




Review

Small Open Reading Frame-Encoded Micro-Peptides: An Emerging Protein World

Xiaoping Dong^{1,2}, Kun Zhang³, Chengfeng Xun^{1,2}, Tianqi Chu^{1,2}, Songping Liang^{1,2}, Yong Zeng^{1,2,3,*} 
and Zhonghua Liu^{1,2,*}

¹ National & Local Joint Engineering Laboratory of Animal Peptide Drug Development, College of Life Sciences, Hunan Normal University, Changsha 410081, China

² Peptide and Small Molecule Drug R&D Platform, Furong Laboratory, Hunan Normal University, Changsha 410081, China

³ The State Key Laboratory of Developmental Biology of Freshwater Fish, College of Life Science, Hunan Normal University, Changsha 410081, China

* Correspondence: yongz@hunnu.edu.cn (Y.Z.); liuzh@hunnu.edu.cn (Z.L.)

Abstract: Small open reading frames (sORFs) are often overlooked features in genomes. In the past, they were labeled as noncoding or “transcriptional noise”. However, accumulating evidence from recent years suggests that sORFs may be transcribed and translated to produce sORF-encoded polypeptides (SEPs) with less than 100 amino acids. The vigorous development of computational algorithms, ribosome profiling, and peptidome has facilitated the prediction and identification of many new SEPs. These SEPs were revealed to be involved in a wide range of basic biological processes, such as gene expression regulation, embryonic development, cellular metabolism, inflammation, and even carcinogenesis. To effectively understand the potential biological functions of SEPs, we discuss the history and development of the newly emerging research on sORFs and SEPs. In particular, we review a range of recently discovered bioinformatics tools for identifying, predicting, and validating SEPs as well as a variety of biochemical experiments for characterizing SEP functions. Lastly, this review underlines the challenges and future directions in identifying and validating sORFs and their encoded micropeptides, providing a significant reference for upcoming research on sORF-encoded peptides.



Citation: Dong, X.; Zhang, K.; Xun, C.; Chu, T.; Liang, S.; Zeng, Y.; Liu, Z. Small Open Reading Frame-Encoded Micro-Peptides: An Emerging Protein World. *Int. J. Mol. Sci.* **2023**, *24*, 10562. <https://doi.org/10.3390/ijms241310562>

Academic Editor: Stanislaw Oldziej

Received: 13 May 2023

Revised: 20 June 2023

Accepted: 21 June 2023

Published: 23 June 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: small (short) open reading frame; sORF; sORF-encoded peptides; SEPs; micro-proteins; peptidome; coding potential prediction

1. Introduction

According to the ENCODE database, about 2–3% of the human genome is composed of protein-coding genes, and more than 80% are viewed as ncRNAs [1,2]. With the advanced development of high-throughput sequencing technology, more and more diverse ncRNAs have been discovered to be involved in numerous essential biological processes, such as genomic regulation [3], environmental responses [4], and body development [5]. Generally, ncRNAs are classified into long non-coding RNAs (lncRNAs), small RNAs (miRNAs), piwi-interacting RNAs (piRNAs), circular RNAs (cirRNAs) and others. They were initially considered “transcriptional noise” [6,7]. However, research has reversed the view that ncRNAs represent “junk” transcription products [8]. One or more short open reading frames (sORFs), which rarely use AUG as the start codon, may be present in these ncRNAs. The majority are initiated by near-homologous codons (meaning codons that differ from AUGs by one nucleotide), such as CUG, GUG, UUG, and ACG [9]. These sORFs may encode small proteins with less than 100 amino acids, and various professional terms have been used to describe these proteins, such as micropeptides, small peptides, microproteins, sORF-encoded peptides (SEPs), etc.

Typically, an ORF is defined as a segment of conserved and non-overlapping nucleotide triplets (codons) that can be translated into a functionally annotated protein [10]. Eukaryotic messenger RNAs (mRNAs) usually contain a main ORF that produces protein-coding regions. However, the traditional genetic rules, such as amino acid conservation and homology, the absolute requirement for a starting codon (methionine), and the minimum translation length, have greatly limited the identification of transcripts with non-canonical protein-coding capabilities. Therefore, we regarded these proteins encoded by previously neglected open reading frames with fewer than 300 nucleotides (nt) as sORF-encoded peptides (SEPs). SEPs are biologically active molecules that range from highly conserved to primate-specific [11], implying that they perform both basal and species-specific functions. To date, SEPs have been found to function in a variety of biological processes, including embryogenesis [12–14], myogenesis [15–17], cellular metabolism [18,19], inflammation [20–22], and carcinogenesis [23–26].

Due to the limitations of the conservation screening mechanism and detection sensitivity, SEPs with a small molecular weight and a low expression abundance are often overlooked, which may lead to many crucial regulatory mechanisms being “hidden”. Therefore, it is challenging to determine the potential functional roles of such micro-proteins. With the increasing interest in SEPs, a large number of new ORF translation products have been identified and validated. In summary, this reflects the diversity of SEPs under different physiological conditions. It is urgent to identify and characterize their functional roles, which may reveal many new molecules involved in regulatory mechanisms.

2. Localities and Characteristics of sORFs and SEPs

In recent years, extensive translations of sORFs at genomic locations in animal, plant, fungal, and bacterial species have been revealed based on high-throughput next-generation sequencing technologies [27–29]. These sORFs can be located within coding transcripts such as 5' UTR (5' untranslated regions), CDS (coding sequences), 3' UTR (3' untranslated regions) or even within non-coding RNAs, such as long non-coding RNAs (lncRNAs), circRNAs, and mitochondrial RNAs (mtRNA) (Figure 1A). sORFs are essentially hidden genomic features in the organism [30]. Therefore, it is possible to find new proteins with interesting functions.

As we all know, non-coding RNAs (ncRNA) include long non-coding RNAs (lncRNAs, longer than 200 nt) and small non-coding RNAs (sncRNAs). Many important physiological processes have been found to be regulated by the micro-peptides translated by lncRNAs [21,31,32]. Traditionally, upstream open reading frames (uORFs) are located upstream of protein-coding genes and are considered as cis-acting elements for downstream expression through a mechanism similar to competitive translation [33]. Beyond these, recent studies have shown that uORFs can encode functional micro-peptides. Like uORFs, small peptides encoded by dORFs (downstream open reading frames) are usually not conserved, and the effects of the dORFs are not dependent on the small peptides, but on the translational activity of the dORFs themselves [34]. CircRNAs often function as miRNA sponges and play roles in transcriptional regulation and protein binding. CircRNAs have been shown to have the ability to translate in recent years [35–37]. In addition, sORF-encoded peptides (SEPs) were discovered in pseudogenes [38] as well as in intergenic regions [39].

Bioinformatic predictions and MS-based proteomics approaches have been used to predict and identify SEPs with different lengths and start codons. Wang et al. identified 1682 peptides from 2544 human sORFs in Hep3B cells using a de novo approach combined with RNA-Seq [40]. Several online sORF databases such as Smprot [41], sORF [42], and OpenProt [43] have been constructed. Unexpectedly, a large proportion of SEPs are translated with non-AUG initiation codons. Usually, alternative start codons only differ from AUG by one nucleotide (e.g., CUG, GUG and UUG). It has been shown that these non-classical start codons are homologous to the classical start codon ATG, which is often located near the Kozak sequence [44]. Another theory suggests that the non-classical start codon of sORFs is derived from the RNA editing of post-transcriptional mRNA, which

converts uracil (U) to cytosine (C) in the transcription product initiation codon AUG by the action of RNA editing enzymes, thus converting the classical start codon to a homologous non-classical codon and regulating the translation efficiency of sORFs [45]. Additionally, it was discovered that SEPs have similar length ranges, but slightly different distributions. A possible explanation for this variation is the use of different scoring algorithms and computational software (Figure 1B,C). However, these SEPs with less than 100 amino acids in length deserve further investigation.

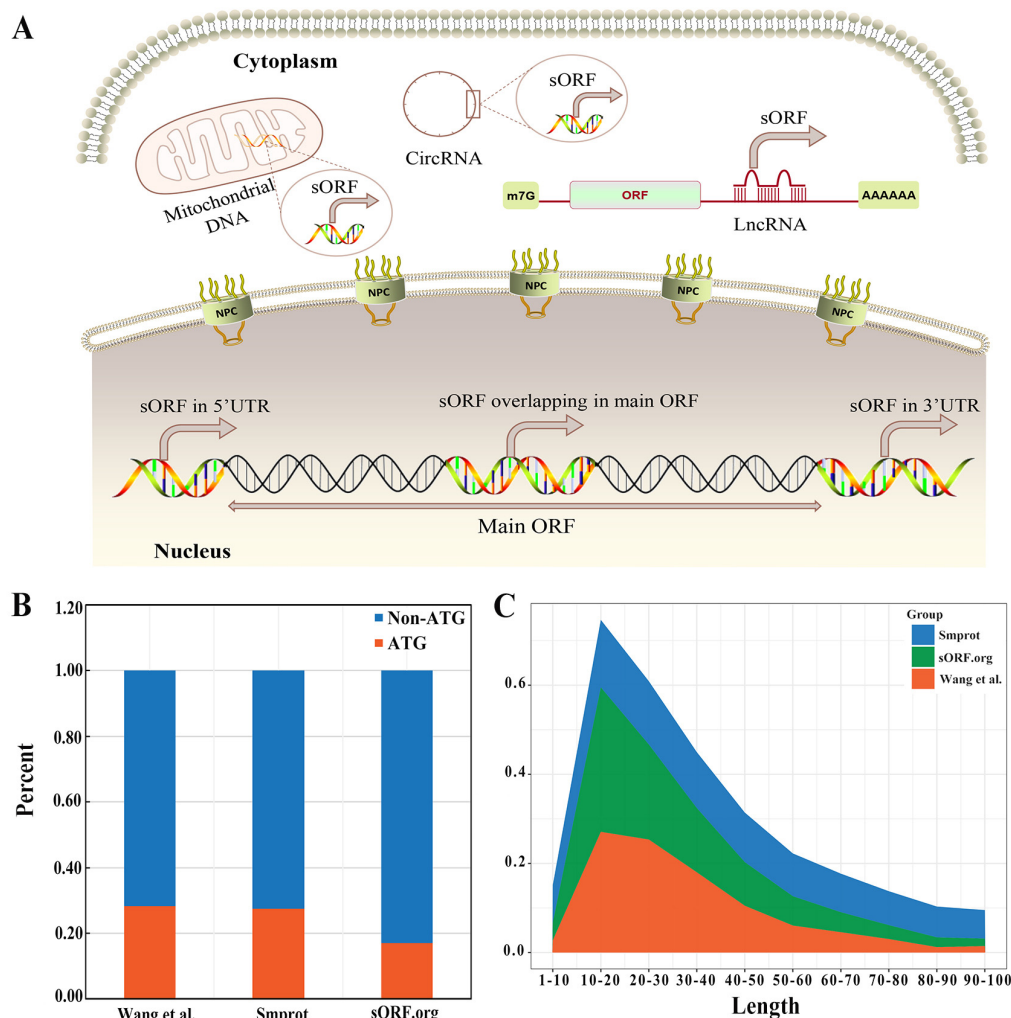


Figure 1. Localities and characteristics of sORFs and SEPs. (A) Examples of sORFs within coding transcripts (5' UTR, CDS, 3'UTR) or even within non-coding RNAs (circRNA, lncRNA, mitochondrial RNA); (B) the percentage of ATG and non-ATG start codons in public databases and a representative study; (C) the AA length distribution of SEPs.

3. Ribosome Profiling (Ribo-Seq) for Identification of SEPs

Ribosome profiling is an emerging technique that uses deep sequencing to monitor *in vivo* translation and provides a systematic method for the experimental annotation of coding regions. The whole workflow is designed to degrade the RNA that is not protected by ribosomes using RNA enzymes before centrifuging to separate the ribosome-protected mRNA fragments. These 30 nt footprints can be directly mapped to the original mRNA by deep sequencing and further used to pinpoint the precise location of the ribosomes during translation (Figure 2). However, Wilson et al. demonstrated that not all sORFs bound to ribosomes are translated [46]. In order to separate the mRNA bound to multiple ribosomes

and distinguish single ribosome–mRNA complexes that are not translated, poly-Ribo-Seq was developed. The technology provides more concrete evidence of active translation.

