


REVIEW

Open Access



Short open reading frames (sORFs) and microproteins: an update on their identification and validation measures

Alyssa Zi-Xin Leong¹, Pey Yee Lee¹, M. Aiman Mohtar¹, Saiful Effendi Syafruddin¹, Yuh-Fen Pung² and Teck Yew Low^{1*} 

Abstract

A short open reading frame (sORFs) constitutes ≤ 300 bases, encoding a microprotein or sORF-encoded protein (SEP) which comprises ≤ 100 amino acids. Traditionally dismissed by genome annotation pipelines as meaningless noise, sORFs were found to possess coding potential with ribosome profiling (RIBO-Seq), which unveiled sORF-based transcripts at various genome locations. Nonetheless, the existence of corresponding microproteins that are stable and functional was little substantiated by experimental evidence initially. With recent advancements in multi-omics, the identification, validation, and functional characterisation of sORFs and microproteins have become feasible. In this review, we discuss the history and development of an emerging research field of sORFs and microproteins. In particular, we focus on an array of bioinformatics and OMICS approaches used for predicting, sequencing, validating, and characterizing these recently discovered entities. These strategies include RIBO-Seq which detects sORF transcripts via ribosome footprints, and mass spectrometry (MS)-based proteomics for sequencing the resultant microproteins. Subsequently, our discussion extends to the functional characterisation of microproteins by incorporating CRISPR/Cas9 screen and protein–protein interaction (PPI) studies. Our review discusses not only detection methodologies, but we also highlight on the challenges and potential solutions in identifying and validating sORFs and their microproteins. The novelty of this review lies within its validation for the functional role of microproteins, which could contribute towards the future landscape of microproteomics.

Keywords: Short open reading frame (sORF), Small open reading frame (smORF), Microproteins, Ribosome profiling (RIBO-Seq), Mass spectrometry, Proteogenomics

Background

Spurred by the first draft of the human genome in 2001, the number of annotated human genes increased drastically in the next decade, with $\sim 20,000$ papers published annually on the protein-coding genes [1, 2]. By decoding the bulk of human DNA, the Human Genome Project (HGP) had empowered genomic research and extended

its impacts to many other species including mouse, rat, fruit-fly and even to plants such as *Arabidopsis thaliana*.

Genome annotation refers to the procedure whereby the locations, coding regions and functions of genes are determined within a sequenced genome. This process is typically performed by automated bioinformatic pipelines based on (i) the innate characteristics and features of genomic sequences, as well as (ii) the sequence homology conserved through evolution [3, 4]. These two principles tie together the possibility of genes acquiring evolutionary selective advantages, thus producing proteins with functionalities essential

*Correspondence: lowteckyew@ppukm.ukm.edu.my

¹ UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, 56000 Kuala Lumpur, Malaysia

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

for survival. Via this process, putative protein-coding genes can be predicted ab initio. This kickstarted the search for protein-coding genes, plateauing in the mid-2000s at ~20,000 protein-coding genes for the human genome [2]. However, this number may be underestimated due to an arbitrary limitation of a 300-base (100-codon) cut-off for transcripts. Such restrictions are imposed on ORF-prediction algorithms by the Functional ANnotation Of the Mammalian Genome (FANTOM) consortium to minimize false positive predictions, especially the mis-classification of non-coding RNAs as mRNAs [5–8].

In the past two decades, the scientific community had nonetheless found mounting experimental evidence for open reading frames (ORFs) comprising <100 codons. These so-called short ORFs (sORFs) or small ORFs (smORFs) can encode functional and stable sORF-encoded protein (SEPs) or microproteins. The first revelation of a functional microprotein came from the discovery of a novel helix-loop-helix protein in 1990 by Benezra et al., known as the “Id” protein. Id functions by inhibiting the trans-activation of MCK gene at the MyoD consensus binding site in myoblasts during muscle differentiation [9]. Whereas for plants, the first discovered microproteins are the LITTLE ZIPPER (ZPR) proteins. Being functional analogues to “Id” proteins, ZPR proteins control stem cell maintenance during plant development [10].

Several recently identified microproteins have been associated with diseases. For example, a 54-amino acid mitochondrial microprotein known as PIGBOS was believed to play a role in stress signalling. PIGBOS is localised in the mitochondrial membrane and mediates signalling events leading to the unfolded protein response (UPR) [11]. This microprotein interacts with CLCC1, an endoplasmic reticulum (ER) protein, forming a connection between the mitochondria and the ER [12]. PIGBOS’s role in stress signalling may indicate a new direction in research concerning high ER stress, including cancer. Another microprotein that is linked to cancer is CYREN, which inhibits non-homologous end joining (NHEJ) repair during S and G2 phases [13]. CYREN ensures accurate homologous recombination, and its dysfunction may destabilise DNA integrity.

In accordance with advances in OMICS research and bioinformatics, functional and stable microproteins have increasingly been identified and characterized. We therefore dedicate this review to discuss the current developments in microproteome research, followed by state-of-the-art strategies that are used to identify and validate them at different biomolecular levels.

Short open reading frames (sORFs)

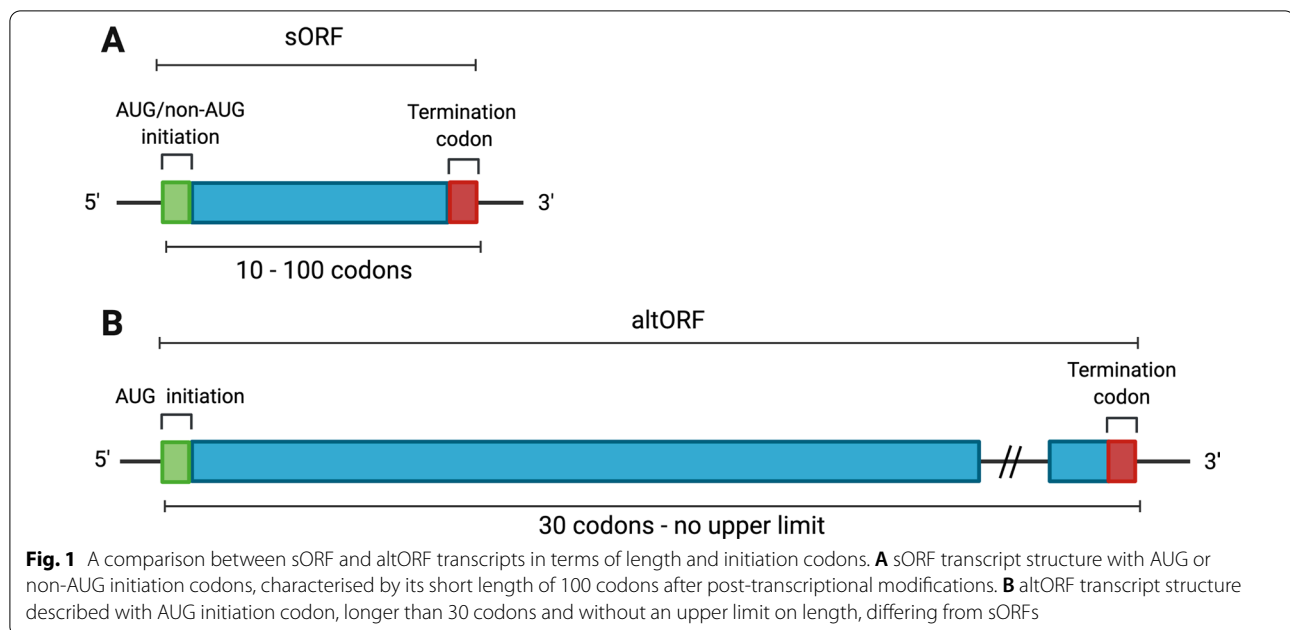
Conventionally, an ORF is defined as a stretch of consecutive and non-overlapping nucleotide triplets (codons) that can be translated into proteins, whereby it should also initiate with an in-frame start codon (AUG), and terminate with one of the three stop codons (UAA, UAG, UGA). In a theoretical manner, Olexiouk et al. (2018) estimated that the probability of randomly generating a start codon within the nucleotide space is 1 out of 64, and that the chances of finding a stop codon within the next 99 codons is ~99%; discounting splice variants, reading frames and GC-rich regions, strandedness and nucleotide biases [14]. Consequently, this means that ~1.5% of the genome may encode ORFs <100 codons [14, 15]. Naturally, this results in an unreasonably large number of putative sORFs, whose chances of being transcribed and translated into functional polypeptides seem far-fetched. Hence, a 300-nucleotides cut-off was introduced, as most of these sORFs were deemed meaningless and random [16, 17].

Another contributing factor for this cut-off is that existing algorithms are so far not ideal for annotating the sORFs. This is due to the propensity of short sequences to score low in evolutionary conservation, an indicator for functionality [18]. Combined with the technical difficulties in delineating sORFs from chance in-frame start and stop codons and to isolate sORF-translated microproteins, these sORFs were either considered noise, occurring by chance, non-coding or unlikely to be translated [19–21]. As such, these sORFs and the resulting microproteins were traditionally ignored by the scientific community.

Ever since its discovery, sORFs, along with its short lengths, were considered unorthodox, as it could be initiated with AUG as well as non-AUG codons [22–24]. As unconventional ORFs, the definition of sORFs overlaps with another distinctive class of ORFs, known as alternative ORFs (altORFs) (Fig. 1). The altORFs yield transcripts that initiate only with AUG codons and are at least 30 codons, but without an upper length limit [16]. AltORFs were also found to encode proteins, an example being AltMRV11 from the 3’ UTR of *MRV11* interacting with BRCA1 in the nucleus [25]. The overlapping definitions of these unconventional ORFs were consequently represented in protein databases, where there is apparent annotation of sORFs under altORFs prior to sole focus of research on sORFs.

The localities and regulatory functions of sORFs

The 300-base cut-off raises a paradox by contradicting the initial aims of HGP to resolve all ambiguities in the human genome [1]. Post cut-off, there is now a limitation



concerning the sORFs and their translated products [4, 20]. There is indefiniteness when it comes to discussing sORFs, and whether they exist putatively or due to random sequence matching to other ORFs. This subsequently leads to obscurity in the methods used for determining microproteins.

