Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins

Dörte Schlesinger1,2 and Simon J. Elsässer1,2

1 Science for Life Laboratory, Division of Genome Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden
2 Ming Wai Lau Centre for Reparative Medicine, Stockholm node, Karolinska Institutet, Stockholm, Sweden

Short ORFs (sORFs), that is, occurrences of a start and stop codon within 100 codons or less, can be found in organisms of all domains of life, outnumbering annotated protein-coding ORFs by orders of magnitude. Even though functional proteins smaller than 100 amino acids are known, the coding potential of sORFs has often been overlooked, as it is not trivial to predict and test for functionality within the large number of sORFs. Recent advances in ribosome profiling and mass spectrometry approaches, together with refined bioinformatic predictions, have enabled a huge leap forward in this field and identified thousands of likely coding sORFs. A relatively low number of small proteins or microproteins produced from these sORFs have been characterized so far on the molecular, structural, and/or mechanistic level. These however display versatile and, in some cases, essential cellular functions, allowing for the exciting possibility that many more, previously unknown small proteins might be encoded in the genome, waiting to be discovered. This review will give an overview of the steadily growing microprotein field, focusing on eukaryotic small proteins. We will discuss emerging themes in the molecular action of microproteins, as well as advances and challenges in microprotein identification and characterization.

Introduction

Microproteins, also termed micropeptides, SEPs (short ORF-encoded polypeptides) or simply ‘small proteins’, are proteins encoded by short/small ORFs (sORFs, smORFs) of < 100 codons. While examples of cellular proteins smaller than 100 amino acids in their mature form (ubiquitin just to mention one) predate the genomics era, these were in most cases known to be processed from longer, canonical ORF products. Thus, the terms microprotein, micropeptide and SEP have been adopted in recent literature to contour a new class of small proteins that are in fact ‘born small’. By this definition, microproteins are distinguishable from peptide hormones and neuropeptides that are small in size due to post-translational processing from larger secreted proproteins, even though microproteins can also contain signal peptides and be subject to processing [1,2]. While microproteins might display some unusual properties due to their size, they may be more adequately viewed as the low end of the spectrum of canonical proteins. In support of this notion, there are
various examples of small proteins which have been discovered long before the term microprotein was coined: the 50 amino acid epsilon subunit of mitochondrial F1-ATPase and the 46 amino acid subunit of RNA Polymerase II RPB10 are two examples of sORF products, which were discovered decades ago alongside larger subunits of these macromolecular complexes [3–6].

Short ORFs and their putative microprotein products have been largely absent from early genome annotations because of the difficulties in defining functional, protein-coding sORFs amongst a myriad of start and stop codons occurring in-frame in the genome by chance [7,8]. Thus, historically, criteria that defined an ORF included a conventional start and stop codon, a length of at least 100 codons and usually the presence of a single ORF per transcript [7–10]. Transcripts containing only smaller ORFs have consequently been deemed noncoding, a concept that has finally been overturned by advances in detection methods and the increased discovery and attention to the fact that even sORF-produced small proteins can exhibit biological functions. Nevertheless, it remains largely unknown what fraction of the enormous number of putative sORFs are translated into functional microproteins [7].

Techniques such as ribosome profiling (Ribo-seq) and mass spectrometry (MS) have narrowed down the number of likely coding sORFs to various thousand. Identification and validation of functional microproteins is only slowly catching up with these large-scale exploratory studies. Nevertheless, the examples existing to date demonstrate that microproteins are ubiquitous in origin, cellular localization, and function. They exist in all domains of life [11–13] and across a variety of species such as plants [14], humans [15], beetles [12], various bacteria [11], and yeast [16], and are also employed by viruses [17–19].

Here, we will review the functional space of sORFs and microproteins known to date, as well as the state-of-the-art technologies used to predict, identify, and validate functional microproteins focusing on eukaryotes. Readers interested in microproteins of prokaryotes are referred to respective reviews [20–23]. We will discuss general themes emerging from the growing knowledge and strategies to overcome challenges in the study of microproteins.

**Microprotein function**

Microproteins known to date are involved in diverse areas of cellular and organismal function, including RNA processing [15,24], protein processing [25], gene expression [26], cellular trafficking [27,28], signaling [29–32], metabolism [32–37], repair mechanisms and stress responses [38–41], cell death [26,42], and cell communication [1,43]. In higher eukaryotes, some key organismal processes, namely embryonic development [1,12,43–50], mitochondrial processes [33–37,42], muscle function [44,51–58], and immunity [27,28,59,60], appear particularly rich in microproteins. While microproteins may have been able to evolve more rapidly in functional niches particularly suited for small proteins (such as the ones we will speculate on below), it may also be possible that the currently observed inclination for the above areas could be the result of the ways in which most microproteins have been discovered and where laboratories in this emerging field have spent most effort to mine for functional sORFs. This would suggest that we may have just scratched the surface of a large functional repertoire and it is conceivable that microproteins are just as versatile as their bigger canonical siblings.

While the molecular structure and mechanism of many recently identified microproteins has yet to be elucidated, long-known small proteins provide valuable examples for the diverse molecular actions of microproteins (Fig. 1): they may function as (a) constitutive or optional subunits of protein complexes, (b) allosteric regulators of enzymes or macromolecular complexes, (c) signaling molecules as ligands of receptors, or finally (d) perform autonomous functions (Fig. 2). Beyond shared molecular features with canonical proteins, it is conceivable that microproteins have provided favorable or even unique evolutionary solutions for some cellular or organismal functions. While the evidence for such specialized functional evolution of microproteins is still scarce, we suggest three emerging themes on the basis of select examples.

**Secrected microproteins**

We speculate that a larger class of microproteins represent (directly) encoded peptide hormones for endocrine, paracrine or autocrine signaling.

A recent study on the translatome of the human heart indicates that many uncharacterized secretory microproteins might exist [61]. Known examples include Apelin [62,63] (Fig. 1) and Toddler/Elabela (also called Apela/Ende) [1,43], which are produced as 77 and 54 amino acid preproteins including a signal peptide, respectively. Zebrafish Toddler/Elabela is a bona-fide ligand for the apelin receptor and acts as a mitogen, required for mesendodermal cell migration during gastrulation and for cardiac development [1,43]. Toddler/Elabela also seems to be involved in
Molecular Actions of Microproteins - Small protein examples in the Protein Data Bank (PDB)

1) Microproteins are essential or optional structural components of macromolecular complexes.

The mammalian mitochondrial F-ATPase epsilon subunit is a 50-65 amino acid protein [3]. It is absent in bacterial or chloroplast ATPases, and even to date, its evolutionary history and functional purpose is not entirely elucidated. Genetic ablation suggests that it functions as an assembly or stabilizing factor of the macromolecular complex [3,5]. The microprotein folds into short helices along the central stalk of ATPase [6], thus potentially regulating assembly of the F0 and F1 subcomplexes. PDB: 6J5J

2) Microproteins can act as allosteric regulators of macromolecular complexes

Next to various recently identified microproteins [54], also two long-known proteins, the homologous transmembrane proteins Sarcolipin and Phospholamban [68-71] regulate activity of the sarcoplasmic reticulum Ca²⁺-ATPase (SERCA) in muscle cells. Through lateral interactions in the ER membrane, Sarcolipin and Phospholamban stabilize intermediate states in the conformational cycle of the ATP-dependent calcium pump, thus slowing down the enzyme turnover rate [71]. It is conceivable from this example, that microproteins can evolve independently of the core enzyme/macromolecular complex, thus providing evolutionary opportunities for new regulatory mechanism without sacrificing core activity. PDB: 4H1W

3) Microproteins can act in signaling cascades as intra- or extracellular messengers.

Apelin is a secreted microprotein synthesized as a 77 amino acid pre-pro-protein with a signal peptide. After cleavage of the signal peptide, further proteolysis can generate a number of isoforms active as ligands for G-protein coupled receptors. A co-crystal structure with the Apelin receptor, exemplifies the extensive interactions made in a deep binding pocket, conferring exquisite specificity to this protein-protein interaction [64]. The same principles of molecular action apply to a wealth of known peptide hormones processed from larger pre-pro-proteins. PDB: 5VBL

4) Microproteins can act in an autonomous fashion to catalyze an enzymatic activity or enable transport, modulate the properties of cellular membranes or other cellular components.

In addition to its role in regulating SERCA activity described above, Phospholamban has been shown to self-assemble pentameric transmembrane pores [69] that appear to have gating properties similar to regular membrane channels. Beyond cellular microproteins, viral membrane-inserting peptides are known to form viroporins [19]. PDB: 1ZLL

The family of high-mobility group proteins, including a 90 amino acid protein, HMGN2, contain large stretches of intrinsically unstructured, Arg/Lys-rich regions with highly basic net charge. Associated with the linker DNA between nucleosomes, relatively little is known about how these proteins contribute to packaging and condensation of chromatin, and regulation of DNA-templated processes such as transcription or replication [78].
cardiovascular development in mice, where the peptide is secreted from the placenta and embryonic tissue, and peptide loss results in defective placental angiogenesis and promotes preeclampsia [48,49]. Toddler/Elabela deficiency can lead to embryonic lethality [1,43,48,49], illustrating that secreted microproteins do not only carry out fine-tuning functions but may play essential roles.

Serving an entirely different biology, the 66 amino acid secretory microprotein SELF-1, is required for self-recognition in *Pristionchus pacificus* [64], a predatory nematode which feeds on larvae. While the

---

**Fig. 2.** The different functions of sORF translation. Next to producing microproteins with independent functions on the protein level (Top left), sORF translation can be a means of regulating translation. For example, uORFs expression is thought to commonly repress the downstream canonical ORF, though in some cases can also lead to upregulation of the canonical ORF (Top right). Further, sORF translation can be utilized as a source of neutrally evolving peptides which, under certain circumstances might acquire function and thus become precursors of novel proteins (Bottom left). Lastly, some sORFs produce peptides, which are utilized for antigen display instead of carrying out individual, independent functions (Bottom right). Finally, it is important to note that these sORF functions are not mutually exclusive, and thus, translation of an sORFs may serve multiple purposes.
molecular mechanism is not understood, SELF-1 helps
to distinguish kin from other closely related species by
recognition of a hypervariable stretch of amino acids
within the microprotein [64].

Excitingly, biophysical properties of some smaller
microproteins appear to be compatible with cellular
secretion and uptake without the need for a signal
peptide: the mitochondrial-derived 16 amino acid
human/mouse MOTS-c [36] and the 11–32 amino acid
cytoplasmic reticulum miR sORF microproteins [45], which are
implicated in insect embryonic pattern formation, have
been shown to act in a non-cell-autonomous manner and
thus must be membrane-diffusible or use yet to be
discovered transport machinery. Together, these ex-
amples foreshadow a wider use of secreted or diffusible
microproteins in cellular, organismal and interspecies
chemical communication.

