MINI REVIEW
published: 06 May 2021
doi: 10.3389/fgene.2021.651485

Minireview: Novel Micropeptide Discovery by Proteomics and Deep Sequencing Methods

Ravi Tharakan[1]* *and Akira Sawa*[2,3]

[1] National Institute on Aging, National Institutes of Health, Baltimore, MD, United States, [2] Departments of Psychiatry, Neuroscience, Biomedical Engineering, and Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, United States, [3] Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

A novel class of small proteins, called micropeptides, has recently been discovered in the genome. These proteins, which have been found to play important roles in many physiological and cellular systems, are shorter than 100 amino acids and were overlooked during previous genome annotations. Discovery and characterization of more micropeptides has been ongoing, often using -omics methods such as proteomics, RNA sequencing, and ribosome profiling. In this review, we survey the recent advances in the micropeptides field and describe the methodological and conceptual challenges facing future micropeptide endeavors.

Keywords: micropeptides, miniproteins, proteogenomics, sORF, ribosome profiling, proteomics, genomics, RNA sequencing

OPEN ACCESS

Edited by:
Liangliang Sun,
Michigan State University,
United States

Reviewed by:
Yanbao Yu,
J. Craig Venter Institute (Rockville),
United States
Hongqiang Qin,
Dalian Institute of Chemical Physics,
Chinese Academy of Sciences, China

****Correspondence:***
Ravi Tharakan
ravi.tharakan@nih.gov

Specialty section:
This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: *09 January 2021*
Accepted: *22 March 2021*
Published: *06 May 2021*

Citation:
Tharakan R and Sawa A (2021)
Minireview: Novel Micropeptide
Discovery by Proteomics and Deep
Sequencing Methods.
Front. Genet. 12:651485.
doi: 10.3389/fgene.2021.651485

INTRODUCTION

The sequencing and publication of complete genomic sequences of many organisms have aided the medical sciences greatly, allowing advances in both human genetics and the biology of human disease, as well as a greater understanding of the biology of human pathogens (Firth and Lipkin, 2013). In particular, the human genome sequence has advanced human disease genetics by allowing genome-wide association studies (Hofker et al., 2014) (GWAS), and knowledge of gene sequences and their chromosomal loci has produced a deeper understanding of the biology of all organisms whose genomes have been sequenced, including human pathogens such as viruses (Lu et al., 2020). Crucial to all of these efforts is genome annotation, which uses genomic, genetic, epigenetic, and other information to find loci in the genome which code functional genes (Salzberg, 2019). While genome annotation has been performed alongside efforts to sequence genomes, the continuing pace of novel gene discovery, aided by advancing technology in molecular biology and biochemistry, suggests that annotation is often incomplete.

During the early stages of genome annotation, when genomes such as the human genome were first sequenced, a lower limit was placed on the length of an open reading frame (ORF) that could be considered a possible gene (Dujon et al., 1994). These limits were set by modeling a biochemically equivalent random genome and determining the ORF length distribution over that random genome, thus producing a length distribution of "random" ORFs. In the case of the human genome, in order to exclude such random ORFs, the minimum ORF length was set to be 100 codons by reasoning about the size distribution of random ORFs (Dujon et al., 1994). However, while a high proportion of ORFs below these limits may be spurious, there are also a substantial number which are real genes, and these are missed by such a filtering step (Basrai et al., 1997). Furthermore, while gene discovery can be accomplished by means other than genome annotation, it appears that these original annotation decisions continue to be reflected in our current understanding of genomes, since some human protein databases contain a disproportionately low number of genes annotated below the 100 codon cutoff (Frith et al., 2006).

Recently, evidence has emerged that many of these short (<100 codons) open reading frames may indeed be protein-coding genes, whose gene products have been named "micropeptides" (Couso, 2015). Evidence for functionality of these short peptides has come from several sources, including bioinformatics, through novel conservation analyses (Crowe et al., 2006), and biochemical approaches, such as expressed sequence tag experiments (Frith et al., 2006), deep sequencing based experiments such as RNA sequencing (Kageyama et al., 2011) and ribosome profiling (Ingolia et al., 2011), as well as proteomics (Slavoff et al., 2013, 2014; Khitun and Slavoff, 2019; Cao et al., 2020). These micropeptides have been found in studies of many model organisms, suggesting that micropeptide genes indeed exist throughout all genomes. Nevertheless, it has been controversial how many micropeptide genes there are, and general estimates have varied by orders of magnitude (Andrews and Rothnagel, 2014). In particular, since all techniques have biases and false positives, there has been continued debate on the extent to which evidence for micropeptides is artifact of the techniques used for discovery (Guttman et al., 2013; Ingolia et al., 2014). Furthermore, the mechanisms by which micropeptides perform their functions in the cell is often difficult to determine; in particular, it is still not clear whether micropeptides, as a class, share a general cellular role, or whether they have diverse functions in the same way large well-annotated proteins do (Couso and Patraquim, 2017). In this review, we will examine some of the recent developments in the field of micropeptide discovery, with a special emphasis on mass spectrometry-based approaches to the study of micropeptides.