In continuance from a bulk of research on protein-coding genes, there was simultaneously an exponential increase in non-protein coding elements discovered [2]. By scanning these non-coding regions, researchers have found embedded sORFs that scatter in different genomic locations [26]. The different localities of sORFs are shown in Fig. 2. These locations include the upstream (uORF) and downstream open reading frames (dORF) within the 5' and 3' untranslated regions (UTRs) of a gene, even overlapping with the main ORF if the sORF is out of frame [27–31]. Examples of small uORFs include two uORFs in *MDM2*, where the translation products repress the main ORF encoding *MDM2* [28]. Other essential genes involved in the developmental process such as *POU5F3* (*Oct4*), *Smad7* and *Nanog* also encode multiple small uORFs, as discovered in zebrafish and human [28, 30]. Whereas, the expression of dORFs was found to enhance the translation of canonical ORFs, where thousands of dORFs have been found translated in human cells and zebrafish embryos [31]. Furthermore, an example of an overlapping sORF is uORF2 in the *ATF4* gene, whereby it represses the transcription of *ATF4* under normal conditions [27, 29]. This suggests the regulatory role of sORF in post-transcriptional control and translational efficiency [20, 32, 33].

In addition, sORFs are found among pseudogenes and intergenic regions too, with the latter being more difficult to identify due to the high false-negative rates of existing gene finding algorithms [32, 34]. As examples, Kalyana-Sundaram et al. (2012) reported the expression of cancer-specific pseudogenes, such as *KLKPI* which encodes a 54-amino acid microprotein in LNCaP, a prostate cancer cell line [35]. Meanwhile, Hanada et al. (2007) discovered that 4282 sORFs are located in the intergenic regions, out of 7159 sORFs in *Arabidopsis thaliana* [36]. Apart from genomic DNA, sORFs are encoded by mitochondrial DNA (mtDNA), where the microprotein products have been shown to play roles in muscle and fat metabolism [37]. Besides, sORFs are found embedded in various RNA transcripts previously believed to be non-coding, such as pri-microRNA, circRNA, lincRNA and lncRNA [8, 20, 26, 38–49].

The emergence of new transcripts from genomic regions previously considered non-coding has presented sORFs as a new source of protein-coding genes, consistent with the depiction of an evolutionary mechanism for generating novel polypeptides. With evolutionary conservation, the continuous expression of certain sORFs may provide selective advantages, hence withstanding time and becoming de novo protein-coding genes [8].

Strategies to detect sORFs with bioinformatics prediction

The need to re-evaluate the coding potentials of sORFs was raised as these minute ORFs were found to be capable of being transcribed and possibly translated. Several

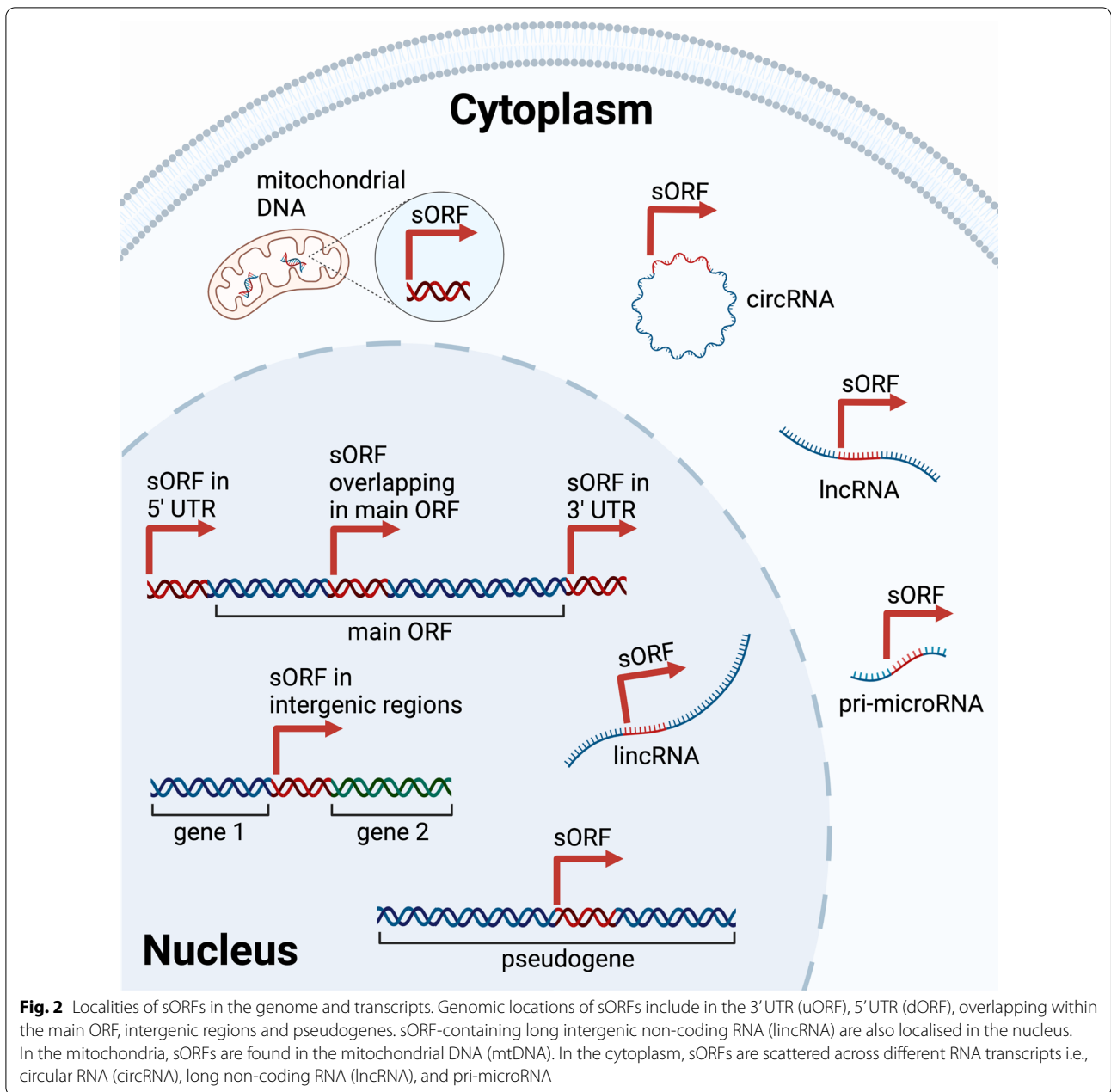


Fig. 2 Localities of sORFs in the genome and transcripts. Genomic locations of sORFs include in the 3' UTR (uORF), 5' UTR (dORF), overlapping within the main ORF, intergenic regions and pseudogenes. sORF-containing long intergenic non-coding RNA (lincRNA) are also localised in the nucleus. In the mitochondria, sORFs are found in the mitochondrial DNA (mtDNA). In the cytoplasm, sORFs are scattered across different RNA transcripts i.e., circular RNA (circRNA), long non-coding RNA (lincRNA), and pri-microRNA

methods based on computational analysis, next-generation sequencing (NGS) and mass spectrometry (MS) have been employed to predict the coding potentials, sequences, and identify these sORFs or their products. These measures confirm that such sORFs contain actual coding sequences (CDSs) and produce functional polypeptides; as some bona fide lincRNAs were found to not encode for microproteins, such as XIST, HOTAIR and NEAT1 [50]. To begin with, one must understand the difficulties in detecting microproteins, whereby their

smaller sizes (< 100 amino acids; < 20 kDa) necessitates the adaptation of existing laboratory techniques.

Bioinformatic predictions of sORFs proved valuable since it does not cost nearly as much as experimental validations. To differentiate expressed elements based on functionality, several aspects are considered i.e., (i) the conservation of a particular sequence through evolution, and (ii) sequence similarity. Sequence conservation weighs in evolutionary selection, indicating that the sequence remains functionally useful throughout

phylogenetic trees [3, 51]. Whereas sequence similarity denotes similar protein motifs or domains aligned over previously identified protein sequences so as to derive coding potentials and potential protein functionalities [4, 17].

Initially, the detection of sORFs was limited by the conservativeness of gene finder algorithms, as only few bioinformatic tools were specifically designed specifically for sORFs. One pioneering sORF prediction tool is sORF finder, which applies coding index (CI) based on nucleotide composition bias in predicting CDSs [52]. This tool successfully enabled the identification of 2376 putative

sORFs in the intergenic regions of *Arabidopsis thaliana* [36]. From this set of sORFs, Hanada et al. (2013) conducted a follow-up study and reported the overexpression of 473 sORFs that were associated with plant morphogenesis [53]. Since then, the training datasets used for such tools have become much larger, realizing higher prediction qualities. Experimental validations of sORFs are now used for homology searches, for ab initio training and for machine learning purposes in the development of better sORF prediction tools [4, 21]. Table 1 provides a list of sORF prediction tools that are applicable to multiple species, and their web addresses. It should

