J. S. Weissman, *Science* **2020**, *367*, 1140-1146.

[68] M. A. Brunet, J. F. Lucier, M. Levesque, S. Leblanc, J. F. Jacques, H. R. H. Al-Saedi, N. Guilloy, F. Grenier, M. Avino, I. Fournier, M. Salzet, A. Ouangraoua, M. S. Scott, F. M. Boisvert, X. Roucou, *Nucleic Acids Res* **2021**, *49*, D380-d388.

[69] V. Olexiouk, W. Van Criekinge, G. Menschaert, *Nucleic Acids Res* **2018**, *46*, D497-d502.

[70] R. R. Hazarika, B. De Coninck, L. R. Yamamoto, L. R. Martin, B. P. Cammue, V. van Noort, *BMC Bioinformatics* **2017**, *18*, 37.

[71] Y. Chen, D. Li, W. Fan, X. Zheng, Y. Zhou, H. Ye, X. Liang, W. Du, Y. Zhou, K. Wang, *Plant Biotechnology Journal* **2020**, *18*, 2158-2160.

[72] S. A. Choteau, A. Wagner, P. Pierre, L. Spinelli, C. Brun, *Database : the journal of biological databases and curation* **2021**, *2021*.

[73] H. Yang, H. Chi, W. F. Zeng, W. J. Zhou, S. M. He, *Bioinformatics (Oxford, England)* **2019**, *35*, i183-i190.

[74] B. Ma, *J Am Soc Mass Spectrom* **2015**, *26*, 1885-1894.

[75] A. Frank, P. Pevzner, *Anal Chem* **2005**, *77*, 964-973.

[76] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, G. Lajoie, *Rapid Commun Mass Spectrom* **2003**, *17*, 2337-2342.

[77] A. Devabhaktuni, S. Lin, L. Zhang, K. Swaminathan, C. G. Gonzalez, N. Olsson, S. M. Pearlman, K. Rawson, J. E. Elias, *Nature Biotechnology* **2019**, *37*, 469-479.

[78] I. O'Bryon, S. C. Jenson, E. D. Merkley, *Protein Sci* **2020**, *29*, 1864-1878.

[79] L. Chen, Y. Zhang, Y. Yang, Y. Yang, H. Li, X. Dong, H. Wang, Z. Xie, Q. Zhao, *J Am Soc Mass Spectrom* **2021**.

[80] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, *BMC Bioinformatics* **2009**, *10*, 421.

[81] T. Muth, B. Y. Renard, *Briefings in bioinformatics* **2018**, *19*, 954-970.

[82] T. Muth, F. Hartkopf, M. Vaudel, B. Y. Renard, *Proteomics* **2018**, *18*, e1700150.

[83] F. Erhard, L. Dölken, B. Schilling, A. Schlosser, *Cancer Immunol Res* **2020**, *8*, 1018-1026.

[84] A. Chugunova, T. Navalayeu, O. Dontsova, P. Sergiev, *Journal of proteome research* **2018**, *17*, 1-11.

[85] M. F. Lin, I. Jungreis, M. Kellis, *Bioinformatics (Oxford, England)* **2011**, *27*, i275-282.

[86] S. Washietl, S. Findeiss, S. A. Müller, S. Kalkhof, M. von Bergen, I. L. Hofacker, P. F. Stadler, N. Goldman, *Rna* **2011**, *17*, 578-594.

[87] A. Skarshewski, M. Stanton-Cook, T. Huber, S. Al Mansoori, R. Smith, S. A. Beatson, J. A. Rothnagel, *BMC Bioinformatics* **2014**, *15*, 36.

[88] A. A. Bazzini, T. G. Johnstone, R. Christiano, S. D. Mackowiak, B. Obermayer, E. S. Fleming, C. E. Vejnar, M. T. Lee, N. Rajewsky, T. C. Walther, A. J. Giraldez, *The EMBO Journal* **2014**, *33*, 981-993.

[89] J. H. Badger, G. J. Olsen, *Mol Biol Evol* **1999**, *16*, 512-524.

[90] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, D. Haussler, *Genome Res* **2005**, *15*, 1034-1050.

[91] K. Hanada, K. Akiyama, T. Sakurai, T. Toyoda, K. Shinozaki, S.-H. Shiu, *Bioinformatics (Oxford, England)* **2009**, *26*, 399-400.

[92] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J Mol Biol* **1990**, *215*, 403-410.

[93] S. R. Eddy, *Proc Int Conf Intell Syst Mol Biol* **1995**, *3*, 114-120.

[94] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. E. Tosatto, R. D. Finn, *Nucleic Acids Res* **2019**, *47*, D427-d432.