Membrane-spanning microproteins

Single-pass transmembrane proteins comprise an extra-
cellular domain, a single hydrophobic transmembrane
helix and an intracellular domain. While the majority
of these proteins range from hundreds to thousands of
amino acids in size, long-known examples of minimal
single-pass transmembrane proteins exist in various
 cellular compartments. For example, the translocase of
the outer mitochondrial membrane (TOM) complex
contains three 50–70 amino acid subunits, TOM5,
TOM6, and TOM7 [65,66] (Fig. 1). Phospholamban
[67,68] and sarcolipin [69,70], 52, and 31 amino acid
proteins, respectively, are regulators of the sarco/endo-
plasmic reticulum Ca\(^{2+}\)-ATPase (SERCA) in muscle
cells (Fig. 1). Recent discoveries have added to this
class of microproteins localized to various cellular
membranes: sarcolamban [54], DWORF [52], myoreg-
ulin [51], endoregulin, and another-regulin [53] joined
the group of SERCA modulators. SPAR resides in the
lysosomal membrane and negatively regulates
mTORC1 activity in stress conditions through v-
ATPase interaction [31]. Pigbos is located in the outer
mitochondrial membrane at contact sites with the ER
and modulates the unfolded protein response [41],
while mitoregulin spans the inner mitochondrial mem-
brale, influencing cellular lipid composition and sup-
porting respiratory activity [34,37]. Hemotin localizes
to endosomes in Drosophila macrophages (hemocytes)
and contributes to functional phagocytosis by indi-
rectly repressing phosphatidylinositol enzymes. Conse-
quently, hemotin dysfunction results in a reduced
ability to clear infections and decreased lifespan [27].
Notably, membrane microproteins have also been
implicated in cell fusion, such as the microprotein
myomixer/minion/myomerger, which functions in myo-
blast fusion and knockout of which leads to motion-
less embryos in mice [44,55,56]. Beyond these
examples, UniProt has curated a further 43 small inte-
gral membrane proteins in the human genome
(SMIM1-SMIM43), most of which have not been
classified.

As far as known from existing structures (Fig. 1),
membrane-spanning microproteins exert their molecular
function via lateral interactions of their transmem-
brane helix. It is easily rationalized that multiple
allosteric regulators, integrating different signals, could
collectively fine-tune activity of membrane-resident
complexes such as channels, pumps, and signaling
receptors via lateral interactions within the lipid
bilayer. Membrane-spanning microproteins thus repre-
sent a minimal highly effective structural theme.
Finally, given that viruses are masters in imitating cel-
lular principles, it does not seem surprising that
amongst the many sORFs encoded in the dense viral
genomes [17,18], a considerable fraction also comprises
predicted transmembrane helices.

Intrinsically disordered microproteins

Intrinsically disordered regions are part of most pro-
teins and, rather than merely spacing folded functional
domains, contribute to flexible, adaptable, and rapid
evolvable protein–protein interactions networks via so-
called linear motifs [71,72]. sORF-produced protein
candidates are enriched in intrinsically disordered pro-
tein (IDP) sequences [73], which are not equivalent to
random sequence and thus indicate natural selection.
We thus speculate that seemingly spurious microprotein
products may hence provide a combinatorial repertoire
of linear motifs that could be suitable for fine-tuning
protein–protein interaction networks. For example, the
54 amino acid protein NoBody, a component of P bod-
ies, is predicted to be disordered except for a short cen-
tral motif that interacts with RNA decapping factor
EDC4 [15]. The p53-induced 54 amino acid IDP Noxa
contains a conserved BH3 motif and responds to stress
conditions by triggering apoptosis [74] (Fig. 1). The 90
amino acid high nuclear mobility group proteins
HMGN2/4 are Arg/Lys-rich and entirely disordered,
yet contain a conserved linear motif that mediates inter-
action with the nucleosome core particle [75]. Interest-
ingly, all of the above examples appear to be relatively
recent evolutionary additions (NoBody arose in mam-
mals, Noxa and HMGN2/4 in vertebrates), supporting
the idea of new functions arising via short linear motifs
(Fig. 2, see also ‘sORFs as a mechanism for protein
evolution’ below).
**Sources of microproteins**

Short ORFs can be found everywhere in the genome, from canonical protein-coding genes to noncoding genes to intergenic transcribed regions (Fig. 3A) [76]. Hitherto deemed ‘noncoding’ transcripts that produce microproteins include long noncoding RNAs (lncRNAs) [24,28,29,31,61,77,78], microRNAs [79], circular RNAs [61,80], or ribosomal RNAs [36,81]. Microproteins may also be encoded by antisense transcripts [82].

Further, many canonical mRNAs exhibit alternative ORFs, which can either display no, full, or partial overlap with the canonical protein-coding ORF [83–95]. Even though eukaryotic transcripts are considered monocistronic by default due to the 5’ scanning mechanism of translation initiation and the disassembly of ribosomes at the termination site [96], there is growing evidence that this mechanism is leaky in allowing alternative start sites and reinitiation after termination sites. Consequently, multiple ORFs per transcript akin polycistronic prokaryotic mRNAs have also been observed in eukaryotes [45,89,90,97] and various microproteins from upstream ORFs (uORFs) in the 5’UTR have been reported [89–95]. Ribosome profiling [85,86], proteomic [83,84,87], and overexpression studies [83] also suggest evidence for peptides translated from the 3’UTR. Thus, downstream ORFs (dORF)-produced active microproteins may be discovered in the future. Finally, peptides may also arise from...

---

**Fig. 3.** Sources and identification of coding sORFs. (A) The transcriptome contains a plethora of sORFs (gray arrows) amongst which are a smaller number of coding sORFs (red arrows). These microprotein-encoding sORFs can be located in the 5’ or 3’UTR, in a different frame from the translated canonical ORF/coding region (CDS) (blue arrow), in previously deemed noncoding RNAs, introns or in transcripts produced from antisense regions. (B) The three main methods for narrowing down the number of coding sORFs are based on bioinformatics, ribosome profiling and MS.
Alternative functions of sORF translation

Despite the many examples of functional sORF products presented above, it is important to appreciate that sORF translation on its own does not imply production of a stable and distinctly functional microprotein. sORF translation may also serve a number of alternative (and not mutually exclusive) purposes as summarized below (Fig. 2).

Regulatory sORFs

Short ORFs translation can be a means for expression regulation of other ORFs. For example, translation of uORFs is commonly observed in ribosome profiling studies, and while it may produce stable microproteins [89–94], uORF translation is thought to predominantly regulate expression of the canonical ORF in cis. Various studies have shown that presence of uORFs in an RNA is correlated with lower levels of the respective RNA and a significant translational repression of the ‘main ORF’ [102–104]. This activity seems to be influenced by a variety of factors such as uORF number, length, and sequence [102–104]. uORF translation can elicit repression of a downstream ORF through a number of mechanisms including sequestering ribosomes away from the ‘main ORF’ or increasing transcript targeting to non-sense-mediated RNA decay pathways [95,105,106]. Though there are examples of uORFs which produce microproteins that in turn regulate canonical ORF expression [92–94], only a small percentage of translated uORFs show amino acid conservation that could be indicative of a functional microprotein product [61,102] (Fig. 2).

Contrary to these examples of translational repression, some recent genome-wide studies have not observed a general anticorrelation between uORF and main ORF translation [107] or have even described a mild positive correlation [61]. Interestingly, in certain stress conditions, uORFs can actually uphold translation of the main ORF while global translation is decreased, a mechanism that has been implicated in cancer initiation of squamous cell carcinomas, in which uORFs are maintaining the translation of cancer genes through usage of alternative translation initiation factors [108,109]. Additionally, it has recently been described that also translation of dORF can regulate and enhance expression of the canonical ORF [110], highlighting how much there is yet to discover about the diverse regulatory functions of alternative ORFs.

sORFs as a source for antigenic peptides

Evidence that translation products of sORFs can be loaded and displayed onto major histocompatibility complex class I molecules (MHC I) (Fig. 2) suggests a utility of sORF translation for the immune system [98,99,111–116]. MHC I complexes present fragments of the intracellular proteome to cells of the immune system, which, in case of foreign or altered antigens indicating disease, can trigger cellular destruction. Originally, it was thought that peptides displayed on MHC class I molecules mostly arise from proteolytic cleavage of canonical proteins, including those of intracellular pathogens [111]. However, in recent years it has been shown that antigenic peptides can also be produced from cryptic translation events including translation of sORFs [98,99,111–116]. These can serve as signals of ‘self’ in normal conditions, but may also provide a source of disease-associated antigens. For example, mRNAs overexpressed in cancer can produce peptides from uORFs or alternative ORFs that are then loaded onto the MHC I complex and elicit immune responses [115,116]. Due to their small size, sORF-produced peptides require no or minimal
processing and might thus also possess a kinetic advantage compared to ‘conventional’ antigens [111].
However, since antigen production is not mutually exclusive with regulatory sORF functions or the production of a distinctly functional microprotein, it is not clear to date if antigen production is the sole or main purpose of these sORFs.

**sORFs as a mechanism for protein evolution**

Finally, pervasive sORF translation resulting in non-functional products may serve as an intermediate for *de novo* protein evolution [117] (Fig. 2). Many sORFs are not conserved across species and do not show any selection pressure. It is thought that these neutrally evolving sORFs provide a pool of sequences which might acquire functionality under new conditions and thus serve as precursors of bioactive microproteins [77,118]. This function would be similar to unstructured regions in proteins that have been proposed to provide opportunities to rapidly evolve new or remodel existing protein-protein interactions [119]. Consequently, one could speculate that microproteins might be enriched in processes that require rapid adaptation to changing environments or external stresses. A fitting example in this context is the secreted taxon-restricted microprotein SELF-1 in nematodes (also discussed above), which provides self-recognition from closely related species through rapid diversification [64].

Alternative to *de novo* evolution, frequent gene duplication events during evolution create a pool of pseudogenes, typically containing fragments of once coding proteins. While often considered nonfunctional, pseudogene-encoded sORFs that can independently evolve from the parental gene may provide another useful source for evolving new protein function. Indeed, it has been shown that peptides from pseudogenes are expressed and underlie purifying selection [120,121]. The evolutionary expansion of Bcl-2 family proteins, master regulators of apoptosis, exemplifies such rapid evolution of new protein–protein interactions: based on the BH3 domain, a relatively short conserved motif, a variety of small so-called BH3-only proteins have evolved either *de novo* or by gene duplication [122]. Noxa (Fig. 1, also discussed above) is a relatively recent addition in evolution. Since no common ancestors with other BH3-only proteins have been identified, the Noxa sORF is thought to have evolved *de novo* as a p53-dependent inducer of apoptosis. In rodent evolution, Noxa ORF has doubled in size through a tandem duplication of the BH3 domain, exemplifying how sORFs can be intermediates in the evolution of larger proteins [74,122,123].