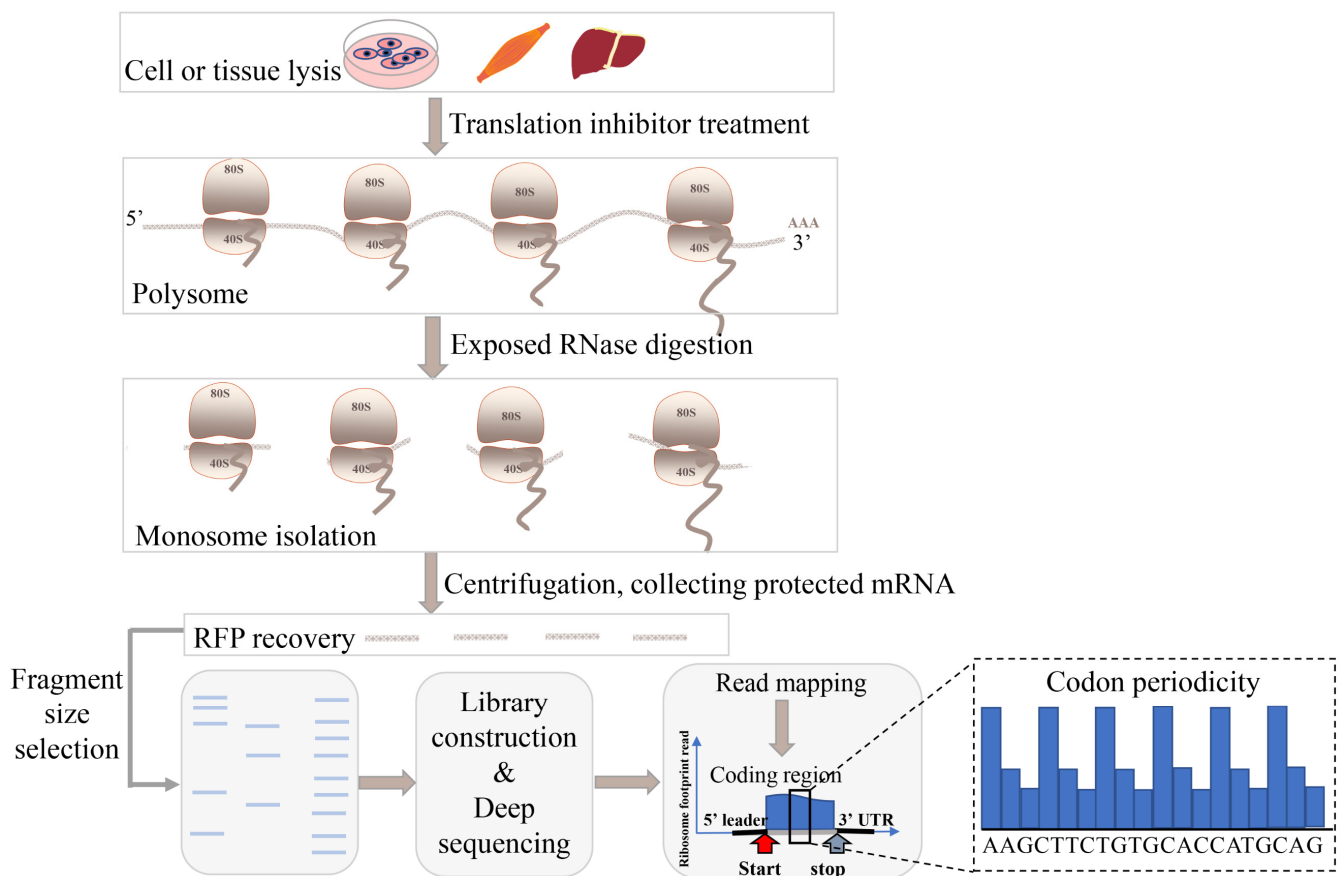


Figure 2. Ribosome profiling process, where ribosome footprints are obtained for deep sequencing. Isolation of ribosome-bound mRNAs is conducted through treatment of non-specific nucleases such as RNase I. Ribosome footprints (showing positioning between start and stop codon of gene) are then used for library generation and deep sequencing.

Ribosome profiling has proven to be a powerful technique to explore the translation potential of sORFs by using multiple pipelines (Table 1). Previous studies have presented an experimental and analytical framework for the systematic identification and quantification of translation based on ORF-RATER [47]. RiboTaper can detect regions of active translation based on three-nucleotide periodicity [48]. Calviello et al. used RiboTaper to identify 218 novel proteins in Chinese hamster tissue and CHO cell lines [49]. As the technology has advanced and matured, more analytical tools such as Ribowave [50], RibORF [51], and RiboCode [52] have been used to support the reference database construction for mining SEPs from MS data.

While ribosome sequencing can provide a landscape of ribosome occupancy throughout the transcriptome, its sequencing data can provide information on where translation occurs and quantitative information, such as how much of the region is occupied by ribosomes. Ribo-Seq does have some limitations. Firstly, Ribo-Seq requires the rapid suppression of translation to capture ribosome snapshots in a specific physiological state, leading to possible inaccuracies in data collection [53]. Secondly, the technique requires inferring the speed of protein synthesis, but it is accurate based on the assumption that all of the ribosomes have completed translation. In fact, translation pauses or discontinuities may also occur under certain conditions, such as starvation [54]. Thirdly, contaminated RNA fragments (including non-coding RNAs or ribosome–protein complexes) may migrate during gradient centrifugation, be found in cDNA libraries, and lead to misreading

in translation. Lastly, the generated RPFs with 30 bp are not easy to map [55]. Because these RPFs are often too short to provide unique mapping information, when these short sequences are aligned to the reference genome or transcriptome, they may align to multiple locations due to the presence of repetitive or highly similar sequences. This makes it difficult to determine the precise location of the mRNA on the ribosome during translation. Moreover, the short length of these sequences can also lead to sequencing errors, which further complicates the mapping process. In addition, these short sequences may not provide enough context to accurately identify the frame of translation, which can affect downstream analysis and interpretation.

Table 1. Evaluation tools/software of sORF translation by ribosomal profiling methods.

Tool	Feature	Availability	Ref.
ORF-RAETER	Translated ORF identification and quantification based on linear regression.	https://github.com/alexfields/ORF-RATER . githhttps://github.com/alexfields/ORF-RATER	[47]
Ribo-TISH	A comprehensive toolkit for analyzing TIs and predicting putative ORFs.	https://github.com/zhpn1024/ribotish.git	[56]
Ribowave	ORF prediction, protein abundance estimation, TE calculation, and ribosomal frameshift identification based on wavelet transform.	https://ribowave.ncrnalab.org/	[50]
RiboORF	ORF identification based on the arrangement of the ribosome A site, the 3 nt periodicity, and the consistency between codons.	https://github.com/zhejilab/RibORF.git	[51]
RiboHMM	ORF identification based on the total abundance and periodic codon structure in RPF.	https://github.com/djf604/RiboHMM.git	[57]
ORFquant	ORF prediction and quantification based on multi-taper method.	https://github.com/lcalviell/ORFquant/releases/tag/1.02	[49]
Ribotricer	ORF prediction based on three-nucleotide periodicity.	https://github.com/smithlabcode/ribotricer/releases/tag/v1.3.3	[58]
PRICE	Resolving overlapping sORFs and noncanonical translation initiation based on machine-learning model.	https://github.com/erhard-lab/price.git	[59]
Ribocode	De novo assembly and annotation for translatoemes based on Wilcoxon signed-rank test.	https://github.com/xryanglab/RiboCode.git	[52]
RiboTaper	ORF identification based on the characteristic three-nucleotide periodicity of Ribo-Seq.	https://ohlerlab.mdc-berlin.de/software/RiboTaper_126/	[48]
RP-BP	ORF prediction based on unsupervised Bayesian approach.	https://github.com/dieterich-lab/rp-bp.git	[60]
SPECTre	ORF prediction based on 3 nt periodicity.	https://github.com/mills-lab/spectre.git	[61]
RiboToolkit	Ribo-Seq web application for analysis and implementation of a full ORF prediction pipeline.	http://rnainformatics.org.cn/RiboToolkit/analysis.php	[62]
GWIPS-Viz	Online web server for visualizing processed	https://gwips.ucc.ie/	[63]
Trips-Viz	Ribo-Seq data.	https://trips.ucc.ie/	[64]

4. Peptidomic-Based Methodology for Identification of SEPs

The MS-based technique is the most direct evidence that sORFs can be translated. As with traditional bottom-up proteomics studies, the identification workflow for SEPs includes sample extraction and enrichment, digestion and separation, MS data collection, and analysis.

4.1. Sample Extraction and Enrichment

The first critical step for SEP identification is extracting SEPs from complex biological matrices while ensuring their integrity (Figure 3). SEPs with a small size and a low molecular weight are difficult for peptidases to hydrolyze, and they may not have any sites for protease digestion or may be covered by undesired protein degradation products [65]. SEP extraction is, therefore, more challenging than that of proteins. Previous studies have tried various methods to ensure the integrity of SEPs, such as heating samples in boiling water or lysis buffers, using an ultrasonic treatment, or adding protease inhibitors to inhibit peptidase and protease activity [66,67]. However, some polypeptides, such as peptidases or protease inhibitors, can interfere with the subsequent analysis of SEPs. Therefore, some studies have proposed alternative methods, such as inducing protein precipitation with hydrochloric acid or acetic acid, which not only effectively prevent the degradation of SEPs, but also do not interact with polypeptide enzymes [67]. Therefore, the treatment of biological samples is a key step in extracting SEPs. The stability of biological samples and the objectives of the research should guide the selection of the appropriate extraction techniques.

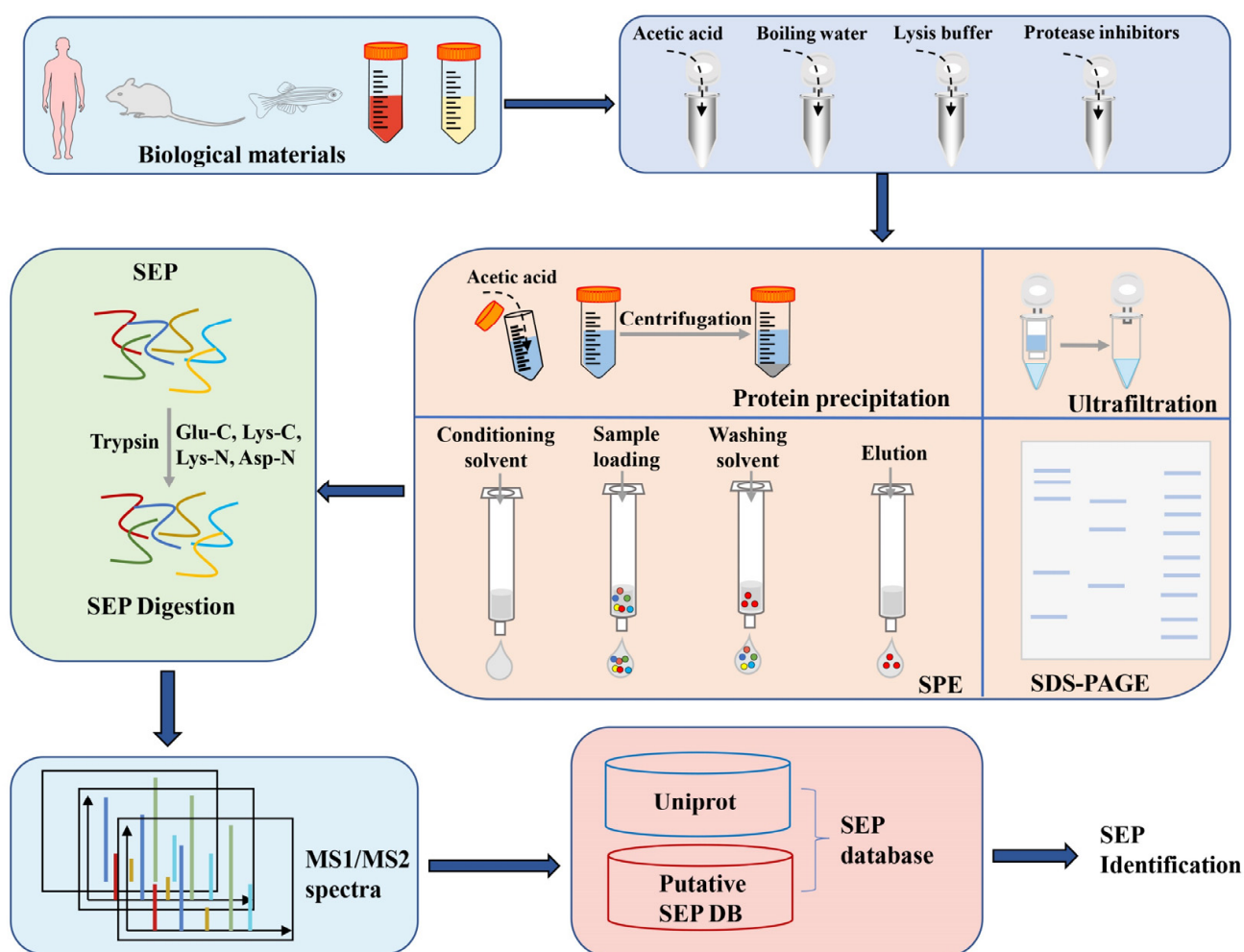


Figure 3. General workflow of a peptidomics approach for identifying SEPs. SEPs are extracted from complex biological samples, enriched by different methods such as protein precipitation, ultrafiltration, and SPE, then digested with trypsin (or multiple enzymes). The tryptic peptides are subjected to fractionation, MS data acquisition, and data analysis to identify SEPs.

