HHS Public Access

Author manuscript

Trends Genet. Author manuscript; available in PMC 2026 February 01.

Published in final edited form as:

Trends Genet. 2025 February ; 41(2): 107–118. doi:10.1016/j.tig.2024.12.001.

Finding functional microproteins

Sikandar Azam[1,*], **Feiyue Yang**[1,*], **Xuebing Wu**[1,#]

[1]Department of Medicine and Department of Systems Biology, Columbia University Irving Medical Center, New York, NY, 10032, USA

Abstract

Genome-wide translational profiling has uncovered the synthesis of thousands of microproteins in human cells, a class of proteins traditionally overlooked in functional studies. Although an increasing number of these microproteins have been found to play critical roles in cellular processes, the functional relevance of the majority remains poorly understood. Studying these low-abundance, often unstable proteins is further complicated by the challenge of disentangling their functions from the noncoding roles of the associated DNA, RNA, and the act of translation. This review highlights recent advances in functional genomics that have led to the discovery of over one thousand human microproteins required for optimal cell proliferation. Ongoing technological innovations will continue to clarify the roles and mechanisms of microproteins in both normal physiology and disease, potentially opening new avenues for therapeutic exploration.

Keywords

Functional microprotein; CRISPR screen; noncanonical ORF; short ORF; lncRNA

Microproteins: hidden treasure of the proteome

Proteins are the workhorses of life, and they come in a wide range of sizes. Human proteins, for instance, vary from as small as 12 amino-acid (aa) to as large as 35,991 aa, with a median size of 431 aa (Figure 1). Historically, proteins shorter than 100 aa —often referred to as **microproteins** (see **Glossary**), micropeptides, or short/small open reading frame (**ORF**)-encoded peptides (SEPs)—have largely been overlooked in functional studies. Gene-finding algorithms often use a threshold of 100 aa to filter out numerous **short ORFs**, which may arise from random sequences and are therefore considered less likely to encode functional proteins [1]. Advances in DNA sequencing technologies have enabled deep surveys of the transcriptome, leading to the discovery of tens of thousands of transcripts initially predicted to be noncoding [2]. While these were mostly annotated as

[#]Correspondence: xuebing.wu@columbia.edu (X. W.).
[*]Equal contribution

Declaration of interests

X.W. is a member of the Scientific Advisory Board for Epitor Therapeutics.

long noncoding RNAs (**lncRNAs**), many contain short ORFs and were later found to be translated, as revealed by ribosome profiling (**Ribo-seq**) [3–5], a sensitive assay that maps translating ribosomes genome-wide [6]. The translation of short ORFs in lncRNAs is not surprising given that there is no known cellular mechanism to measure ORF length without first translating it. In addition to lncRNA ORFs, translation has also been found in several other types of noncanonical ORFs (**ncORFs**) (Figure 2), including upstream ORFs (**uORFs**) in 5' UTRs of canonical mRNAs [4,7] and **circORFs** in **circular RNAs (circRNAs)** [8–10].

The functional relevance of most microproteins remains unknown. Even among the most reliably curated lists of human proteins, such as MANE Select [11], 11 of the 13 proteins shorter than 30 aa are uncharacterized, despite their high sequence conservation. This lack of functional characterization underscores the presence of major challenges in investigating microproteins. However, this also presents an exciting opportunity for new discoveries, as many essential proteins—including seven ribosomal proteins—are technically microproteins (25–98 aa, e.g., the 25-aa RPL41). In recent years, large-scale functional screens, particularly CRISPR/Cas9-mediated loss-of-function screens, have facilitated the systematic discovery of microproteins that are likely crucial for optimal cell growth [7,12–16]. Detailed characterization of top candidates has unveiled novel mechanisms by which microproteins exert their functions [7,12–16]. In this review, we will discuss some of the major challenges in studying microproteins, summarize recent advancements enabled by genetic screens, and discuss future directions for research.

The expanding catalog of microproteins

A comprehensive catalog of microproteins is a crucial first step toward systematic functional studies aimed at understanding these molecules. Recent advances in omics approaches, including **mass spectrometry (MS)**, Ribo-seq, and associated bioinformatics tools, have dramatically increased the number of putative microproteins identified in cells (see reviews [17–19]). While MS is the gold standard for high-throughput protein detection, the direct detection of microproteins using MS is challenging due to their small size, low abundance, and often poor annotations. Even with customized, sensitive MS pipelines, each study often detected less than a hundred microproteins [7,20].

Instead, the presence of most microproteins has been inferred from Ribo-seq data, which quantifies the footprints of ribosomes synthesizing microproteins encoded by short ORFs. A recent meta-analysis of 669 human Ribo-seq samples identified 58,383 translated ncORFs, over 90% of which encode putative microproteins [21]. Most of these are uORFs in mRNA 5' UTRs (28,981) and short ORFs in lncRNAs (13,047), compared to 33,251 canonical ORFs detected in the same analysis (Figure 2). Approximately one-third of these ncORFs use non-AUG start codons, in contrast to less than 0.5% of canonical ORFs. Other similar meta-analyses of Ribo-seq data have identified even more translated short ORFs in humans and other species, exceeding 100,000 [22,23].

A key challenge in the field is to generate a standardized, high-quality reference annotation of microproteins for each species, starting with humans. A comprehensive reference set would greatly facilitate the functional characterization of microproteins. In the absence of

such a consensus, functional screens have targeted different sets of microproteins, resulting in limited overlap among identified functional microproteins, even in screens conducted within the same cell line (see discussion below). One major barrier to creating a reference microprotein annotation is the limited overlap in identified translated short ORFs across different computational tools [24]. For example, only about 2% of short ORFs were identified by all five commonly used tools when applied to the same high-resolution Ribo-seq dataset [25].

To address this, a community-led effort has compiled a consolidated catalog of 7,264 Ribo-seq supported ncORFs from seven previous studies [26]. The majority (95%) of these translated ncORFs encode microproteins. Subsequent studies have further characterized this reference set of microproteins, including their evolutionary origins [27], peptide evidence [28], and interacting partners [27]. Large-scale functional genetic screens have yet to be conducted for this set of ORFs.