Taken together, it is important to appreciate that sORF functions are diverse and that translation of an sORF does not necessarily equate to production of a distinctly functional microprotein.

**Identifying coding sORFs**

The concept that sORF could encode active microproteins was proposed several decades ago [7]. However, we still know very little about the functionality of the vast majority of sORFs products, since it is not trivial to distinguish coding from noncoding sORFs. Many protein identification methods possess biases against small proteins and many sORFs contain nonconventional start codons which further complicates their identification [124]. However, in recent years, technological advances, particular in ‘omics’ methods, have propelled this field forward. Here, we give a brief overview of these methods, which can be divided into *in silico*, ribosome profiling-, and MS-based techniques (Fig. 3B).

**In Silico-based approaches**

There are a variety of studies that have predicted microproteins predominantly through bioinformatic tools [9,125–140], which are mostly based on detecting purifying selection or similarity with known proteins or domains (Table 1, Fig. 3B). One of the common ways to assess purifying selection is to test for nonsynonymous versus synonymous mutation ratios (dN/dS) by utilizing homologous sequence alignments [139]. Since this methodology can be problematic for small proteins as sequences might be too short for successful alignment and limited number of codons might result in less robust substitution rates [117,139], tools like PhyloCSF [139] additionally take phylogenetic models into account and have been shown to perform better on short sequences than (dN/dS) alone [139]. If an sORF is not conserved between species and thus no alignment can be performed, it is also possible to utilize single nucleotide variants to investigate for selection pressure [117,118]. Another selection-based approach is assessment of nucleotide composition with tools such as PhastCons [130]. While it is important to consider that a sequence might also be conserved for reasons other than translation of the sORF (e.g., regulatory regions) [117], an advantage of this metric is that it takes into account the sORF context (e.g., favorable translation initiation sites) instead of only focusing on conservation across the protein-coding region [117]. Further, many emerging tools, such as DeepCPP and RNAsamba [141,142] utilize machine learning approaches and possess the advantage of only
rnasamba Machine learning Camargo et al.
RNAcode Codon substitution, Kalkhof et al.
micPDPa Codon substitution, Bazzini et al.
PhyloCSF Codon substitution, Lin et al.
PFAM Similarity to protein motifs, El-Gebali
PhastCons Nucleotide composition, Siepel
BLAST Sequence similarity, Altschul et al.
sORF finder a nucleotide composition, Eddy et al.
CRITICA Codon substitution, Badger et al.
DeepCPP Machine learning, Zhang et al.
ELM Similarity to linear protein motifs, Eddy et al.
HMMER Similarity to protein domains, El-Gebali et al.
PhastCons Nucleotide composition, Sniegel et al.
PhyloCSF Codon substitution, Lin et al.
RNAcode Codon substitution, Kaikhoof et al.
RNAsamba Machine learning, Camargo et al.
ufEPeroni Codon substitution, cross-species conservation, Skarshewski et al.

*a* sORF-specific tools.

| Table 1. Selection of bioinformatic tools that aid in predicting coding potential. |
|---------------------------------|------------------|------------------|
| **Tool**                        | **Assessment principle** | **Source**       |
| micPDPa                        | Codon substitution    | Bazzini et al., 2014 [86] |
| miPepida                      | Machine learning      | Zhu et al., 2019 [136]  |
| sORF finder                    | Codon substitution, nucleotide composition | Hanada et al., 2009 [131] |
| BLAST                          | Sequence similarity   | Altschul et al., 1990 [126] |
| CRITICA                        | Codon substitution, nucleotide composition | Badger et al., 1999 [138] |
| DeepCPP                        | Machine learning      | Zhang et al., 2020 [141]  |
| ELM                            | Similarity to linear protein motifs | Eddy et al., 1995 [128] |
| HMMER                          | Sequence similarity   | Eddy et al., 1995 [128] |
| PFAM                           | Similarity to protein domains | El-Gebali et al., 2019 [129] |
| PhastCons                      | Nucleotide composition | Sniegel et al., 2005 [130] |
| PhyloCSF                       | Codon substitution    | Lin et al., 2011 [139]  |
| RNAcode                        | Codon substitution    | Kaikhoof et al., 2011 [178] |
| RNAsamba                       | Machine learning      | Camargo et al., 2020 [142] |
| uPEPeroni                      | Codon substitution, cross-species conservation | Skarshewski et al., 2014 [140] |

A second commonly utilized way of testing for coding potential is to investigate whether the predicted sORF displays similarity to any known proteins, which might also help to indicate potential functionality of the microprotein. Tools to test for sequence similarity include BLAST [126] and HMMER [128], while resources like ELM [127] and PFAM [129] can identify linear motifs or functional domains, respectively. Staying with the molecular themes discussed above, signal peptide and transmembrane prediction algorithms should also be useful to classify microprotein candidates.

While these approaches, particularly if applied together, may greatly raise confidence in predicting potentially coding sORFs, they may still miss out on small or poorly conserved microproteins [117]. For identification of those peptides, it is therefore crucial to complement in silico methods with experimental omics technologies, such as ribosome profiling.

**Ribosome profiling-based approaches**

Ribosome profiling (Ribo-seq) was first described by Ingolia et al. [143] and identifies ribosome-bound RNAs through sequencing of RNA fragments that are protected from nuclease-digestion by the ribosome (Fig. 3B), thereby indicating translated regions of the transcriptome [143]. In recent years, Ribo-seq with the help of additional software (see description below and Table 2) has been crucial in identifying thousands of potential microproteins [61,85,86,97,107,144–147] in tissues and cells of many organisms including Drosophila [107], human [61,85,144], zebrafish [86], and *E. coli* [147]. It is important to consider, also given the many roles of sORFs, that Ribo-seq signals are not sufficient to imply production of a microprotein, but can instead arise from spurious and regulatory translation [148]. In an effort to increase signal reliability, a variety of practical improvements have been made, including utilizing biotin pulldown of tagged ribosomes instead of sedimentation methods [149] and arresting ribosomes at initiation sites, via treatment with harringtonine [150] and lactimidomycin [151], which helps to assess sORFs start sites, including the non-AUG start codons contained in many microproteins [152]. Further, other techniques such as Poly-Ribo-Seq have since been developed, allowing for sequencing of fragments bound by polysomes and thus excluding more sporadic, single ribosome binding events [107]. To more reliably assess protein-coding potential, Ribo-seq data are additionally often coupled with other metrics (Fig. 3B). For example, ribosomes possess a specific footprint and translated sORFs should be bound mostly within the same frame and throughout the whole ORF body with an increase in ribosome density at the start site and decrease after the stop codon. Thus, consideration of RNA fragment lengths, 3nt-periodicity, uniformity of ribosome signal, and ribosome density [85,86,97,144,146,148,149,153–158] have been employed to corroborate evidence for thousands of sORFs. A selection of tools exploiting these parameters is listed in Table 2, and platforms such as RiboToolkit have been developed in an effort to integrate various approaches and simplify Ribo-Seq analysis [159].

**Mass spectrometry-based approaches**

A direct mean of discovering microproteins is through MS-based approaches (Fig. 3B), which have aided in the identification of hundreds of novel microproteins [83,84,160–168]. However, it is not trivial to detect microproteins through MS. For example, sample preparation usually involves tryptic digestion and, due to their small size, microproteins might contain none or very few suitable tryptic peptide fragments biasing them against larger and more stable proteins [152].
Thus, more targeted means of identifying microproteins include utilization of different proteases or top-down analysis of native microproteins [169,170] and optimized extraction buffers [10,87,165], size selection [10,84,87,165], or fractionation protocols [171] to increase detection sensitivity. Many sORFs are not contained in the typical protein repositories that are utilized as reference protein databases [84]. Therefore, proteogenomics or transcriptomics (that is coupling MS to genomic or transcriptomic data of the same source) has been increasingly employed (Fig. 3B): Various studies have identified potentially coding sORFs by mapping the MS-generated spectra to in silico 6-frame or 3-frame translations of RNA-sequencing data of the relevant tissues (Fig. 3B) [10,83,84,160–167,172–175]. While these approaches enable discovery of previously nonannotated sORFs, they also significantly increase the search space, leading to decreased sensitivity and reliability [163]. Thus, they are ideally combined with additional in silico filters [175], fractionation protocols [84,163], and/or smaller sORF-specific (Ribo-seq) datasets [166,174]. Overall, identification of microproteins has immensely profited from the use of MS, and advances in sensitivity and throughput will further increase its discovery potential.

Combining methods

Bioinformatic, translaticomic, and proteomic methods are most powerful when used in combination: Various software tools including Proteoformer [174] or PinStripe [168] have been developed to aid in integrating the different types of data (Table 3) [168,174,176], and a number of studies have already used combinatorial approaches to identify coding sORFs [61,86,90,134,135,152,158,170,172,177,178]. For example, Mackowiak et al. [134] identified sORFs in five different organisms by creating a support vector machine based on PhyloCSF and PhastCons scores and overlapped the resulting catalogue of predicted coding sORFs with previously published Ribo-Seq and MS datasets. In this way, the authors were able to identify 2000 novel sORFs of which 100 possessed Ribo-seq evidence and 70 contained MS evidence [134]. Another extensive study performed by van Heesch et al. [61] performed RNA-seq and Ribo-seq experiments in human cardiac, kidney, and liver tissue. Subsequent assessment of conservation and MS evidence of the microproteins predicted through Ribo-seq, allowed for identification of candidates which were finally tested for translation, localization, and interaction partners with low throughput experimental methods. Finally, Martinez et al. [152] utilized de novo

Table 2. Selection of Ribo-seq analysis tools.

| Tool         | Assessment principle                  | Prediction of Source                                      |
|--------------|--------------------------------------|-----------------------------------------------------------|
| FLOSS        | Read length                          | True ribosomal footprints                                 |
| ORF-RATER    | Read density                         | Coding ORFs                                                |
| ORFscore     | 3nt-periodicity, read density, read length probability | Coding ORFs                                                |
| PRICE        | 3nt-periodicity, read density, read length probability | Coding sORFs                                                |
| PTS          | 3nt-periodicity                      | Overlapping coding ORFs                                   |
| RibCode      | 3nt-periodicity                      | Coding ORFs                                                |
| riboHMIM     | 3nt-periodicity, read density         | Coding ORFs                                                |
| RibORF       | 3nt-periodicity, read uniformity      | Coding ORFs                                                |
| RiboTaper    | 3nt-periodicity                      | Coding ORFs                                                |
| Ribo-TISH    | 3nt-periodicity, read density, read length probability | Coding ORFs, translation initiation sites                  |
| RiboWave     | 3nt-periodicity, read density         | Coding ORFs                                                |
| RP-BP        | 3nt-periodicity, read length          | Coding ORFs                                                |
| RRS          | Read density (translation termination)| Coding ORFs                                                |
| SPECtre      | 3nt-periodicity                      | Coding ORFs                                                |
| TOC          | Read density, read uniformity,        | Coding ORFs                                                |
|              | translation efficiency                |                                                           |

Table 3. Selection of tools that aid in integrating different methodologies.

| Tool         | Intersected methods                  | Source                                      |
|--------------|--------------------------------------|---------------------------------------------|
| FSPP         | MS, Ribo-seq, RNA-Seq                | Li et al., 2018 [176]                       |
| PinStripe    | Bioinformatics, MS, RNA-seq          | Gascoigne et al., 2012 [168]                |
| Proteoformer | MS, Ribo-seq                         | Crappé et al., 2015 [174]                  |
transcriptome assembly (RNA-seq) in combination with in silico 3-frame translation to generate an sORFs catalogue which could then be tested for Ribo-seq evidence and conservation. This allowed for the identification of 7554 sORFs in three human cell lines, of which 2689 sORFs were detected in at least two cell lines. Combining this with MS data from human leukocyte antigen I immunoprecipitation, Martinez et al. [152] could identify 320 microproteins with proteomic evidence of which 131 also possessed Ribo-seq evidence in two cell lines [152].