The enrichment of SEPs is mainly used to separate the target peptides from other proteins in the same sample, thus reducing the complexity of the sample. These separation

methods frequently depend on various physical properties of the sample, such as the size, hydrophobicity and charge. Organic solvent (acetonitrile [40], acetone, methanol [68], trichloroacetic acid, and acetic acid [67]) precipitation could effectively retain low-molecular weight proteins in the supernatant liquid. In another endeavor, sequential precipitation and dehydration (SPD) based on a methyl tert-butyl ether/methanol/water system was used to successfully detect 129 proteins smaller than 30 kDa from human plasma, showing a good sensitivity and reproducibility. Size exclusion approaches have also been extensively used for protein isolation. High-molecular weight proteins can be kept on filter by using 10 or 30 kDa molecular weight cut-off (MWCO) ultrafiltration membranes [69,70]. However, this membrane-based technique has several drawbacks, including the potential blockage of membrane pores due to concentrated macromolecules, the non-specific binding of small proteins to hydrophobic surfaces, and time-consuming processes. C8-SPE is another method based on hydrophilic and hydrophobic properties of SEPs [66]. It was reported that a combination of these methods may be able to identify more SEPs [71].

4.2. Digestion of Samples for Mass Spectrometry

Sample digestion is a crucial component as well. SEPs tend to be short peptides with less than 100 amino acids and fewer arginine and lysine residues than other peptides. The single trypsin may cause cleavage failure or produce fewer trypsin peptides, reducing the sequence coverage and making it impossible to be detected by MS [72]. Due to the sequential digestion and complementary cleavage specificity, multi-protease digestion combined with trypsin and other proteolytic enzymes such as Glu-C (endoproteinase Glu-C), Lys-C (endoproteinase Lys-C), Lys-N (endoproteinase Lys-N), Asp-N (endoproteinase Asp-N), Arg-C (endoproteinase Arg-C), and chymotrypsin has been shown to enhance micro-peptide recognition effectively [73,74].

To date, mass spectrometry is still the only method available for the direct detection and quantification of SEPs. Data dependent acquisition (DDA) is the most widely used for MS acquisition analyses. In the past five years, thousands of SEPs have been identified using DDA from different species, including humans [75], *E. coli* [76], and plants [77]. The method is suitable for peptides ranging from 8 to 25 amino acids, but SEPs cannot produce fragments in this range due to the absence of required cleavage sites [78]. On the other hand, due to the small size of micro-peptides, only one peptide can be used for a peptide spectrum match (PSM) [79,80], which may increase the false detection rate in SEP identification [81,82]. Fortunately, it was discovered that targeted proteomics is a promising method with higher confidence. The expression of SEPs was tracked using parallel reaction monitoring (PRM) and data independent acquisition (DIA) in different biological samples [83]. In addition to simultaneously breaking up all precursor ions, DIA also preserves data that can be analyzed repeatedly *in silico* using various spectral libraries. Pak et al. [84] reported that the number of immune peptides identified had increased by almost three-fold using DIA. By selectively detecting particular peptides, parallel reaction monitoring (PRM) aims to achieve the relative or absolute quantification of a target protein or peptide [85]. These approaches are expected to benefit substantially from further improvements in analytical pipelines.

4.3. Database Construction for SEPs

With the accumulation of encoded sORFs and their corresponding SEPs, numerous publicly accessible repositories devoted to sORFs have been developed for SEP identification (Table 2).

Table 2. Commonly used databases for micro-peptide research.

Database	Resource	URL	Function
SmProt		http://bigdata.ibp.ac.cn/SmProt/index.html	A reliable repository with a comprehensive annotation of small proteins derived from ribosome profiling, literature, mass spectrometry (MS), etc.
sORF.org		http://www.sorfs.org (The website link is temporarily unavailable)	A public repository for sORFs identified both from experiments and in silico (based on various bioinformatics tools) to allow researchers to examine individual sORFs or to perform searches.
Openprot	Human and other species	https://www.openprot.org/	Contains all known proteins (RefProts), novel predicted isoforms (Isoforms) and novel predicted proteins from alternative ORFs (AltProts), supporting the annotation of thousands of predicted ORFs.
nORF.org		https://norfs.org/home	Combines existing databases such as sORFs.org, OpenProt, and openCB to provide more comprehensive information.
MetamORF		https://metamorf.hb.univ-amu.fr/	Provides unique sORFs identified in the human and mouse genomes with both experimental and computational approaches.
FuncPEP		https://bioinformatics.mdanderson.org/Supplements/FuncPEP	A new database of functional ncRNA encoded peptides, containing all experimentally validated and functionally characterized ncPEPs.
PsORF	Plant	http://psorf.whu.edu.cn/	A web collection resource of small open reading frames (sORFs) for 35 plant species to provide translation evidence and information on the evolutionary conservation of those small ORFs.
ARA-PEPs		https://github.com/rashmihazarika/ARA-PEPs.git	Specific to all putative sORF-encoded peptides in <i>Arabidopsis thaliana</i> .

Both SmProt [86] and sORF.org [42] are well known to researchers. SmProt collects small proteins from eight species, including *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Danio rerio*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Escherichia coli*, that have been identified through ribosomal analysis data, the literature, and mass spectrometry (MS) [86]. SmProt also includes information about the sequences, genomic locations, coding potential assessment, function, and other characteristics of the collected small proteins [41]. sORFs.org and OpenProt assess the identity of protein sequences based on BLASTp scores. Specific details about the target micro-peptide, such as the species, chromosome number, starting codon, and sORF attributes, can be requested through sORFs.org [42]. While OpenProt proposes a comprehensive annotation of predicted coding sequences on all transcripts, it provides obvious evidence for the expression of novel protein products [43]. The combination of these public databases could speed up the identification of micro-proteins. MetamORF only contains sORF data for *Homo sapiens* and *Mus musculus* [87]. ARA-PEPs [88] and PsORF [88] are comprehensive web servers dedicated to searching, browsing, visualizing, and downloading plant sORF-encoded peptides. These resources make it simple to construct reference databases for identifying and analyzing SEPs.

The combination of private databases and public databases is also a good choice. Generally, custom databases rely on a six-frame translation of the genome sequence to produce a reference database with all potential SEPs. Many researchers have combined custom databases and public databases such as Ensembl, RefSeq, and UniProtKB [44,89] for mining new SEPs [73,90–92]. However, it is undeniable that these databases contain a large number of pseudo-sequences, which reduces the confidence of peptide profile matching (PSM) and makes it challenging to detect SEPs with a low abundance [83].

The de novo sequencing of MS data is a library-independent method that deciphers protein or peptide sequences only from the spectrum without any genomic reference information [93]. Chen et al. and Wang et al. identified hundreds of SEPs using PEAK [94] and pNovo3 [40], respectively. However, it should be noted that many de novo peptides cannot be matched to any ORFs using the algorithms available today. This may be due to rare starting codons, mutations, or splicing, or it might require improved gene mapping algorithms to dock de novo sequencing.

4.4. Bioinformatic Tools for sORF and SEP Predictions

With the advance of high-throughput sequencing technology, many functional SEPs have been found. It is necessary to re-evaluate the coding potential of sORFs. However, the identification and prediction of sORFs with coding capabilities have become more complex due to the relative lack of consensus features. A wide variety of computational tools have been developed for predicting and distinguishing non-coding and coding transcripts based on nucleotide composition, codon substitution, machine-learning algorithms, and others (Table 3).

Table 3. Some bioinformatic resources for coding potential of sORF.

Tools	Data Requirement	Principle Utilized	URL	Ref.
CRITICA	Whole genome	Nucleotide sequence composition	http://rdpwww.life.uiuc.edu (The website link is temporarily unavailable)	[95]
CPC/CPC2	RNA-seq	Nucleotide composition, sequence similarity	http://cpc2.gao-lab.org/	[96]
PLEK	RNA-seq	Kmer composition of sequence	https://sourceforge.net/projects/plek	[97]
micPDP	RNA-seq	Codon conservation	https://github.com/Drmirdeep/micpdp.git	[98]
PhlyoCSF	RNA-seq	Codon substitution, nucleotide sequence similarity	http://compbio.mit.edu/PhlyoCSF	[99]
PhastCons	Whole genome	Nucleotide composition	http://compugen.cshl.edu/phast/	[100]
sORF finder	Any nucleotide sequence	Nucleotide composition	http://evolver.psc.riken.jp/ (The website link is temporarily unavailable)	[101]
RNAcode	RNA-seq	Codon conservation	https://github.com/ViennaRNA/RNAcode.git	[102]
CNCI	RNA-seq	Nucleotide composition	https://github.com/www-bioinfo-org/CNCI.git	[103]
CPAT	RNA-seq	Permutation-free logistic regression model	https://wlcblab.uci.edu/cpat	[104]
CNIT	RNA-seq	Nucleotide composition	https://github.com/www-bioinfo-org/CNCI.git	[105]
ORF finder	Whole genome	Nucleotide composition	https://www.ncbi.nlm.nih.gov/orffinder/	[106]
iSeeRNA	RNA-seq	Machine-learning model	http://www.myogenesisdb.org/iSeeRNA (The website link is temporarily unavailable)	[107]
COME	RNA-seq	Machine-learning model	https://github.com/lulab/COME.git	[108]
LncRNA-ID	RNA-seq	Machine-learning model	https://github.com/zhangy72/LncRNA-ID.git	[109]
lncRNA-MFDL	RNA-seq	Deep-learning classification algorithms	http://compgenomics.utsa.edu/lncRNA_MDFL/	[110]
uPEPperoni	RNA-seq	Codon substitution	http://u pep-scmb.biosci.uq.edu.au	[111]
DeepCPP	RNA-seq	Machine-learning algorithms	https://github.com/yuuuuzhang/DeepCPP.git	[112]
RNASamba	RNA-seq	Similarity to known proteins	https://rnasamba.lge.ibi.unicamp.br/	[113]
MipepID	Whole genome	Machine-learning algorithms	https://github.com/MindAI/MiPepid.git	[114]

4.5. Prediction of Coding Potential and Sequence Conversion of sORFs

As an original tool, the coding region identification tool invoking comparative analysis (CRITICA) compares genomic regions across multiple species to identify conserved non-coding regions that are likely to contain functional sORFs [95]. Another computational tool, the coding potential calculator (CPC), calculates a coding potential score based on features such as ORF length, ORF coverage, and conservation [96]. Therefore, it may miss some functional sORFs that are not conserved across species. Y. et al. used CPC to predict the coding potential of the lncRNA DLEU1 and found that DLEU1 encodes a membrane-channel small peptide that affects glioma cell development, invasion, and metastasis [115]. Other tools such as RNAcode [102], micPDP [98], and phyloCSF [99] make use of a different principle known as codon substitution. The criteria used by PhyloCSF to identify sORFs include the presence of an ORF with a length of at least 30 nucleotides and evidence of purifying selection across multiple species. Mackowiak et al. predicted 354 conserved sORFs in the lncRNAs based on Ribo-Seq and PhyloCSF, and validated 22 peptides using MS spectral data [116]. ORF finder is a tool that identifies ORFs in nucleotide sequences. It does not specifically predict sORF coding potential, but rather identifies all ORFs, including potentially functional and non-functional ones. Growing evidence points to machine-learning (ML) algorithms as another options for sORF coding potential prediction, such as DeepCPP [112] and MipepID [114]. In particular, MipepID is designed specifically to predict the coding potential of sORFs. Fesenko, Igor et al. identified thousands of evolutionarily conserved smORFs in *Physcomitrium patens* using MipepID [117]. Therefore, CRITICA is the most effective tool for identifying functional sORFs in some species, while PhyloCSF and CPC/CPC2 may be better suited for identifying conserved and novel sORFs, respectively. ORF finder is a useful tool for identifying all ORFs in a sequence, but may identify many false positive sORFs. The emergence and development of these tools reflects the endeavor to study sORFs.