Challenges in functional studies of microproteins

In addition to the sheer number of putative microproteins to be characterized and the lack of a widely accepted reference annotation, several unique challenges complicate the functional studies of microproteins. These include their small size, low abundance, short half-life, and low conservation. Another major issue is the need to unequivocally assign functions to microproteins themselves, rather than attributing observed effects to noncoding functions of the underlying RNA, DNA, or even the act of translation (Figure 3).

Too small to study

While the small size of microproteins may enable them to perform functions that larger proteins cannot, it also poses significant challenges for their study. For instance, microproteins produce fewer peptides detectable by MS. Computationally, accurately assessing sequence conservation and coding potential for short ORFs is more difficult. Experimentally, protein tags (e.g., GFP, HA, FLAG) are commonly used to study the localization, interactions, and dynamics of proteins. However, even a small tag like FLAG can significantly increase the size of a microprotein, potentially disrupting its normal function and regulation. Additionally, the small size of ORFs limits the number of CRISPR gRNAs available for loss-of-function studies, particularly for CRISPR systems with strong PAM dependencies. Even when efficient gRNAs induce frameshifting indels, the resulting frameshift may lengthen the protein rather than truncating it, increasing the risk of gain-of-function effects and false positive results.

Low abundance and rapid degradation

Most microproteins encoded by Ribo-seq-annotated short ORFs remain undetected in proteomics studies [28,29]. Several mechanisms suppress the production and accumulation of proteins derived outside of canonical coding sequences [30]. One such mechanism is BAG6-mediated noncoding translation mitigation, which targets hydrophobic C-termini commonly found in proteins encoded by ncORFs. This often leads to the proteasomal degradation of these proteins immediately after translation [31,32]. Nascent proteins

degraded through the BAG6/proteasome pathway serve as a significant source of **MHC-I/HLA-I antigen peptides** [33]. Additionally, some microproteins may be degraded by other cellular proteases or through the lysosomal pathway [21]. Reflecting the typically short half-life of most microproteins, only 66 (0.9%) of the 7,264 reference Ribo-seq ncORFs [26] are supported by high-quality peptide evidence from a meta-analysis of 3.5 billion MS spectra [28]. In contrast, 1,785 (24.6%) of these ncORFs were detected in a smaller HLA immunopeptidomics dataset (0.24 billion MS spectra), with 94.3% being HLA-I peptides [28]. These findings suggest that the vast majority of microproteins are rapidly degraded, likely via the BAG6/proteasome pathway, and subsequently processed into HLA-I peptides. Although a short half-life does not exclude the possibility of functional roles for microproteins, it does pose challenges for functional and mechanistic studies.

Lack of conservation

Most microprotein-encoding ORFs lack sequence conservation, limiting the utility of conservation as a reliable filter for functional microproteins. Additionally, when conservation is present, it may reflect selective pressure on the RNA or DNA level, rather than on the microprotein itself. Codon-level conservation scores, which examine protein-coding signatures such as synonymous codon substitutions, are frequently employed to identify conserved microproteins. However, many functional microproteins have evolved recently and thus exhibit little to no conservation [7]. For instance, none of the 44 known functional human microproteins that originated *de novo* from noncoding sequences [34] were predicted as coding by PhyloCSF [35], a widely used tool for assessing coding potential based on multiple sequence alignment.

Translation-independent functions

Detecting translation activity in an ORF does not necessarily indicate functional translation or a functional protein product, nor does it exclude translation-independent functions of the underlying RNA or DNA. For instance, a highly translated lncRNA was shown to be essential for pancreatic endocrine cell development, yet it retained its functionality even when the translated ORFs within it were deleted or frameshifted [36]. To definitively establish that translation is necessary for an ORF's function, it is essential to rule out any potential noncoding roles of the underlying RNA or DNA (Figure 3).

Microproteins are encoded by UTRs, lncRNAs, and circRNAs, all of which primarily function as RNA molecules, often independent of whether translation occurs. The majority of functional lncRNAs and circRNAs execute their functions without being translated, instead influencing gene expression and cellular processes through RNA-RNA, RNA-DNA, and RNA-protein interactions [37–39] (Figure 3A). Both 5' and 3' UTRs play crucial roles in regulating mRNA stability, localization, and the translation of the main ORF by recruiting RNA-binding proteins and regulatory RNAs (e.g., microRNAs) [40–42] (Figure 3B).

In addition to RNA-based functions, the genomic sequences encoding lncRNAs and UTRs frequently serve as functional DNA elements. For instance, many lncRNA loci act as enhancers, which loop to gene promoters to activate target gene transcription [43,44] (Figure 3A). The DNA sequences encoding 5' UTRs are located near the promoter and transcription

start site of their associated coding genes, allowing them to regulate key transcriptional processes such as nucleosome positioning, transcription initiation, and promoter-proximal pausing and release [45] (Figure 3B).

Disruption of the genomic locus of a short ORF can impact both DNA- and RNA-dependent functions, which must be carefully ruled out to attribute the observed phenotype specifically to ORF translation or to the microprotein. A common approach to exclude translation-independent effects is a phenotypic rescue experiment in which the start codon of the short ORF is mutated. A successful rescue with a translation-deficient construct would suggest that the observed effects are not due to the microprotein or translation but rather to other regulatory functions of the DNA or RNA.

Translation-dependent but protein-independent functions

The act of translation can serve a functional role even when the resulting protein is rapidly degraded. For instance, translation of uORFs in the 5’ UTR often inhibits the translation of the downstream main ORF on the same mRNA, typically independent of the specific sequence of the microprotein encoded by the uORF [46]. Additionally, translation of uORFs in mRNAs and short ORFs in lncRNAs has been shown to trigger RNA degradation via the **nonsense-mediated RNA decay (NMD) pathway** [47–51]. The rapid degradation of most microproteins [28] also supports the idea that the act of translation, rather than the microprotein itself, may be the primary functional event. Translation in other types of ncORFs found in mRNAs (Figure 2) may similarly regulate the host mRNA. To determine whether the microprotein is responsible for the observed phenotype, one can attempt a rescue of the knockout phenotype by exogenously expressing the microprotein without its native sequence context. For example, uORFs can be tested without the main ORF, or lncRNA ORFs without the remainder of the lncRNA sequence. The ORF can be further recoded using synonymous codons, enabling modifications to the underlying DNA and RNA sequence without affecting the amino acid sequence of the microprotein. A success rescue would support a role of the microprotein.