[95] a) Y. Zhang, C. Jia, M. J. Fullwood, C. K. Kwoh, *Briefings in bioinformatics* **2020**, *22*, 2073-2084; b) M. Zhu, M. Gribskov, *BMC Bioinformatics* **2019**, *20*, 559.

[96] I. P. Ivanov, A. E. Firth, A. M. Michel, J. F. Atkins, P. V. Baranov, *Nucleic Acids Res.* **2011**, *39*, 4220-4234.

[97] a) P. Puntervoll, R. Linding, C. Gemünd, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D. M. Martin, G. Ausiello, B. Brannetti, A. Costantini, F. Ferrè, V. Maselli, A. Via, G. Cesareni, F. Diella, G. Superti-Furga, L. Wyrwicz, C. Ramu, C. McGuigan, R. Gudavalli, I. Letunic, P. Bork, L. Rychlewski, B. Küster, M. Helmer-Citterich, W. N. Hunter, R. Aasland, T. J. Gibson, *Nucleic Acids Res* **2003**, *31*, 3625-3630; b) M. Gouw, S. Michael, H. Sámano-Sánchez, M. Kumar, A. Zeke, B. Lang, B. Bely, L. B. Chemes, N. E. Davey, Z. Deng, F. Diella, C. M. Gürth, A. K. Huber, S. Kleinsorg, L. S. Schlegel, N. Palopoli, K. V. Roey, B. Altenberg, A. Reményi, H. Dinkel, T. J. Gibson, *Nucleic Acids Res* **2018**, *46*, D428-d434; c) M. Kumar, M. Gouw, S. Michael, H. Sámano-Sánchez, R. Pancsa, J. Glavina, A. Diakogianni, J. A. Valverde, D. Bukirova, J. Čalyševa, N. Palopoli, N. E. Davey, L. B. Chemes, T. J. Gibson, *Nucleic Acids Res.* **2019**, *48*, D296-D306.

[98] A. P. Camargo, V. Sourkov, Gonçalo A G. Pereira, Marcelo F. Carazzolle, *NAR Genomics and Bioinformatics* **2020**, *2*.

[99] a) Y. Pang, Z. Liu, H. Han, B. Wang, W. Li, C. Mao, S. Liu, *Journal of Hepatology* **2020**, *73*, 1155-1169; b) X. L. Li, L. Pongor, W. Tang, S. Das, B. R. Muys, M. F. Jones, S. B. Lazar, E. A. Dangelmaier, C. C. R. Hartford, I. Grammatikakis, Q. Hao, Q. Sun, A. Schetter, J. L. Martindale, B. Tang, L. M. Jenkins, A. I. Robles, R. L. Walker, S. Ambs, R. Chari, S. A. Shabalina, M. Gorospe, S. P. Hussain, C. C. Harris, P. S. Meltzer, K. V. Prasanth, M. I. Aladjem, T. Andresson, A. Lal, *eLife* **2020**, *9*, e53734; c) S. Begum, A. Yiu, J. Stebbing, L. Castellano, *Oncogene* **2018**, *37*, 4055-4057; d) I. Unk, I. Hajdú, K. Fátyol, B. Szakál, A. Blastyák, V. Bermudez, J. Hurwitz, L. Prakash, S. Prakash, L. Haracska, *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 18107-18112; e) M. Zhang, K. Zhao, X. Xu, Y. Yang, S. Yan, P. Wei, H. Liu, J. Xu, F. Xiao, H. Zhou, X. Yang, N. Huang, J. Liu, K. He, K. Xie, G. Zhang, S. Huang, N. Zhang, *Nature Communications* **2018**, *9*, 4475; f) Y. Yang, X. Gao, M. Zhang, S. Yan, C. Sun, F. Xiao, N. Huang, X. Yang, K. Zhao, H. Zhou, S. Huang, B. Xie, N. Zhang, *J Natl Cancer Inst* **2018**, *110*, 304-315.

[100] a) Y. Chen, L. Ho, V. Tergaonkar, *Cancer Letters* **2021**, *500*, 263-270; b) D. M. Anderson, C. A. Makarewich, K. M. Anderson, J. M. Shelton, S. Bezprozvannaya, R. Bassel-Duby, E. N. Olson, *Science Signaling* **2016**, *9*, ra119-ra119; c) B. R. Nelson, C. A. Makarewich, D. M. Anderson, B. R. Winders, C. D. Troupes, F. Wu, A. L. Reese, J. R. McAnally, X. Chen, E. T. Kavalali, S. C. Cannon, S. R. Houser, R. Bassel-Duby, E. N. Olson, *Science* **2016**, *351*, 271-275; d) A. Matsumoto, A. Pasut, M. Matsumoto, R. Yamashita, J. Fung, E. Monteleone, A. Saghatelian, K. I. Nakayama, J. G. Clohessy, P. P. Pandolfi, *Nature* **2017**, *541*, 228-232; e) P. Bi, A. Ramirez-Martinez, H. Li, J. Cannavino, J. R. McAnally, J. M. Shelton, E. Sánchez-Ortiz, R. Bassel-Duby, E. N. Olson, *Science* **2017**, *356*, 323-327; f) B. Bonilauri, B. Dallagiovanna, *Front Physiol* **2020**, *11*, 567614.