Additionally, a number of sORF databases such as sORFs.org [179,180], smProt [181], and ARA-PEP [182] exist (Table 4), which aggregate various sources of microprotein evidence.

The development of new methodologies and advances in omics technologies thus have enabled large-scale discovery of sORFs. One of the next challenges is to further annotate, characterize, and validate the function of their candidate microproteins products.

### Functional screens for microproteins

While the identification of coding sORFs has seen significant advances in recent years, the further systematic investigation of functional microproteins is lagging behind. A successful strategy for identifying microproteins has been to rationally narrow down the sORF search space, yielding a manageable and promising candidate list, which subsequently can be experimentally tested and validated for functionality. Common approaches include combining differential gene expression data (e.g., from different developmental stages, tissues, or biological conditions) with the sORF identification methods described above, or focusing on a genomic region of special functional interest (such as the mitochondrial genome) for sORF coding potential, both of which led to the discovery of various functional microproteins [1,12,24,29,36,43–50,80,82].

In contrast, only a limited number of sORF-specific large-scale functional studies have been carried out to date. The first of such screens performed systematic gene deletions and designed 140 mutant sORF strains in yeast, identifying five candidates which were required for growth [16]. Another study in *A. thaliana* [14] carried out an overexpression screen, testing 473 potentially coding sORF, which were selected through integrating conservation and expression analysis, and identified 49 hits that elicited alterations in plant morphology, growth rate and flowering time [183]. In a third study, Guo et al. [184] performed a DNA variant library screen for sORFs in yeast. This screen utilized Cas9-sgRNAs-mediated double stranded breaks and homology-directed repair to introduce respective donor templates, which contained a small deletion around the sORF start codon, into the sgRNA-targeted locus, thus essentially knocking out sORF translation. In this way, the authors designed a library targeting 315 sORFs in the yeast genome and identified 68 sORF candidates whose mutations seemed to induce a fitness phenotype. While the precise targeting allowed a confident association of the phenotype to the respective sORF, the candidates were not further validated for activity on the protein level.

### Table 4. Selection of sORF-integrating databases.

| Database     | Database content                                                        | Organisms                   | Source                                |
|--------------|-------------------------------------------------------------------------|-----------------------------|--------------------------------------|
| ARA-PEP      | sORFs from Tiling Arrays, RNA-seq Integrates motif and conservation searches | *Arabidopsis thaliana*      | Hazarika et al., 2017 [182]          |
| psORF        | sORFs from literature, databases, Ribo-seq, MS Integrates conservation searches | Five core plant species and 30 additional plant species based on homology | Chen et al., 2020 [198]              |
| SmProt       | sORFs from literature, databases, Ribo-seq, MS Integrates conservation searches | Nematode, Bacteria, Yeast, Fruit fly, Human, Mouse, Rat, Zebrafish | Hao et al., 2018 [181]               |
| sORFs.org    | sORFs from Ribo-seq Integrates MS evidence and conservation searches | Mouse, Rat, Zebrafish, Human, Nematode, Fruit fly | Olexiouk et al., 2015; Olexiouk et al., 2018 [179,180] |
| OpenProt     | Alternative ORFs from 3-frame in silico translation Integrates Ribo-seq, MS, functional domain and conservation searches | Human, Mouse, Rat, Chimpanzee, Cattle, Sheep, Zebrafish, Fruit fly, Nematode, Yeast | Brunet et al., 2019; Brunet et al., 2021 [199,200] |
| uORFdb       | uORF literature Integrates isoform and conservation searches            | Various                     | Wethmar et al., 2014 [201]           |
| TISdb        | Alternative initiation sites Integrates isoform and conservation searches | Human, Mouse               | Wan et al., 2014 [202]               |

*sORF-specific tools."
Finally, a recent study screened thousands of noncanonical ORF candidates via a nonhomologous end joining-based CRISPR/Cas9 knockout screen [185]. While a previous investigation had identified the muscle microprotein myomixer in this way with a standard genome-wide library [44], Chen et al. [185] for the first time utilized a noncanonical ORF-specific library. The authors performed Riboseq experiments from which 2596 candidates, the majority being sORFs, were selected to carry out an essentiality CRISPR/Cas9 knockout screen in induced human pluripotent stem cells and a human leukemia cell line. Chen et al. [185] discovered over 500 candidates which displayed a knockout phenotype. Subsequently, the authors chose 163 candidates based on conservation and phenotypic readout, for further analysis via single-cell RNA sequencing and showed that upon knockdown of specific ORFs, a wide variety of cellular pathways seemed to be affected. Through subsequent rescue experiments of 16 microprotein candidates, Chen et al. [185] could validate that these indeed seemed to act on the protein level since rescue with the sORF but not with a start codon mutant could alleviate the phenotype. Eventually, via fusion to split mNeonGreen and subsequent microscopy and pulldown, Chen et al. [185] investigated microprotein localization and interaction partners [185].

In another recent comprehensive study, Prensner et al. [186] surveyed 553 curated noncanonical ORFs (386 of which were sORFs), for translatability and stability by expressing reporter fusions as well as carrying out in vitro translation. Additionally, the authors tested for ORF action by investigating differential RNA expression upon ectopic ORF expression and performing CRISPR/Cas9 knockout screening in eight different cell lines. Overexpression of 237 out of 553 ORFs displayed robust gene expression changes, while 57 ORF candidates exhibited viability phenotypes upon sgRNA knockout.

The latter examples highlight the immense potential of discovering functional sORFs with targeted CRISPR libraries. The combination with cDNA screens promises a more high-throughput validation of function on the microprotein level, albeit some phenotypes may be associated with unphysiologically high levels of a microprotein in such overexpression screens. CRISPRa/CRISPRi-based screens target entire transcripts and thus preclude the direct discovery of sORFs. Nevertheless, the wealth of large-scale genetic screening data acquired with these libraries as well as standard CRISPR libraries across various cell types, tissues, and organisms may still prove useful for guiding more targeted screens for sORFs in the future.

**Microprotein characterization**

**Validation of functionality on the protein level**

Irrespective of the screening approach and in context of functional complexity, additional experiments are typically required to attribute a phenotype or function unequivocally to the translation product rather than to a noncoding function of the underlying RNA sequence. This is commonly done by performing experiments with frameshift or start-codon mutants. For example, Anderson et al. [51] tested the function of myoregulin, a microprotein involved in SERCA pump regulation, by validating that overexpression of the sORF and sORF-containing lncRNA but not of a frameshift-mutated construct could elicit the phenotype [51]. Similarly, various studies have generated knockout phenotypes and tested rescue with either wild-type or frameshifted constructs [1,45,187], while van Heesch et al. mutated the endogenous AUG in the UPPERHAND lncRNA and could show that this elicited the same phenotype as siRNA knockout of the lncRNA, indicating that the RNA can likely act on the protein level [61]. Another approach of testing that an sORF functions as a microprotein could further be to supply a synthetic peptide [1,79]. This was successfully employed for example by Pauli et al. [1] where it could be shown that the synthetic Toddler peptide elicits the same phenotype as Toddler mRNA overexpression.

**Validation of protein-coding potential**

Beyond proving microprotein function, the exact start of the sORF (particularly when relying on noncanonical start codons), the extent of translation from the endogenous AUG in the UPPERHAND lncRNA and how microproteins often need to be validated case-by-case. Detecting the endogenous microprotein itself in cells or tissues typically requires the use of antibodies [24,29,31,36,52,82]. However, raising appropriate microprotein antibodies can be challenging since the choice of epitopes for immunization is limited [188]. Thus, a common way of testing sORF translation in cells has been to express a transgene encoding an epitope-tagged variant of the native sORF sequence [29,52,82]. Finally, genome editing to introduce fluorescent or epitope tags to the endogenous microprotein has been increasingly used to demonstrate endogenous translation of the sORF [28,51,189].

**Tagging of microproteins**

One of the most common methods for microprotein characterization has been to fuse the microprotein of
choice to conventional protein tags. The choice of tag is important in context of the relatively small size of the microprotein, since additional peptide or protein moieties may alter the microprotein’s biophysical properties, localization, or protein-protein interactions. While a number of studies have successfully generated APEX or GFP fusions [24,80,90,190], in other cases even a smaller tag such as FLAG tag disrupted the native function of the microprotein [64]. Split-fluorescent proteins have been used more recently and can provide a good compromise between size and versatility [185]. A minimal approach consists of tagging the microprotein with the use of a noncanonical amino acid [191–194]. Overall, it should be decided and validated from case to case which method is the most favorable and least-disruptive for studying a microprotein of interest and given the increase in microprotein research, new approaches for characterizing small sORFs will likely emerge in the future.

Discussion

Advances in a number of omics techniques such as MS, ribosome profiling, and genetic screening have propelled the microprotein field forward in the last few years, and resulted in the realization that a larger fraction of the human genome is translated into functional proteins than previously assumed. Combined with the observation that a transcript may perform more than one separable coding and/or noncoding function, the true complexity and dense information content of eukaryotic genomes becomes apparent.

So far, microprotein investigations were mostly focused on model organisms and cell lines. While a few studies [61,170] have already been performed on human tissues, it is expected that this trend will continue, entailing studies in a variety of organs as well as disease conditions. Indeed, microproteins have been implicated in diseases such as cancers [24,29,30,80], heart failure [195], and preeclampsia [49], and it is likely that through study of their molecular action, we will also increasingly discover how microproteins might be dysregulated in disease. Consequently, it is conceivable that microproteins, particularly if secreted, could be exploited as diagnostic biomarkers. Questions and challenges remain: How large a fraction of sORFs still encode hidden, undiscovered microproteins? Can we exclude the possibility that translated sORF products that do not show an apparent function are still relevant for an organism in some specific yet to be discovered context, that is, in a specialized cell type, specific window of development, under certain stress condition, and so on? Can we learn from the growing list of well-characterized microproteins how to infer function for the large pool of yet uncharacterized candidates?