MICROPEPTIDE DISCOVERY IN LOWER-ORDER ORGANISMS

In prokaryotes, annotation of sORF-encoded micropeptides has been of increasing interest. In the bacterium *Mycoplasma pneumoniae*, micropeptides translated from ncRNA were discovered by mass spectrometry, defining the total complement of micropeptides in that organism's genome at 67, approximately 5% of all coding genes (Lluch-Senar et al., 2015). A transposon-based essentiality screen found that 53% of these micropeptides were essential for the bacterium's growth, indicating that many micropeptides are not only functional but essential for bacterial growth. Since *M. pneumoniae* has a relatively small genome of 816 kb which is likely to be well annotated, such an essentiality study also suggests that many micropeptides are proteins essential for a minimal organism and thus essential for life (Lluch-Senar et al., 2015). Finally, in the widely used model bacterium *Escherichia coli*, one study discovered 44 micropeptides, many of which also perform basic functions in the cell (Hemm et al., 2008), while a more recent study found 36, using an epitope tagging method (VanOrsdel et al., 2018).

Micropeptide discovery has also continued apace in eukaryotic model systems. A study in *Saccharomyces cerevisiae* by ribosome profiling found many cases of translation of micropeptides, which appear to serve some purpose during the meiotic process (Brar et al., 2012). During meiosis, translation of some 9,989 unannotated ORFs was found, and these novel genes appeared to have their translation increased in a regulated fashion during meiosis, indicating once again that they are functional, although the functions of these many putative novel genes have not been defined yet (Brar et al., 2012). A bioinformatics screen of the yeast genome searching for micropeptides similarly yielded 184 yeast sORFs conserved across many species, suggesting that they may be functional (Kastenmayer et al., 2006). These novel genes were then validated as to function by development of deletion strains for 140 of them, of which nine gave clearly observable phenotypes. The latter study provides a model of how novel micropeptide genes can be validated as to function; furthermore, the large increase in the number of micropeptides observed between the two studies demonstrates the role of advancing technology in the rapid advance of micropeptide annotation.

HUMAN-RELEVANT MICROPEPTIDE DISCOVERY IN MODEL SYSTEMS

Studies in higher-order animals have also found several significant novel micropeptides. Very relevantly for human health, a screen in *Danio rerio* discovered a novel micropeptide translated from a transcript annotated as a long non-coding transcript (lncRNA), naming the micropeptide *Toddler* (Pauli et al., 2014). This micropeptide was discovered by a series of screens, in which developing zebrafish were subjected to RNA sequencing to discover novel lncRNAs (Chew et al., 2013), and the same material was used for a ribosome profiling study (Pauli et al., 2012). The latter ribosome profiling study showed that 399 of the putative lncRNAs were indeed translated, of which one produced a micropeptide confirmed by mass spectrometry. GFP tagging of the novel micropeptide to find tissue distribution and a knockout animal model showed that the micropeptide appears to function as an extracellular secreted ligand for the Apelin receptor, essential for cell migration during embryonic development (Pauli et al., 2014). Importantly, the same micropeptide was found to be essential for regulation of the human cardiovascular system, again functioning as a ligand for the Apelin receptor (Yang et al., 2017). This micropeptide has been found to be involved in pre-eclampsia pathology in a mouse model (Ho et al., 2017). As GPCRs are very frequently targeted by drugs, the discovery of this novel ligand could eventually produce novel therapies, as has been recently proposed (Kuba et al., 2019). Indeed, while no other micropeptides have so far been found to be GPCR ligands, the *Toddler* peptide shows that it may be productive to screen sets of detected micropeptides for GPCR activity. For example, hits from mass spectrometry searches could be tested for activity against GPCRs, as has been done for peptide libraries before (Zhang et al., 2015; Yaginuma et al., 2019). Furthermore, this series of studies thus shows that discoveries of micropeptides in animals cannot only be relevant for human biology, they can also almost immediately lead to novel therapies.