CRISPR screens for functional microproteins

Over the past five years, six CRISPR/Cas9-mediated knockout screens have been conducted to systematically assess the functional impact of thousands of microproteins in various human cell lines [7,12–16] (Table 1). These studies have identified more than 1,000 putative functional microproteins and provided valuable insights into their mechanisms of action. For a subset of top hits, carefully designed experiments have ruled out potential RNA, DNA, or translation-dependent noncoding functions. These studies also highlight certain limitations of current methodologies and suggest avenues for future improvements.

CRISPR/Cas9 knockout screens

CRISPR/Cas9-mediated knockout screens have become a widely used tool for systematically identifying proteins involved in specific phenotypes [52–55]. In commonly used pooled CRISPR screens, a library of gRNAs designed to target either all or a subset of genes is introduced into cells via lentiviral vectors, with most cells stably expressing

either one or no gRNA. Untransduced cells are removed using antibiotics or sorting, and the remaining cells are then selected based on phenotypes such as cell growth or drug resistance. After selection, sequencing is used to measure the enrichment or depletion of gRNAs, which reflects the impact of knocking out the corresponding gene or ORF on the phenotype. For instance, gRNAs targeting essential genes will be depleted during cell culture. Thousands of pooled CRISPR screens, primarily genome-wide, have been performed, significantly advancing the functional annotation of canonical coding genes. This approach is now being applied to discover functional microproteins in an unbiased and systematic manner.

Selecting microproteins to screen

In the absence of widely accepted reference annotations, the six studies each screened a unique set of ncORFs, with the number of ORFs ranging from 553 to 11,776 (Table 1) [7,12–16]. The study that targeted the largest number of ORFs (11,776) focused exclusively on short ORFs encoding microproteins [16], while the other five studies screened ncORFs, a small fraction of which encoded proteins longer than 100 aa. Most studies concentrated on ORFs that were translated in the specific cell type used for screening. For example, Chen et al. targeted 2,353 ncORFs annotated using Ribo-seq data from induced pluripotent stem cells (iPSCs) and several other cell lines, conducting their screens in iPSCs and K562 cells [7]. Hofman et al. performed screens in multiple medulloblastoma cell lines, using a gRNA library that targeted 2,019 ncORFs, a subset of previously annotated ncORFs [13,26] that showed Ribo-seq evidence of translation in 32 medulloblastoma samples [12]. Zheng et al. identified 758 lncRNA ORFs through Ribo-seq in MCF7 cells and conducted their screen in the same cell type [14]. A similar approach was used in another study to screen 1,046 lncRNA ORFs in HCT116 cells [15]. The remaining two studies, by Prensener et al. [13] and Schlesinger et al. [16], curated ORFs from publicly available data and performed their screens in multiple cell lines (Table 1). On average, each ORF was targeted by 4 to 8 gRNAs.

Identifying functional microproteins

All six studies used cell growth as the phenotypic readout and identified microproteins essential for optimal cell survival and proliferation, i.e., knocking out a microprotein impairs these processes. When applied to canonical ORFs, this type of screening is often referred to as an “essential gene screen.” The number of functional microproteins identified in each of the six studies varied significantly, ranging from 13 to 703 (Table 1). This variation is not solely due to differences in the total number of ORFs screened, as the percentage of ORFs identified as functional also varied widely, from 0.8% to 36% (Table 1). While the high percentage of hits in some studies (e.g., 36% [7]) supports the idea that microproteins have widespread functional roles, other studies suggest a more limited scope of microprotein functionality.

In addition, despite using the same cell growth phenotype, there is limited overlap in the functional microproteins identified across studies (Figure 4). In total, 1,049 functional ORFs encoding microproteins have been reported by at least one study. However, only 109 (10.4%) are supported by more than one study, and no ORF has been identified as functional in more than three studies. This lack of substantial overlap is likely attributed

to differences in the target ORF pools and the distinct cell lines used in the screens. Microproteins may exhibit cell-type-specific functions, and cancer cell lines can vary in their dependency on the same functional pathways. For example, only 50 ORFs were shared between the 553 ORFs targeted by Prensner et al. [13] and the 11,776 ORFs targeted by Schlesinger et al. [16]. Even though both studies used the same cell line (A375), only one of these 50 ORFs screened in both studies was identified as a common hit, and it failed to be validated in a secondary screen [16]. In contrast to the variation between studies, results within the same study tend to show strong correlations between replicates, even when conducted in different cell lines. For instance, in the study by Chen et al. [7], 70% of the hits from the iPSC screen were also hits in the K562 screen. More analysis is needed to understand the inconsistencies across studies, which underscores the need for a consensus reference annotation and standardized CRISPR screening protocols. This includes the use of standardized positive and negative controls to assess and compare the sensitivity and specificity of the screens.

Validating microprotein functions

As discussed earlier, the knockout phenotype of a microprotein-encoding ORF may result from the loss of noncoding functions rather than the loss of the microprotein itself (Figure 3). To address this, most studies have conducted additional experiments to validate a subset of hits and determine which functions can be definitively attributed to the microprotein. For example, Chen et al. generated individual knockout cell lines for nine uORF hits and confirmed cell growth defects in all cases [7]. Importantly, the growth defects were at least partially rescued by ectopic expression of the wild-type uORF but not by a mutant with the start codon mutated, indicating that the phenotype depends on the translation of these uORFs [7]. Similar validations were conducted in other studies. Consistent with translation-dependent functions, three studies performed gRNA tiling screens on positive hits, finding that gRNAs targeting the ORF itself tend to cause stronger growth defects than those targeting the flanking regions of the ORF [7,12,13].

It is more challenging to determine whether the observed knockout phenotype of short ORFs is associated with the microprotein or with the act of translation itself. This is particularly problematic for uORFs, as their translation is known to impact the translation of the main ORF [46]. In a screen focusing on 964 pairs of uORFs and their corresponding main ORFs, Hofman et al. found that 7.2% uORFs exhibited a stronger knockout phenotype than the main ORF [12]. For most uORFs it is difficult to assess the extent to which their function is independent of the main ORF. Similar observations were made by Chen et al. when comparing their uORF screen data to published main ORF screens [7]. It remains to be determined to what extent translation-dependent but protein-independent functions contribute to the knockout phenotypes of ORF hits in these screens.