We envision that the separation of microproteins and proteins may blur as a more representative and contiguous picture of the universe of small proteins emerges. Yet, microproteins might be favored and sometimes uniquely suited for certain tasks in the cell or organism. Thus, we may only appreciate the entire range of microprotein function if we dare to think beyond the limits of canonical proteins.

Acknowledgements

We thank L. Lafranchi and the reviewers for helpful feedback on the manuscript. DS was funded by a Boehringer Ingelheim Fonds PhD Fellowship. SJE acknowledges project funding from Ragnar Söderberg Fellowship in Medicine and Ming Wai Lau Center for Reparative Medicine. Figures were created with BioRender.com.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

DS and SE wrote the review.

References

1 Pauli A, Norris ML, Valen E, Chew G-L, Gagnon JA, Zimmerman S, Mitchell A, Ma J, Dubrulle J, Reyon D et al. (2014) Toddler: an embryonic signal that promotes cell movement via apelin receptors. Science 343, 1248636.
2 Herberg S, Gert KR, Schleiffer A & Pauli A (2018) The Ly6 / uPAR protein Bouncer is necessary and sufficient for species-specific fertilization. Science 361, 1029–1033.
3 Guélin E, Chevallier J, Rigoulet M, Guérin B & Velours J (1993) ATP synthase of yeast mitochondria. Isolation and disruption of the ATP epsilon gene. J Biol Chem 268, 161–167.
4 Woychik NA & Young RA (1990) RNA polymerase II subunit RPB10 is essential for yeast cell viability. J Biol Chem 265, 17816–17819.
5 Walker JE, Fearnley IM, Gay NJ, Gibson BW, Northrop FD, Powell SJ, Runswick MJ, Saraste M & Tybulewicz VL (1985) Primary structure and subunit stoichiometry of F1-ATPase from bovine mitochondria. J Mol Biol 184, 677–701.
6 Gu J, Zhang L, Zong S, Guo R, Liu T, Yi J, Wang P, Zhuo W & Yang M (2019) Cryo-EM structure of the mammalian ATP synthase tetramer bound with inhibitory protein IF1. *Science* **364**, 1068–1075.

7 Basrai MA, Hieter P & Boeke JD (1997) Small open reading frames: beautiful needles in the haystack. *Genome Announc* **7**, 768–771.

8 Saghatelian A & Couso JP (2015) Discovery and characterization of smORF-encoded polycistronic apolipoproteins. *Nat Chem Biol* **11**, 909–916.

9 Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, Carninci P, Hayashizaki Y, Bailey TL & Grimmond SM (2006) The abundance of short proteins in the mammalian proteome. *PLoS Genet* **2**, 515–528.

10 He C, Jia C, Zhang Y & Xu P (2018) Enrichment-based proteogenomics identifies microproteins, missing proteins, and novel smORFs in *Saccharomyces cerevisiae*. *J Proteome Res* **17**, 2335–2344.

11 Sbero H, Fremin BJ, Zlitni S, Edfors F, Greenfield N, Savard J, Marques-Souza H, Aranda M & Tautz D (2014) A segmentation gene in tribolium produces a polycistronic mRNA that codes for multiple conserved proteins encoded by spRNAs in *Methanosarcina mazei*. *Biochimie* **117**, 138–148.

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

13 Prasse D, Thomsen J, De Santis R, Muntel J, Becher D & Schmitz RA (2015) First description of small proteins encoded by spRNAs in *Saccharomyces cerevisiae*. *Biochimie* **117**, 134–148.

14 Kastenmayer JP, Ni L, Chu A, Kitchen LE, Au W, Yang H, Carter CD, Wheeler D, Davis RW, Boeke JD et al. (2006) Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res* **16**, 365–373.

15 Stern-Ginossar N, Weishurdb B, Michalski A, Le VTK, Hein MY, Huang S-X, Ma M, Shen B, Qian S-B, Hengel H et al. (2012) Decoding human cytomegalovirus. *Science* **338**, 1088–1093.

16 Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Morgenstern D, Yahalom-Ronen Y, Tamir H, Achdout H, Stein D, Israeli O et al. (2021) The coding capacity of SARS-CoV-2. *Nature* **589**, 125–130.

17 Nieto-Torres JL, Verdiá-Báguena C, Castaño-Rodriguez C, Aguilera VM & Enjuanes L (2015) Relevance of viroporin ion channel activity on viral replication and pathogenesis. *Viruses* **7**, 3552–3573.

18 Storz G, Wolf YI & Ramamurthi KS (2014) Small proteins can no longer be ignored. *Annu Rev Biochem* **83**, 753–777.

19 Hemm MR, Weaver J & Storz G (2020) *Escherichia coli* small proteome. *EcoSal Plus*. doi:10.1128/ecosalplus.ESP-0031-2019.

20 Duval M & Cossart P (2017) Small bacterial and phagic proteins: an updated view on a rapidly moving field. *Curr Opin Microbiol* **39**, 81–88.

21 Garai P & Blanc-Potard A (2020) Uncovering small membrane proteins in pathogenic bacteria: regulatory functions and therapeutic potential. *Mol Microbiol* **114**, 710–720.

22 Huang JZ, Chen M, Chen D, Gao XC, Zhu S, Huang H, Hu M, Zhu H & Yan GR (2017) A peptide encoded by a putative lncRNA HOXB-AS3 suppresses colon cancer growth. *Mol Cell* **68**, 171–184.e6.

23 Ray S, Rosenberg ML, Schwertner B, Toubiana W, Auman T, Schnellhammer I, Teuscher M, Valenti P, Khila A, Klingler M et al. (2019) The mlpt / Ubr3 / Svb module comprises an ancient developmental switch for embryonic patterning. *eLife* **8**, e39748.

24 Guo B, Zhai D, Cabezas E, Welsh K & Nouraini S (2003) Humanin peptide suppresses apoptosis by interfering with Bax activation. *Nature* **425**, 456–461.

25 Pueyo JJ, Magny EG, Sampson CJ, Amin U, Evans IR, Bishop SA & Couso JP (2016) Hemotin, a regulator of phagocytosis encoded by a small ORF and conserved across metazoans. *PLoS Biol* **14**, e1002395.

26 Niu L, Lou F, Sun Y, Sun L, Cai X, Liu Z, Zhou H, Wang H, Wang Z, Bai J et al. (2020) A micropeptide encoded by lncRNA MIR155HG suppresses autoimmune inflammation via modulating antigen presentation. *Sci Adv* **6**, eaaaz2059.

27 Guo B, Wu S, Zhu X, Zhang L, Deng J, Li F, Wang Y, Zhang S, Wu R, Lu J et al. (2020) Micropeptide CIP 2A- BP encoded by LINC 00665 inhibits triple-negative breast cancer progression. *EMBO J* **39**, e102190.

28 Wang Y, Wu S, Zhu X, Zhang L, Deng J, Li F, Guo B, Zhang S, Wu R, Zhang Z et al. (2020) LncRNA-encoded polypeptide ASRPS inhibits triple-negative breast cancer angiogenesis. *Cell Mol Biol* **217**, e20190950.

29 Matsumoto A, Pasut A, Matsumoto M, Yamashita R, Fung J, Monteleone E, Saghatelian A, Nakayama KI, Clohessy JG & Pandolfi PP (2017) lncRNAs encode ASR polypeptide ASRPS inhibits triple-negative breast cancer progression. *EMBO J* **39**, e102190.

30 Polycarpou-Schwarz M, Groß M, Mestdagh P, Schott J, Grund SE, Hildenbrand C, Rom J, Aulmann S, Jo
34 Stein CS, Jadiya P, Zhang X, McLendon JM, Abouassaly GM, Witmer NH, Anderson EJ, Elrod JW & Boudreau RL (2018) Mitoregulin: a lncRNA-encoded microprotein that supports mitochondrial supercomplexes and respiratory efficiency. Cell Rep 23, 3710–3709.

35 Zhang S, Reliji B, Liang C, Kerouanton B, Francisco JC, Peh JH, Mary C, Jagannathan NS, Olexiouk V, Tang C et al. (2020) Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. Nat Commun 11, 1312.

36 Lee C, Zeng J, Drew BG, Sallam T, Martin-Montalvo A, Wan J, Kim SJ, Hevener AL, De Cabo R et al. (2015) The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. Cell Metab 21, 443–454.

37 Chugunova A, Loseva E, Mazin P, Mitina A, Navalayeuy T & Bilan D (2019) LINC00116 codes for a mitochondrial peptide linking respiration and lipid metabolism. Proc Natl Acad Sci USA 116, 4940–4945.

38 Arnoult N, Correia A, Ma J, Merlo A, Garcia-Gomez S, Marie C, Tognetti M, Benner CW, Boulton SJ, Saghatelain A et al. (2017) Regulation of DNA repair pathway choice in S and G2 phases by the NHEJ inhibitor CYREN. Nature 549, 548–552.

39 Hung PJ, Johnson B, Chen B, Zha S, Tyler JK, Sleekman BP, Hung PJ, Johnson B, Chen B, Byrum AK et al. (2018) MRI is a DNA damage response adaptor during classical non-homologous end joining article MRI is a DNA damage response adaptor during classical non-homologous end joining. Mol Cell 71, 332–342.e8.

40 Slavoff SA, Heo J, Budnik BA, Hanakahi LA & Saghatelain A (2014) A human short open reading frame (sORF)-Encoded polypeptide that stimulates DNA end joining. J Biol Chem 289, 10950–10957.

41 Chu Q, Martinez TF, Novak SW, Donaldson CJ, Tan D, Vaughan JM, Chang T, Diedrich JK, Andrade L, Kim A et al. (2019) Regulation of the ER stress response by a mitochondrial microprotein. Nat Commun 10, 4883.

42 Hashimoto Y, Niikura T, Tajima H, Yasukawa T, Sudo H, Ito Y, Kita Y, Kawasumi M, Kouyama K, Doyu M et al. (2001) A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer’s disease genes and Aβ. Proc Natl Acad Sci USA 98, 6336–6341.

43 Chng SC, Ho L, Tian J & Reversade B (2013) ELABELA: a hormone essential for heart development signals via the apelin receptor. Dev Cell 27, 672–680.

44 Bi P, Ramirez-Martinez A, Li H, Cannavino J, McAnally JR, Shelton JM, Sánchez-Ortiz E, Bassel-Duby R & Olson EN (2017) Control of muscle formation by the fusogenic micropeptide myomixer. Science 356, 323–327.