Table 1 sORF prediction tools

Prediction tool	References	Website	Description
Coding Non-Coding Identifying Tool (CNIT)	[126]	http://cnit.noncode.org/CNIT/	Distinguishes between coding and non-coding regions based on intrinsic sequence compositions
Coding Region Identification Tool Invoking Comparative Analysis (CRITICA)	[127]	http://rdpwww.life.uiuc.edu/	Analyses nucleotide sequence composition and conservation at the amino acid level
Coding Potential Calculator (CPC)/CPC2	[128, 129]	http://cpc.cbi.pku.edu.cn http://cpc2.gao-lab.org/	Assess protein-coding potential based on important features (ORF size, coverage, integrity); CPC2 improves run speed and accuracy
Coding Potential Predictor (CPPred)	[130]	http://www.rnabinding.com/CPred/	Predicts the coding potential of RNA transcript
CPPred-sORF	[131]	http://www.rnabinding.com/CPred-sORF/	Addition of 2 new features from CCPred i.e., GCcount, mRNN-11 codons and CUG, GUG start codons
MicroPeptide Tool (MiPepid)	[21]	https://github.com/MindAI/MiPepid	Identifies coding sORFs based on existing microproteins subpopulation set
sORF Finder	[52]	http://evolver.psc.riken.jp/	Identifies sORF with high coding potential based on nucleotide composition bias and potential functional constraint at the amino acid level
smORFunction	[132]	https://www.cuilab.cn/smorfunction/home	Provides function prediction of sORFs/microproteins
miPFinder	[133]	https://github.com/DaStraub/miPFinder	Identifies and evaluates microproteins functionality using information on size, domain, protein interactions and evolutionary origin
PhastCons	[134]	http://compugen.cshl.edu/phast/	Based on conservation scoring and identification of conserved elements
PhyloCSF	[135]	http://compbio.mit.edu/PhyloCSF	Determines a conserved protein-coding region based on formal statistical comparison of phylogenetic codon models
uPEPPERoni	[136]	http://u pep-scmb.biosci.uq.edu.au/	Specifically for 5'UTR sORFs, based on conservation
AnABLAST	[34]	http://www.bioinfocabd.upo.es/ab/	Identifies putative protein-coding regions in DNA regardless of ORF length and reading frame shifts
Small Peptide Alignment Discovery Application (SPADA)	[137]	https://github.com/orionzhou/SPADA	Homology-based gene prediction programme
Deep Neural Network for coding potential prediction (DeepCPP)	[138]	https://github.com/yuuuzhang/DeepCPP	Effective on RNA coding potential prediction, specifically sORF mRNA prediction

This table shows prediction tools that can be used for putative sORF detection based on sequence homology and similarity in all genomes. CNIT and CPPred utilises a positive set of normal-sized proteins and may not be optimised for sORF and microprotein detection. CPPred-sORF is an improved version of CPPred for sORF detection. MiPepid, sORF Finder, miPFinder and smORFunction are designed especially for sORF detection, identification, and function prediction. PhastCons, PhyloCSF, SPADA and uPEPPERoni utilise conservation analyses for prediction, with the latter designed specifically for sORFs in the upstream region. DeepCPP is based on a deep learning method to evaluate RNA coding potential and demonstrated high performance in sORF data

be noted that since microproteins are distinct from normal-sized proteins in terms of biophysical properties, some of these tools may not be optimised for the detection of sORFs.

Ribosome profiling (RIBO-Seq)

Although the coding potentials of sORFs can be predicted via computational tools, this does not provide sufficient evidence that these sORFs were transcribed and translated. Particularly, many recently discovered putative sORFs were not considered de novo protein-coding genes due to their low levels of evolutionary conservation (e.g., PhyloCSF signals) since these newly found ORFs are evolutionarily very young [54]. The lack of evolutionary constraint renders these sORFs impossible to identify without supporting experimental techniques such as ribosome profiling (RIBO-seq) [54, 55].

Translatomics, the measurement of cellular translational activity, provides not only a snapshot of the translation process, but also on how translation regulates proteome composition. Translatomics experiments were

enabled by ribosome profiling or RIBO-Seq, an NGS-based tool developed by Ingolia et al. (2009) for measuring ribosome-protected mRNA fragments (Fig. 3) [56]. RIBO-Seq has been applied to identify novel sORFs to explore the protein-coding potentials of RNAs in various models, including mouse embryonic stem cells, budding yeast and *Drosophila* [56–58].

On average, a ribosome can bind and protect 31 nucleotides of mRNA during translation, forming a ribosome footprint. Isolating and identifying ribosome footprints help unveil the translation of polypeptides by systematically recording the exact positions at which translation is halted, after the addition of a chosen protein synthesis inhibitor. Through ribosome footprinting, novel and translatable sORFs can be detected and annotated to their respective genomic coding regions [59, 60]. Treatment with translation inhibitors such as harringtonine or lactimidomycin prior to deep sequencing stalls actively translating ribosomes at the sites of translation initiation [20, 59]. With RIBO-Seq, not only can the translation start sites be identified, but also affinity-based isolation of

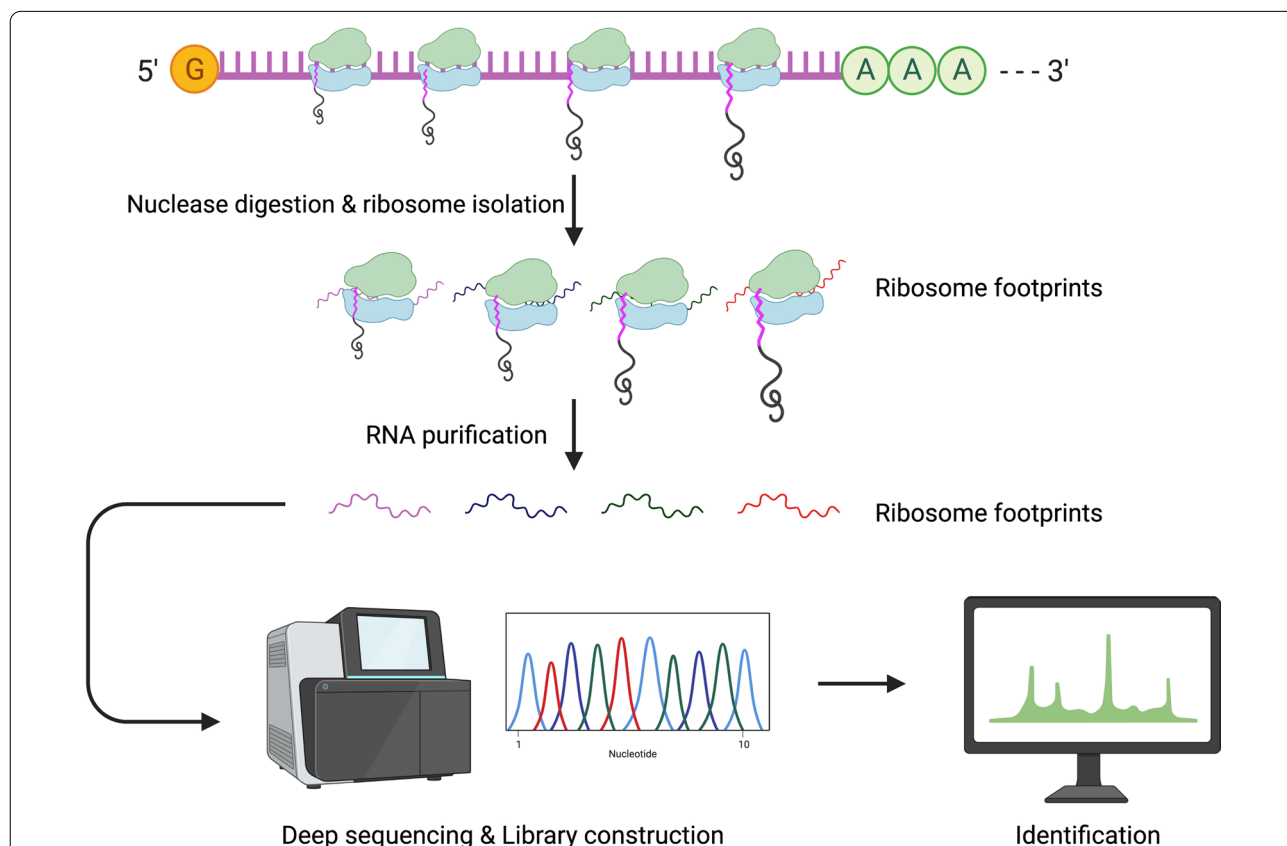


Fig. 3 Ribosome profiling process where ribosome footprints are obtained for deep sequencing. Isolation of ribosome-bound mRNAs is conducted through treatment of non-specific nucleases such as RNase I or micrococcal nuclease). Ribosome footprints (showing positioning between start and stop codon of gene) are then used for library generation and deep sequencing. Identification of novel small peptides made possible by isolation of actively translated regions of the transcript, which is directly mapped back to genomic coding regions

translating ribosomes is possible, to avoid co-purification of other ribonucleoproteins with the ribosomes [61–63]. Moreover, conventional RIBO-Seq can be modified into Poly-RIBO-Seq (isolation of polysomes), where multiple clusters of ribosomes on a transcript are isolated, providing a more concrete proof of active translation and thereby reducing the number of false positives. Aspden et al. (2014) utilised Poly-RIBO-Seq and reported identification of two types of sORF in *Drosophila* in a genome-wide assessment of sORF translation, i.e. (i) longer sORFs (~80 amino acids) with resemblance to canonical proteins and (ii) less conserved shorter sORFs (~20 amino acids) without functional characterisation from existing bioinformatic pipelines [44].

The most widely used information from ribosome footprints is the 3-nucleotide periodicity, where the codon-wise ribosome movement enables detection of frameshift events and overlapping ORF translations [64, 65]. Besides, the rate of synthesis for a particular peptide could be deduced based on the density of protected fragments obtained. The empirical measurement of protein identity is possible by determining the position of the said footprints [60]. A key advantage of applying RIBO-Seq in the identification of microproteins is the ability to identify sORFs with both AUG and non-AUG initiation codons [22]. Several studies have discovered and confirmed that half of translated sORFs are initiated with a non-AUG codon [23, 24].

Nevertheless, RIBO-Seq must be seen in the light of some limitations. These include experimentally induced distortions due to the need for rapid inhibition of ribosomes to reflect a particular physiological state, thus leading to possible inaccuracies in data collection [59]. Inferring protein synthesis rates from RIBO-Seq builds on the assumption that all ribosomes complete the translation process. However, it would be inaccurate to assume so, since regulated translation pausing and abortion can occur in different physiological conditions, such as starvation [66]. Besides, the transcript capture process by ribosomes can be non-specific and transient whereby no functional polypeptides are produced [67, 68]. The ability of RIBO-Seq to identify protein-coding capacity is limited, as a seminal paper from Guttman et al. raised the idea that ribosome occupancy alone is not a reliable indicator of classifying a transcript as coding or non-coding. Possible explanations include protection of RNA molecule by non-ribosomal RNA–protein complexes, or some of the observed fragments were not from 80S ribosomal footprints [69]. On top of that, some microproteins derived from overlapping ORFs or alternatively spliced transcripts may also avoid detection by RIBO-Seq [70]. Consequently, RIBO-Seq, as useful as it is in providing information on the translation of sORF-containing

transcripts, requires another complementary method for further confirming the completed products of sORF translation.