Mechanistic dissection of microprotein functions

Detailed analysis of top hits from CRISPR screens has elucidated how microproteins contribute to cellular functions at a molecular level. For instance, Chen et al. tagged several microproteins with a 16-aa split-fluorescent protein tag, which enabled localization studies through fluorescent imaging and interaction analyses via co-immunoprecipitation

(IP) and MS [7]. Their findings revealed that many microproteins localize to membranes, interact with membrane proteins, and, notably, about half of uORF-encoded microproteins interact with the protein encoded by the corresponding main ORF [7]. However, a potential limitation is that the split fluorescent tag, despite being small, may alter the microprotein's localization and interactions, particularly when overexpressed. Future studies using antibodies specific to the endogenous microproteins will be necessary to confirm these findings. Zheng et al. similarly employed IP-MS and other assays to show that a microprotein named SMIMP binds to the cohesion complex via SMC1A and represses tumor-suppressive cell cycle regulators CDKN1A and CDKN2B, thereby explaining the observed cell growth defects in their CRISPR screen [15]. In addition to the detailed mechanistic dissection of individual top hits, high-throughput assays such as Perturb-seq [7] and large-scale ORF overexpression followed by RNA-seq [13] have been employed to uncover the genes and pathways affected by the perturbation of a large number of hits. These approaches provide valuable insights into how each microprotein impacts cell growth [13–15].

Concluding Remarks and Future Perspectives

Our genome continually surprises us with new functional elements, including introns, microRNAs, lncRNAs, and now, microproteins. The discovery of a vast number of translated short ORFs presents both a challenge and an opportunity: to explore hidden functional microproteins and uncover novel biological mechanisms. Advances in high-throughput genomics, particularly pooled CRISPR screens, have enabled the systematic discovery and functional characterization of these elusive microproteins. However, several significant challenges remain (see Outstanding Questions).

First, there is limited consensus from large-scale screens on the prevalence of functional microproteins, with estimates ranging widely from less than 1% [16] to 36% [7]. These discrepancies likely stem from variations in short ORF annotations, screening protocols, statistical methods, and cell lines used. Establishing a consensus reference annotation and standardizing CRISPR screening protocols are essential to address this uncertainty. Although progress has been made toward creating a unified annotation based on Ribo-seq evidence [26], the lack of conventional MS peptide support for 99% of these Ribo-seq ORFs hinders downstream studies [28]. With the growing catalog of MS-supported microproteins, it may be time to shift focus from Ribo-seq-annotated ORFs to microproteins detected by MS. For example, the SmProt database (v2.0) includes evidence for 1,177 human microproteins validated by MS [22]. The catalog of MS-supported microproteins can be further expanded by applying refined MS-based discovery pipelines [20,28] across a broader range of cell lines and tissues. We envision future CRISPR screens focusing on MS-supported microproteins to yield more reliable and functionally significant discoveries.

Given the identification of over 1,000 putative functional microproteins across various studies (Table 1), there is a critical need for orthogonal high-throughput methods to eliminate false positives. These includes cases where ORFs are falsely identified as functional, or where the observed effects stem from the underlying DNA, RNA, or the act of translation itself rather than the microprotein. This challenge is particularly pressing

because most of these microproteins cannot be reliably detected at the protein level. For lncRNA-encoded microproteins, comparing knockout screens with orthogonal knockdown approaches, such as CRISPRi (transcriptional shutdown) [56] and CRISPR/Cas13 (RNA degradation) [57] screens, could help identify false positives and elucidate their mechanisms of action. Notably, most lncRNAs encoding putative functional microproteins identified in CRISPR knockout screens did not exhibit significant growth defects in CRISPRi screens [7]. While incomplete knockdown by CRISPRi may explain part of this discrepancy, additional studies—such as comparisons with Cas13 screens [57]—could offer further insights. Moreover, current studies often fail to adequately differentiate between translation-dependent but protein-independent functions and bona fide microprotein activities. These gaps represent major limitation in existing validation practices and calls for the development of high-throughput methods capable of unambiguously distinguishing between coding and noncoding functions encoded by the same DNA sequence.

Large-scale screens have predominately focused on identifying microproteins essential for cell proliferation *in vitro*. While microproteins involved in regulating cell growth may contribute to tumorigenesis, their roles in other diseases and normal physiological processes remain largely unexplored. Expanding functional screens to investigate diverse cellular functions—such as cell differentiation, stress responses, and drug resistance—could uncover a broader range of microprotein functions. Additionally, adapting these screens for use in primary cells and *in vivo* models, as has been done with canonical coding gene screens [58,59], could provide critical insights into physiologically relevant microproteins.

Lastly, existing CRISPR knockout screens cannot target short ORFs that overlap with canonical ORFs (Figure 2) or circORFs, which also frequently overlap with canonical ORFs in linear mRNAs. In such cases, precise editing is required to specifically disrupt the short ORF without affecting the canonical ORF in a different reading frame. Techniques like base editing [60] or prime editing [61] offer potential solutions by introducing a premature stop codon selectively within the short ORF, leaving the canonical ORF largely intact.

Microproteins are at the forefront of expanding our understanding of the genome, offering insights into previously uncharted dimensions of gene regulation and cellular function. By refining our tools and approaches, we are poised to unlock the full potential of microproteins, shedding light on their contributions to biology and their implications for health and disease.

Acknowledgments

X.W. is supported by NIH Director’s New Innovator Award (1DP2GM140977), NIH/NHLBI grant 1R01HL171664–01, Pershing Square Sohn Prize for Cancer Research, Pershing Square Foundation MIND Prize, Pew-Stewart Scholar for Cancer Research Award, Glenn Foundation Discovery Award, Impetus Longevity Grants, Million Dollar Bike Ride (MDBR) Grant, Ines Mandl Research Foundation (IMRF) grant, and a grant from the Cure Alzheimer’s Fund.