45 Kondo T, Hashimoto Y, Kato K, Inagaki S, Hayashi S & Kageyama Y (2007) Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nat Cell Biol 9, 660–665.

46 Pueyo JI & Couso JP (2008) The 11-aminoacid long Tarsal-less peptides trigger a cell signal in Drosophila leg development. Dev Biol 324, 192–201.

47 Kondo T, Zanet J, Benrabah E, Valenti P, Hashimoto Y, Kobayashi S & Payre F (2010) Small peptides switch the transcriptional activity of shavenbaby during drosophila embryogenesis. Science 329, 336–339.

48 Freyer L, Hsu C-W, Nowotschin S, Pauli A, Ishida J, Kuba K, Fukamizu A, Schier AF, Hoodless PA, Dickinson ME et al. (2017) Loss of apela peptide in mice causes low penetrance embryonic lethality and defects in early mesodermal derivatives. Cell Rep 20, 2116–2130.

49 Ho L, Van DM, Tan S, Chye J, Messerschmidt DM, Chng SC, Ong S, Yi JK, Boussata S, Goh GH et al. (2017) ELABELA deficiency promotes preeclampsia and cardiovascular malformations in mice. Science 357, 707–713.

50 Zanet J, Benrabah E, Li T, Pelissier-Monier A, Chanut-Delalande H, Ronsin B, Bellen HJ, Payre F & Plaza S (2015) Pri sORF peptides induce selective proteasome-mediated protein processing. Science 349, 1356–1358.

51 Anderson DM, Anderson KM, Chang CL, Makarewich CA, Nelson BR, McAnally JR, Kasaragod P, Shelton JM, Liou J, Bassel-Duby R et al. (2015) A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell 160, 595–606.

52 Nelson BR, Makarewich CA, Anderson DM, Winders BR, Troupes CD, Wu F, Reese AL, McAnally JR, Chen X, Kavalali ET et al. (2016) A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. Science 351, 271–275.

53 Anderson DM, Makarewich CA, Anderson KM, Shelton JM, Bezprozvannaya S, Bassel-Duby R & Olson EN (2016) Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. Sci Signal 9, ra119.
54 Magny EG, Pueyo JI, Pearl FMG, Cespedes MA, Niven JE, Bishop SA & Couso JP (2013) Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. Science 341, 1116–1120.
55 Zhang Q, Vashisht AA, O’Rourke J, Corbel SY, Moran R, Romero A, Miraglia L, Zhang J, Durrant E, Schmedt C et al. (2017) The microprotein Minion controls cell fusion and muscle formation. Nat Commun 8, 15664.
56 Quinn ME, Goh Q, Kurosa M, Gamage DG, Petrany MJ, Prasad V & Millay DP (2017) Myomerger induces fusion of non-fusogenic cells and is required for skeletal muscle development. Nat Commun 8, 15665.
57 Shi J, Bi P, Pei J, Li H, Grishin NV, Bassel-Duby R, Chen EH & Olson EN (2017) Requirement of the fusogenic micropeptide myomixer for muscle formation in zebrafish. Proc Natl Acad Sci USA 114, 11950–11955.
58 Bi P, Mccanally JR, Shelton JM, Sánchez-Ortiz E, Bassel-Duby R & Olson EN (2018) Fusogenic micropeptide Myomixer is essential for satellite cell fusion and muscle regeneration. Proc Natl Acad Sci USA 115, 3864–3869.
59 Bhatta A, Atianand M, Jiang Z, Crabtree J, Blin J & Fitz Gerald KA (2020) A mitochondrial micropeptide is required for activation of the Nlrp3 inflamasome. J Immunol 204, 428–437.
60 Diao MQ, Li C, Xu JD, Zhao XF & Wang JX (2019) RPS27, a sORF-encoded polypeptide, functions antivirally by activating the NF-κB pathway and interacting with viral envelope proteins in shrimp. Front Immunol 10, doi:10.3389/fimmu.2019.02763.
61 van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Diao MQ, Li C, Xu JD, Zhao XF & Wang JX (2019). The microprotein Minion encoded in small open reading frames. FEBS Journal (2021) Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. Science 341, 1116–1120.
62 Tatemoto K, Hosoya M, Habata Y, Fujii R, Kagawa T, Zhou MX, Kawamata Y, Fukusumi S, Hinuma S, Kitada C et al. (1998) Isolation and characterization of a novel endogenous peptide ligand for the human APJ receptor. Biochem Biophys Res Commun 251, 471–476.
63 Ma Y, Yue Y, Ma Y, Zhang Q, Zhou Q, Song Y, Shen Y, Li X, Ma X, Li C et al. (2017) Structural basis for apelin control of the human apelin receptor. Structure 25, 858–866.e4.
64 Lightfoot JW, Wilecki M, Rödelsperger C, Moreno E, Susoy V, Witte H & Sommer RJ (2019) Small peptide-mediated self-recognition prevents cannibalism in predatory nematodes. Science 364, 86–89.
65 Walther DM & Rapaport D (2009) Biogenesis of mitochondrial outer membrane proteins. Biochim Biophys Acta 1793, 42–51.
66 Dembowksi M, Kunkele KP, Nargang FE, Neupert W & Rapaport D (2001) Assembly of Tom6 and Tom7 into the TOM core complex of Neurospora crassa. J Biol Chem 276, 17679–17685.
67 Kirchberger MA, Tada M & Katz AM (1975) Phospholamban: a regulatory protein of the cardiac sarcoplasmic reticulum. Recent Adv Stud Cardiac Struct Metab 5, 103–115.
68 Oxenoid K & Chou JJ (2005) The structure of phospholamban pentamer reveals a channel-like architecture in membranes. Proc Natl Acad Sci USA 102, 10870–10875.
69 Wawrzynow A, Theibert JL, Murphy C, Jona I, Martinoson A & Collins JH (1992) Sarcolipin, the “proteolipid” of skeletal muscle sarcoplasmic reticulum, is a unique, amphipathic, 31-residue peptide. Arch Biochem Biophys 298, 620–623.
70 Winther A-ML, Bublitz M, Karlsen JL, Möller JV, Hansen JB, Nissen P & Buch-Pedersen MJ (2013) The sarcolipin-bound calcium pump stabilizes calcium sites exposed to the cytoplasm. Nature 495, 265–269.
71 Meyer K, Kirchner M, Uyar B, Cheng J-Y, Russo G, Hernandez-Miranda LR, Szymborska A, Zauber H, Rudolph I-M, Willnow TE et al. (2018) Mutations in disordered regions can cause disease by creating dileucine motifs. Cell 175, 239–253.e17.
72 Babu MM, van der Lee R, de Groot NS & Gsponer J (2011) Intrinsically disordered proteins: regulation and disease. Curr Opin Struct Biol 21, 432–440.
73 Erady C, Boxall A, Puntambekar S, Suhas Jagannathan N, Chauhan R, Chong D, Meena N, Kulkarni A, Kasabe B, Prathivadi Bhayankaram K et al. (2021) Pan-cancer analysis of transcripts encoding novel open-reading frames (nORFs) and their potential biological functions. npj Genomic Med 6, doi:10.1038/s41525-020-00167-4.
74 Ploner C, Kofler R & Villunger A (2008) Noxa: at the tip of the balance between life and death. Oncogene 27, S84–S92.
75 Hock R, Furusawa T, Ueda T & Bustin M (2007) HMG chromosomal proteins in development and disease. Trends Cell Biol 17, 72–79.
76 Couso J & Patraquim P (2017) Classification and function of small open reading frames. Nat Rev Mol Cell Biol 18, 575–589.
77 Ruiz-Oreja J, Meseguer X, Subirana JA & Alba MM (2014) Long non-coding RNAs as a source of new peptides. Elife 3, e03523.
78 Cai B, Li Z, Ma M, Wang Z & Han P (2017) LncRNA-Six1 encodes a micropeptide to activate Six1 in Cis and is involved in cell proliferation and muscle growth. Front Physiol 8, 1–13.
79 Laressergues D, Couzigou J-M, San Clemente H, Martinez Y, Dunand C, Becard G, Comberi JP, Clemente HS, Martinez Y, Dunand C et al. (2015)
Primary transcripts of microRNAs encode regulatory peptides. *Nature* **520**, 90–93.

80. Zhang M, Zhao K, Xu X, Yang Y, Yan S, Wei P, Liu H, Xu J, Xiao F, Zhou H *et al.* (2018) A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. *Nat Commun* **9**, 4475.

81. Cobb J, Lee C, Xiao J, Yen K, Wong RG, Nakamura HK, Mehta H, Gao Q, Ashur C, Huffman DM *et al.* (2016) Naturally occurring mitochondrial - derived peptides are age - dependent regulators of apoptosis, insulin sensitivity, and inflammatory markers. *Aging* **8**, 796–809.

82. Denli AM, Narvaiz A, Kerman BE, Pena M, Benner C, Marchetto MCN, Diedrich JK, Aslanian A, Ma J, Moresco JJ *et al.* (2015) Primate-specific ORF0 contributes to retrotransposon-mediated diversity. *Cell* **163**, 833–893.

83. Vandenberghe B, Lucier JF, Bissonnette C, Motard J, Tremblay G, Vandenberghe S, Wiszorski M, Salzet M, Boisvert FM & Roucou X (2013) Direct detection of alternative open reading frame translations products in human significantly expands the proteome. *PLoS One* **8**, e70698.

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

85. Ji Z, Song R, Regev A & Struhl K (2015) Many lncRNAs, 5′ UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4**, e08990.

86. Bazzini AA, Johnstone TG, Christiano R, MacKowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC *et al.* (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**, 981–993.

87. Ma J, Diedrich JK, Jungreis I, Donaldson C, Vaughan J, Kellis M, Yates JR & Saghatelian A (2016) Improved identification and analysis of small open reading frame encoded polypeptides. *Anal Chem* **88**, 3967–3975.

88. Vandenberghe B, Staskevicius AB, Tremblay G, McCoy M, O’Neill MA, Cashman NR & Roucou X (2011) An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. *FASEB J* **25**, 2373–2386.

89. Delcourt V, Roy AV, Salzet M, Fournier I & Roucou X (2018) 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 Proteomics* **17**, 2402–2411.

90. Samandi S, Roy AV, Delcourt V, Lucier JF, Gagnon J, Beaudoin MC, Vandenberghe B, Breton MA, Motard J, Jacques JF *et al.* (2017) Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *Elife* **6**, e27860.

91. Liu J, Yosten GL, Ji H, Zhang D, Zheng W, Speth RC, Samson WK & Sandberg K (2014) Selective inhibition of angiotensin receptor signaling through Erk1/2 pathway by a novel peptide. *AJP Regul Integr Comp Physiol* **306**, R619–R626.