Mass spectrometry-based approaches

RIBO-Seq alone does not provide sufficient evidence for the expression of sORF at the protein level although it does demonstrate the translatability of a selected sORF [32, 44, 71–74]. Whereas MS-based proteomics remains indispensable in proteomics for sequencing and quantifying proteins and peptides. Thus, MS strategy can be incorporated in sORF research because such procedure directly analyses the pool of microproteins. Notwithstanding, MS-based profiling of the microproteome requires prior optimisation and modification mainly due to the low abundance and small sizes of microproteins. Notably, a key procedure is to reduce sample complexity using pre-fractionation/enrichment approaches, as shown in Fig. 4.

Size exclusion approaches have been widely employed in peptidomics studies to filter for low-molecular-weight peptides from total lysate [75, 76]. For instance, molecular weight cut off (MWCO) filters are commonly used to retain high molecular weight proteins on the filter, resulting in the enrichment of microproteins in the filtrate. However, Ma et al. (2016) had reported that alternative enrichment procedures such as acid precipitation and C8 reverse phase solid phase extraction (SPE) cartridges could result in higher enrichment of microproteins, and therefore a combination of both enrichment procedures was recommended to maximize the recovery of microproteins [75]. Another pre-fractionation measure is electrostatic repulsion-hydrophilic interaction chromatography (ERLIC), that allows charge-driven, orthogonal fractionation of peptides prior to LC–MS/MS [70]. In addition to ERLIC, pre-fractionation with high resolution isoelectric focusing (Hi-RIEF) has also been shown to improve the detectability of microproteins by MS in a highly reproducible manner [77].

To assign peptide sequences with high confidence, high quality MS/MS spectra are crucial, with two main aspects to consider i.e., high sequence coverage and low background noise. To obtain high sequence coverage, Ma et al. (2016) compared MS/MS of microprotein-derived peptides between Collision Induced Dissociation (CID) and High-energy Collisional Dissociation (HCD) on Fusion Tribrid MS and Q-Exactive MS [75]. They found that the latter yielded improvements in peptide sequence coverage and lower background noise [75]. Apart from the direct identification and quantification of translated microproteins, MS-based approaches help decipher post-translational modifications (PTMs), such as

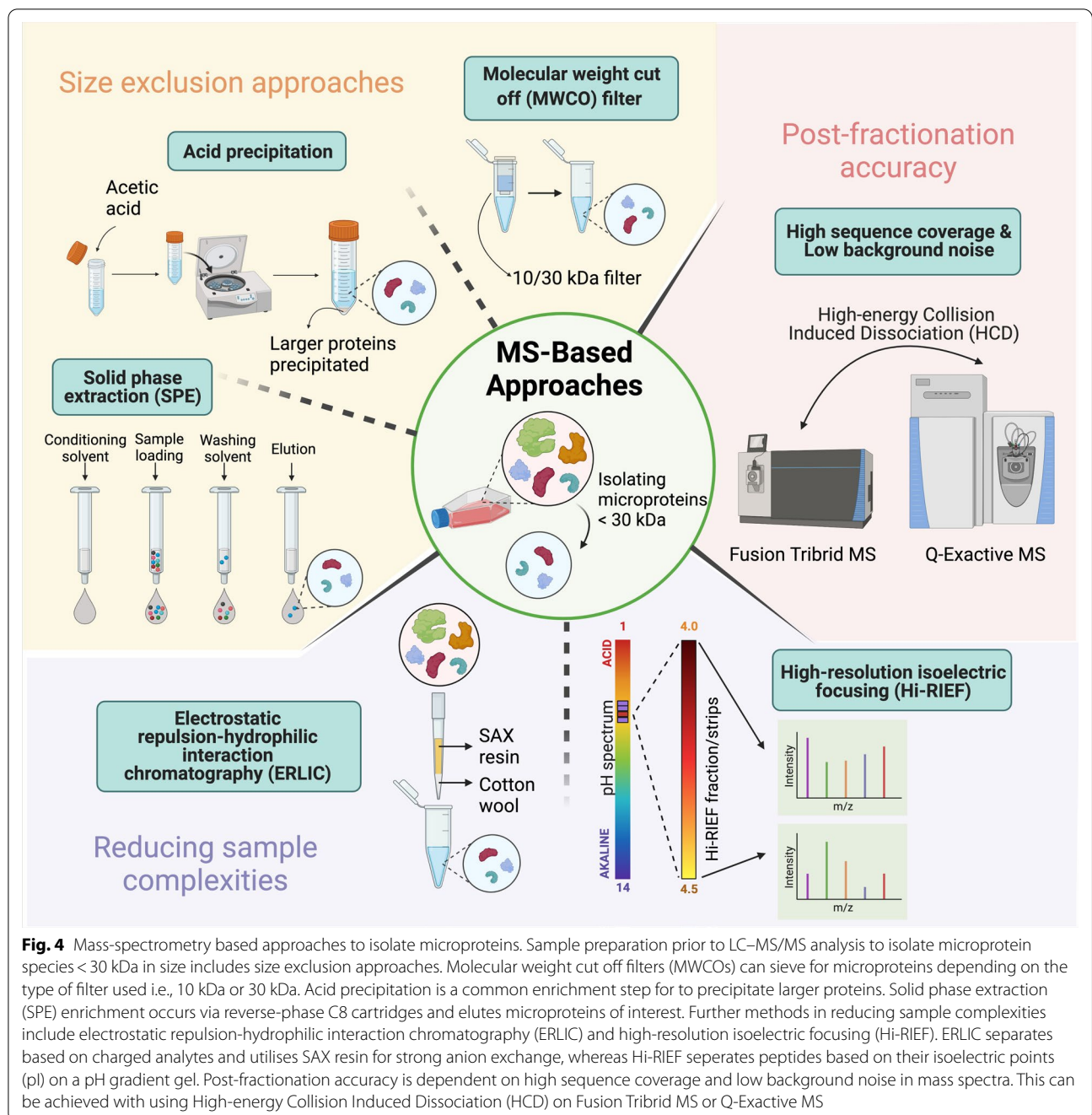


Fig. 4 Mass-spectrometry based approaches to isolate microproteins. Sample preparation prior to LC-MS/MS analysis to isolate microprotein species < 30 kDa in size includes size exclusion approaches. Molecular weight cut off filters (MWCOs) can sieve for microproteins depending on the type of filter used i.e., 10 kDa or 30 kDa. Acid precipitation is a common enrichment step for to precipitate larger proteins. Solid phase extraction (SPE) enrichment occurs via reverse-phase C8 cartridges and elutes microproteins of interest. Further methods in reducing sample complexities include electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) and high-resolution isoelectric focusing (Hi-RIEF). ERLIC separates based on charged analytes and utilises SAX resin for strong anion exchange, whereas Hi-RIEF separates peptides based on their isoelectric points (pI) on a pH gradient gel. Post-fractionation accuracy is dependent on high sequence coverage and low background noise in mass spectra. This can be achieved with using High-energy Collision Induced Dissociation (HCD) on Fusion Tribrid MS or Q-Exactive MS

phosphorylation of microproteins to infer insights in biological functions and signalling pathways [78–80].

Nonetheless, there are several challenges associated with MS-based microproteomics. First, the small size of microproteins render them under-detected by MS due to the low number of tryptic peptides generated [70]. Besides, smaller peptides tend to contain fewer arginine and lysine residues, resulting in non-cleavage or reduced tryptic cleavages. A solution is to replace or to combine

trypsin with other proteolytic enzymes, such as Glu-C, Lys-C, Lys-N, Asp-N, Arg-C and chymotrypsin [81–83]. Not to mention, sequential digestion incorporating proteases with complementary cleavage specificities is beneficial for enhancing microprotein identification [84, 85]. Still, microproteins usually lack stable secondary structures, leading to rapid degradation during extraction [86].

Data-dependent acquisition (DDA), the most widely-adopted MS acquisition methods in microproteomics,

is prone to under-sampling as MS is limited by sample complexity and sequencing speed [51, 87, 88]. Furthermore, DDA-based detection of microproteins is “amplification-free” and thus limited by the sensitivity and dynamic range of the MS [89]. This leads to a long-standing preferential detection of higher abundance proteins and proteins containing peptides with higher ionisation efficiencies. Conventional bottom-up/shotgun proteomics using DDA is biased towards the detection of proteins with >10 kDa, which represents more than 90% of the annotated proteome [89]. In addition to this, conventional MS studies using DDA report statistically significant under-representation of experimentally identified small proteins without the inclusion of enrichment procedures for microproteins [89]. The challenge with bottom-up proteomics for microprotein detection lies in several aspects, i.e., (i) during experimental sample preparation, where the lack of necessary cleavage sites in microproteins restricts its digestion; (ii) the need for alternative proteases, due to lack of generating MS-detectable peptides; (iii) under-representation of sORFs in conventional databases, where most gene algorithms apply a length restriction threshold to avoid annotating spurious sORFs; and (iv) in orthodox peptide spectrum matching (PSM) requiring detected microproteins to be matched against two unique peptides for higher confidence in protein identification, whereby microproteins are often only identified by only a single peptide owing to their small sizes [6, 90–93]. This results in an elevated false discovery rate (FDR) in protein identification [92, 93]. Hence, conventional MS should be modified to allow for a more reliable and efficient discovery of small proteins.

In lieu of DDA-based MS detection, targeted proteomics is a promising candidate for a higher confidence identification of microproteins. Selected reaction monitoring (SRM) and data-independent acquisition (DIA) can be used to monitor changes in microprotein expression across different biological samples [87]. In particular, DIA is able to validate, quantify and provide a more complete picture of the microprotein expression as compared to DDA [94, 95]. However, these approaches require specific a priori knowledge of known targets and thus are not suitable for microprotein discovery.