In another animal model system, *Drosophila melanogaster*, progress in annotation of micropeptides has been even more rapid. Several screens of this system have produced a number of novel micropeptides which regulate the cardiovascular system (Magny et al., 2013), developmental regulation through proteasomal function (Zanet et al., 2015), and control of RNA polymerase (Hanyu-Nakamura et al., 2008). In the cardiovascular system, the micropeptides Sarcolamban A and B were initially found in a search for functional short ORFs (sORFs) among the set of putative ncRNAs, a screen which found two possible micropeptides of 28 and 29 amino acids long on a single transcript (Magny et al., 2013). *In vivo* translation and GFP tagging confirmed translation localized to the sarcoplasmic reticulum, and a null mutant showed a cardiac arrhythmia phenotype with dysregulated calcium transients, suggesting a novel micropeptide involved in regulation of the SERCA pump. Strikingly, the micropeptides were found to be highly conserved throughout evolution, including in humans (Magny et al., 2013). In *Drosophila* development, the *mlpt/tal/pri* gene, discovered independently by several groups, contains four sORFs which appear to code for peptides (Zanet et al., 2016). Null mutants of this gene show dramatically dysregulated development, apparently due to disruption of a transcription factor (Zanet et al., 2015). In this case, the micropeptides appear in some fashion to regulate the ubiquitination and proteasomal degradation of the transcription factor. These examples, unlike the above *Toddler* example, show cases of intracellular micropeptides regulating protein-protein interactions. Finally, a genome-wide study by ribosome profiling has increased the number of candidate micropeptides in the *Drosophila* genome to ∼285, although most of these are of unknown function (Aspden et al., 2014; Zanet et al., 2016).

In mouse and human, finally, there has also been progress in identifying biologically relevant micropeptides, including the case of the *Toddler* peptide described above. Next discovered was the *Myoregulin* micropeptide, which followed on from the discovery of *Sarcolamban* in *Drosophila* (Anderson et al., 2015). Like *Sarcolamban*, the *Myoregulin* micropeptide regulates activity of the SERCA pump and thereby calcium transients, and similarly, it was found by bioinformatically screening newly discovered non-coding RNA transcripts for short ORFs which could encode putative micropeptides. The peptide was found by labeling experiments to interact with the SERCA pump and have some homology to *Sarcolamban*, as well as to the human peptides *Phospholamban* and *Sarcolipin* (Anderson et al., 2015). Further extending work on the *Myoregulin* micropeptide, three more micropeptide members of the same family were found by screening the peptide-binding motif of the family, discovering the micropeptides *Dworf* (Nelson et al., 2016), *Endoregulin*, and *Another-regulin* (Anderson et al., 2016). These peptides all regulate SERCA, and their tissue distribution is substantially different, including tissues beyond muscle (Anderson et al., 2016), indicating that micropeptide-SERCA interactions are a widespread and perhaps fundamental system for regulating calcium transients in mouse and human.

Besides bioinformatics screens of ORFs in ncRNAs, the human and mouse systems have also been substantially probed for micropeptides using ribosome profiling and mass spectrometry. A ribosome profiling study in mouse embryonic stem cells found that many short ORFs are translated in these cells, and the majority of lncRNAs have translation levels comparable with annotated protein-coding genes, suggesting that many lncRNAs in fact encode micropeptides (Ingolia et al., 2011). The same study also observed widespread translation in upstream ORFs (uORFs) of known coding genes, which translation was downregulated during embryoid body formation. However, these latter observations cannot necessarily be taken to indicate functional micropeptides, since ribosome presence on a transcript means only that the ribosome is bound, but not that a functional peptide is produced. For example, ribosome presence on upstream ORFs has canonically been interpreted as a regulatory process by which the ribosome's access to the main coding ORF is blocked, thus impeding translation of the main ORF (Hinnebusch, 2014). These uORF peptides, however, have been recently found to be presented on MHC molecules, suggesting a potential function in human immunity (Starck et al., 2016). Mass spectrometry-based discovery of novel micropeptides has been successfully performed by peptidomics studies on human cell lines. In this approach, small peptides are purified by size exclusion from the larger proteome, and these small peptides are then analyzed by mass spectrometry (Slavoff et al., 2013). This approach was successfully applied to find 90 novel micropeptides, including NoBody (D'Lima et al., 2017), a small peptide that appears to downregulate mRNA processing granule formation and participate in RNA decapping; and MRI-2, which appears to participate in DNA repair (Slavoff et al., 2014).

Finally, a very large-scale search for micropeptides in human cells used a combination of ribosome profiling and mass spectrometry to generate potentially translated micropeptides and CRISPR knockouts to validate the micropeptides as functional (Chen et al., 2020), resulting in some 570 novel micropeptides. In this study, ribosome profiling was performed on several cell lines in order to find novel coding regions, resulting in 3,455 novel coding regions, as well as 2,466 extensions of known coding regions. Very few of the peptides encoded by these novel regions were detected by mass spectrometry, but the authors then developed a CRISPR-based screen to validate the peptides functionally, constructing an sgRNA library targeting 2,353 of the putative novel regions, and testing for the effect of the CRISPR knockout on cell growth. Several hundred of the micropeptide knockouts showed phenotypes, suggesting functional micropeptides. CRISPR knockouts disrupt the DNA, and so it is not possible to distinguish between a functional non-coding RNA and a functional micropeptide, because both the RNA and protein level will be disrupted. A solution for this is to use CRISPR to mutate only the start codon of the micropeptide, which will allow a non-coding RNA to function but block expression of the micropeptide (Chen et al., 2020).