Glossary

Microprotein

Proteins shorter than 100 amino acids. Also called micropeptides or short/small open reading frame-encoded peptides (SEPs)

Open reading frame (ORF)
A span of DNA sequence between a start codon and an in-frame stop codon

Short ORF
An ORF encoding a peptide shorter than 100 amino acids

lncRNA
Long noncoding RNA (>200 nucleotides)

Ribo-seq
A sequencing-based assay that maps ribosome footprints genome-wide. Also known as ribosome profiling

ncORF
Noncanonical ORF, which refers to an ORF other than the main ORF in canonical mRNAs

uORF
Upstream ORF located in the 5’ UTR of canonical mRNAs

Circular RNA (circRNA)
A type of RNA whose 5’ end and 3’ end are covalently linked, forming a circular structure

circORF
ORFs found within circular RNAs

Mass spectrometry (MS)
An analytic technique used to measure the mass-to-charge ratio of ions, commonly used for high-throughput protein identification and quantification

MHC-I/HLA-I antigen peptides
Short peptides derived from intracellular protein degradation subsequently presented on the on the cell surface by major histocompatibility complex class I (MHC-I) or human leukocyte antigen class I (HLA-I) molecules, involved in immune recognition

Nonsense-mediated RNA decay (NMD)
A surveillance pathway that degrades translated RNAs containing premature stop codons or a long 3’ UTR

References

1. Harrison PM et al. (2002) A question of size: the eukaryotic proteome and the problems in defining it. Nucleic Acids Res 30, 1083–1090. 10.1093/nar/30.5.1083 [PubMed: 11861898]
2. Djebali S et al. (2012) Landscape of transcription in human cells. Nature 489, 101–108. 10.1038/nature11233 [PubMed: 22955620]
3. Ingolia NT et al. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147, 789–802. 10.1016/j.cell.2011.10.002 [PubMed: 22056041]

4. Ingolia Nicholas T. et al. (2014) Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. Cell Reports 8, 1365--1379. 10.1016/j.celrep.2014.07.045 [PubMed: 25159147]
5. Ji Z et al. (2015) Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife 4, e08890. 10.7554/eLife.08890 [PubMed: 26687005]
6. Ingolia NT et al. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324, 218–223. 10.1126/science.1168978 [PubMed: 19213877]
7. Chen J et al. (2020) Pervasive functional translation of noncanonical human open reading frames. Science 367, 1140–1146. 10.1126/science.aay0262 [PubMed: 32139545]
8. Pamudurti NR et al. (2017) Translation of CircRNAs. Mol Cell 66, 9–21 e27. 10.1016/j.molcel.2017.02.021 [PubMed: 28344080]
9. Sun P and Li G (2019) CircCode: A Powerful Tool for Identifying circRNA Coding Ability. Front Genet 10, 981. 10.3389/fgene.2019.00981 [PubMed: 31649739]
10. Fan X et al. (2022) Pervasive translation of circular RNAs driven by short IRES-like elements. Nat Commun 13, 3751. 10.1038/s41467-022-31327-y [PubMed: 35768398]
11. Morales J et al. (2022) A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature 604, 310–315. 10.1038/s41586-022-04558-8 [PubMed: 35388217]
12. Hofman DA et al. (2024) Translation of non-canonical open reading frames as a cancer cell survival mechanism in childhood medulloblastoma. Mol Cell 84, 261–276 e218. 10.1016/j.molcel.2023.12.003 [PubMed: 38176414]
13. Prensner JR et al. (2021) Noncanonical open reading frames encode functional proteins essential for cancer cell survival. Nat Biotechnol 39, 697–704. 10.1038/s41587-020-00806-2 [PubMed: 33510483]
14. Zheng C et al. (2023) CRISPR/Cas9 screen uncovers functional translation of cryptic lncRNA-encoded open reading frames in human cancer. J Clin Invest 133. 10.1172/JCI159940
15. Zheng C et al. (2023) CRISPR-Cas9-based functional interrogation of unconventional translatome reveals human cancer dependency on cryptic non-canonical open reading frames. Nat Struct Mol Biol 30, 1878–1892. 10.1038/s41594-023-01117-1 [PubMed: 37932451]
16. Schlesinger D et al. (2023) A large-scale sORF screen identifies putative microproteins and provides insights into their interaction partners, localisation and function. bioRxiv, 2023.2006.2013.544808. 10.1101/2023.06.13.544808
17. Valdivia-Francia F and Sendoel A (2024) No country for old methods: New tools for studying microproteins. iScience 27, 108972. 10.1016/j.isci.2024.108972 [PubMed: 38333695]
18. Mohsen JJ et al. (2023) Microproteins-Discovery, structure, and function. Proteomics 23, e2100211. 10.1002/pmic.202100211 [PubMed: 37603371]
19. Schlesinger D and Elsasser SJ (2022) Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. FEBS J 289, 53–74. 10.1111/febs.15769 [PubMed: 33595896]
20. Slavoff SA et al. (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol 9, 59–64. 10.1038/nchembio.1120 [PubMed: 23160002]
21. Yang H et al. (2024) Widespread stable noncanonical peptides identified by integrated analyses of ribosome profiling and ORF features. Nat Commun 15, 1932. 10.1038/s41467-024-46240-9 [PubMed: 38431639]
22. Li Y et al. (2021) SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling. Genomics Proteomics Bioinformatics 19, 602–610. 10.1016/j.gpb.2021.09.002 [PubMed: 34536568]
23. Leblanc S et al. (2024) OpenProt 2.0 builds a path to the functional characterization of alternative proteins. Nucleic Acids Res 52, D522–D528. 10.1093/nar/gkad1050 [PubMed: 37956315]
24. Chothani S et al. (2023) Discovering microproteins: making the most of ribosome profiling data. RNA Biol 20, 943–954. 10.1080/15476286.2023.2279845 [PubMed: 38013207]
25. Tong G et al. (2024) Comparison of software packages for detecting unannotated translated small open reading frames by Ribo-seq. Brief Bioinform 25. 10.1093/bib/bbae268