[101] M. Koh, I. Ahmad, Y. Ko, Y. Zhang, T. F. Martinez, J. K. Diedrich, Q. Chu, J. J. Moresco, M. A. Erb, A. Saghatelian, P. G. Schultz, M. J. Bollong, *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, e2021943118.

[102] J. Li, C. Liu, *Frontiers in Genetics* **2019**, *10*.

[103] a) J.-Z. Huang, M. Chen, D. Chen, X.-C. Gao, S. Zhu, H. Huang, M. Hu, H. Zhu, G.-R. Yan, *Molecular Cell* **2017**, *68*, 171-184.e176; b) V. Delcourt, J. Franck, E. Leblanc, F. Narducci, Y.-M. Robin, J.-P. Gimeno, J. Quanico, M. Wisztorski, F. Kobeissy, J.-F. Jacques, X. Roucou, M. Salzet, I. Fournier, *EBioMedicine* **2017**, *21*, 55-64.

[104] a) R. Jackson, L. Kroehling, A. Khitun, W. Bailis, A. Jarret, A. G. York, O. M. Khan, J. R. Brewer, M. H. Skadow, C. Duizer, C. C. D. Harman, L. Chang, P. Bielecki, A. G. Solis, H. R. Steach, S. Slavoff, R. A. Flavell, *Nature* **2018**, *564*, 434-438; b) Q. Chu, T. F. Martinez, S. W. Novak, C. J. Donaldson, D. Tan, J. M. Vaughan, T. Chang, J. K. Diedrich, L. Andrade, A. Kim, T. Zhang, U. Manor, A. Saghatelian, *Nature Communications* **2019**, *10*, 4883.

[105] P. Wu, Y. Mo, M. Peng, T. Tang, Y. Zhong, X. Deng, F. Xiong, C. Guo, X. Wu, Y. Li, X. Li, G. Li, Z. Zeng, W. Xiong, *Molecular Cancer* **2020**, *19*, 22.

[106] S. W. Choi, H. W. Kim, J. W. Nam, *Briefings in bioinformatics* **2019**, *20*, 1853-1864.

[107] Y. Yan, R. Tang, B. Li, L. Cheng, S. Ye, T. Yang, Y.-C. Han, C. Liu, Y. Dong, L.-H. Qu, K. O. Lui, J.-H. Yang, Z.-P. Huang, *Molecular Therapy* **2021**, *29*, 2253-2267.

[108] A. Khitun, T. J. Ness, S. A. Slavoff, *Molecular Omics* **2019**, *15*, 108-116.

[109] S. Kong, M. Tao, X. Shen, S. Ju, *Cancer Letters* **2020**, *483*, 59-65.

[110] C. C. R. Hartford, A. Lal, *Molecular and Cellular Biology* **2020**, *40*, e00528-00519.

[111] N. R. Pamudurti, O. Bartok, M. Jens, R. Ashwal-Fluss, C. Stottmeister, L. Ruhe, M. Hanan, E. Wyler, D. Perez-Hernandez, E. Ramberger, S. Shenzis, M. Samson, G. Dittmar, M. Landthaler, M. Chekulaeva, N. Rajewsky, S. Kadener, *Molecular Cell* **2017**, *66*, 9-21.e27.

[112] X. Gao, X. Xia, F. Li, M. Zhang, H. Zhou, X. Wu, J. Zhong, Z. Zhao, K. Zhao, D. Liu, F. Xiao, Q. Xu, T. Jiang, B. Li, S.-Y. Cheng, N. Zhang, *Nature Cell Biology* **2021**, *23*, 278-291.

[113] D. P. Bartel, *Cell* **2018**, *173*, 20-51.

[114] a) L. Niu, F. Lou, Y. Sun, L. Sun, X. Cai, Z. Liu, H. Zhou, H. Wang, Z. Wang, J. Bai, Q. Yin, J. Zhang, L. Chen, D. Peng, Z. Xu, Y. Gao, S. Tang, L. Fan, H. Wang, *Science Advances* **2020**, *6*, eaaz2059; b) L. Adams, *Nature Reviews Genetics* **2017**, *18*, 145-145.

[115] G. M. Rice, V. Shivashankar, E. J. Ma, J. L. Baryza, R. Nutiu, *Molecular Cell* **2020**, *80*, 892-902.e894.