4.6. Prediction Tools Related to SEP Structure

In addition to predicting the coding potential of sORFs, it is necessary to perform structurally related predictions of their functional micro-peptides. Currently, several tools, such as TMHMM [118], SignalP [119], ProtScale [120], and AlphaFold2 [121], also have been used to predict the localization, transmembrane regions and protein structure of the target micro-peptides. TMHMM is currently the most effective and best-performing method for the prediction of transmembrane segments of micropeptides [118]. SignalP 5.0 predicts the presence of signal peptides and the location of their cleavage sites, which helps researchers to understand the mode of action of micro-peptides. A tool called ProtScale makes it possible to compute and represent the profile produced by any amino acid scale, and it serves as a guide for the identification of micro-peptide transmembrane regions. Additionally, SWISS-MODEL [122] and AlphaFold2 can be applied to generate reliable 3D protein models, which can enable an in-depth exploration of the biological functions and structural features of micro-peptides. Zhou et al. used several functional tools, including IAMPE [123], Phobius [124], Pfam [125], TMHMM, and ProtScale, to analyze these candidate micro-peptides, indicating that an SEP (SEP068184) may regulate oxidative resistance through involving metabolic pathways and interacting with cytoplasmic proteins in *Deinococcus radiodurans* [126]. Moreover, there are additional resources available for researchers to investigate the specific physical and chemical properties or functions of SEPs, including ProtParam, BUSCA [127], and SOPMA [128].

5. Experimental Validation of Micro-Peptide Coding Potential and Function

Recent studies have identified thousands of additional components of the proteome. The majority of these components are micro-peptides that sORFs in noncoding regions translate. Although Ribo-Seq, bioinformatics prediction and peptidomics are mostly sufficient for the requirements of high-throughput micro-peptide screening and discovery, corresponding biochemical experiments are necessary to prove their true existence.

5.1. Validation of Translation of sORFs from Putative SEPs

Firstly, antibodies can specifically recognize a target protein and are a direct and highly sensitive method for detecting the endogenous expression of SEPs in tissues or cells. Li et al. detected the endogenous expression of MIAC by preparing monoclonal antibodies to MIAC [129]. However, these SEPs have a low antigenicity and contain transmembrane structural domains, which largely limit the selection of immune epitopes and make it still extremely challenging to produce specific and effective antibodies against peptides [19,30,130,131].

For SEPs without corresponding antibodies, an epitope tag is another option for detecting the endogenous expression of SEPs. In order to create a fusion protein that contains both SEPs and protein tags, fluorescence (GFP) or epitope tags can be inserted into the candidate SEP sequence using CRISPR/Cas9-mediated gene-editing techniques. The presence of the SEP is then confirmed by Western blotting and the immunoprecipitation of these fusion proteins [17,132–135]. To determine whether the sORF in the CASIMO1 transcript was translated into a micro-peptide, Schwarz et al. inserted a FLAG tag at the C-terminus of the CASIMO1 coding sequence and detected the expression of CASIMO1-FLAG by an anti-FLAG antibody [136]. Nevertheless, if these SEPs are relatively small, additional peptide fractions may alter their physiological properties, localization, or protein interactions [137]. There are a variety of different epitope tags, including FLAG [136], APEX [137], HA [138], V5, fluorescent proteins, etc. It is essential to choose the appropriate epitope tag according to the characteristics of the SEP.

In addition to micro-peptide validations based on antibody studies, the sORF coding potential may also be determined using *in vitro* translation assays [15,138,139]. The experiment requires additional experiments to verify the release of the SEP [130], such as the introduction of frameshift mutations, which are used as negative controls to further verify the results.

5.2. Demonstration and Validation of Biological Relevance for SEPs

The above experiments validated the capacity of sORFs for translation; however, molecular experiments are needed to determine the potential function of the identified SEPs. Most of these methods are similar to determining the common protein function, but are relatively complex. The CRISPR/Cas9 technique is frequently employed to detect the effects of SEPs on phenotypes [11,15,140]. Special vectors for SEPs, such as loss-of-function (e.g., knockdown or knockout) or gain-of-function (e.g., overexpression or activation) vectors, can be designed for cell transfection to observe the effect on the phenotype, further inferring the function of SEPs. Fu et al. identified a highly conserved transmembrane micro-peptide called NEMEP by CRISPR/Cas9, providing a clear example of the direct functional effect of altered glucose metabolism on cell fate decisions [138]. However, not all SEPs can benefit equally from the functional validation experiments of CRISPR/Cas9. When translatable sORFs exist in lncRNAs, the validation experiments often need to be achieved using frameshift or start codon mutations, which not only selectively inhibit micro-peptide expression, but also have no impact on lncRNAs [30,141].

Synthesizing the corresponding peptides is another way to confirm the function of SEPs. Pauli et al. successfully applied this method to demonstrate that the synthesized toddler peptide has the same phenotype as mRNA overexpression [142]. In addition, rescue experiments can be performed to verify whether the sORFs or SEPs are responsible for regulatory functions [143]. After the functions of SEPs are certified, the underlying regulatory mechanism behind these SEPs becomes an urgent issue for subsequent research. MS and immunoprecipitation can be used to identify specific protein complexes. The function or pathway of the co-purification protein can then be used to deduce the function of the micro-peptides [132].

The functional verification of the SEPs encoded by UTR regions is relatively difficult. To characterize the biological relevance of uORF-encoded micro-peptides, uORF perturbations may affect the stability of the main ORF, further confusing the process for revealing the uORF function. In a previous study, antisense oligonucleotides (ASOs) against uORF were

used to up-regulate the CDS expression, which was a more novel strategy [144]. Although the underlying regulatory mechanism is unclear, uORF-targeted ASO has been used to restore downstream gene expression by regulating the efficiency of ribosome initiation [145]. Thus, ASO is suitable as a functional tool to assess the effect of a given uORF on the CDS expression.

6. Biological Functions of sORF-Encoded Peptides: Relevant Examples

To date, many SEPs have been identified and characterized, and they are involved in a variety of physiological processes, such as calcium homeostasis, metabolism, muscle development, substance degradation, gene transcription and translation regulation, and cancer development.

For example, the lncRNA MIR155HG was the subject of extensive research for its contribution of miRNA products (miR-155) in inflammation and adaptive immune responses. It was reported that the human lncRNA MIR155HG encoded the 17-amino acid micro-peptide miPEP155 (P155). MIR155HG is highly expressed by inflamed antigen-presenting cells, leading to the discovery that P155 interacts with the adenosine 5'-triphosphate binding domain of heat shock cognate protein 70 (HSC70), a chaperone required for antigen trafficking and presentation in dendritic cells (DCs). P155 modulates major histocompatibility complex class II-mediated antigen presentation and T cell priming by disrupting the HSC70-HSP90 machinery [21]. Here, a summary of more SEPs and their biological functions is provided (Table 4 and Figure 4).

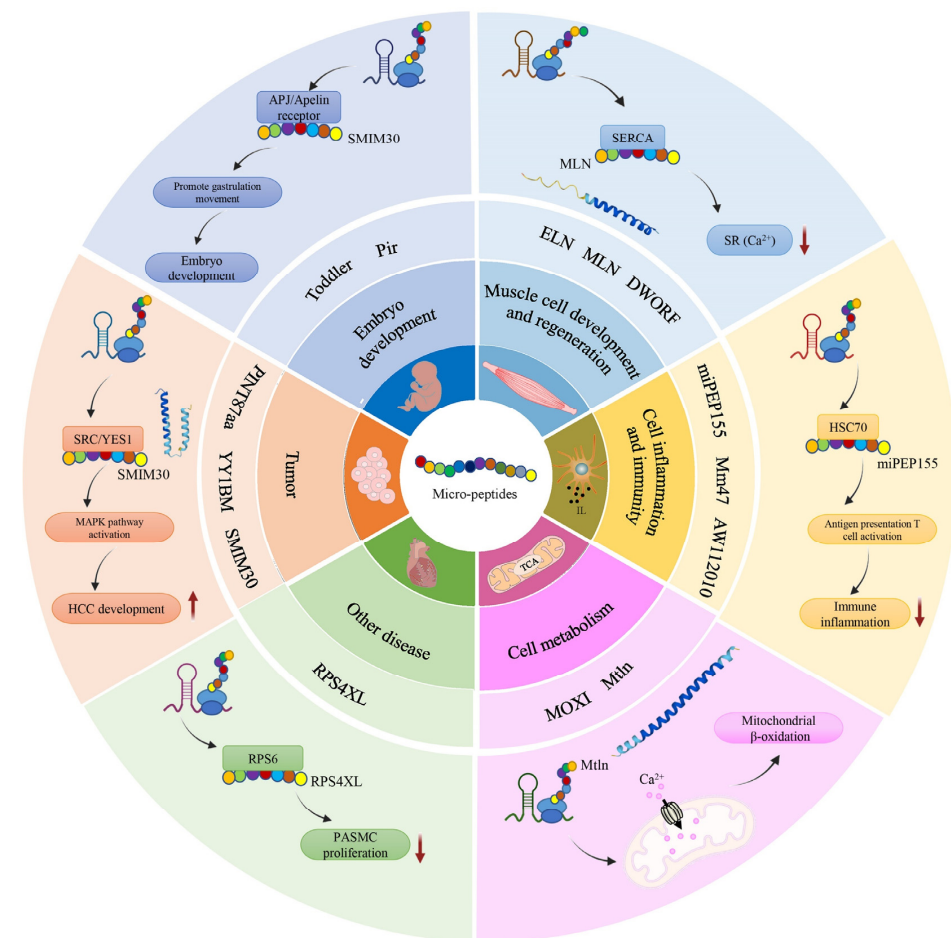


Figure 4. Various biological roles of micro-peptides encoded by putative ncRNAs. From the outer layer to the inner layer are the specific molecular mechanisms of representative SEPs, representative SEPs and their corresponding biological functions, respectively. Upward arrows indicate promotion, downward arrows indicate inhibition.

Table 4. Association between SEP expression and diverse biological function.

Species	Symbol	Micro-Peptide	Length (aa)	Function	Category	Ref.
Human	LINC00948	MLN	46			[15]
Human	110017F19Rik /SMIM6	ELN	56	Inhibits SERCA and regulates Ca ²⁺ transport		[146]
Human	1810037I17Rik	ALN	65			[146]
Human	LOC100507537	DWORF	34	Activates SERCA and regulates Ca ²⁺ transport		[131]
Mouse	LOC101929726	Minion	84	Promotes myoblast fusion and muscle development	Micro-peptide related muscle	[147]
Mouse	LOC101929726	Myomixer	84			[16,148]
Fruit Fly	Scl	SclB	<30	Regulates Ca ²⁺ transport		[149]
Chicken	Six ORF2	SIX1	74	Promotes cell proliferation and is involved in muscle growth		[142]
Mouse/ Zerbreifh	MyolncR4/ 1500011K16RIK	LEMP	56	Promotes muscle formation and regeneration in mice		[150]
Human	SPAR	LINC00961	75	Promotes muscle development		[17]
Human	miR-155HG	miPEP155 (P155)	17	Modulates antigen transport and presentation of antigen presenting cells		[21]
Mouse	lncRNA 1810058I24Rik	Mm47	47	Controls innate immunity	Micro-peptide related inflammation and immunity	[20]
Mouse	lncRNA Aw112010	Aw112010		Drives IL-12p40 production and mediates innate immune response		[22]
Human		IMP	–	Regulates inflammatory gene expression		[151]
Human /Mouse	LINC00116/ 1500011K16Rik	MOXI /Mitoregulin	56	Enhances mitochondrial β -oxidation	Micro-peptide related metabolism	[19]
Mouse	LINC00116	Mtln	56	Supports mitochondrial super-complexes and respiratory efficiency		[18]
Human	LINC-PINT	PINT87aa	87	Tumor suppressor in glioblastoma		[152]
Human	LINC00278	YY1BM	21	Promotes apoptosis and downregulates the survival rate of ESCC cells		[153]
Human	circFNDC3B	circFNDC3B-218aa	218	Inhibits the expression of oncogene Snail and promotes CRC		[154]
Human	circPPP1R12A	PPP1R12A-C	73	Activates Hippo-YAP signaling pathway and inhibits CRC		[155]
Human	Meloe	MELOE-1	46	Involved in T cell immune surveillance; optimal T cell target for melanoma immunotherapy	Micro-peptides related tumors	[156]
Human		MELOE-2	39			[157]
Human	LINC00665	CIP2A-BP	52	Inhibits tumor invasion and metastasis		[158]
Human	LINC00998	SMIM30	59	Promotes cell proliferation and migration		[139]
Human	NCBP2-AS2	KRASIM	99	Inhibits carcinogenic signaling in hepatocellular carcinoma cells		[159]
Human	UBAP1-AST6	BAP1-AST6 (aa)	-	Promotes cell proliferation		[160]
Human	HOXB-AS3	HOXB-AS3	53	Inhibits cell proliferation, invasion, and metastasis		[161]
Human	CTD-2256P15.2	PACMP	44	Regulates cancer progression and drug resistance by modulating DNA damage response		[26]
Human	Rps41	RPS4XL	–	Inhibition of hypoxia-induced proliferation of pulmonary artery smooth muscle cells	Other diseases	[162]