The sequence of a microprotein may closely resemble or overlap with that of motifs derived from larger proteins. In the context of MS-based microprotein discovery, this overlap in the form of shared peptides presents an impediment in protein inference when differentiating between a microprotein and other homologous sequences. In another aspect, microproteins tend to score low in conservation due to their short lengths, which in turn, leads to a high rate of false positives in

computational methods [17, 19]. This misalignment of information may also be present in reference protein databases. Consequently, these limitations in MS methods call for an integrated strategy i.e., the proteogenomics approach.

Proteogenomics approach

Proteogenomics is a comprehensive approach where MS data is coupled to genomic, transcriptomic or translational data from the same source, providing an alternative for further validation of low-abundant microproteins [96, 97]. Recent developments in both proteomic and deep sequencing have rapidly established proteogenomics as a reliable technique for studying unexplored or partially sequenced genomes [98–100]. As microproteins fulfil both the criteria of being low-abundant and a relatively new and uncharacterised class of proteins, applying proteogenomics techniques is a reasonable and coherent measure.

This approach involves generating a customised protein sequence database from genomic and transcriptomic sequences to be matched against MS spectra. By compiling predicted, novel peptide sequences and their variants, proteogenomics also refines protein sequence databases. Several proteogenomic studies had mapped MS spectra to RNA-seq data for detecting unannotated sORFs [76, 101, 102]. However, this strategy often suffers from reduced sensitivity and reliability because of the inflated search space. This is because a bloated protein database that is *in silico* translated from RNA-Seq data also comes with increased false positive peptide-spectral matches [70]. Proposed measures to remedy this problem includes incorporating (i) protein fractionation, (ii) *in silico* filters and (iii) RIBO-Seq data instead of RNA-Seq data, so that higher sORF-specificity and selectivity can be achieved [103, 104].

Microprotein databases

Until now, the number of sORFs and microproteins which they encode have been accumulating. In Table 2, we compiled several publicly available repositories specialized for sORFs. Incorporating a computational pipeline to corroborate with experimental data obtained would boost the credibility when characterising annotated and unannotated small peptides. A substantial number of studies have integrated this combinatorial approach in identify microproteins, proving its advantage when addressing the technical issues on validating coding sORFs [51, 71, 73, 102, 105, 106].

A few of the databases mentioned in Table 2 are specifically tailored for storing sORF information, such as sORFs.org, SmProt, ARA-PEPs, PsORF and Meta-mORF. However, between these databases, PsORF and

Table 2 Online repositories tailored for sORF identification

Database	References	Website	Type	Description
sORFs.org	[14]	http://www.sorfs.org	sORF repository	Obtains experimental data from RIBO-seq with conservation analyses and rescanning MS data from PRIDE for updated small peptide validation
SmProt	[109]	http://bioinfo.ibp.ac.cn/SmProt/	sORF repository	Database on small proteins specifically from lncRNA, obtains data from RIBO-seq, literature mining and MS data, integrates conservation analyses
OpenProt	[111, 112]	https://www.openprot.org/	altORF resource	Contains information on protein isoforms and altORFs with experimental evidence, integrates RIBO-seq, MS, conservation analyses and functional domains
ARA-PEPs	[108]	http://www.biw.kuleuven.be/CSB/ARA-PEPs	sORF repository	Repository of putative sORF-encoded peptides specifically in <i>Arabidopsis thaliana</i> , data obtained from in-house Tiling arrays and RNA-seq data
PsORF	[107]	http://psorf.whu.edu.cn/	sORF repository	Database of sORF across different plant species, incorporating genomic, transcriptomic, RIBO-Seq and MS data
MetamORF	[110]	http://metamorf.hb.univ-amu.fr/	sORF repository	A repository of unique sORFs in <i>H. sapiens</i> and <i>M. musculus</i> genomes by experimental and computational methods
nORFs.org	[113]	https://norfs.org/	novel ORF (nORF) repository	Provides aggregated information from databases such as sORFs.org, OpenProt and OpenCB

This table shows the databases available publicly for sORF identification. sORFs.org and OpenProt evaluate protein sequence identity based on BLASTp score, whereas SmProt provides a BLAST alignment search for manual evaluation of protein sequence identity. OpenProt annotates sORFs but under the label of altORFs that are longer than 30 codons and originating from ncRNAs, pseudogenes or has multiple ORFs per transcript, hence the limits set during search identification should be noted. ARA-PEPs were developed specifically from *A. thaliana* sORF experimental data, and PsORF aimed to store a more complete record of plant sORF. A large bulk of both MetamORF and nORFs.org data was obtained from sORFs.org and OpenProt. nORFs.org provides additional protein sequence viewer, OpenCB variants and customises annotation metrics functions

ARA-PEPs are resources specially for plants, with the latter storing sORF data exclusively for *Arabidopsis thaliana* [107, 108]. For databases with higher coverage over sORFs datasets in multiple organisms, sORFs.org and SmProt are more suitable since they store sORF data over a range of organisms, such as human, mouse, rat, zebrafish, nematode and fruit fly [14, 109]. In addition to that, SmProt also saves sORF data from bacteria and yeast [109]. For detailed query, sORFs.org incorporates a Biomart function, where one can tailor the search for microproteins by customising their queries according to species, chromosome number, start codons and sORFs attributes [14]. On the other hand, MetamORF is more limited in the sense that it only contains sORF data from *Homo sapiens* and *Mus musculus*, but its data has been experimentally and computationally verified, where one can browse the database according to gene locus, ORF, and transcript [110]. For OpenProt, since it is a large database, sORFs are stored under the heading of altORFs, where there is annotation of sORFs within ncRNA and pseudogene transcripts, hence some data downloaded from OpenProt in FASTA files may contain sequences longer than 100 codons [111, 112]. Finally, nORFs.org compiled data from sORFs.org and OpenProt, whilst also incorporating genomic information from OpenCB. This

allows for nORFs.org to complement information from both databases and incorporate OpenCB variants in its database [113]. Choosing which database to use when annotating experimental microproteins is highly dependent on the sample used, as well as the variety of information and query method that is provided by each database.

Validation of microproteins and their biological functions

Whilst the aforementioned methods are efficient at detecting microproteins, a validation step is required to elucidate the exact biological functions of these identified microproteins. However, since it is unnecessarily complicated to validate all possible microproteins, a rational step is to narrow down these sORF candidates into a simplified list [96]. One strategy is to analyse differential expression data of the transcripts or proteins. As an example, Cao et al. (2020) investigated 16 potential microproteins from leukaemia cell lines K562 and MOLT4; and found that 4 out of 16 of these microproteins were differentially expressed, demonstrating their potential roles in leukaemia [24]. Another approach is to focus on specific genomic regions which are likely to possess high coding potentials, and thus capable of coding for microproteins of interest. As demonstrated

by Lee, Kim and Cohen (2016) and H. Lu et al. (2019), a study was conducted on a selected mtDNA genomic region which encodes MOTS-c, a microprotein which plays a role in muscle and fat metabolism [37, 114].

To explore the biological roles of microproteins, one must first understand their modes of actions. Microproteins exert their cellular functions either by forming a complex with larger canonical proteins, or by acting autonomously [10, 96]. Slavoff et al. (2013) proposed that microproteins need to exist at biologically relevant concentrations to exert their physiological function [23]. Thus, by logic, genome-wide CRISPR/Cas9-based screens would be advantageous in detecting the extent to which these microproteins have an influence on the phenotype [45, 73, 115–118]. For direct observation of the targeted microprotein, loss-of-function (e.g., knockdown, knockout) or gain-of-function (e.g., overexpression or activation) assays can be performed and the function of the microprotein can be deduced based on the resulting phenotype. The scale of how altered the phenotype is from the wildtype can infer insights on how influential the microprotein is in that process, and at which concentrations they produce significant effects.

To elaborate, Stein et al. (2018) validated the function of a 56-amino-acid microprotein mitoregulin (MtlN) in supporting mitochondrial super-complexes and respiratory efficiency [116]. They reported disturbances in mitochondrial respiratory super-complex formation, reduced fatty acid oxidation, TCA cycle enzymes and calcium ion retention capacity in MtlN-knockout mice models, whereas MtlN overexpression in HeLa cells led to increased mitochondrial respiratory and calcium ion buffering capacities, whilst decreasing generation of reactive oxygen species [116]. Another study came from in-vivo work by Matsumoto, Clohessy and Pandolfi (2017) [115]. They uncovered a novel microprotein SPAR that played a role in regulation of mTORC1 recruitment that resulted in reduced muscle regeneration, through Spar-deficient mice models [115]. In addition, Lu et al. (2019) reported functional characterisation of lncRNA-encoded microprotein UBAP1-AST6 in A549 lung cancer cell line by overexpression and knockout models, where the prior significantly promoted cell proliferation and clonogenic property [102]. Nonetheless, since microproteins may reside in lncRNA instead of mRNA, this presents an obstacle when applying CRISPR/Cas9 editing, as CRISPR targeting disrupts the expression of both lncRNA and mRNA, impeding clear interpretation of genome editing data. One solution would be to selectively block the expression of microprotein by mutating or knocking in the start codon, while allowing lncRNA expression at the same time [86, 118].