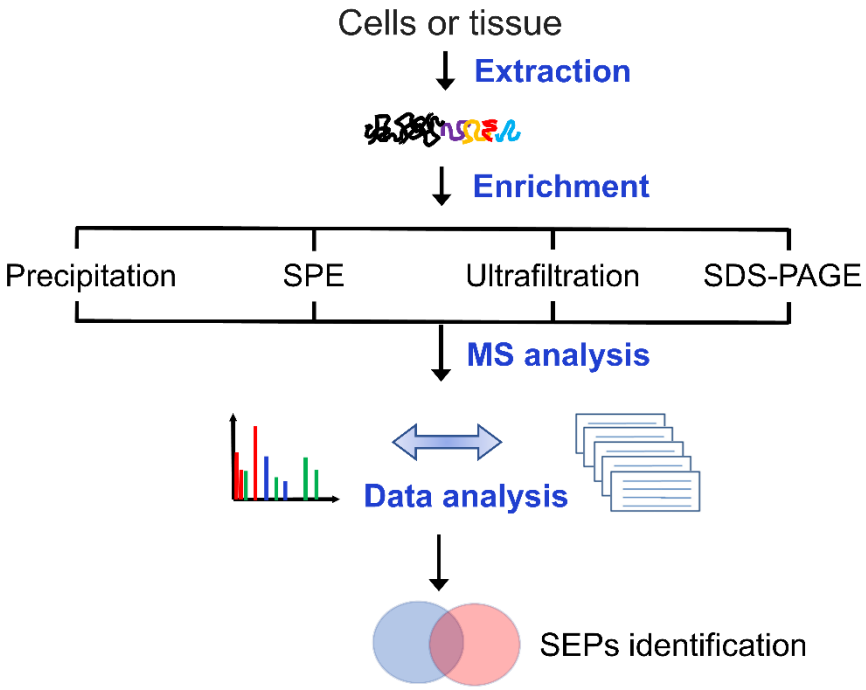
[116] J. Fang, S. Morsalin, V. N. Rao, E. S. P. Reddy, *Journal of Pharmaceutical Sciences and Pharmacology* **2017**, *3*, 23-27.

[117] a) E. Duvallet, M. Boulpicante, T. Yamazaki, C. Daskalogianni, R. Prado Martins, S. Baconnais, B. Manoury, R. Fahraeus, S. Apcher, *Oncoimmunology* **2016**, *5*, e1198865; b) S. Apcher, G. Millot, C. Daskalogianni, A. Scherl, B. Manoury, R. Fåhraeus, *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 17951-17956; c) S. R. Starck, N. Shastri, *Cell Mol Life Sci* **2011**, *68*, 1471-1479; d) S. R. Schwab, K. C. Li, C. Kang, N. Shastri, *Science* **2003**, *301*, 1367-1371; e) S. A. Rosenberg, P. Tong-On, Y. Li, J. P. Riley, M. El-Gamil, M. R. Parkhurst, P. F. Robbins, *Journal of immunology (Baltimore, Md. : 1950)* **2002**, *168*, 2402-2407; f) A. O. Weinzierl, D. Maurer, F. Altenberend, N. Schneiderhan-Marra, K. Klingel, O. Schoor, D. Wernet, T. Joos, H. G. Rammensee, S. Stevanović, *Cancer research* **2008**, *68*, 2447-2454.

[118] C. M. Laumont, K. Vincent, L. Hesnard, É. Audemard, É. Bonneil, J.-P. Laverdure, P. Gendron, M. Courcelles, M.-P. Hardy, C. Côté, C. Durette, C. St-Pierre, M. Benhammadi, J. Lanoix, S. Vobecky, E. Haddad, S. Lemieux, P. Thibault, C. Perreault, *Science Translational Medicine* **2018**, *10*, eaau5516.

[119] T. Ouspenskaia, T. Law, K. R. Clauser, S. Klaeger, S. Sarkizova, F. Aguet, B. Li, E. Christian, B. A. Knisbacher, P. M. Le, C. R. Hartigan, H. Keshishian, A. Apffel, G. Oliveira, W. Zhang, Y. T. Chow, Z. Ji, S. A. Shukla, P. Bachireddy, G. Getz, N. Hacohen, D. B. Keskin, S. A. Carr, C. J. Wu, A. Regev, *bioRxiv* **2020**, 2020.2002.2012.945840.

Entry for the Table of Contents



Small open reading frames (sORFs) were historically annotated as noncoding or even junk sequences. Accumulating evidence suggests the existence of sORFs-encoded peptides (SEPs) in recent years. Here we discuss the latest advance in methodologies for identifying SEPs with mass spectrometry, as well as the progress on functional studies of SEPs.