7. Conclusions and Future Perspectives

Tradition dictates that genes encode only one protein and that transcripts without a main ORF are non-coding. In this review, we revealed a new research area: ncRNAs that can encode peptides or small proteins. We elaborated on the location of sORFs in the genome, the identification of encoded peptides, and the analytical procedures and subsequent methods for the validation of biological function mechanisms, revealing previously unrecognized complexity in the proteome. In recent years, SEPs have been found to exist and play important biological regulatory roles in most species, including humans, mice, rats, zebrafish, flies, yeast, and *Escherichia coli*. In addition, with a relatively small size, a tissue-specific expression pattern, and a low cytotoxicity, SEPs will be a new resource pool for screening anti-tumor peptides or protein drugs, and they will play an important role in accurate diagnoses, precise classifications, precise treatments and tumor prognoses. So far, SEPs have been found to have significant antitumor functions by inhibiting cancer metabolic reprogramming, oncogenic protein stability, and oncogenic-related pathways, making them new therapeutic targets for clinical applications. However, SEPs are characterized by short peptide fragments, a small molecular weight, and a low expression abundance, which may cause difficulties in the extraction and synthesis of micro-peptide drugs and inaccurate identification of relevant detection technologies. Therefore, continued advancements in the field will depend on clever experimental designs and further optimization of the relevant technology.

Although many SEPs with coding potential have been characterized in the last few years, the following crucial and urgent questions still need to be answered: (1) How can a sufficient number of SEP samples be obtained for a more thorough investigation? The small molecular weight and low expression abundance of SEPs make it difficult to obtain active samples via genetic engineering; (2) The annotation of SEPs is primarily based on phylogenetically conserved analyses, but how else can new peptides be validated in the absence of sequence conservation? How do the different SEPs work? (3) Given the growing evidence that not all peptides initiate translation by AUG, how do we begin to validate the true translation initiation codons with the current genome annotations of uORFs and main ORFs? Do initiation codons other than AUG codons employ a different mechanism? (4) Only the human and a few animal models are included in the current database of species annotated for SEPs. The inter-species differences have led to many databases being insufficient to meet the requirements of micro-peptide research at this stage, so the establishment of functional annotation databases is particularly important. There is no doubt that the mechanism of sORF-encoded micro-peptides will spark a new research boom and advance the life sciences; they will provide new insights for future investigations to unravel intricate physiological processes and diagnose diseases in living organisms.

Author Contributions: X.D.: conceptualization, project administration, writing—original draft. K.Z., C.X. and T.C.: investigation, methodology, validation. S.L.: funding acquisition, project administration, Z.L. and Y.Z.: conceptualization, supervision, funding acquisition, project administration, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Science Foundation of China (32273111, 32271329, 32071262, 32171271) and the Science and Technology Innovation Program of Hunan Province (2020RC4023).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: <http://bigdata.ibp.ac.cn/SmProt/index.html> and <http://www.sorfs.org>.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, Y.; Ho, L.; Tergaonkar, V. sORF-Encoded MicroPeptides: New players in inflammation, metabolism, and precision medicine. *Cancer Lett.* **2021**, *500*, 263–270. [[CrossRef](#)]
2. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74. [[CrossRef](#)]
3. Houseley, J.; Rubbi, L.; Grunstein, M.; Tollervey, D.; Vogelauer, M. A ncRNA modulates histone modification and mRNA induction in the yeast GAL gene cluster. *Mol. Cell* **2008**, *32*, 685–695. [[CrossRef](#)]
4. Li, D.; Tolleson, W.H.; Yu, D.; Chen, S.; Guo, L.; Xiao, W.; Tong, W.; Ning, B. Regulation of cytochrome P450 expression by microRNAs and long noncoding RNAs: Epigenetic mechanisms in environmental toxicology and carcinogenesis. *J. Environ. Sci. Health Part C Environ. Carcinog. Ecotoxicol. Rev.* **2019**, *37*, 180–214. [[CrossRef](#)]
5. Landgraf, P.; Rusu, M.; Sheridan, R.; Sewer, A.; Iovino, N.; Aravin, A.; Pfeffer, S.; Rice, A.; Kamphorst, A.O.; Landthaler, M.; et al. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **2007**, *129*, 1401–1414. [[CrossRef](#)]
6. Chew, C.L.; Conos, S.A.; Unal, B.; Tergaonkar, V. Noncoding RNAs: Master Regulators of Inflammatory Signaling. *Trends Mol. Med.* **2018**, *24*, 66–84. [[CrossRef](#)]
7. Khalili-Tanha, G.; Moghbeli, M. Long non-coding RNAs as the critical regulators of doxorubicin resistance in tumor cells. *Cell. Mol. Biol. Lett.* **2021**, *26*, 39. [[CrossRef](#)]
8. Xing, J.; Liu, H.; Jiang, W.; Wang, L. LncRNA-Encoded Peptide: Functions and Predicting Methods. *Front. Oncol.* **2020**, *10*, 622294. [[CrossRef](#)]
9. Orr, M.W.; Mao, Y.; Storz, G.; Qian, S.B. Alternative ORFs and small ORFs: Shedding light on the dark proteome. *Nucleic Acids Res.* **2020**, *48*, 1029–1042. [[CrossRef](#)]
10. Couso, J.P.; Patraquim, P. Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.* **2017**, *18*, 575–589. [[CrossRef](#)]
11. van Heesch, S.; Witte, F.; Schneider-Lunitz, V.; Schulz, J.F.; Adami, E.; Faber, A.B.; Kirchner, M.; Maatz, H.; Blachut, S.; Sandmann, C.L.; et al. The Translational Landscape of the Human Heart. *Cell* **2019**, *178*, 242–260.e229. [[CrossRef](#)]
12. Pauli, A.; Norris, M.L.; Valen, E.; Chew, G.L.; Gagnon, J.A.; Zimmerman, S.; Mitchell, A.; Ma, J.; Dubrulle, J.; Reyon, D.; et al. Toddler: An embryonic signal that promotes cell movement via Apelin receptors. *Science* **2014**, *343*, 1248636. [[CrossRef](#)]
13. Savard, J.; Marques-Souza, H.; Aranda, M.; Tautz, D. A segmentation gene in tribolium produces a polycistronic mRNA that codes for multiple conserved peptides. *Cell* **2006**, *126*, 559–569. [[CrossRef](#)]
14. Kondo, T.; Hashimoto, Y.; Kato, K.; Inagaki, S.; Hayashi, S.; Kageyama, Y. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat. Cell Biol.* **2007**, *9*, 660–665. [[CrossRef](#)]
15. Anderson, D.M.; Anderson, K.M.; Chang, C.L.; Makarewich, C.A.; Nelson, B.R.; McAnally, J.R.; Kasaragod, P.; Shelton, J.M.; Liou, J.; Bassel-Duby, R.; et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* **2015**, *160*, 595–606. [[CrossRef](#)]
16. Bi, P.; Ramirez-Martinez, A.; Li, H.; Cannavino, J.; McAnally, J.R.; Shelton, J.M.; Sanchez-Ortiz, E.; Bassel-Duby, R.; Olson, E.N. Control of muscle formation by the fusogenic micropeptide myomixer. *Science* **2017**, *356*, 323–327. [[CrossRef](#)]
17. Matsumoto, A.; Pasut, A.; Matsumoto, M.; Yamashita, R.; Fung, J.; Monteleone, E.; Saghatelian, A.; Nakayama, K.I.; Clohessy, J.G.; Pandolfi, P.P. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* **2017**, *541*, 228–232. [[CrossRef](#)]
18. Stein, C.S.; Jadia, P.; Zhang, X.; McLendon, J.M.; Abouassaly, G.M.; Witmer, N.H.; Anderson, E.J.; Elrod, J.W.; Boudreau, R.L. Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. *Cell Rep.* **2018**, *23*, 3710–3720.e3718. [[CrossRef](#)]
19. Makarewich, C.A.; Baskin, K.K.; Munir, A.Z.; Bezprozvannaya, S.; Sharma, G.; Khemtong, C.; Shah, A.M.; McAnally, J.R.; Malloy, C.R.; Szweda, L.I.; et al. MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid beta-Oxidation. *Cell Rep.* **2018**, *23*, 3701–3709. [[CrossRef](#)]
20. Bhatta, A.; Atianand, M.; Jiang, Z.; Crabtree, J.; Blin, J.; Fitzgerald, K.A. A Mitochondrial Micropeptide Is Required for Activation of the Nlrp3 Inflammasome. *J. Immunol.* **2020**, *204*, 428–437. [[CrossRef](#)]
21. Niu, L.; Lou, F.; Sun, Y.; Sun, L.; Cai, X.; Liu, Z.; Zhou, H.; Wang, H.; Wang, Z.; Bai, J.; et al. A micropeptide encoded by lncRNA MIR155HG suppresses autoimmune inflammation via modulating antigen presentation. *Sci. Adv.* **2020**, *6*, eaaz2059. [[CrossRef](#)]
22. Jackson, R.; Kroehling, L.; Khitun, A.; Bailis, W.; Jarret, A.; York, A.G.; Khan, O.M.; Brewer, J.R.; Skadow, M.H.; Duizer, C.; et al. The translation of non-canonical open reading frames controls mucosal immunity. *Nature* **2018**, *564*, 434–438. [[CrossRef](#)]
23. Wang, J.; Zhu, S.; Meng, N.; He, Y.; Lu, R.; Yan, G.R. ncRNA-Encoded Peptides or Proteins and Cancer. *Mol. Ther. J. Am. Soc. Gene Ther.* **2019**, *27*, 1718–1725. [[CrossRef](#)]
24. Li, X.L.; Pongor, L.; Tang, W.; Das, S.; Muys, B.R.; Jones, M.F.; Lazar, S.B.; Dangelmaier, E.A.; Hartford, C.C.; Grammatikakis, I.; et al. A small protein encoded by a putative lncRNA regulates apoptosis and tumorigenicity in human colorectal cancer cells. *Elife* **2020**, *9*, e53734. [[CrossRef](#)]
25. Huang, N.; Li, F.; Zhang, M.; Zhou, H.; Chen, Z.; Ma, X.; Yang, L.; Wu, X.; Zhong, J.; Xiao, F.; et al. An Upstream Open Reading Frame in Phosphatase and Tensin Homolog Encodes a Circuit Breaker of Lactate Metabolism. *Cell Metab.* **2021**, *33*, 128–144.e129. [[CrossRef](#)]
26. Zhang, C.; Zhou, B.; Gu, F.; Liu, H.; Wu, H.; Yao, F.; Zheng, H.; Fu, H.; Chong, W.; Cai, S.; et al. Micropeptide PACMP inhibition elicits synthetic lethal effects by decreasing CtIP and poly(ADP-ribosylation). *Mol. Cell* **2022**, *82*, 1297–1312.e1298. [[CrossRef](#)]
27. Galindo, M.I.; Pueyo, J.I.; Fouix, S.; Bishop, S.A.; Couso, J.P. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol.* **2007**, *5*, e106. [[CrossRef](#)]