On the contrary, another application for CRISPR/Cas9 is to knock-in an epitope sequence into the native sORF sequence so that the expression and localization of said sORF sequence can be monitored by corresponding antibodies that bind to the epitope tags. [96]. Choices of epitopes include FLAG, APEX or fluorescent proteins and split-fluorescent proteins, among many others, depending on which characteristic of the microprotein that is being scrutinised [24, 45, 119–121]. Apart from assessing the expression levels of tagged proteins, antibody capture of epitope tags can be applied to harvest *bona-fide* prey proteins that interact in a specific manner with the epitope-tagged baits. This form of protein–protein interaction (PPI) assay is named co-immunoprecipitation (coIP) or affinity purification (AP) [120]. CoIP can be applied to functionally validate microproteins, based on the understanding that like most proteins, microproteins exert their functions via the formation of microprotein-protein assemblies. Thus, by co-purifying and identifying *bona-fide* protein binding partners, the functions of these microproteins can subsequently be elucidated based on the functions or the pathways of the co-purified partners, akin to guilt-by-association. The working principles and the different MS-based strategies for elucidating PPIs have recently been reviewed in detail by Low et al. so it will not be discussed further here [122]. In the context of microproteins, Rodrigues et al. (2018) performed coIP with FLAG-GFP transgenic plants and confirmed microprotein miP1a is a part of a DNA binding complex in *FT* promoter, regulating the floral transition in *Arabidopsis* [123]. In another study, the interaction between HA-tagged micropeptide myoregulin (MLN) and membrane pumps SERCA1, SERCA2a and SERCA2b on sarcoplasmic reticulum (SR) was visualised through coIP of its stable complex, thus proving MLN's role in regulating muscle performance by impeding uptake of calcium ions into the SR [45]. Furthermore, the cytoprotective property of mitochondrial-derived peptide humanin (HN) was demonstrated through coIP of HN-GFP with pro-apoptotic protein BimEL, leading to the conclusion that HN has the capacity to inhibit BimEL-induced activation of Bak and thus inhibiting apoptosis [124]. These studies show the usefulness of coIP in validating microprotein functions, based on the logic of guilt-by-association.

Conclusions

In the last two decades, sORFs and microproteins represent an expanding landscape, with multiple technological innovations advancing the speed of their discovery and annotations. Exploration of this part of the genome that was disregarded have yielded results in terms of sORF transcription and microprotein

functionality. The current measures in identification and validation were successful in unveiling the microproteome, yet progress must be made to address the caveats of these techniques. Functional characterisation of existing microproteins remains to be improved, and the investigation for small peptides in new sample types. There has been some exploration into the potential of microproteins as therapeutic targets. For instance, Ontak (denileukin) was developed as toxins, targeted for tumour-specific microproteins in the treatment for T cell lymphoma [125]. However, through exhaustive literature mining, it appears that this field is not yet extensively studied, thus providing a lead on where future research could perhaps focus on. Previously neglected in the genome, further progress in this field would shed light on sORF and microproteins, with the opportunity to raise new knowledge on a hidden sub-proteome and its role in biological function.

Abbreviations

AltORF: Alternate open reading frame; AP: Affinity purification; CDS: Coding sequences; CI: Coding Index; CID: Collision induced dissociation; circRNA: Circular ribonucleic acid; CoIP: Co-immunoprecipitation; CRISPR/Cas 9: Clustered regularly interspaced short palindromic repeats and CRISPR-associated protein 9; DDA: Data-dependent acquisition; DIA: Data-independent acquisition; DNA: Deoxyribonucleic acid; dORF: Downstream open reading frame; ER: Endoplasmic reticulum; ERLIC: Electrostatic repulsion-hydrophilic interaction chromatography; FANTOM: Functional ANnotation Of the Mammalian Genome Consortium; GFP: Green fluorescent protein; HCD: High-energy collisional dissociation; HGP: Human Genome Project; Hi-RIEF: High Resolution Isoelectric Focusing; LC-MS/MS: Liquid Chromatography with Tandem Mass Spectrometry; lincRNA: Long intergenic non-coding ribonucleic acid; lncRNA: Long non-coding ribonucleic acid; pri-microRNA: Precursor micro-ribonucleic acid; mRNA: Messenger ribonucleic acid; MS: Mass spectrometry; MS/MS: Tandem mass spectrometry; mtDNA: Mitochondrial deoxyribonucleic acid; MWCO: Molecular weight cut off; ncRNA: Non-coding ribonucleic acid; NGS: Next generation sequencing; NHEJ: Non-homologous end joining; ORF: Open reading frame; PhyloCSF: Phylogenetic codon substitution frequencies; PPI: Protein-protein interaction; PTM: Post translational modifications; RIBO-Seq: Ribosome profiling; RNA: Ribonucleic acid; RNA-Seq: Ribonucleic acid sequencing; SEP: Short open reading frame-encoded proteins; smORF: Small open reading frame; sORF: Short open reading frame; SPE: Solid phase extraction; SR: Sarco-plasmic reticulum; TCA: The citric acid cycle; uORF: Upstream open reading frame; UPR: Unfolded protein response; UTR: Untranslated region.

Acknowledgements

We would like to extend our gratitude to Dr. Sebaastian van Heesch from Princess Máxima Center for his insightful comments on the manuscript. All figures were created with BioRender.com.

Authors' contributions

ALZX and TYL drafted the manuscript and wrote most of the article. PYL, MAM, SES and YFP participated in discussion and helped to write the manuscript. TYL oversaw the whole project. All authors read and approved the final manuscript.

Funding

This review paper received funding from the National University of Malaysia (UKM) Research Grant (Geran Universiti Penyelidikan; GUP-2020-078).

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors have declared no conflict of interest.

Author details

¹UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, 56000 Kuala Lumpur, Malaysia. ²Division of Biomedical Science, School of Pharmacy, University of Nottingham Malaysia, Semenyih, 43500 Selangor, Malaysia.

Received: 31 December 2021 Accepted: 9 March 2022

Published online: 17 March 2022

References

- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409:860–921.
- Gates AJ, Gysi DM, Kellis M, Barabási A-L. A wealth of discovery built on the Human Genome Project—by the numbers. *Nature*. 2021;590:212–5.
- Skovgaard M, Jensen LJ, Brunak S, Ussery D, Krogh A. On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet*. 2001 [cited 2021 Apr 15]. p. 425–8. <https://linkinghub.elsevier.com/retrieve/pii/S0168952501023721>. Accessed 15 Apr 2021.
- Cheng H, Soon Chan W, Li Z, Wang D, Liu S, Zhou Y. Small open reading frames: current prediction techniques and future prospect. *Curr Protein Pept Sci*. 2011;12:503–7.
- Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*. 2002;420:563–73.
- Dinger ME, Pang KC, Mercer TR, Mattick JS. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol*. 2008;4: e1000176. <https://doi.org/10.1371/journal.pcbi.1000176>.
- Merino-Valverde I, Greco E, Abad M. The microproteome of cancer: From invisibility to relevance. *Exp Cell Res*. 2020;392(1): <https://doi.org/10.1016/j.yexcr.2020.111997>.
- Ruiz-Orera J, Villanueva-Cañas JL, Albà MM. Evolution of new proteins from translated sORFs in long non-coding RNAs. *Exp Cell Res*. 2020;391: 111940. <https://doi.org/10.1016/j.yexcr.2020.111940>.
- Benezra R, Davis RL, Lockshon D, Turner DL, Weintraub H. The protein Id: a negative regulator of helix-loop-helix DNA binding proteins. *Cell*. 1990;61:49–59.
- Bhati KK, Blaakmeer A, Paredes EB, Dolde U, Eggen T, Hong SY, et al. Approaches to identify and characterize microProteins and their potential uses in biotechnology. *Cell Mol Life Sci*. 2018;75:2529–36. <https://doi.org/10.1007/s00018-018-2818-8>.
- Makarewicz CA. The hidden world of membrane microproteins. *Exp Cell Res*. 2020;388: 111853. <https://doi.org/10.1016/j.yexcr.2020.111853>.
- Chu Q, Martinez TF, Novak SW, Donaldson CJ, Tan D, Vaughan JM, et al. Regulation of the ER stress response by a mitochondrial microprotein. *Nat Commun*. 2019;10:1–13. <https://doi.org/10.1038/s41467-019-12816-z>.
- Arnoult N, Correia A, Ma J, Merlo A, Garcia-Gomez S, Maric M, et al. Regulation of DNA repair pathway choice in S and G2 phases by the NHEJ inhibitor CYREN. *Nature*. 2017;549:548–52.
- Olexiuk 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:D497–502.
- Brown TA. Understanding a genome sequence. Wiley-Liss; 2002; <https://www.ncbi.nlm.nih.gov/books/NBK21136/>. Accessed 28 Sep 2021.