92. Rahmani F, Hummel M, Schuermann J, Wiesekingberg A, Smeenk S & Hanson J (2009) Sucrose control of translation mediated by an upstream open reading frame-encoded peptide. *Plant Physiol* **150**, 1356–1367.

93. Ebina I, Takemoto-tsutsumi M, Watanabe S, Koyama H, Endo Y, Kimata K, Igarashi T, Murakami K, Kudo R, Ohsumi A *et al.* (2015) Identification of novel *Arabidopsis thaliana* upstream open reading frames that control expression of the main coding sequences in a peptide sequence-dependent manner. *Nucleic Acids Res* **43**, 1562–1576.

94. Jousse C, Bruhat A, Carraro V, Urano F, Ferrara M, Ron D & Fafournoux P (2001) Inhibition of CHOP translation by a peptide encoded by an open reading frame localized in the chop 5′UTR. *Nucleic Acids Res* **29**, 4341–4351.

95. Renz PF, Valdivia Francisca F & Sendoel A (2020) Some like it translated: small ORFs in the 5′UTR. *Exp Cell Res* **396**, 112229.

96. Hinnebusch AG (2014) The scanning mechanism of eukaryotic translation initiation. *Annu Rev Biochem* **83**, 779–812.

97. Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF & Baranov PV (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res* **22**, 2219–2229.

98. Apcher S, Millot G, Daskalogianni C, Scherl A, Manoury B & Fahraeus R (2013) Translation of pre-spliced RNAs in the nuclear compartment generates peptides for the MHC class I pathway. *Proc Natl Acad Sci USA* **110**, 17951–17956.

99. Duvallet E, Boulpicante M, Yamazaki T, Atkins JF & Baranov PV (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res* **22**, 2219–2229.

100. Yu X, Zhang Y, Li T, Ma Z, Jia H, Chen Q, Zhao Y, Zhai L, Zhong R, Li C *et al.* (2017) Long non-coding RNA Linc-RAM enhances myogenic differentiation by interacting with MyoD. *Nat Commun* **8**, 14016.

101. Lewandowski JP, Dumbovič G, Watson AR, Hwang T, Jacobs-Palmer E, Chang N, Much C, Turner K, Kirby C, Rubinstein ND *et al.* (2020) The Tug1 locus is essential for male fertility. *Genome Biol* **21**, 237.
102 Johnstone TG, Bazzini AA & Giraldez AJ (2016) Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J* **35**, 706–723.

103 Calvo SE, Pagliarini DJ & Mootha VK (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci USA* **106**, 7507–7512.

104 Chew G-L, Pauli A & Schier AF (2016) Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nat Commun* **7**, 11663.

105 Cabrera-Quio LE, Herberg S & Pauli A (2016) Decoding sORF translation – from small proteins to gene regulation. *RNA Biol* **13**, 1051–1059.

106 Tanaka M, Sotta N, Yamazumi Y, Yamashita Y, Miwa K, Murota K, Chiba Y, Hirai MY, Akiyama T, Onouchi H *et al.* (2016) The Minimum open reading frame, AUG-stop, induces boron-dependent ribosome stalling and mRNA degradation. *Plant Cell* **28**, 2830–2849.

107 Aspen JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M & Couso J (2014) Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *Elife* **3**, e03528.

108 Sendoel A, Dunn JG, Rodriguez EH, Naik S, Gomez NC, Hurwitz B, Leverage J, Dill BD, Schramek D, Molina H *et al.* (2017) Translation from unconventional S' start sites drives tumor initiation. *Nature* **541**, 494–499.

109 Andreev DE, O’Connor PBF, Fahey C, Kenny EM, Terenin IM, Dmitriev SE, Cormican P, Morris DW, Shatsky IN & Baranov PV (2015) Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *Elife* **4**, e03971.

110 Wu Q, Wright M, Gogol MM, Bradford WD, Zhang N & Bazzini AA (2020) Translation of small downstream ORFs enhances translation of canonical main open reading frames. *EMBO J* **39**, e104763.

111 Starck SR & Shastri N (2011) Non-conventional sources of peptides presented by MHC class I. *Cell Mol Life Sci* **68**, 1471–1479.

112 Schwab SR, Li KC, Kang C & Shastri N (2003) Constitutive display of cryptic translation products by MHC class I molecules. *Science* **301**, 1367–1371.

113 Starck SR, Jiang V, Pavon-Eternod M, Prasad S, McCarthy B, Pan T & Shastri N (2012) Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC Class I. *Science* **319**, 1719–1724.

114 Starck SR, Ow Y, Jiang V, Tokuyama M, Rivera M, Qi X, Roberts RW & Shastri N (2008) A distinct translation initiation mechanism generates cryptic peptides for immune surveillance. *PloS One* **3**, e3460.

115 Rosenberg SA, Tong-On P, Li Y, Riley JP, El-gamil M, Parkhurst MR, Robbins PF, Robbins PF *et al.* (2002) Identification of BING-4 cancer antigen translated from an alternative open reading frame of a gene in the extended MHC class II region using lymphocytes from a patient with a durable complete regression following immunotherapy. *J Immunol* **168**, 2402–2407.

116 Weinzierl AO, Maurer D, Altenberend F, Schneiderhan-Marra N, Klingel K, Schoor O, Wernet D, Joos T, Rammensee HG & Stevanovic S (2008) A cryptic vascular endothelial growth factor T-cell epitope: identification and characterization by mass spectrometry and T-cell assays. *Cancer Res* **68**, 2447–2454.

117 Ruiz-orera J & Albà MM (2019) Translation of small open reading frames: roles in regulation and evolutionary innovation. *Trends Genet* **35**, 186–198.

118 Ruiz-Orrera J, Verdaguer-Grau P, Villanueva-Cañas JL, Messeguer X & Albà MM (2018) Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat Ecol Evol* **2**, 890–896.

119 Babu MM (2016) The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem Soc Trans* **44**, 1185–1200.

120 Xu J & Zhang J (2016) Are human translated pseudogenes functional? *Mol Biol Evol* **33**, 755–760.

121 Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S *et al.* (2014) A draft map of the human proteome. *Nature* **509**, 575–581.

122 Aouacheria A, Rech de Laval V, Combet C & Hardwick JM (2013) Evolution of Bcl-2 homology motifs: homology versus homoplasy. *Trends Cell Biol* **23**, 103–111.

123 Saito H, Kashida S, Inoue T & Shiba K (2007) The role of peptide motifs in the evolution of a protein network. *Nucleic Acids Res* **35**, 6357–6366.

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

125 Crowe M, Wang X-Q & Rothnagel J (2006) Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. *BMC Genom* **7**, 6.

126 Alschesul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**, 403–410.

127 Gouw M, Michael S, Sámano-Sánchez H, Kumar M, Zeke A, Lang B, Bely B, Chemes LB, Davey NE, Deng Z *et al.* (2018) The eukaryotic linear motif resource - 2018 update. *Nucleic Acids Res* **46**, D428–D434.

128 Eddy SR (1995) Multiple alignment using hidden Markov models. *Proc Third Int Conf Intell Syst Mol Biol* **3**, 114–120.
129 El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A et al. (2019) The Pfam protein families database in 2019. *Nucleic Acids Res* 47, D427–D432.

130 Siepel A, Bejerano G, Pinhasi R, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LDW, Richards S et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15, 1034–1050.

131 Hanada K, Akiyama K, Sakurai T, Toyoda T, Shinozaki K & Shiu SH (2009) sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics* 26, 399–400.

132 Hemm MR, Paul BJ, Schneider TD, Storz G & Rudd KE (2008) Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol Microbiol* 70, 1487–1501.

133 Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A & Couso J (2011) Hundreds of putatively functional small open reading frames in Drosophila. *Genome Biol* 12, R118.

134 Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, Mastrobuoni G, Rajewsky N, Kempa S, Selbach M et al. (2015) Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol* 16, 179.

135 Mudge JM, Jungreis I, Hunt T, Gonzalez JM, Wright JC, Kay M, Davidson C, Fitzgerald S, Seal R, Tweedie S et al. (2019) Discovery of high-confidence human protein-coding genes and exons by whole-genome PhylLoCsf helps elucidate 118 GWAS loci. *Genome Res* 29, 2073–2087.

136 Zhu M & Gribskov M (2019) MiPepid: micropeptide identification tool using machine learning. *BMC Bioinformatics* 20, 559.

137 Kozema E, Estrada Girona G, Paci G, Lemke EA, Kele P, Chin JW, Lemke EA, Kele P, Normanno D, Singer RH et al. (2017) Bioorthogonal double-fluorogenic siliconrhodamine probes for intracellular super-resolution microscopy. *Chem Commun* 53, 6696–6699.

138 Badger JH & Olsen GJ (1999) CRITICA: coding region identification tool involving comparative analysis. *Mol Biol Evol* 16, 512–524.

139 Lin MF, Jungreis I & Kellis M (2011) PhylLoCsf: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 27, 275–282.

140 Skarszewski A, Stanton-Cook M, Huber T, Al Mansoori S, Smith R, Beatson SA & Rothenagel JA (2014) uPEPperoni: an online tool for upstream open reading frame location and analysis of transcript conservation. *BMC Bioinformatics* 15, 36.

141 Zhang Y, Jia C, Fullwood MJ & Kwoh CK (2020) DeepCPP: a deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for RNA coding potential prediction. *Brief Bioinform*, bbaa039.

142 Camargo AP, Sourkov V, Pereira GAG & Carazzolle MF (2020) RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences. *NAR Genomics Bioinforma* 2, lqz024.

143 Ingolia NT, Ghaemmaghami S, Newman JRS & Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218–223.

144 Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, Landthaler M, Obermayer B & Ohler U (2016) Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods* 13, 165–170.

145 Crapppé J, Van Criekinge W, Trooskens G, Hayakawa E, Luyten W, Baggerman G & Menshchert G (2013) Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genom* 14, 648.

146 Fields AP, Rodriguez EH, Jovanovic M, Stern-Ginossar N, Haas BJ, Mertins P, Raychowdhury R, Haochen N, Carr SA, Ingolia NT et al. (2015) A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol Cell* 60, 816–827.

147 Weaver J, Mohammad F, Buskirk AR & Storz G (2019) Identifying small proteins by ribosome profiling with stalled initiation complexes. *MBio* 10, e02819-18.

148 Guttmann M, Russell P, Ingolia NT, Weissman JS & Lander ES (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154, 240–251.

149 Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Tallhaourne GJS, Jackson SE, Wills MR & Weissman JS (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* 8, 1365–1379.

150 Ingolia NT, Lareau LF & Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147, 789–802.

151 Lee S, Liu B, Lee S, Huang S, Shen B & Qian S (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci USA* 109, E2424–E2432.