28. Rohrig, H.; Schmidt, J.; Miklashevichs, E.; Schell, J.; John, M. Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 1915–1920. [[CrossRef](#)]
29. Chu, Q.; Martinez, T.F.; Novak, S.W.; Donaldson, C.J.; Tan, D.; Vaughan, J.M.; Chang, T.; Diedrich, J.K.; Andrade, L.; Kim, A.; et al. Regulation of the ER stress response by a mitochondrial microprotein. *Nat. Commun.* **2019**, *10*, 4883. [[CrossRef](#)]
30. Chen, J.; Brunner, A.D.; Cogan, J.Z.; Nuñez, J.K.; Fields, A.P.; Adamson, B.; Itzhak, D.N.; Li, J.Y.; Mann, M.; Leonetti, M.D.; et al. Pervasive functional translation of noncanonical human open reading frames. *Science* **2020**, *367*, 1140–1146. [[CrossRef](#)]
31. Kang, M.; Tang, B.; Li, J.; Zhou, Z.; Liu, K.; Wang, R.; Jiang, Z.; Bi, F.; Patrick, D.; Kim, D.; et al. Identification of miPEP133 as a novel tumor-suppressor microprotein encoded by miR-34a pri-miRNA. *Mol. Cancer* **2020**, *19*, 143. [[CrossRef](#)]
32. Bartel, D.P. MicroRNAs: Target recognition and regulatory functions. *Cell* **2009**, *136*, 215–233. [[CrossRef](#)]
33. Somers, J.; Pöyry, T.; Willis, A.E. A perspective on mammalian upstream open reading frame function. *Int. J. Biochem. Cell Biol.* **2013**, *45*, 1690–1700. [[CrossRef](#)]
34. Wu, Q.; Wright, M.; Gogol, M.M.; Bradford, W.D.; Zhang, N.; Bazzini, A.A. Translation of small downstream ORFs enhances translation of canonical main open reading frames. *EMBO J.* **2020**, *39*, e104763. [[CrossRef](#)]
35. Zhang, M.; Huang, N.; Yang, X.; Luo, J.; Yan, S.; Xiao, F.; Chen, W.; Gao, X.; Zhao, K.; Zhou, H.; et al. A novel protein encoded by the circular form of the SHPRH gene suppresses glioma tumorigenesis. *Oncogene* **2018**, *37*, 1805–1814. [[CrossRef](#)]
36. Liang, W.C.; Wong, C.W.; Liang, P.P.; Shi, M.; Cao, Y.; Rao, S.T.; Tsui, S.K.; Wayne, M.M.; Zhang, Q.; Fu, W.M.; et al. Translation of the circular RNA circ β -catenin promotes liver cancer cell growth through activation of the Wnt pathway. *Genome Biol.* **2019**, *20*, 84. [[CrossRef](#)]
37. Gao, X.; Xia, X.; Li, F.; Zhang, M.; Zhou, H.; Wu, X.; Zhong, J.; Zhao, Z.; Zhao, K.; Liu, D.; et al. Circular RNA-encoded oncogenic E-cadherin variant promotes glioblastoma tumorigenicity through activation of EGFR-STAT3 signalling. *Nat. Cell Biol.* **2021**, *23*, 278–291. [[CrossRef](#)]
38. Kalyana-Sundaram, S.; Kumar-Sinha, C.; Shankar, S.; Robinson, D.R.; Wu, Y.M.; Cao, X.; Asangani, I.A.; Kothari, V.; Prensner, J.R.; Lonigro, R.J.; et al. Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell* **2012**, *149*, 1622–1634. [[CrossRef](#)]
39. Hanada, K.; Zhang, X.; Borevitz, J.O.; Li, W.H.; Shiu, S.H. A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. *Genome Res.* **2007**, *17*, 632–640. [[CrossRef](#)]
40. Wang, B.; Wang, Z.; Pan, N.; Huang, J.; Wan, C. Improved Identification of Small Open Reading Frames Encoded Peptides by Top-Down Proteomic Approaches and De Novo Sequencing. *Int. J. Mol. Sci.* **2021**, *22*, 5476. [[CrossRef](#)]
41. Li, Y.; Zhou, H.; Chen, X.; Zheng, Y.; Kang, Q.; Hao, D.; Zhang, L.; Song, T.; Luo, H.; Hao, Y.; et al. SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling. *Genom. Proteom. Bioinform.* **2021**, *19*, 602–610. [[CrossRef](#)]
42. 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. [[CrossRef](#)]
43. Brunet, M.A.; Lucier, J.F.; Levesque, M.; Leblanc, S.; Jacques, J.F.; Al-Saedi, H.R.H.; Guilloy, N.; Grenier, F.; Avino, M.; Fournier, I.; et al. OpenProt 2021: Deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Res.* **2021**, *49*, D380–D388. [[CrossRef](#)]
44. Slavoff, S.A.; Mitchell, A.J.; Schwaid, A.G.; Cabili, M.N.; Ma, J.; Levin, J.Z.; Karger, A.D.; Budnik, B.A.; Rinn, J.L.; Saghatelian, A. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **2013**, *9*, 59–64. [[CrossRef](#)]
45. Wedekind, J.E.; Dance, G.S.; Sowden, M.P.; Smith, H.C. Messenger RNA editing in mammals: New members of the APOBEC family seeking roles in the family business. *Trends Genet.* **2003**, *19*, 207–216. [[CrossRef](#)]
46. Hornstein, N.; Torres, D.; Das Sharma, S.; Tang, G.; Canoll, P.; Sims, P.A. Ligation-free ribosome profiling of cell type-specific translation in the brain. *Genome Biol.* **2016**, *17*, 149. [[CrossRef](#)]
47. Fields, A.P.; Rodriguez, E.H.; Jovanovic, M.; Stern-Ginossar, N.; Haas, B.J.; Mertins, P.; Raychowdhury, R.; Hacohen, N.; Carr, S.A.; Ingolia, N.T.; et al. A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol. Cell* **2015**, *60*, 816–827. [[CrossRef](#)]
48. Calviello, L.; Mukherjee, N.; Wyler, E.; Zauber, H.; Hirsekorn, A.; Selbach, M.; Landthaler, M.; Obermayer, B.; Ohler, U. Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* **2016**, *13*, 165–170. [[CrossRef](#)]
49. Calviello, L.; Hirsekorn, A.; Ohler, U. Quantification of translation uncovers the functions of the alternative transcriptome. *Nat. Struct. Mol. Biol.* **2020**, *27*, 717–725. [[CrossRef](#)]
50. Xu, Z.; Hu, L.; Shi, B.; Geng, S.; Xu, L.; Wang, D.; Lu, Z.J. Ribosome elongating footprints denoised by wavelet transform comprehensively characterize dynamic cellular translation events. *Nucleic Acids Res.* **2018**, *46*, e109. [[CrossRef](#)]
51. Ji, Z. RibORF: Identifying Genome-Wide Translated Open Reading Frames Using Ribosome Profiling. *Curr. Protoc. Mol. Biol.* **2018**, *124*, e67. [[CrossRef](#)]
52. Xiao, Z.; Huang, R.; Xing, X.; Chen, Y.; Deng, H.; Yang, X. De novo annotation and characterization of the translome with ribosome profiling data. *Nucleic Acids Res.* **2018**, *46*, e61. [[CrossRef](#)]
53. Ingolia, N.T.; Brar, G.A.; Rouskin, S.; McGeachy, A.M.; Weissman, J.S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **2012**, *7*, 1534–1550. [[CrossRef](#)]
54. Subramaniam, A.R.; Zid, B.M.; O’Shea, E.K. An integrated approach reveals regulatory controls on bacterial translation elongation. *Cell* **2014**, *159*, 1200–1211. [[CrossRef](#)]

55. Kim, M.S.; Pinto, S.M.; Getnet, D.; Nirujogi, R.S.; Manda, S.S.; Chaerkady, R.; Madugundu, A.K.; Kelkar, D.S.; Isserlin, R.; Jain, S.; et al. A draft map of the human proteome. *Nature* **2014**, *509*, 575–581. [[CrossRef](#)]
56. Zhang, P.; He, D.; Xu, Y.; Hou, J.; Pan, B.F.; Wang, Y.; Liu, T.; Davis, C.M.; Ehli, E.A.; Tan, L.; et al. Genome-wide identification and differential analysis of translational initiation. *Nat. Commun.* **2017**, *8*, 1749. [[CrossRef](#)]
57. Raj, A.; Wang, S.H.; Shim, H.; Harpak, A.; Li, Y.I.; Engelmann, B.; Stephens, M.; Gilad, Y.; Pritchard, J.K. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Life* **2016**, *5*, e13328. [[CrossRef](#)]
58. Choudhary, S.; Li, W.; Smith, A.D. Accurate detection of short and long active ORFs using Ribo-seq data. *Bioinformatics* **2020**, *36*, 2053–2059. [[CrossRef](#)]
59. Erhard, F.; Halenius, A.; Zimmermann, C.; L'Hernault, A.; Kowalewski, D.J.; Weekes, M.P.; Stevanovic, S.; Zimmer, R.; Dölken, L. Improved Ribo-seq enables identification of cryptic translation events. *Nat. Methods* **2018**, *15*, 363–366. [[CrossRef](#)]
60. Malone, B.; Atanassov, I.; Aeschmann, F.; Li, X.; Großhans, H.; Dieterich, C. Bayesian prediction of RNA translation from ribosome profiling. *Nucleic Acids Res.* **2017**, *45*, 2960–2972. [[CrossRef](#)]
61. Ingolia, N.T.; Brar, G.A.; Stern-Ginossar, N.; Harris, M.S.; Talhouarne, G.J.; Jackson, S.E.; Wills, M.R.; Weissman, J.S. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* **2014**, *8*, 1365–1379. [[CrossRef](#)]
62. Liu, Q.; Shvarts, T.; Sliz, P.; Gregory, R.I. RiboToolKit: An integrated platform for analysis and annotation of ribosome profiling data to decode mRNA translation at codon resolution. *Nucleic Acids Res.* **2020**, *48*, W218–W229. [[CrossRef](#)]
63. Michel, A.M.; Fox, G.; Kiran, A.M.; De Bo, C.; O'Connor, P.B.; Heaphy, S.M.; Mullan, J.P.; Donohue, C.A.; Higgins, D.G.; Baranov, P.V. GWIPS-viz: Development of a ribo-seq genome browser. *Nucleic Acids Res.* **2014**, *42*, D859–D864. [[CrossRef](#)]
64. Kiniry, S.J.; O'Connor, P.B.F.; Michel, A.M.; Baranov, P.V. Trips-Viz: A transcriptome browser for exploring Ribo-Seq data. *Nucleic Acids Res.* **2019**, *47*, D847–D852. [[CrossRef](#)]
65. Khitun, A.; Slavoff, S.A. Proteomic Detection and Validation of Translated Small Open Reading Frames. *Curr. Protoc. Chem. Biol.* **2019**, *11*, e77. [[CrossRef](#)]
66. Zhang, Q.; Wu, E.; Tang, Y.; Cai, T.; Zhang, L.; Wang, J.; Hao, Y.; Zhang, B.; Zhou, Y.; Guo, X.; et al. Deeply Mining a Universe of Peptides Encoded by Long Noncoding RNAs. *Mol. Cell. Proteom. MCP* **2021**, *20*, 100109. [[CrossRef](#)]
67. Ma, J.; Diedrich, J.K.; Jungreis, I.; Donaldson, C.; Vaughan, J.; Kellis, M.; Yates, J.R., 3rd; Saghatelian, A. Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. *Anal. Chem.* **2016**, *88*, 3967–3975. [[CrossRef](#)]
68. 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. [[CrossRef](#)]
69. Ma, J.; Ward, C.C.; Jungreis, I.; Slavoff, S.A.; Schwaid, A.G.; Neveu, J.; Budnik, B.A.; Kellis, M.; Saghatelian, A. Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J. Proteome Res.* **2014**, *13*, 1757–1765. [[CrossRef](#)]
70. 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. [[CrossRef](#)]
71. Chen, L.; Yang, Y.; Zhang, Y.; Li, K.; Cai, H.; Wang, H.; Zhao, Q. The Small Open Reading Frame-Encoded Peptides: Advances in Methodologies and Functional Studies. *Chembiochem A Eur. J. Chem. Biol.* **2022**, *23*, e202100534. [[CrossRef](#)]
72. Huesgen, P.F.; Lange, P.F.; Rogers, L.D.; Solis, N.; Eckhard, U.; Kleifeld, O.; Goulas, T.; Gomis-Rüth, F.X.; Overall, C.M. LysargiNase mirrors trypsin for protein C-terminal and methylation-site identification. *Nat. Methods* **2015**, *12*, 55–58. [[CrossRef](#)]
73. Bartel, J.; Varadarajan, A.R.; Sura, T.; Ahrens, C.H.; Maaß, S.; Becher, D. Optimized Proteomics Workflow for the Detection of Small Proteins. *J. Proteome Res.* **2020**, *19*, 4004–4018. [[CrossRef](#)]
74. Kaulich, P.T.; Cassidy, L.; Bartel, J.; Schmitz, R.A.; Tholey, A. Multi-protease Approach for the Improved Identification and Molecular Characterization of Small Proteins and Short Open Reading Frame-Encoded Peptides. *J. Proteome Res.* **2021**, *20*, 2895–2903. [[CrossRef](#)]
75. D'Lima, N.G.; Ma, J.; Winkler, L.; Chu, Q.; Loh, K.H.; Corpuz, E.O.; Budnik, B.A.; Lykke-Andersen, J.; Saghatelian, A.; Slavoff, S.A. A human microprotein that interacts with the mRNA decapping complex. *Nat. Chem. Biol.* **2017**, *13*, 174–180. [[CrossRef](#)]
76. Hemm, M.R.; Weaver, J.; Storz, G. *Escherichia coli* Small Proteome. *EcoSal Plus* **2020**, *9*, 1–16. [[CrossRef](#)]
77. Fesenko, I.; Kirov, I.; Kniazev, A.; Khazigaleeva, R.; Lazarev, V.; Kharlampieva, D.; Grafkaia, 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. [[CrossRef](#)]
78. Ahrens, C.H.; Wade, J.T.; Champion, M.M.; Langer, J.D. A Practical Guide to Small Protein Discovery and Characterization Using Mass Spectrometry. *J. Bacteriol.* **2022**, *204*, e0035321. [[CrossRef](#)]
79. Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc* **2016**, *11*, 2301–2319. [[CrossRef](#)]
80. Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A. The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol. Cell. Proteom. MCP* **2004**, *3*, 531–533. [[CrossRef](#)]
81. Deutsch, E.W.; Mendoza, L.; Shteynberg, D.; Farrar, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10*, 1150–1159. [[CrossRef](#)]
82. Ludwig, C.; Claassen, M.; Schmidt, A.; Aebersold, R. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry. *Mol. Cell. Proteom. MCP* **2012**, *11*, M111.013987. [[CrossRef](#)]
83. Fabre, B.; Combier, J.P.; Plaza, S. Recent advances in mass spectrometry-based peptidomics workflows to identify short-open-reading-frame-encoded peptides and explore their functions. *Curr. Opin. Chem. Biol.* **2021**, *60*, 122–130. [[CrossRef](#)]