16. Brunet MA, Leblanc S, Roucou X. Reconsidering proteomic diversity with functional investigation of small ORFs and alternative ORFs. *Exp Cell Res*. 2020;393:112057. <https://doi.org/10.1016/j.yexcr.2020.112057>.
17. Peeters MKR, Menschaert G. The hunt for sORFs: A multidisciplinary strategy. *Exp Cell Res*. 2020;391(1). <https://doi.org/10.1016/j.yexcr.2020.111923>.
18. Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP. Hundreds of putatively functional small open reading frames in *Drosophila*. *Genome Biol*. 2011;12:1–17. <https://doi.org/10.1186/gb-2011-12-11-r118>.
19. Couso JP, Patraquim P. Classification and function of small open reading frames. *Nat Rev Mol Cell Biol*. 2017;18(9):575–89. <https://doi.org/10.1038/nrm.2017.58>.
20. Chugunova A, Navalayeu T, Dontsova O, Sergiev P. Mining for small translated ORFs. *J Proteome Res*. 2018;17:1–11.
21. Zhu M, Gribskov M. MiPepid: MicroPeptide identification tool using machine learning. *BMC Bioinform*. 2019;20:1–11. <https://doi.org/10.1186/s12859-019-3033-9>.
22. Kearsle MG, Wilusz JE. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev*. 2017;31:1717. <https://doi.org/10.1101/gad.305250.117>.
23. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol*. 2013;9:59–64.
24. Cao X, Khitun A, Na Z, Dumitrescu DG, Kubica M, Olatunji E, et al. Comparative proteomic profiling of unannotated microproteins and alternative proteins in human cell lines. *J Proteome Res Am Chem Soc*. 2020;19:3418–26.
25. Vanderperre B, Lucier J-F, Bissonnette C, Motard J, Tremblay G, Vanderperre S, et al. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS ONE*. 2013;8:70698.
26. Hellens RP, Brown CM, Chisnall MAW, Waterhouse PM, Macknight RC. The emerging world of small ORFs. *Trends Plant Sci*. 2016;21:317–28. <https://doi.org/10.1016/j.tplants.2015.11.005>.
27. Harding HP, Novoa I, Zhang Y, Zeng H, Wek R, Schapira M, et al. Regulated translation initiation controls stress-induced gene expression in mammalian cells. *Mol Cell Cell Press*. 2000;6:1099–108.
28. Jin X, Turcott E, Englehardt S, Mize GJ, Morris DR. The two upstream open reading frames of oncogene *mdm2* have different translational regulatory properties*. *J Biol Chem*. 2003;278:25716–21.
29. Vatter KM, Wek RC. Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc Natl Acad Sci*. 2004;101:11269–74.
30. Johnstone TG, Bazzini AA, Giraldez AJ. Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J*. 2016;35:706.
31. Wu Q, Wright M, Gogol MM, Bradford WD, Zhang N, Bazzini AA. Translation of small downstream ORFs enhances translation of canonical main open reading frames. *EMBO J*. 2020;39: e104763. <https://doi.org/10.15252/embj.2020104763>.
32. Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife*. 4. <https://elifesciences.org/articles/08890>. Accessed 4 Feb 2021.
33. Renz PF, Valdivia Francia F, Sandoel A. Some like it translated: small ORFs in the 5'UTR. *Exp Cell Res*. 2020;396: 112229. <https://doi.org/10.1016/j.yexcr.2020.112229>.
34. Casimiro-Soriguer CS, Rigual MM, Brokate-Llanos AM, Muñoz MJ, Garzón A, Pérez-Pulido AJ, et al. Using AnAblast for intergenic sORF prediction in the *Caenorhabditis elegans* genome. *Bioinformatics*. 2020;36:4827–32.
35. Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, Robinson DR, Wu Y-M, Cao X, et al. Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell*. 2012;149:1622.
36. Hanada K, Zhang X, Borevitz JO, 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.
37. Lee C, Kim KH, Cohen P. MOT5-c: a novel mitochondrial-derived peptide regulating muscle and fat metabolism. *Free Radic Biol Med*. 2016. <https://doi.org/10.1016/j.freeradbiomed.2016.05.015>.
38. Zheng X, Chen L, Zhou Y, Wang Q, Zheng Z, Xu B, 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. <https://doi.org/10.1186/s12943-019-1010-6>.
39. Choi SW, Kim HW, Nam JW. The small peptide world in long noncoding RNAs. *Brief Bioinform*. 2019;20(5):1853–64. <https://doi.org/10.1093/bib/bby055>.
40. Hartford CCR, Lal A. When long noncoding becomes protein coding. *Mol Cell Biol*. 2020. <https://doi.org/10.1128/MCB.00528-19>.
41. Wu P, Mo Y, Peng M, Tang T, Zhong Y, Deng X, et al. Emerging role of tumor-related functional peptides encoded by lncRNA and circRNA. *Mol Cancer*. 2020. <https://doi.org/10.1186/s12943-020-1147-3>.
42. Yadav A, Sanyal I, Rai SP, Lata C. An overview on miRNA-encoded peptides in plant biology research. *Genomics*. 2021;113:2385–91.
43. Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. *Cell*. 2013;154:46.
44. Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M, et al. Extensive translation of small open reading frames revealed by poly-ribo-seq. *Elife*. 2014;3:1–19.
45. Anderson DM, Anderson KM, Chang CL, Makarewich CA, Nelson BR, McAnally JR, et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell*. 2015;160:595–606.
46. Cabrera-Quio LE, Herberg S, Pauli A. Decoding sORF translation—from small proteins to gene regulation. *RNA Biol*. 2016;13:1051–9.
47. Plaza S, Menschaert G, Payre F. In search of lost small peptides. *Annu Rev Cell Dev Biol*. 2017;33:391–416. <https://doi.org/10.1146/annurev-cellbio-100616-060516>.
48. Zhu S, Wang J, He Y, Meng N, Yan GR. Peptides/proteins encoded by non-coding RNA: a novel resource bank for drug targets and biomarkers. *Front Pharmacol*. 2018. <https://doi.org/10.3389/fphar.2018.01295>.
49. Yeasmin F, Yada T, Akimitsu N. Micropeptides encoded in transcripts previously identified as long noncoding RNAs: a new chapter in transcriptomics and proteomics. *Front Genet*. 2018. <https://doi.org/10.3389/fgene.2018.00144>.
50. Zlotorynski E. The functions of short ORFs and their microproteins. *Nat Rev Mol Cell Biol*. 2020;21:252–3. <https://doi.org/10.1038/s41580-020-0239-7>.
51. Martinez TF, Chu Q, Donaldson C, Tan D, Shokhirev MN, Saghatelian A. Accurate annotation of human protein-coding small open reading frames. *Nat Chem Biol*. 2021;16:458–68. <https://doi.org/10.1038/s41589-019-0425-0>.
52. 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. <https://doi.org/10.1093/bioinformatics/btp688>.
53. Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, Nakaminami K, et al. Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc Natl Acad Sci U S A*. 2013;110:2395–400. <https://doi.org/10.1073/pnas.1213958110>.
54. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Gonzalez JM, Magrane M, et al. A community-driven roadmap to advance research on translated open reading frames detected by Ribo-seq. *bioRxiv*. 2021;2021.06.10.447896. <http://biorxiv.org/content/early/2021/06/10/2021.06.10.447896.abstract>
55. McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. *Nat Rev Genet*. 2016;17:567–78. <https://doi.org/10.1038/nrg.2016.78>.
56. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*. 2009;324:218.
57. Kondo T, Plaza S, Zanet J, Benrabah E, Valenti P, Hashimoto Y, et al. Small peptides switch the transcriptional activity of shavenbaby during *drosophila* embryogenesis. *Science* (-80). 2010;329:336–9.
58. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*. 2011;147:789–802.
59. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc*. 2012;7:1534–50.
60. Brar GA, Weissman JS. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol*. 2015;16:651–64.

61. Heiman M, Schaefer A, Gong S, Peterson J, Day M, Ramsey K, et al. A translational profiling approach for the molecular characterization of CNS cell types. *Cell*. 2008;135:738–48.
62. Sanz E, Yang L, Su T, Morris DR, McKnight GS, Amieux PS. Cell-type-specific isolation of ribosome-associated mRNA from complex tissues. *Proc Natl Acad Sci*. 2009;106:13939–44.
63. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJS, Jackson SE, et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep*. 2014;8:1365.
64. Gerashchenko MV, Gladyshev VN. Ribonuclease selection for ribosome profiling. *Nucleic Acids Res*. 2017;45:e6–e6.
65. Chung BY, Hardcastle TJ, Jones JD, Irigoyen N, Firth AE, Baulcombe DC, et al. The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis. *RNA*. 2015;21:1731.
66. Subramaniam AR, Zid BM, O'Shea EK. An integrated approach reveals regulatory controls on bacterial translation elongation. *Cell*. 2021;159:1200–11.
67. Ingolia NT. Ribosome footprint profiling of translation throughout the genome. *Cell*. 2016;165:22–33.
68. Raj A, Wang SH, Shim H, Harpak A, Li Yi, Engelmann B, et al. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife*. 2016. <https://doi.org/10.7554/eLife.13328>.
69. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. Ribosome profiling provides evidence that large non-coding RNAs do not encode proteins. *Cell*. 2013;154:240.
70. Khitun A, Slavoff SA. Proteomic detection and validation of translated small open reading frames. *Curr Protoc Chem Biol*. 2021;11: e77. <https://doi.org/10.1002/cpch.77>.
71. Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J*. 2014;33:981–93. <https://doi.org/10.1002/emboj.201488411>.
72. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, et al. Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods*. 2016;13:165–70.
73. van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, et al. The translational landscape of the human heart. *Cell*. 2021;178:242–260. <https://doi.org/10.1016/j.cell.2019.05.010>.
74. Weaver J, Mohammad F, Buskirk AR, Storz G. Identifying small proteins by ribosome profiling with stalled initiation complexes. *MBio*. 2019. <https://doi.org/10.1128/mBio.02819-18>.
75. Ma J, Diedrich JK, Jungreis I, Donaldson C, Vaughan J, Kellis M, et al. Improved identification and analysis of small open reading frame encoded polypeptides. *Anal Chem*. 2016;88:3967–75.
76. He C, Jia C, Zhang Y, Xu P. Enrichment-based proteogenomics identifies microproteins, missing proteins, and novel smORFs in *Saccharomyces cerevisiae*. *J Proteome Res Am Chem Soc*. 2018;17:2335–44.
77. Branca RMM, Orre LM, Johansson HJ, Granholm V, Huss M, Pérez-Bercoff A, et al. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat Methods*. 2013;11:59–62.
78. López E, Wang X, Madero L, López-Pascual J, Latterich M. Functional phosphoproteomic mass spectrometry-based approaches. *Clin Transl Med*. 2012. <https://doi.org/10.1186/2001-1326-1-20>.
79. Kosako H, Nagano K. Expert review of proteomics quantitative phosphoproteomics strategies for understanding protein kinase-mediated signal transduction pathways. 2014; <https://www.tandfonline.com/action/journalInformation?journalCode=ieru20>. Accessed 9 Jun 2021.
80. Low TY, Mohtar MA, Lee PY, Omar N, Zhou H, Ye M. Widening the bottleneck of phosphoproteomics: evolving strategies for phosphopeptide enrichment. *Mass Spectrom Rev*. 2021;40:309–33.
81. Tsiatsiani L, Heck AJR. Proteomics beyond trypsin. *FEBS J*. 2021;282:2612–26. <https://doi.org/10.1111/febs.13287>.
82. Giansanti P, Tsiatsiani L, Low TY, Heck AJ. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat Protoc*. 2016;11:993–1006.
83. Low TY, van Heesch S, van den Toorn H, Giansanti P, Cristobal A, Toonen P, et al. Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Rep*. 2013;5:1469–78.
84. Swaney DL, Wenger CD, Coon JJ. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J Proteome Res*. 2010;9:1323–9.
85. Dau T, Bartolomucci G, Rappsilber J. Proteomics using protease alternatives to trypsin benefits from sequential digestion with trypsin. *Anal Chem*. 2020;92:9523–7. <https://doi.org/10.1021/acs.analchem.0c00478>.
86. Tharakan R, Sawa A. Minireview: novel micropeptide discovery by proteomics and deep sequencing methods. *Front Genet*. 2021;12:536.
87. Fabre B, Combi JP, 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;60:122–30. <https://doi.org/10.1016/j.cbpa.2020.12.002>.
88. Ahrens CH, Wade JT, Champion MM, Langer JD. A practical guide to small protein discovery and characterization using mass spectrometry. *J Bacteriol*. 2022. <https://doi.org/10.1128/jb.00353-21>.
89. Becher D, Bartel J, Varadarajan AR, Sura T, Ahrens CH, Maaß S. Optimized proteomics workflow for the detection of small proteins. *J Proteome Res*. 2020;19:4004–18.
90. Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc*. 2022;11:2301–19.
91. 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 Proteomics*. 2004;3:531–2.
92. Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, et al. A guided tour of the trans-proteomic pipeline. *Proteomics*. 2010;10:1150–9.
93. Ludwig C, Claassen M, Schmidt A, Aebersold R. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry. *Mol Cell Proteomics*. 2012. <https://doi.org/10.1074/mcp.M111.013987>.
94. Bruderer R, Bernhardt OM, Gandhi T, Xuan Y, Sondermann J, Schmidt M, et al. Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol Cell Proteomics*. 2017;16:2296–309.
95. Fabre B, Korona D, Mata CI, Parsons HT, Deery MJ, Hertog MLATM, et al. Spectral libraries for SWATH-MS assays for *Drosophila melanogaster* and *Solanum lycopersicum*. *Proteomics*. 2017. <https://doi.org/10.1002/pmic.201700216>.
96. Schlesinger D, Elsässer SJ. Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. *FEBS J*. 2021. <https://doi.org/10.1111/febs.15769>.
97. Low TY, Heck AJ. Reconciling proteomics with next generation sequencing. *Curr Opin Chem Biol*. 2016;30:14–20.
98. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods*. 2014. p. 1114–25. <https://www.nature.com/articles/nmeth.3144>. Accessed 10 Jun 2021.
99. Low TY, Mohtar MA, Ang MY, Jamal R. Connecting proteomics to next-generation sequencing: proteogenomics and its current applications in biology. *Proteomics*. 2019. <https://doi.org/10.1002/pmic.201800235>.
100. Ang MY, Low TY, Lee PY, Nazarie WFWM, Guryev V, Jamal R. Proteogenomics: from next-generation sequencing (NGS) and mass spectrometry-based proteomics to precision medicine. *Clin Chim Acta*. 2019;498:38–46.
101. Zhu Y, Orre LM, Johansson HJ, Huss M, Boekel J, Vesterlund M, et al. Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nat Commun*. 2018. <https://doi.org/10.1038/s41467-018-03311-y>.
102. Lu S, Zhang J, Lian X, Sun L, Meng K, Chen Y, et al. A hidden human proteome encoded by “non-coding” genes. *Nucleic Acids Res*. 2019;47:8111–25.
103. Koch A, Gawron D, Steyaert S, Ndah E, Crappé J, De Keulenaer S, et al. A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics*. 2014;14:2688–98.
104. Crappé J, Ndah E, Koch A, Steyaert S, Gawron D, De Keulenaer S, et al. PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res*. 2015. <https://doi.org/10.1093/nar/gku1283>.

105. Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, et al. Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.* 2015;16:179. <https://doi.org/10.1186/s13059-015-0742-x>.
106. Budamgunta H, Olexiouk V, Luyten W, Schildermans K, Maes E, Boonen K, et al. Comprehensive peptide analysis of mouse brain striatum identifies novel sORF-encoded polypeptides. *Proteomics.* 2018. <https://doi.org/10.1002/pmic.201700218>.
107. Chen Y, Li D, Fan W, Zheng X, Zhou Y, Ye H, et al. PsORF: a database of small ORFs in plants. *Plant Biotechnol J.* 2021;18:2158–60. <https://doi.org/10.1111/pbi.13389>.
108. Hazarika RR, De Coninck B, Yamamoto LR, Martin LR, Cammue BPA, van Noort V. ARA-PEPs: a repository of putative sORF-encoded peptides in *Arabidopsis thaliana*. *BMC Bioinform.* 2017;18:1–9. <https://doi.org/10.1186/s12859-016-1458-y>.
109. Hao Y, Zhang L, Niu Y, Cai T, Luo J, He S, et al. SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief Bioinform.* 2017;19:bbx005. <https://doi.org/10.1093/bib/bbx005>.
110. Choteau SA, Wagner A, Pierre P, Spinelli L, Brun C. MetaORF: a repository of unique short open reading frames identified by both experimental and computational approaches for gene and metagenome analyses. *Database (Oxford).* 2021;2021:baab032. <https://doi.org/10.1093/database/baab032>.
111. Brunet MA, Lucier JF, Levesque M, Leblanc S, Jacques JF, Al-Saedi HRH, et al. OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Res.* 2021;49:D380–8.
112. Brunet MA, Brunelle M, Lucier JF, Delcourt V, Levesque M, Grenier F, et al. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.* 2019;47:D403–10.
113. Neville MDC, Kohze R, Erady C, Meena N, Hayden M, Cooper DN, et al. A platform for curated products from novel open reading frames prompts reinterpretation of disease variants. *Genome Res.* 2021;31:327–36.
114. Lu H, Wei M, Zhai Y, Li Q, Ye Z, Wang L, et al. MOT5-c peptide regulates adipose homeostasis to prevent ovariectomy-induced metabolic dysfunction. *J Mol Med.* 2019;97:473–85. <https://doi.org/10.1007/s00109-018-01738-w>.
115. Matsumoto A, Clohessy JG, Pandolfi PP. SPAR, a lncRNA encoded mTORC1 inhibitor. *Cell Cycle.* 2017;16:815–6. <https://doi.org/10.1080/15384101.2017.1304735>.
116. Stein CS, Jadia P, Zhang X, McLendon JM, Abouassaly GM, Witmer NH, et al. Mitoregulin: a lncRNA-encoded microprotein that supports mitochondrial supercomplexes and respiratory efficiency. *Cell Rep.* 2018;23:3710–37208.
117. Zhang M, Zhao K, Xu X, Yang Y, Yan S, Wei P, et al. A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. *Nat Commun.* 2018;9:1–17.
118. Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, et al. Pervasive functional translation of non-canonical human open reading frames. *Science (80-).* 2020;367:1140–6.
119. Slavoff SA, Heo J, Budnik BA, Hanakahi LA, Saghatelian A. A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *J Biol Chem.* 2014;289:10950–7.
120. Chu Q, Rathore A, Diedrich JK, Donaldson CJ, Yates JR, Saghatelian A. Identification of microprotein–protein interactions via APEX tagging. *Biochemistry.* 2017;56:3299–306.
121. Niu L, Lou F, Sun Y, Sun L, Cai X, Liu Z, et al. A micropeptide encoded by lncRNA MIR155HG suppresses autoimmune inflammation via modulating antigen presentation. *Sci Adv.* 2020;6:2059.
122. Low TY, Syafruddin SE, Mohtar MA, Vellaichamy A, Rahman NSA, Pung Y-F, et al. Recent progress in mass spectrometry-based strategies for elucidating protein–protein interactions. *Cell Mol Life Sci.* 2021;78:5325–39.
123. Rodrigues VL, Dolde U, Straub D, Eguen T, Botterweg-Paredes E, Sun B, et al. Dissection of the microProtein miP1 floral repressor complex in *Arabidopsis*. *bioRxiv.* 2018;258228. <https://www.biorxiv.org/content/https://doi.org/10.1101/258228v1>. Accessed 30 Sep 2021.
124. Luciano F, Zhai D, Zhu X, Bailly-Maitre B, Ricci JE, Satterthwait AC, et al. Cytoprotective peptide humanin binds and inhibits proapoptotic Bcl-2/Bax family protein BimEL. *J Biol Chem.* 2005;280:15825–35.
125. Dang L, Van Damme EJM. Toxic proteins in plants. *Phytochemistry.* 2015. <https://doi.org/10.1016/j.phytochem.2015.05.020>.
126. Guo JC, Fang SS, Wu Y, Zhang JH, Chen Y, Liu J, 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:W516–22.
127. Badger JH, Olsen GJ. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol.* 1999;16:512–24.
128. Kong L, Zhang Y, Ye Z-Q, Liu X-Q, Zhao S-Q, Wei L, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 2007;35:W345.
129. Kang Y-J, Yang D-C, Kong L, Hou M, Meng Y-Q, Wei L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* 2017;45:W12–6.
130. Tong X, Liu S. CPPred: coding potential prediction based on the global description of RNA sequence. *Nucleic Acids Res.* 2019;47:43.
131. Tong X, Hong X, Xie J, Liu S. CPPred-sORF: coding potential prediction of sORF based on non-AUG. *bioRxiv.* 2020. <https://doi.org/10.1101/2020.03.31.017525v1>.
132. Ji X, Cui C, Cui Q. smORFfunction: a tool for predicting functions of small open reading frames and microproteins. *BMC Bioinform.* 2020;21:1–13.
133. Straub D, Wenkel S. Cross-species genome-wide identification of evolutionary conserved microproteins. *Genome Biol Evol.* 2021;9:777–89.
134. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 2005;15:1034.
135. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics.* 2011;27:i275–82.
136. Skarshewski A, Stanton-Cook M, Huber T, Al Mansoori S, Smith R, Beatson SA, et al. uPEPPERoni: an online tool for upstream open reading frame location and analysis of transcript conservation. *BMC Bioinform.* 2014;15:1–6. <https://doi.org/10.1186/1471-2105-15-36>.
137. Zhou P, Silverstein KA, Gao L, Walton JD, Nallu S, Guhlin J, et al. Detecting small plant peptides using SPADA (Small Peptide Alignment Discovery Application). *BMC Bioinform.* 2013;14:1–16. <https://doi.org/10.1186/1471-2105-14-335>.
138. Zhang Y, Jia C, Fullwood MJ, Kwok CK. DeepCPP: a deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for RNA coding potential prediction. *Brief Bioinform.* Oxford Academic; 2021 [cited 2021 Sep 7];22:2073–84. <https://academic.oup.com/bib/article/22/2/2073/5813257>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