Since ribosome profiling and mass spectrometry have so far been the most successful methods by which mammalian micropeptides have been discovered, we have recently proposed combining the two methods (Tharakan et al., 2020), in a so-called proteogenomics approach (Nesvizhskii, 2014). In these studies, a combination of RNA sequencing and proteomics data are used, where the RNA sequencing database is used to assemble a transcriptome, which is then translated in six frames to yield a database of all possible proteins and peptides. Several groups have attempted this approach in human samples, most notably in the TCGA project (Cancer Genome Atlas Network, 2012; Wang and Zhang, 2013) and other projects (Wang et al., 2019). The central problem with these databases is their extremely large size, usually on the order of several million candidate proteins. Databases with a very large size will decrease the sensitivity of the search. To solve this problem, we propose using ribosome profiling data to filter the candidate protein list before the final analysis. This proteogenomics approach, of combining RNA sequencing, ribosome profiling, and mass spectrometry (Tharakan et al., 2020), can also be used to identify so-called tumor "neoantigens," which are mutated proteins produced by tumors (Schumacher and Schreiber, 2015). Once again, exome sequencing or RNAseq of tumors produces databases that are too large, which can be reduced by filtering through ribosome profiling.

Several micropeptides involved in the regulation of human metabolism have been found. Using mass spectrometry, the micropeptide SPAR, small regulatory polypeptide of amino acid response, was discovered by a proteomic analysis of a human cell line (Matsumoto et al., 2017), and found to regulate mTOR function. Similarly, the micropeptides mitoregulin (Stein et al., 2018) and MOXI (Makarewich et al., 2018) which both regulate mitochondrial function, were found through bioinformatics methods coupled with experimental validation. Functional micropeptides have also been found in the human mitochondrial genome. The *Humanin* peptide, encoded by a short ORF from the mitochondrial DNA, was discovered years ago by a purely functional cDNA library screen for genes inhibiting apoptosis (Hashimoto et al., 2001). More recently, the MOTS-c peptide has been discovered in an sORF on the 12s rRNA gene in the mitochondrial genome by a bioinformatics screen of sORFs in mtDNA, showing a conserved 51nt ORF with a strong Kozak context (Lee et al., 2015). Treatment of HEK293 cells with a synthetic peptide substantially regulated gene expression of enzymes involved in cellular metabolism, and treatment of mice fed a high-fat diet prevented obesity, suggesting that the peptide is an extracellular signaling molecule.

VIRAL GENOME ANNOTATION BY MICROPEPTIDE ANALYSIS

Finally, searches of viral genomes for micropeptides have also yielded some interesting results. Due to the small size of many of their genomes, historically, viral genomes were not usually annotated with explicit lower limits on the lengths of open reading frames, although the vaccinia virus, which has a large genome, had a lower limit of 65 codons for its original annotation (Goebel et al., 1990). Even without explicit minima during annotation, however, many viral micropeptides were simply overlooked because of systematic assumptions of how large genes should be, as genomes were annotated in an *ad hoc* fashion (Ratner et al., 1985). Micropeptides in viral genomes have been discovered in a similarly *ad hoc* manner, with several micropeptides found in influenza virus, human immunodeficiency virus, papillomavirus, poxviruses, and paramyxoviruses (DiMaio, 2014). These micropeptides are often dominated by a single alpha-helical transmembrane domain, which allows them to be inserted into lipid bilayers, in which context they can interact with and regulate host cell proteins (i.e., HIV-1 Vpu), form ion-selective pores (i.e., influenza M2), allow binding and entry into host cells (i.e., poxvirus O3L), or perform other functions crucial for the viral life cycle (DiMaio, 2014). A genome-wide screen of human cytomegalovirus by ribosome profiling and mass spectrometry found ∼484 novel ORFs shorter than 80 codons, including 245 shorter than 20 codons which were actively translated by the ribosome (Stern-Ginossar et al., 2012). Similarly, a recent study has applied ribosome profiling methods to the novel coronavirus SARS-CoV-2 and was able to find evidence of 23 novel proteins, beyond the 37 proteins already annotated for the virus (Finkel et al., 2020). These micropeptides must be validated by functional studies, but the high number identified here shows that there may be a substantial number of micropeptides to be found in all viruses.

HOW MANY NOVEL MICROPEPTIDES REMAIN TO BE DISCOVERED?