84. Pak, H.; Michaux, J.; Huber, F.; Chong, C.; Stevenson, B.J.; Müller, M.; Coukos, G.; Bassani-Sternberg, M. Sensitive Immunopeptidomics by Leveraging Available Large-Scale Multi-HLA Spectral Libraries, Data-Independent Acquisition, and MS/MS Prediction. *Mol. Cell. Proteom. MCP* **2021**, *20*, 100080. [[CrossRef](#)]
85. Delcourt, V.; Brunelle, M.; Roy, A.V.; Jacques, J.F.; Salzet, M.; Fournier, I.; Roucou, X. The Protein Coded by a Short Open Reading Frame, Not by the Annotated Coding Sequence, Is the Main Gene Product of the Dual-Coding Gene MIEF1. *Mol. Cell. Proteom. MCP* **2018**, *17*, 2402–2411. [[CrossRef](#)]
86. 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. [[CrossRef](#)]
87. Choteau, S.A.; Wagner, A.; Pierre, P.; Spinelli, L.; Brun, C. MetamORF: A repository of unique short open reading frames identified by both experimental and computational approaches for gene and metagene analyses. *Database J. Biol. Databases Curation* **2021**, *2021*, baab032. [[CrossRef](#)]
88. Hazarika, R.R.; De Coninck, B.; Yamamoto, L.R.; Martin, L.R.; Cammue, B.P.; van Noort, V. ARA-PEPs: A repository of putative sORF-encoded peptides in *Arabidopsis thaliana*. *BMC Bioinform.* **2017**, *18*, 37. [[CrossRef](#)]
89. Ruggles, K.V.; Krug, K.; Wang, X.; Clauser, K.R.; Wang, J.; Payne, S.H.; Fenyö, D.; Zhang, B.; Mani, D.R. Methods, Tools and Current Perspectives in Proteogenomics. *Mol. Cell. Proteom. MCP* **2017**, *16*, 959–981. [[CrossRef](#)]
90. Deng, Y.; Bamigbade, A.T.; Hammad, M.A.; Xu, S.; Liu, P. Identification of small ORF-encoded peptides in mouse serum. *Biophys. Rep.* **2018**, *4*, 39–49. [[CrossRef](#)]
91. Cai, T.; Zhang, Q.; Wu, B.; Wang, J.; Li, N.; Zhang, T.; Wang, Z.; Luo, J.; Guo, X.; Ding, X.; et al. lncRNA-encoded microproteins: A new form of cargo in cell culture-derived and circulating extracellular vesicles. *J. Extracell. Vesicles* **2021**, *10*, e12123. [[CrossRef](#)]
92. 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. [[CrossRef](#)]
93. Szalay, T.; Golovchenko, J.A. De novo sequencing and variant calling with nanopores using PoreSeq. *Nat. Biotechnol.* **2015**, *33*, 1087–1091. [[CrossRef](#)]
94. Chen, L.; Zhang, Y.; Yang, Y.; Li, H.; Dong, X.; Wang, H.; Xie, Z.; Zhao, Q. An Integrated Approach for Discovering Noncanonical MHC-I Peptides Encoded by Small Open Reading Frames. *J. Am. Soc. Mass. Spectrom.* **2021**, *32*, 2346–2357. [[CrossRef](#)]
95. Badger, J.H.; Olsen, G.J. CRITICA: Coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.* **1999**, *16*, 512–524. [[CrossRef](#)]
96. Kang, Y.J.; Yang, D.C.; Kong, L.; Hou, M.; Meng, Y.Q.; Wei, L.; Gao, G. CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* **2017**, *45*, W12–W16. [[CrossRef](#)]
97. Li, A.; Zhang, J.; Zhou, Z. PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinform.* **2014**, *15*, 311. [[CrossRef](#)]
98. Bazzini, A.A.; Johnstone, T.G.; Christiano, R.; Mackowiak, S.D.; Obermayer, B.; Fleming, E.S.; Vejnar, C.E.; Lee, M.T.; Rajewsky, N.; Walther, T.C.; et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **2014**, *33*, 981–993. [[CrossRef](#)]
99. Lin, M.F.; Jungreis, I.; Kellis, M. PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **2011**, *27*, i275–i282. [[CrossRef](#)]
100. Siepel, A.; Bejerano, G.; Pedersen, J.S.; Hinrichs, A.S.; Hou, M.; Rosenbloom, K.; Clawson, H.; Spieth, J.; Hillier, L.W.; Richards, S.; et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **2005**, *15*, 1034–1050. [[CrossRef](#)]
101. Hanada, K.; Akiyama, K.; Sakurai, T.; Toyoda, T.; Shinozaki, K.; Shiu, S.H. sORF finder: A program package to identify small open reading frames with high coding potential. *Bioinformatics* **2010**, *26*, 399–400. [[CrossRef](#)]
102. Washietl, S.; Findeiss, S.; Müller, S.A.; Kalkhof, S.; von Bergen, M.; Hofacker, I.L.; Stadler, P.F.; Goldman, N. RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* **2011**, *17*, 578–594. [[CrossRef](#)]
103. Sun, L.; Luo, H.; Bu, D.; Zhao, G.; Yu, K.; Zhang, C.; Liu, Y.; Chen, R.; Zhao, Y. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* **2013**, *41*, e166. [[CrossRef](#)]
104. Wang, T.; Cui, Y.; Jin, J.; Guo, J.; Wang, G.; Yin, X.; He, Q.Y.; Zhang, G. Translating mRNAs strongly correlate to proteins in a multivariate manner and their translation ratios are phenotype specific. *Nucleic Acids Res.* **2013**, *41*, 4743–4754. [[CrossRef](#)]
105. Guo, J.C.; Fang, S.S.; Wu, Y.; Zhang, J.H.; Chen, Y.; Liu, J.; Wu, B.; Wu, J.R.; Li, E.M.; Xu, L.Y.; et al. CNIT: A fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. *Nucleic Acids Res.* **2019**, *47*, W516–W522. [[CrossRef](#)]
106. Sayers, E.W.; Beck, J.; Bolton, E.E.; Bourexis, D.; Brister, J.R.; Canese, K.; Comeau, D.C.; Funk, K.; Kim, S.; Klimke, W.; et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2021**, *49*, D10–D17. [[CrossRef](#)]
107. Sun, K.; Chen, X.; Jiang, P.; Song, X.; Wang, H.; Sun, H. iSeeRNA: Identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genom.* **2013**, *14* (Suppl. S2), S7. [[CrossRef](#)]
108. Hu, L.; Xu, Z.; Hu, B.; Lu, Z.J. COME: A robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. *Nucleic Acids Res.* **2017**, *45*, e2. [[CrossRef](#)]
109. Achawanantakun, R.; Chen, J.; Sun, Y.; Zhang, Y. lncRNA-ID: Long non-coding RNA IDentification using balanced random forests. *Bioinformatics* **2015**, *31*, 3897–3905. [[CrossRef](#)]
110. Fan, X.N.; Zhang, S.W. lncRNA-MFDL: Identification of human long non-coding RNAs by fusing multiple features and using deep learning. *Mol. BioSyst.* **2015**, *11*, 892–897. [[CrossRef](#)]