26. Mudge JM et al. (2022) Standardized annotation of translated open reading frames. Nat Biotechnol 40, 994–999. 10.1038/s41587-022-01369-0 [PubMed: 35831657]
27. Sandmann CL et al. (2023) Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. Mol Cell 83, 994–1011 e1018. 10.1016/j.molcel.2023.01.023 [PubMed: 36806354]
28. Deutsch EW et al. (2024) High-quality peptide evidence for annotating non-canonical open reading frames as human proteins. bioRxiv, 2024.2009.2009.612016. 10.1101/2024.09.09.612016
29. Martinez TF et al. (2023) Profiling mouse brown and white adipocytes to identify metabolically relevant small ORFs and functional microproteins. Cell Metab 35, 166–183 e111. 10.1016/j.cmet.2022.12.004 [PubMed: 36599300]
30. Kesner JS and Wu X Mechanisms suppressing noncoding translation. Trends in Cell Biology. 10.1016/j.tcb.2024.09.004
31. Kesner JS et al. (2023) Noncoding translation mitigation. Nature 617, 395–402. 10.1038/s41586-023-05946-4 [PubMed: 37046090]
32. Muller MBD et al. (2023) Mechanisms of readthrough mitigation reveal principles of GCN1-mediated translational quality control. Cell 186, 3227–3244 e3220. 10.1016/j.cell.2023.05.035 [PubMed: 37339632]
33. Minami R et al. (2010) BAG-6 is essential for selective elimination of defective proteasomal substrates. J Cell Biol 190, 637–650. 10.1083/jcb.200908092 [PubMed: 20713601]
34. Vakirlis N et al. (2022) De novo birth of functional microproteins in the human lineage. Cell Rep 41, 111808. 10.1016/j.celrep.2022.111808 [PubMed: 36543139]
35. Lin MF et al. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and noncoding regions. Bioinformatics 27, i275–282. 10.1093/bioinformatics/btr209 [PubMed: 21685081]
36. Gaertner B et al. (2020) A human ESC-based screen identifies a role for the translated lncRNA LINC00261 in pancreatic endocrine differentiation. Elife 9. 10.7554/eLife.58659
37. Mattick JS et al. (2023) Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol 24, 430–447. 10.1038/s41580-022-00566-8 [PubMed: 36596869]
38. Statello L et al. (2021) Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol 22, 96–118. 10.1038/s41580-020-00315-9 [PubMed: 33353982]
39. Kristensen LS et al. (2019) The biogenesis, biology and characterization of circular RNAs. Nat Rev Genet 20, 675–691. 10.1038/s41576-019-0158-7 [PubMed: 31395983]
40. Mayr C (2019) What Are 3′ UTRs Doing? Cold Spring Harbor Perspectives in Biology 11. 10.1101/cshperspect.a034728
41. Leppek K et al. (2018) Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat Rev Mol Cell Biol 19, 158–174. 10.1038/nrm.2017.103 [PubMed: 29165424]
42. Ryczek N et al. (2023) The Functional Meaning of 5'UTR in Protein-Coding Genes. Int J Mol Sci 24. 10.3390/ijms24032976
43. Kim T-K et al. (2015) Enhancer RNAs: A Class of Long Noncoding RNAs Synthesized at Enhancers. Cold Spring Harbor Perspectives in Biology 7. 10.1101/cshperspect.a018622
44. Natoli G and Andrau J-C (2012) Noncoding transcription at enhancers: general principles and functional models. Annual review of genetics 46, 1--19. 10.1146/annurev-genet-110711-155459
45. Core L and Adelman K (2019) Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. Genes Dev 33, 960–982. 10.1101/gad.325142.119 [PubMed: 31123063]
46. Dever TE et al. (2023) Translational regulation by uORFs and start codon selection stringency. Genes Dev 37, 474–489. 10.1101/gad.350752.123 [PubMed: 37433636]
47. Hurt JA et al. (2013) Global analyses of UPF1 binding and function reveal expanded scope of nonsensemediated mRNA decay. Genome research 23, 1636--1650. 10.1101/gr.157354.113 [PubMed: 23766421]

48. Smith JE et al. (2014) Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. Cell Rep 7, 1858–1866. 10.1016/j.celrep.2014.05.023 [PubMed: 24931603]

49. Tani H et al. (2013) The RNA degradation pathway regulates the function of GAS5 a non-coding RNA in mammalian cells. PLoS One 8, e55684. 10.1371/journal.pone.0055684 [PubMed: 23383264]

50. Carlevaro-Fita J et al. (2016) Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. RNA 22, 867–882. 10.1261/rna.053561.115 [PubMed: 27090285]

51. Chew G-L et al. (2013) Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. Development (Cambridge, England) 140, 2828--2834. 10.1242/dev.098343 [PubMed: 23698349]

52. Bock C et al. (2022) High-content CRISPR screening. Nat Rev Methods Primers 2. 10.1038/s43586-022-00098-7

53. Shalem O et al. (2015) High-throughput functional genomics using CRISPR-Cas9. Nat Rev Genet 16, 299–311. 10.1038/nrg3899 [PubMed: 25854182]

54. Doench JG (2018) Am I ready for CRISPR? A user's guide to genetic screens. Nat Rev Genet 19, 67–80. 10.1038/nrg.2017.97 [PubMed: 29199283]

55. Wang T et al. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. Science (New York, N.Y.) 343, 80--84. 10.1126/science.1246981 [PubMed: 24336569]

56. Liu SJ et al. (2017) CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science (New York, N.Y.) 355, eaah7111. 10.1126/science.aah7111

57. Liang WW et al. (2024) Transcriptome-scale RNA-targeting CRISPR screens reveal essential lncRNAs in human cells. Cell. 10.1016/j.cell.2024.10.021

58. Chow RD and Chen S (2018) Cancer CRISPR Screens In Vivo. Trends Cancer 4, 349–358. 10.1016/j.trecan.2018.03.002 [PubMed: 29709259]

59. Dong MB et al. (2019) Systematic Immunotherapy Target Discovery Using Genome-Scale In Vivo CRISPR Screens in CD8 T Cells. Cell 178, 1189–1204 e1123. 10.1016/j.cell.2019.07.044 [PubMed: 31442407]

60. Rees HA and Liu DR (2018) Base editing: precision chemistry on the genome and transcriptome of living cells. Nat Rev Genet 19, 770–788. 10.1038/s41576-018-0059-1 [PubMed: 30323312]

61. Chen PJ and Liu DR (2023) Prime editing for precise and highly versatile genome manipulation. Nat Rev Genet 24, 161–177. 10.1038/s41576-022-00541-1 [PubMed: 36344749]

Highlights

- Thousands of microproteins with unclear functional significance are synthesized in human cells and other organisms.
- Large-scale CRISPR knockout screens in human cells have identified over a thousand candidate functional microproteins essential for optimal cell growth *in vitro*.
- Validating these candidate functional microproteins and uncovering their molecular mechanisms and physiological relevance remain significant challenges.