It has been widely shown by a variety of methods, therefore, that despite underrepresentation in genome annotations, sORFs encode functional micropeptides in nearly all genomes studied, thus suggesting that there may be many micropeptides produced both from the human genome and from human pathogens that may be of relevance to human health. However, how many micropeptides there may be in genomes is still controversial. Many types of genome-wide approaches have given a wide range for how many coding micropeptides there are in the human genome, from tens of thousands to a few dozen (Andrews and Rothnagel, 2014). These problems reflect the issues identifying sORFs which caused them to be overlooked. In particular, it is difficult to detect evolutionary conservation of these ORFs, because their small size disrupts the statistical assumptions of homology detection algorithms such as BLAST or PhyloCSF (Couso, 2015). Thus, genome-wide searches for conservation of sORFs often incorrectly show them to be unconserved. Secondly, very short peptides were thought not to have stable secondary structure, and thus, if biological function is assumed to be entirely dependent on the structure of the protein, short peptides could be assumed not to have a function (Ingolia et al., 2014; Wright and Dyson, 2015). This lack of secondary structure can also be a problem for genome-wide description of sORF-encoded micropeptides, for example by mass spectrometry, since peptides without stable conformations may be rapidly degraded when cells are lysed for extraction (Hackett et al., 1986).

If these are reasons for micropeptide numbers to be underestimated historically, questions of artifact in newer methods may cause the number of functional micropeptides to be overestimated. In particular, each genome-wide method for micropeptide discovery can be prone to false positives for various reasons. Central methods which have been used to discover micropeptides genome-wide are ribosome profiling, mass spectrometry, RNA sequencing, and direct searches for conservation. In ribosome profiling, polysomes are extracted from cells, then treated with RNase to destroy RNA which is unprotected by ribosomes (Ingolia et al., 2009). The ribosome footprints are then sequenced, and these are used to map the positions of the ribosomes. However, failure of RNase to digest a given section of a transcript may be due to factors besides ribosome content; in particular, RNA secondary structure may also block digestion, and RNA may also be underdigested due to sub-optimal reaction conditions. RNA-binding proteins may also block digestion. Furthermore, the mere presence of ribosomes in a given ORF does not necessarily demonstrate translation of that ORF. Although certain features of ribosome profiling datasets, such as codon periodicity, can be used to determine “genuine” translation, this question continues to be controversial and as yet there is no definitive metric by which ORFs can be determined to be translated (Guttman et al., 2013; Ingolia et al., 2014; Calviello et al., 2016).

In RNA sequencing, the length of the transcript is sequenced, and the high sensitivity of this method has allowed many novel transcripts to be discovered. In particular, RNA sequencing allows the study of a large set of RNAs called long non-coding RNAs (Sun et al., 2013; Ulitsky and Bartel, 2013; Hart and Goff, 2016). These RNAs were originally believed to be non-coding because they contain no long ORF, but subsequent conservation analyses of short ORFs have revealed many of these lncRNAs to indeed be micropeptide encoding (Pauli et al., 2014; Anderson et al., 2015). Thus, lncRNAs, which continue to be discovered by improved deep sequencing methods, may therefore be a large source of novel micropeptides. However, for the reasons mentioned above, most mainstream conservation analyses cannot be well applied to sORFs, and thus, screening for conserved sORFs in lncRNAs to search for micropeptides likely underestimates the total number of sORFs. In an attempt to address this problem, a novel conservation detection method was developed, and some 2,000 novel ORFs were found in the genome (Mackowiak et al., 2015). However, only a small percentage of these could be confirmed by mass spectrometry or ribosomal profiling data. Thus, the number of micropeptides produced by newly discovered lncRNAs continues to be an open question. For these kinds of reasons, RNA sequencing experiments have not been able to provide a clear picture of how many micropeptides there may be in genomes.

Mass spectrometry for discovery of micropeptides also seems to run afoul of the underestimation problem. Firstly, as mentioned above, mass spectrometry attempts to directly detect micropeptides, and if it is true that micropeptides are rapidly degraded by proteolytic enzymes after cell lysis, mass spectrometry may have problems with the speed of degradation of micropeptides. Secondly, lncRNAs are known to be very low abundance in any given tissue (Cabili et al., 2015). Furthermore, very short peptides, shorter than 5 amino acids will not be detected by mass spectrometry, and very long peptides are also difficult to detect, although the latter problem may be solved by performing a trypsin digestion. There may also be biochemical issues with particular micropeptide sequences; for example, if the peptide contains no basic amino acids, it becomes difficult to detect by mass spectrometry. Thus, detecting micropeptides translated from lncRNAs by mass spectrometry may have problems with sufficient sensitivity. Indeed, across the literature, one finds a mismatch between deep-sequencing experiments, such as ribosome profiling, and mass spectrometry, with deep-sequencing based results generally producing much higher numbers of micropeptides detected than mass spectrometry. There are two possible explanations for this; first, mass spectrometry may underestimate, or deep sequencing may overestimate, the number of micropeptides in the genome, or both; second, ribosomes may bind to many ORFs promiscuously, but these peptides are either not translated or quickly degraded and non-functional. The latter implies that there is no clear evidence that there is any substantial number of micropeptides in the genome, but the former implies that there are an undetermined number of micropeptides remaining in the genome to discover.