152 Martinez TF, Chu Q, Donaldson C, Tan D, Shokhirev MN & Saghatelian A (2019) Accurate annotation of human protein-coding small open reading frames. *Nat Chem Biol* 16, 458–468.

153 Chew G-L, Pauli A, Rinn JL, Regev A, Schier AF & Valen E (2013) Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development* 140, 2828–2834.
Identifying functional microproteins

D. Schlesinger and S. J. Elsässer

154 Erhard F, Halenius A, Zimmermann C, Lhernault A, Kowalewski D, Weekes MP, Stevanovic S, Zimmer R & Lars D (2018) Improved Ribo-seq enables accurate and validated identification of cryptic translation events. Nat Methods 15, 363–366.
155 Xiao Z, Huang R, Xing X, Chen Y, Deng H & Yang X (2018) De novo annotation and characterization of the translamate with ribosome profiling data. Nucleic Acids Res 46, e61.
156 Chun SY, Rodriguez CM, Todd PK & Mills RE (2016) SPECtre: a spectral coherence-based classifier of actively translated transcripts from ribosome profiling sequence data. BMC Bioinformatics 17, 482.
157 Zhang P, He D, Xu Y, Hou J, Pan BF, Wang Y, Liu T, Davis CM, Ehli EA, Tan L et al. (2017) Genome-wide identification and differential analysis of translational initiation. Nat Commun 8, 1749.
158 Raj A, Wang SH, Shim H, Harpak A, Li YI, Engelmann B, Stephens M, Gilad Y & Pritchard JK (2016) Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. Elife 5, e13328.
159 Liu Q, Shvarts T, Sliz P & Gregory RI (2020) RiboToolkit: an integrated platform for analysis and annotation of ribosome profiling data to decode mRNA translation at codon resolution. Nucleic Acids Res 48, W218–W229.
160 Wright JC, Mudge J, Weisser H, Barzine MP, Gonzalez JM, Brazma A, Choudhary JS & Harrow J (2016) Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. Nat Commun 7, 11778.
161 Oyama M, Itagaki C, Hata H, Suzuki Y, Izumi T, Natsume T, Isebe T & Sugano S (2004) Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. Genome Res 14, 2048–2052.
162 Schwaid AG, Shannon DA, Ma J, Slavoff SA, Levin JZ, Weerapana E & Saghatelian A (2013) Chemoproteomic discovery of cysteine-containing human short open reading frames. J Am Chem Soc 135, 16750–16753.
163 Zhu Y, Orre LM, Johansson HJ, Huss M, Boekel J, Vesterlund M, Fernandez-Woodbridge A, Branca RMM & Lehtio J (2018) Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nat Commun 9, 903.
164 Ma J, Saghatelian A & Shokhirev MN (2018) The influence of transcript assembly on the proteogenomics discovery of microproteins. PLoS One 13, e0194518.
165 Ma J, Ward CC, Jungreis I, Slavoff SA, Schwaid AG, Neveu J, Budnik BA, Kellis M & Saghatelian A (2014) Discovery of human sORF-encoded polyptides (SEPs) in cell lines and tissue. J Proteome Res 13, 1757–1765.
166 Koch A, Gawron DG, Steyaert S, Ndaah E, Crappe J, De KS, De ME, Ma M, Shen B, Gevaert K et al. (2014) A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. Proteomics 14, 2688–2698.
167 Oyama M, Kozuka-Hata H, Suzuki Y, Sembka K, Yamamoto T & Sugano S (2007) Diversity of translation start sites may define increased complexity of the human short ORFeome. Mol Cell Proteomics 6, 1000–1006.
168 Gascoigne DK, Cheetham SW, Cattenoz PB, Clark MB, Amaral PP, Tafj RJ, Wilhelm D, Dinger ME & Mattick JS (2012) Psinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. Bioinformatics 28, 3042–3050.
169 Fesenko I, Kirov I, Kniazee A, Khazigaleeva R, Lazarev V, Kharlampiev D, Grafskaia E, Zgoda V, Butenko I, Arapidi G et al. (2019) Distinct types of short open reading frames are translated in plant cells. Genome Res 29, 1464–1477.
170 Budamgunta H, Olexiuk V, Luyten W, Schildermans K, Maes E, Boonen K, Menschaert G & Baggerman G (2018) Comprehensive peptide analysis of mouse brain striatum identifies novel sORF-encoded polyptides. Proteomics 18, e1700218.
171 Branca RMM, Orre LM, Johansson HJ, Granholm V, Huss M, Perez-Bercoff A, Forshed J, Käll L & Lehtio J (2014) HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nat Methods 11, 59–62.
172 Lu S, Zhang J, Lian X, Sun L, Meng K, Chen Y, Sun Z, Yin X, Li Y, Zhao J et al. (2019) A hidden human proteome encoded by “non-coding” genes. Nucleic Acids Res 47, 8111–8125.
173 Heo H, Lee S, Min J, Ya Y, Young H & Oh SJ (2010) Theoretical small open reading frames (ORFs) database and massProphet: peptide Mass Fingerprinting (PMF) tool for unknown small functional ORFs. Biochem Biophys Res Commun 397, 120–126.
174 Crappe J, Ndah E, Koch A, Steyaert S, Gawron D, De Keulenaer S, De Meester E, De Meyer T, Van Criekinge W, Van Damme P et al. (2015) PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. Nucleic Acids Res 43, e29.
175 Guruceaga E, Garin-Muga A & Segura V (2020) MiTPeptideDB: a proteogenomic resource for the discovery of novel peptides. Bioinformatics 36, 205–211.
176 Li H, Xiao L, Zhang L, Wu J, Wei B, Sun N & Zhao Y (2018) FSPP: a tool for genome-wide prediction of...
Identifying functional microproteins

smORF-encoded peptides and their functions. Front Genet 9, doi:10.3389/fgene.2018.00096.

177 Malone B, Atesanov I, Aeschimann F, Li X, Großhans H & Dieterich C (2017) Bayesian prediction of RNA translation from ribosome profiling. Nucleic Acids Res 45, 2960–2972.

178 Kalkhof S, Bergen MVON, Washietl S, Findeiß S, Mu SA & Goldman N (2011) RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. RNA 17, 578–594.

179 Olexiouk V, Crappe J, Verbruggen S, Verhegen K, Martens L & Menschaert G (2015) SORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res 44, D324–D329.

180 Olexiouk V, Crappe J, Verbruggen S, Verhegen K, Martens L & Menschaert G (2018) An update on SORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res 46, D497–D502.

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

182 Hazarika RR, De Coninck B, Yamamoto LR, Martin LR, Cammue BPA & Van Noort V (2017) ARA-PEPs: a repository of putative SORF-encoded peptides in Arabidopsis thaliana. BMC Bioinformatics 18, doi:10.1186/s12859-016-1458-y.

183 Hanada K, Zhang X, Borevitz JO, Li W-H & Shiu S-H (2007) 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 17, 632–640.

184 Guo X, Chavez A, Tung A, Chan Y, Kaas C, Yin Y, Cecchi R, Garnier SL, Kelsic ED, Schubert M et al. (2018) High-throughput creation and functional profiling of DNA sequence variant libraries using CRISPR-Cas9 in yeast. Nat Biotechnol 36, 540–546.

185 Chen J, Brunner A, Cogan JZ, Nunez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD et al. (2020) Pervasive functional translation of noncanonical human open reading frames. Science 367, 1140–1146.

186 Prensner JR, Enache OM, Luria V, Krug K, Clauer KR, Dempster JM, Karger A, Wang L, Stumbraite K, Wang VM et al. (2021) Noncanonical open reading frames encode functional proteins essential for cancer cell survival. Nat Biotechnol. doi:10.1038/s41587-020-00806-2.

187 Galindo MI, Pueyo JI, Fouix S, Bishop SA & Couso JP (2007) Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol 5, 1052–1062.

188 Makarewich CA & Olson EN (2017) Mining for micropeptides. Trends Cell Biol 27, 685–696.

189 Young DJ, Guydosh NR, Zhang F, Hinnebusch AG & Green R (2015) Rli1/ABCE1 recycles terminating ribosomes and controls translation reinitiation in 3’UTRs in vivo. Cell 162, 872–884.

190 Chu Q, Rathore A, Diedrich JK, Donaldson CJ, Yates JR & Saghatelian A (2017) Identification of microprotein-protein interactions via APEX Tagging. Biochemistry 56, 3299–3306.

191 Peng T & Hang HC (2016) Site-specific bioorthogonal labeling for fluorescence imaging of intracellular proteins in living cells. J Am Chem Soc 138, 14423–14433.

192 Lafranchi L, Schlesinger D, Kimler KJ & Elsässer SJ (2020) Universal single-residue terminal labels for fluorescent live cell imaging of microproteins. J Am Chem Soc 142, 20080–20087.

193 Cao X, Khitun A, Luo Y, Na Z, Phoodokmai T, Sappakhw K, Olatunji E, Uttamapinant C & Slavoff SA (2021) Alt-RPL36 downregulates the P13K-AKT-mTOR signaling pathway by interacting with TMEM24. Nat Commun 12, 508.

194 Koh M, Ahmad I, Ko Y, Zhang Y, Martinez TF, Diedrich JK, Chu Q, Moresco JJ, Erb MA, Saghatelian A et al. (2021) A short ORF-encoded transcriptional regulator. Proc Natl Acad Sci USA 118, e2021943118.

195 Makarewich CA, Munir AZ, Schiattarella GG, Bezprozvannaya S, Raguimova ON, Cho EE, Vidal AH, Robia SL, Bassel-duby R & Olson EN (2018) The DWORF micropeptide enhances contractility and prevents heart failure in a mouse model of dilated cardiomyopathy. Elife 7, e38319.

196 Punttovoll P, Linding R, Gemünd C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DMA, Ausiello G, Brannetti B, Costantini A et al. (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 31, 3625–3630.

197 Xu Z, Hu L, Shi B, Geng S, Xu L, Wang D & Lu ZJ (2018) Ribosome elongating footprints denoised by wavelet transform comprehensively characterize dynamic cellular translation events. Nucleic Acids Res 46, e109.

198 Chen Y, Li D, Fan W, Zheng X, Zhou Y, Ye H, Liang X, Du W, Zhou Y & Wang K (2020) PsORF: a database of small ORFs in plants. Plant Biotechnol J 18, 2158–2160.

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

200 Brunet MA, Lucier JF, Levesque M, Leblanc S, Jacques JF, Al-Saedi HRH, Guillot N, Grenier F, Avino M,
Fournier I et al. (2021) OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Res* **49**, D380–D388.

201 Wethmar K, Barbosa-Silva A, Andrade-Navarro MA & Leutz A (2014) UORFdb - A comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Res* **42**, D60–D67.

202 Wan J & Qian SB (2014) TISdb: a database for alternative translation initiation in mammalian cells. *Nucleic Acids Res* **42**, 845–850.