Outstanding Questions Box

1. How many microproteins are synthesized in human cells? How do we generate a consensus annotation considering the variation in the quality and coverage of Ribo-seq data, the low overlap between different ORF annotation tools, and the lack of peptide evidence for most ORFs?
2. How can we validate and characterize the 1049 putative functional microproteins identified through CRISPR knockout screens but lack direct protein evidence?
3. How can we rule out the noncoding functions of the underlying RNA, DNA, or even the act of translation in a high-throughput manner for all hits from large-scale CRISPR screens?
4. Can CRISPR screens targeting microproteins with peptide evidence uncover a higher proportion of functional microproteins?
5. Can base editing or prime editing be used for functional screens of circORFs and other short ORFs overlapping with canonical ORFs?

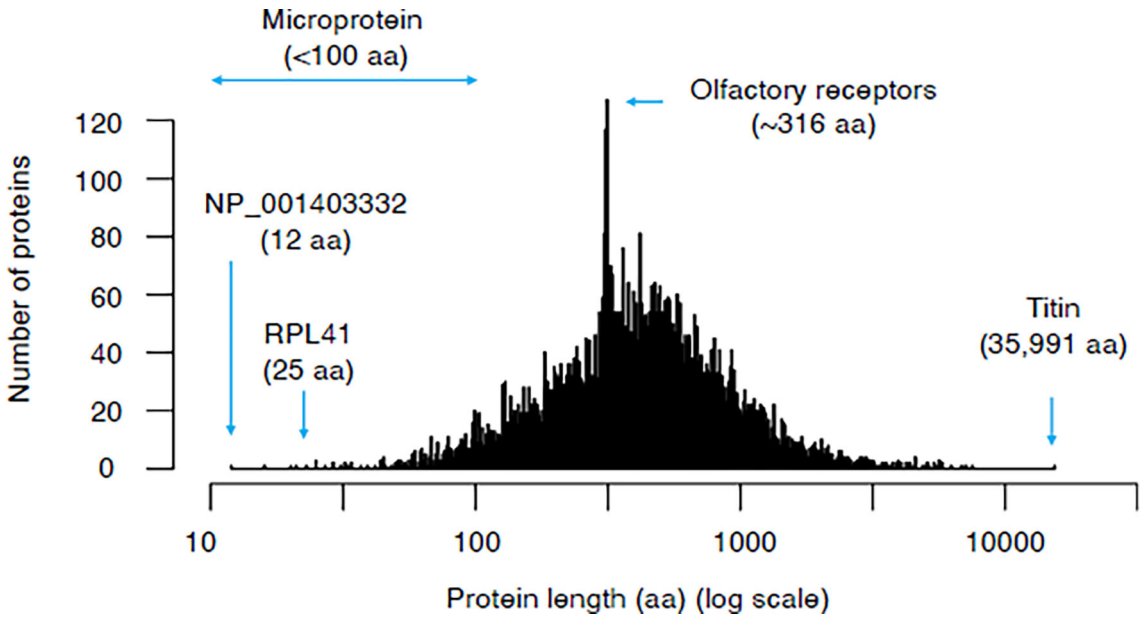

Figure 1: Size distribution of canonical human proteins.

Shown is a histogram plot for the length (aa) of 19,352 human proteins included in the NCBI/EMBL-EBI MANE Select set (v1.3). Note that the x-axis is in log scale.

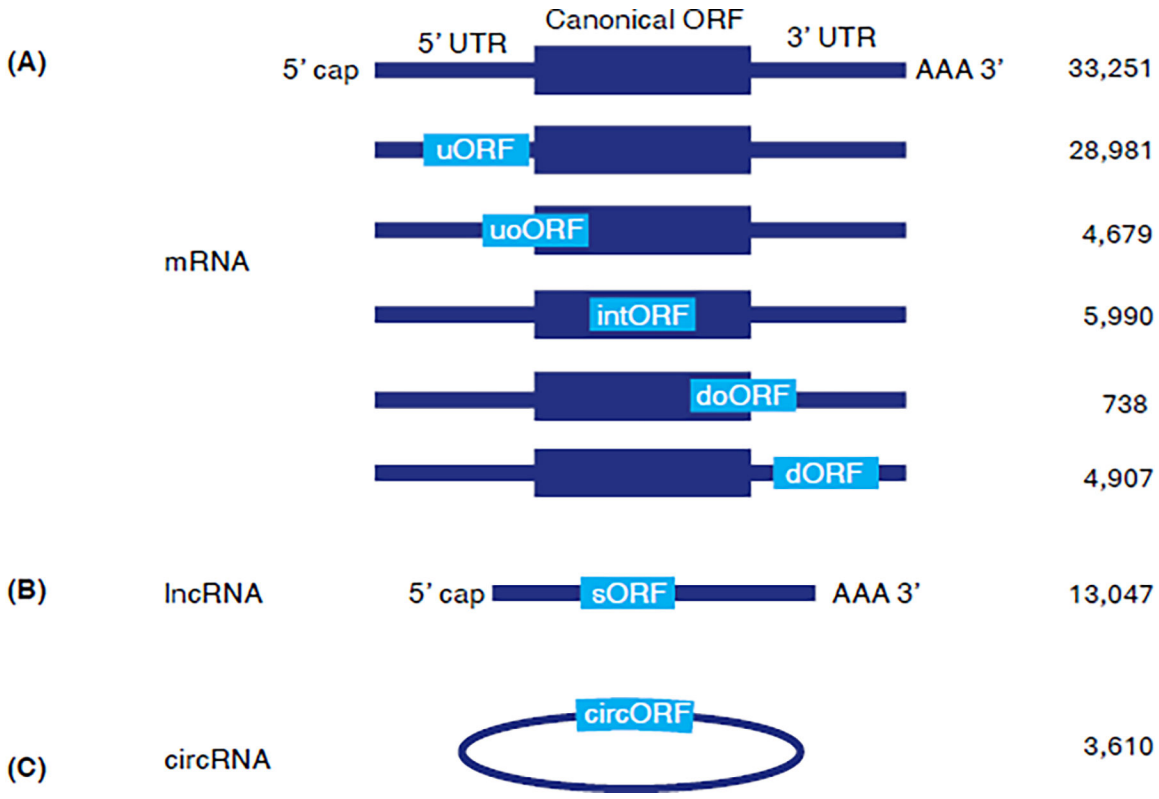

Figure 2: Most microproteins are encoded by noncanonical ORFs (ncORFs).