111. Skarshewski, A.; Stanton-Cook, M.; Huber, T.; Al Mansoori, S.; Smith, R.; Beatson, S.A.; Rothnagel, J.A. uPEPPER: An online tool for upstream open reading frame location and analysis of transcript conservation. *BMC Bioinform.* **2014**, *15*, 36. [[CrossRef](#)]
112. Zhang, Y.; Jia, C.; Fullwood, M.J.; Kwok, C.K. DeepCPP: A deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for RNA coding potential prediction. *Brief. Bioinform.* **2021**, *22*, 2073–2084. [[CrossRef](#)]
113. Camargo, A.P.; Sourkov, V.; Pereira, G.A.G.; Carazzolle, M.F. RNAsamba: Neural network-based assessment of the protein-coding potential of RNA sequences. *NAR Genom. Bioinform.* **2020**, *2*, lqz024. [[CrossRef](#)]
114. Zhu, M.; Gribskov, M. MiPeptide: MicroPeptide identification tool using machine learning. *BMC Bioinform.* **2019**, *20*, 559. [[CrossRef](#)]
115. Cao, Y.; Yang, R.; Lee, I.; Zhang, W.; Sun, J.; Meng, X.; Wang, W. Prediction of LncRNA-encoded small peptides in glioma and oligomer channel functional analysis using in silico approaches. *PLoS ONE* **2021**, *16*, e0248634. [[CrossRef](#)]
116. Choi, S.W.; Kim, H.W.; Nam, J.W. The small peptide world in long noncoding RNAs. *Brief. Bioinform.* **2019**, *20*, 1853–1864. [[CrossRef](#)]
117. Fesenko, I.; Shabalina, S.A.; Mamaeva, A.; Knyazev, A.; Glushkevich, A.; Lyapina, I.; Ziganshin, R.; Kovalchuk, S.; Kharlampieva, D.; Lazarev, V.; et al. A vast pool of lineage-specific microproteins encoded by long non-coding RNAs in plants. *Nucleic Acids Res.* **2021**, *49*, 10328–10346. [[CrossRef](#)]
118. Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **2001**, *305*, 567–580. [[CrossRef](#)]
119. Almagro Armenteros, J.J.; Tsirigos, K.D.; Sønderby, C.K.; Petersen, T.N.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **2019**, *37*, 420–423. [[CrossRef](#)]
120. Duvaud, S.; Gabella, C.; Lisacek, F.; Stockinger, H.; Ioannidis, V.; Durinx, C. ExPasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Res.* **2021**, *49*, W216–W227. [[CrossRef](#)]
121. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [[CrossRef](#)]
122. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F.T.; de Beer, T.A.P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **2018**, *46*, W296–W303. [[CrossRef](#)]
123. Kavousi, K.; Bagheri, M.; Behrouzi, S.; Vafadar, S.; Atanaki, F.F.; Lotfabadi, B.T.; Ariaeenejad, S.; Shockravi, A.; Moosavi-Movahedi, A.A. IAMPE: NMR-Assisted Computational Prediction of Antimicrobial Peptides. *J. Chem. Inf. Model.* **2020**, *60*, 4691–4701. [[CrossRef](#)]
124. Käll, L.; Krogh, A.; Sonnhammer, E.L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **2004**, *338*, 1027–1036. [[CrossRef](#)]
125. Finn, R.D.; Coghill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; et al. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **2016**, *44*, D279–D285. [[CrossRef](#)]
126. Zhou, C.; Wang, Q.; Huang, Y.; Chen, Z.; Chen, S.; Zhao, Y.; Jia, C. Probing the sORF-Encoded Peptides of *Deinococcus radiodurans* in Response to Extreme Stress. *Mol. Cell. Proteom. MCP* **2022**, *21*, 100423. [[CrossRef](#)]
127. Savojardo, C.; Martelli, P.L.; Fariselli, P.; Profiti, G.; Casadio, R. BUSCA: An integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res.* **2018**, *46*, W459–W466. [[CrossRef](#)]
128. Geourjon, C.; Deléage, G. SOPMA: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci.* **1995**, *11*, 681–684. [[CrossRef](#)]
129. Li, M.; Li, X.; Zhang, Y.; Wu, H.; Zhou, H.; Ding, X.; Zhang, X.; Jin, X.; Wang, Y.; Yin, X.; et al. Micropeptide MIAC Inhibits HNSCC Progression by Interacting with Aquaporin 2. *J. Am. Chem. Soc.* **2020**, *142*, 6708–6716. [[CrossRef](#)]
130. Makarewich, C.A.; Olson, E.N. Mining for Micropeptides. *Trends Cell Biol.* **2017**, *27*, 685–696. [[CrossRef](#)]
131. Naval, B.R.; Makarewich, C.A.; Anderson, D.M.; Winders, B.R.; Troupes, C.D.; Wu, F.; Reese, A.L.; McAnally, J.R.; Chen, X.; Kavalali, E.T.; et al. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* **2016**, *351*, 271–275. [[CrossRef](#)]
132. Sousa, M.E.; Farkas, M.H. Micropeptide. *PLoS Genet.* **2018**, *14*, e1007764. [[CrossRef](#)]
133. Zhang, S.; Reljic, B.; Liang, C.; Kerouanton, B.; Francisco, J.C.; Peh, J.H.; Mary, C.; Jagannathan, N.S.; Olexioux, V.; Tang, C.; et al. Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. *Nat. Commun.* **2020**, *11*, 1312. [[CrossRef](#)]
134. Na, Z.; Luo, Y.; Schofield, J.A.; Smelyansky, S.; Khitun, A.; Muthukumar, S.; Valkov, E.; Simon, M.D.; Slavoff, S.A. The NBDY Microprotein Regulates Cellular RNA Decapping. *Biochemistry* **2020**, *59*, 4131–4142. [[CrossRef](#)]
135. Ge, Q.; Jia, D.; Cen, D.; Qi, Y.; Shi, C.; Li, J.; Sang, L.; Yang, L.J.; He, J.; Lin, A.; et al. Micropeptide ASAP encoded by LINC00467 promotes colorectal cancer progression by directly modulating ATP synthase activity. *J. Clin. Investig.* **2021**, *131*, e152911. [[CrossRef](#)]
136. Polycarpou-Schwarz, M.; Groß, M.; Mestdagh, P.; Schott, J.; Grund, S.E.; Hildenbrand, C.; Rom, J.; Aulmann, S.; Sinn, H.P.; Vandesompele, J.; et al. The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. *Oncogene* **2018**, *37*, 4750–4768. [[CrossRef](#)]
137. Chu, Q.; Rathore, A.; Diedrich, J.K.; Donaldson, C.J.; Yates, J.R., 3rd; Saghatelian, A. Identification of Microprotein-Protein Interactions via APEX Tagging. *Biochemistry* **2017**, *56*, 3299–3306. [[CrossRef](#)]
138. Fu, H.; Wang, T.; Kong, X.; Yan, K.; Yang, Y.; Cao, J.; Yuan, Y.; Wang, N.; Kee, K.; Lu, Z.J.; et al. A Nodal enhanced micropeptide NEMEP regulates glucose uptake during mesendoderm differentiation of embryonic stem cells. *Nat. Commun.* **2022**, *13*, 3984. [[CrossRef](#)]
139. Pang, Y.; Liu, Z.; Han, H.; Wang, B.; Li, W.; Mao, C.; Liu, S. Peptide SMIM30 promotes HCC development by inducing SRC/YES1 membrane anchoring and MAPK pathway activation. *J. Hepatol.* **2020**, *73*, 1155–1169. [[CrossRef](#)]
140. Matsumoto, A.; Clohessy, J.G.; Pandolfi, P.P. SPAR, a lncRNA encoded mTORC1 inhibitor. *Cell Cycle* **2017**, *16*, 815–816. [[CrossRef](#)]

141. Tharakan, R.; Sawa, A. Minireview: Novel Micropeptide Discovery by Proteomics and Deep Sequencing Methods. *Front. Genet.* **2021**, *12*, 651485. [[CrossRef](#)]
142. Cai, B.; Li, Z.; Ma, M.; Wang, Z.; Han, P.; Abdalla, B.A.; Nie, Q.; Zhang, X. LncRNA-Six1 Encodes a Micropeptide to Activate Six1 in Cis and Is Involved in Cell Proliferation and Muscle Growth. *Front. Physiol.* **2017**, *8*, 230. [[CrossRef](#)]
143. Zhu, S.; Wang, J.Z.; Chen, D.; He, Y.T.; Meng, N.; Chen, M.; Lu, R.X.; Chen, X.H.; Zhang, X.L.; Yan, G.R. An oncopeptide regulates m(6)A recognition by the m(6)A reader IGF2BP1 and tumorigenesis. *Nat. Commun.* **2020**, *11*, 1685. [[CrossRef](#)]
144. Liang, X.H.; Shen, W.; Sun, H.; Migawa, M.T.; Vickers, T.A.; Crooke, S.T. Translation efficiency of mRNAs is increased by antisense oligonucleotides targeting upstream open reading frames. *Nat. Biotechnol.* **2016**, *34*, 875–880. [[CrossRef](#)]
145. Liang, X.H.; Sun, H.; Shen, W.; Wang, S.; Yao, J.; Migawa, M.T.; Bui, H.H.; Damle, S.S.; Riney, S.; Graham, M.J.; et al. Antisense oligonucleotides targeting translation inhibitory elements in 5' UTRs can selectively increase protein levels. *Nucleic Acids Res.* **2017**, *45*, 9528–9546. [[CrossRef](#)]
146. Anderson, D.M.; Makarewich, C.A.; Anderson, K.M.; Shelton, J.M.; Bezprozvannaya, S.; Bassel-Duby, R.; Olson, E.N. Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. *Sci. Signal.* **2016**, *9*, ra119. [[CrossRef](#)]
147. Zhang, Q.; Vashisht, A.A.; O'Rourke, J.; Corbel, S.Y.; Moran, R.; Romero, A.; Miraglia, L.; Zhang, J.; Durrant, E.; Schmedt, C.; et al. The microprotein Minion controls cell fusion and muscle formation. *Nat. Commun.* **2017**, *8*, 15664. [[CrossRef](#)]
148. Shi, J.; Bi, P.; Pei, J.; Li, H.; Grishin, N.V.; Bassel-Duby, R.; Chen, E.H.; Olson, E.N. Requirement of the fusogenic micropeptide myomixer for muscle formation in zebrafish. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 11950–11955. [[CrossRef](#)]
149. Magny, E.G.; Pueyo, J.I.; Pearl, F.M.; Cespedes, M.A.; Niven, J.E.; Bishop, S.A.; Couso, J.P. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* **2013**, *341*, 1116–1120. [[CrossRef](#)]
150. Wang, L.; Fan, J.; Han, L.; Qi, H.; Wang, Y.; Wang, H.; Chen, S.; Du, L.; Li, S.; Zhang, Y.; et al. The micropeptide LEMP plays an evolutionarily conserved role in myogenesis. *Cell Death Dis.* **2020**, *11*, 357. [[CrossRef](#)]
151. van Solingen, C.; Sharma, M.; Bijkerk, R.; Afonso, M.S.; Koelwyn, G.J.; Scacalossi, K.R.; Holdt, L.M.; Maegdefessel, L.; van Zonneveld, A.J.; Moore, K.J. Abstract 544: A Novel Micropeptide, IMP, Directs Inflammation Through Interaction with Transcriptional Co-activators. *Arterioscler. Thromb. Vasc. Biol.* **2019**, *39*, A544.
152. Xiang, X.; Fu, Y.; Zhao, K.; Miao, R.; Zhang, X.; Ma, X.; Liu, C.; Zhang, N.; Qu, K. Cellular senescence in hepatocellular carcinoma induced by a long non-coding RNA-encoded peptide PINT87aa by blocking FOXM1-mediated PHB2. *Theranostics* **2021**, *11*, 4929–4944. [[CrossRef](#)]
153. Wu, S.; Zhang, L.; Deng, J.; Guo, B.; Li, F.; Wang, Y.; Wu, R.; Zhang, S.; Lu, J.; Zhou, Y. A Novel Micropeptide Encoded by Y-Linked LINC00278 Links Cigarette Smoking and AR Signaling in Male Esophageal Squamous Cell Carcinoma. *Cancer Res.* **2020**, *80*, 2790–2803. [[CrossRef](#)]
154. Pan, Z.; Cai, J.; Lin, J.; Zhou, H.; Peng, J.; Liang, J.; Xia, L.; Yin, Q.; Zou, B.; Zheng, J.; et al. A novel protein encoded by circFNDC3B inhibits tumor progression and EMT through regulating Snail in colon cancer. *Mol. Cancer* **2020**, *19*, 71. [[CrossRef](#)]
155. Zheng, X.; Chen, L.; Zhou, Y.; Wang, Q.; Zheng, Z.; Xu, B.; Wu, C.; Zhou, Q.; Hu, W.; Wu, C.; et al. A novel protein encoded by a circular RNA circPPP1R12A promotes tumor pathogenesis and metastasis of colon cancer via Hippo-YAP signaling. *Mol. Cancer* **2019**, *18*, 47. [[CrossRef](#)]
156. Godet, Y.; Moreau-Aubry, A.; Guilloux, Y.; Vignard, V.; Khammari, A.; Dreno, B.; Jotereau, F.; Labarriere, N. MELOE-1 is a new antigen overexpressed in melanomas and involved in adoptive T cell transfer efficiency. *J. Exp. Med.* **2008**, *205*, 2673–2682. [[CrossRef](#)]
157. Godet, Y.; Moreau-Aubry, A.; Mompelat, D.; Vignard, V.; Khammari, A.; Dreno, B.; Lang, F.; Jotereau, F.; Labarriere, N. An additional ORF on meloe cDNA encodes a new melanoma antigen, MELOE-2, recognized by melanoma-specific T cells in the HLA-A2 context. *Cancer Immunol. Immunother.* **2010**, *59*, 431–439. [[CrossRef](#)]
158. Guo, B.; Wu, S.; Zhu, X.; Zhang, L.; Deng, J.; Li, F.; Wang, Y.; Zhang, S.; Wu, R.; Lu, J.; et al. Micropeptide CIP2A-BP encoded by LINC00665 inhibits triple-negative breast cancer progression. *EMBO J.* **2020**, *39*, e102190. [[CrossRef](#)]
159. Xu, W.; Deng, B.; Lin, P.; Liu, C.; Li, B.; Huang, Q.; Zhou, H.; Yang, J.; Qu, L. Ribosome profiling analysis identified a KRAS-interacting microprotein that represses oncogenic signaling in hepatocellular carcinoma cells. *Sci. China Life Sci.* **2020**, *63*, 529–542. [[CrossRef](#)]
160. Lu, S.; Zhang, J.; Lian, X.; Sun, L.; Meng, K.; Chen, Y.; Sun, Z.; Yin, X.; Li, Y.; Zhao, J.; et al. A hidden human proteome encoded by 'non-coding' genes. *Nucleic Acids Res.* **2019**, *47*, 8111–8125. [[CrossRef](#)]
161. Huang, J.Z.; Chen, M.; Chen, D.; Gao, X.C.; Zhu, S.; Huang, H.; Hu, M.; Zhu, H.; Yan, G.R. A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. *Mol. Cell* **2017**, *68*, 171–184.e6. [[CrossRef](#)]
162. Li, Y.; Zhang, J.; Sun, H.; Chen, Y.; Li, W.; Yu, X.; Zhao, X.; Zhang, L.; Yang, J.; Xin, W.; et al. Inc-Rps4l-encoded peptide RPS4XL regulates RPS6 phosphorylation and inhibits the proliferation of PSMCs caused by hypoxia. *Mol. Ther.* **2021**, *29*, 1411–1424. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.