CONCLUSION

Micropeptide research has been a growing field for the past several years, and advances in technology have made it possible to investigate the nature of this so-called hidden genome within the known genome. However, much work remains to be done, both in functionally characterizing the many micropeptides that have already been found through various methods, and in searching for micropeptides in new sample types. Technologies will also need to continue to be improved in order to find all micropeptides that exist. Recent progress in this field, however, raises the possibility of an entirely new understanding of genome function.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was supported by the Intramural Research Program of the National Institutes of Health (to RT). This work was also supported by MH-094268 (to AS), MH-105660 (to AS), and MH-107730 (to AS); foundation grants from Stanley (to AS) and RUSK/S-R (to AS).

REFERENCES

Anderson, D. M., Anderson, K. M., Chang, C. L., Makarewich, C. A., Nelson, B. R., McAnally, J. R., et al. (2015). A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* 160, 595–606. doi: 10.1016/j.cell.2015.01.009

Anderson, D. M., Makarewich, C. A., Anderson, K. M., Shelton, J. M., Bezprozvannaya, S., Bassel-Duby, R., et al. (2016). Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. *Sci. Signal.* 9:ra119. doi: 10.1126/scisignal.aaj1460

Andrews, S. J., and Rothnagel, J. A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* 15, 193–204. doi: 10.1038/nrg3520

Aspden, J. L., Eyre-Walker, Y. C., Phillips, R. J., Amin, U., Mumtaz, M. A., Brocard, M., et al. (2014). Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *Elife* 3:e03528.

Basrai, M. A., Hieter, P., and Boeke, J. D. (1997). Small open reading frames: beautiful needles in the haystack. *Genome Res.* 7, 768–771. doi: 10.1101/gr.7.8.768

Brar, G. A., Yassour, M., Friedman, N., Regev, A., Ingolia, N. T., and Weissman, J. S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* 335, 552–557. doi: 10.1126/science.1215110

Cabili, M. N., Dunagin, M. C., McClanahan, P. D., Biaesch, A., Padovan-Merhar, O., Regev, A., et al. (2015). Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* 16:20.

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., et al. (2016). Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* 13, 165–170. doi: 10.1038/nmeth.3688

Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330–337. doi: 10.1038/nature11252

Cao, X., Khitun, A., Na, Z., Dumitrescu, D. G., Kubica, M., Olatunji, E., et al. (2020). Comparative proteomic profiling of unannotated microproteins and alternative proteins in human cell lines. *J. Proteome Res.* 19, 3418–3426. doi: 10.1021/acs.jproteome.0c00254

Chen, J., Brunner, A. D., Cogan, J. Z., Nuñez, J. K., Fields, A. P., Adamson, B., et al. (2020). Pervasive functional translation of noncanonical human open reading frames. *Science* 367, 1140–1146. doi: 10.1126/science.aay0262

Chew, G. L., Pauli, A., Rinn, J. L., Regev, A., Schier, A. F., and Valen, E. (2013). Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. *Development* 140, 2828–2834. doi: 10.1242/dev.098343

Couso, J. P. (2015). Finding smORFs: getting closer. *Genome Biol.* 16:189.

Couso, J. P., and Patraquim, P. (2017). Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.* 18, 575–589. doi: 10.1038/nrm.2017.58

Crowe, M. L., Wang, X. Q., and Rothnagel, J. A. (2006). Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. *BMC Genomics* 7:16. doi: 10.1186/1471-2164-7-16

D'Lima, N. G., Ma, J., Winkler, L., Chu, Q., Loh, K. H., Corpuz, E. O., et al. (2017). A human microprotein that interacts with the mRNA decapping complex. *Nat. Chem. Biol.* 13, 174–180. doi: 10.1038/nchembio.2249

DiMaio, D. (2014). Viral miniproteins. *Annu. Rev. Microbiol.* 68, 21–43. doi: 10.1146/annurev-micro-091313-103727

Dujon, B., Alexandraki, D., Andre, B., Ansorge, W., Baladron, V., Ballesta, J. P., et al. (1994). Complete DNA sequence of yeast chromosome XI. *Nature* 369, 371–378.

Finkel, Y., Mizrahi, O., Nachshon, A., Weingarten-Gabbay, S., Morgenstern, D., Yahalom-Ronen, Y., et al. (2020). The coding capacity of SARS-CoV-2. *Nature* 589, 125–130.

Firth, C., and Lipkin, W. I. (2013). The genomics of emerging pathogens. *Annu. Rev. Genomics Hum. Genet.* 14, 281–300. doi: 10.1146/annurev-genom-091212-153446

Frith, M. C., Forrest, A. R., Nourbakhsh, E., Pang, K. C., Kai, C., Kawai, J., et al. (2006). The abundance of short proteins in the mammalian proteome. *PLoS Genet.* 2:e52. doi: 10.1371/journal.pgen.0020052

Goebel, S. J., Johnson, G. P., Perkus, M. E., Davis, S. W., Winslow, J. P., and Paoletti, E. (1990). The complete DNA sequence of vaccinia virus. *Virology* 179, 247–266. doi: 10.1016/0042-6822(90)90294-2

Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., and Lander, E. S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154, 240–251. doi: 10.1016/j.cell.2013.06.009

Hackett, P. B., Petersen, R. B., Hensel, C. H., Albericio, F., Gunderson, S. I., Palmenberg, A. C., et al. (1986). Synthesis in vitro of a seven amino acid peptide encoded in the leader RNA of Rous sarcoma virus. *J. Mol. Biol.* 190, 45–57. doi: 10.1016/0022-2836(86)90074-4

Hanyu-Nakamura, K., Sonobe-Nojima, H., Tanigawa, A., Lasko, P., and Nakamura, A. (2008). Drosophila Pgc protein inhibits P-TEFb recruitment to chromatin in primordial germ cells. *Nature* 451, 730–733. doi: 10.1038/nature06498

Hart, R. P., and Goff, L. A. (2016). Long noncoding RNAs: central to nervous system development. *Int. J. Dev. Neurosci.* 55, 109–116. doi: 10.1016/j.ijdevneu.2016.06.001

Hashimoto, Y., Niikura, T., Tajima, H., Yasukawa, T., Sudo, H., Ito, Y., et al. (2001). A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Abeta. *Proc. Natl. Acad. Sci. U.S.A.* 98, 6336–6341. doi: 10.1073/pnas.101133498

Hemm, M. R., Paul, B. J., Schneider, T. D., Storz, G., and Rudd, K. E. (2008). Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol. Microbiol.* 70, 1487–1501. doi: 10.1111/j.1365-2958.2008.06495.x

Hinnebusch, A. G. (2014). The scanning mechanism of eukaryotic translation initiation. *Annu. Rev. Biochem.* 83, 779–812. doi: 10.1146/annurev-biochem-060713-035802

Ho, L., van Dijk, M., Chye, S. T. J., Messerschmidt, D. M., Chng, S. C., Ong, S., et al. (2017). ELABELA deficiency promotes preeclampsia and cardiovascular malformations in mice. *Science* 357, 707–713. doi: 10.1126/science.aam6607

Hofker, M. H., Fu, J., and Wijmenga, C. (2014). The genome revolution and its role in understanding complex diseases. *Biochim. Biophys. Acta* 1842, 1889–1895. doi: 10.1016/j.bbadis.2014.05.002

Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J., Jackson, S. E., et al. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* 8, 1365–1379. doi: 10.1016/j.celrep.2014.07.045

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218–223. doi: 10.1126/science.1168978

Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147, 789–802. doi: 10.1016/j.cell.2011.10.002

Kageyama, Y., Kondo, T., and Hashimoto, Y. (2011). Coding vs non-coding: translatability of short ORFs found in putative non-coding transcripts. *Biochimie* 93, 1981–1986. doi: 10.1016/j.biochi.2011.06.024

Kastenmayer, J. P., Ni, L., Chu, A., Kitchen, L. E., Au, W. C., Yang, H., et al. (2006). Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. *Genome Res.* 16, 365–373. doi: 10.1101/gr.4355406

Khitun, A., and Slavoff, S. A. (2019). Proteomic detection and validation of translated small open reading frames. *Curr. Protoc. Chem. Biol.* 11:e77.

Kuba, K., Sato, T., Imai, Y., and Yamaguchi, T. (2019). Apelin and Elabela/Toddler; double ligands for APJ/Apelin receptor in heart development, physiology, and pathology. *Peptides* 111, 62–70. doi: 10.1016/j.peptides.2018.04.011

Lee, C., Zeng, J., Drew, B. G., Sallam, T., Martin-Montalvo, A., Wan, J., et al. (2015). The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. *Cell Metab.* 21, 443–454. doi: 10.1016/j.cmet.2015.02.009

Lluch-Senar, M., Delgado, J., Chen, W. H., Lloréns-Rico, V., O'Reilly, F. J., Wodke, J. A., et al. (2015). Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Mol. Syst. Biol.* 11:780. doi: 10.15252/msb.20145558

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 395, 565–574. doi: 10.1016/s0140-6736(20)30251-8

Mackowiak, S. D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., et al. (2015). Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.* 16:179.

Magny, E. G., Pueyo, J. I., Pearl, F. M., Cespedes, M. A., Niven, J. E., Bishop, S. A., et al. (2013). Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* 341, 1116–1120. doi: 10.1126/science.1238802

Makarewich, C. A., Baskin, K. K., Munir, A. Z., Bezprozvannaya, S., Sharma, G., Khemtong, C., et al. (2018). MOXI is a mitochondrial micropeptide that enhances fatty acid beta-oxidation. *Cell Rep.* 23, 3701–3709. doi: 10.1016/j.celrep.2018.05.058

Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E., et al. (2017). mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* 541, 228–232. doi: 10.1038/nature21034

Nelson, B. R., Makarewich, C. A., Anderson, D. M., Winders, B. R., Troupes, C. D., Wu, F., et al. (2016). A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* 351, 271–275. doi: 10.1126/science.aad4076

Nesvizhskii, A. I. (2014). Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* 11, 1114–1125. doi: 10.1038/nmeth.3144

Pauli, A., Norris, M. L., Valen, E., Chew, G. L., Gagnon, J. A., Zimmerman, S., et al. (2014). Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science* 343:1248636. doi: 10.1126/science.1248636

Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., et al. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.* 22, 577–591. doi: 10.1101/gr.133009.111

Ratner, L., Haseltine, W., Patarca, R., Livak, K. J., Starcich, B., Josephs, S. F., et al. (1985). Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature* 313, 277–284.

Salzberg, S. L. (2019). Next-generation genome annotation: we still struggle to get it right. *Genome Biol.* 20:92.

Schumacher, T. N., and Schreiber, R. D. (2015). Neoantigens in cancer immunotherapy. *Science* 348, 69–74.

Slavoff, S. A., Heo, J., Budnik, B. A., Hanakahi, L. A., and Saghatelian, A. (2014). A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *J. Biol. Chem.* 289, 10950–10957. doi: 10.1074/jbc.c113.533968

Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., et al. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* 9, 59–64. doi: 10.1038/nchembio.1120

Starck, S. R., Tsai, J. C., Chen, K., Shodiya, M., Wang, L., Yahiro, K., et al. (2016). Translation from the 5′ untranslated region shapes the integrated stress response. *Science* 351:aad3867. doi: 10.1126/science.aad3867

Stein, C. S., Jadiya, P., Zhang, X., McLendon, J. M., Abouassaly, G. M., Witmer, N. H., et al. (2018). Mitoregulin: a lncRNA-encoded microprotein that supports mitochondrial supercomplexes and respiratory efficiency. *Cell Rep.* 23, 3710–3720.e8.

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V. T., Hein, M. Y., Huang, S. X., et al. (2012). Decoding human cytomegalovirus. *Science* 338, 1088–1093.

Sun, L., Goff, L. A., Trapnell, C., Alexander, R., Lo, K. A., Hacisuleyman, E., et al. (2013). Long noncoding RNAs regulate adipogenesis. *Proc. Natl. Acad. Sci. U.S.A.* 110, 3387–3392.

Tharakan, R., Kreimer, S., Ubaida-Mohien, C., Lavoie, J., Olexiouk, V., Menschaert, G., et al. (2020). A methodology for discovering novel brain-relevant peptides: combination of ribosome profiling and peptidomics. *Neurosci. Res.* 151, 31–37. doi: 10.1016/j.neures.2019.02.006

Ulitsky, I., and Bartel, D. P. (2013). lincRNAs: genomics, evolution, and mechanisms. *Cell* 154, 26–46. doi: 10.1016/j.cell.2013.06.020

VanOrsdel, C. E., Kelly, J. P., Burke, B. N., Lein, C. D., Oufiero, C. E., Sanchez, J. F., et al. (2018). Identifying new small proteins in *Escherichia coli*. *Proteomics* 18:e1700064.

Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D. P., et al. (2019). A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* 15:e8503.

Wang, X., and Zhang, B. (2013). customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* 29, 3235–3237. doi: 10.1093/bioinformatics/btt543

Wright, P. E., and Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 16, 18–29. doi: 10.1038/nrm3920

Yaginuma, K., Aoki, W., Miura, N., Ohtani, Y., Aburaya, S., Kogawa, M., et al. (2019). High-throughput identification of peptide agonists against GPCRs by co-culture of mammalian reporter cells and peptide-secreting yeast cells using droplet microfluidics. *Sci. Rep.* 9:10920.

Yang, P., Read, C., Kuc, R. E., Buonincontri, G., Southwood, M., Torella, R., et al. (2017). Elabela/Toddler is an endogenous agonist of the apelin APJ receptor in the adult cardiovascular system, and exogenous administration of the peptide compensates for the downregulation of its expression in pulmonary arterial hypertension. *Circulation* 135, 1160–1173. doi: 10.1161/circulationaha.116.023218

Zanet, J., Benrabah, E., Li, T., Pélissier-Monier, A., Chanut-Delalande, H., Ronsin, B., et al. (2015). Pri sORF peptides induce selective proteasome-mediated protein processing. *Science* 349, 1356–1358. doi: 10.1126/science.aac5677

Zanet, J., Chanut-Delalande, H., Plaza, S., and Payre, F. (2016). Small peptides as newcomers in the control of *Drosophila* development. *Curr. Top. Dev. Biol.* 117, 199–219. doi: 10.1016/bs.ctdb.2015.11.004

Zhang, H., Sturchler, E., Zhu, J., Nieto, A., Cistrone, P. A., Xie, J., et al. (2015). Autocrine selection of a GLP-1R G-protein biased agonist with potent antidiabetic effects. *Nat. Commun.* 6:8918.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.