(A) ncORFs in canonical mRNAs: uORF – upstream ORF contained entirely in the 5' UTR; uoORF/ouORF – upstream overlapping ORF that begins in the 5' UTR but ends within the canonical ORF in a different reading frame; intORF – ORF within the canonical ORF in a different reading frame; doORF – downstream overlapping ORF that begins in the canonical ORF but ends in the 3' UTR in a different reading frame; dORF – downstream ORF contained entirely in the 3' UTR. (B) Many lncRNAs contain a short ORF (sORF or smORF). (C) Some circular RNAs (circRNAs) contain an ORF (circORF/cORF) that can be translated in a cap-independent manner. The numbers on the right indicate the number of ORFs supported by Ribo-seq data in a previous study [21], except for circRNAs [9].

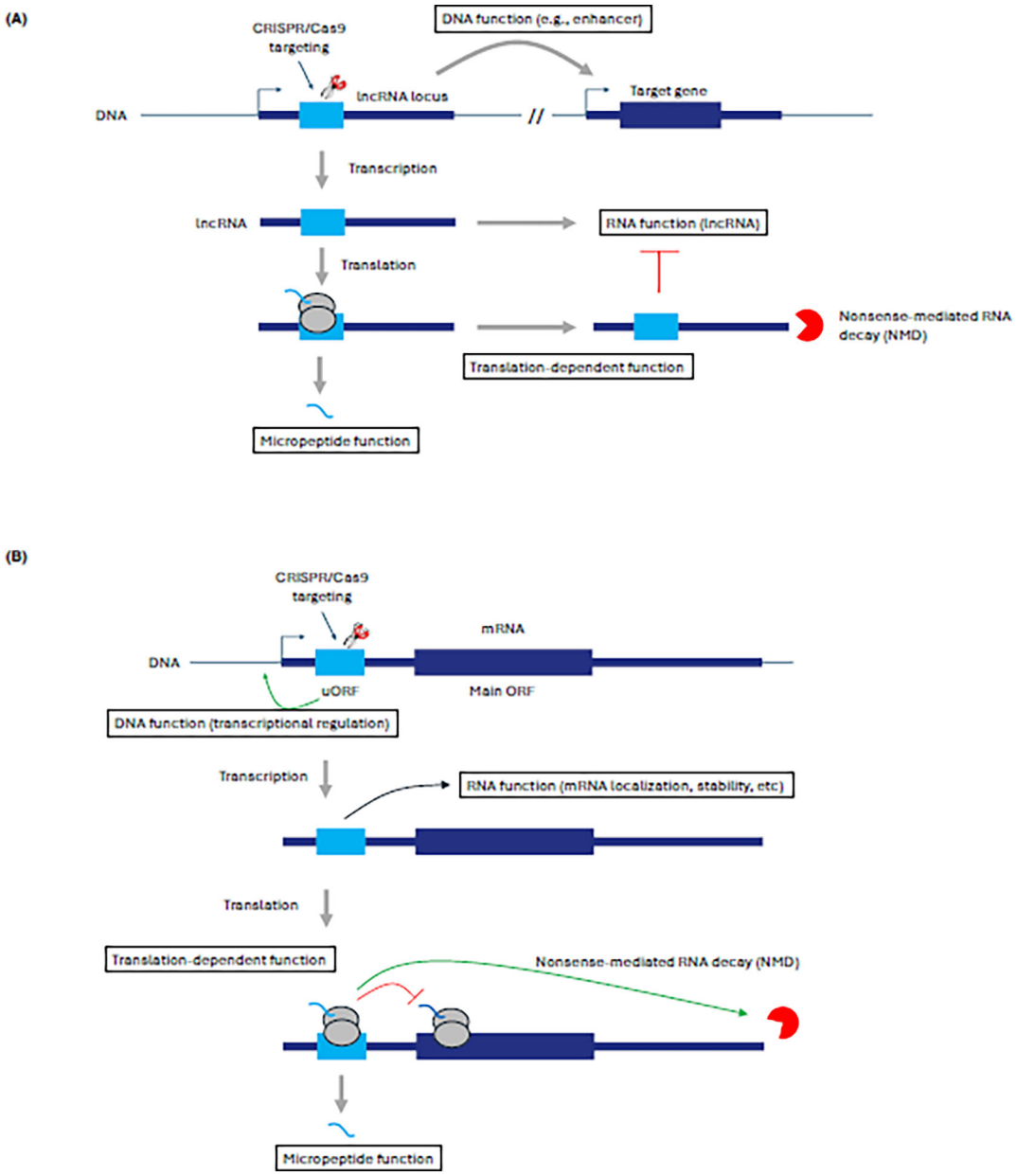

Figure 3: Potential coding and noncoding functions of short ORFs.

(A) lncRNA: At the DNA level, lncRNA loci can function as enhancers, regulating the transcription of neighboring genes. CRISPR-mediated targeting of the short ORFs within lncRNAs could disrupt this enhancer activity. At the RNA level, most lncRNAs exert their functions without being translated. However, translation of short ORFs within lncRNAs may activate nonsense-mediated decay (NMD), leading to the degradation of the lncRNA. Consequently, disrupting the translation of these short ORFs could alter the lncRNA's stability and affect its RNA-mediated noncoding functions. Additionally, the microproteins produced from these lncRNA-encoded ORFs may also have functional roles as proteins. (B) uORF: At the DNA level, uORFs, due to their proximity to the promoter, may influence the transcription of the host mRNA gene when disrupted by CRISPR. At the RNA level,

uORF sequences can regulate the stability and localization of the host mRNA. Translation of uORFs frequently inhibits the translation of the main ORF and can also trigger NMD, leading to the degradation of the host mRNA.

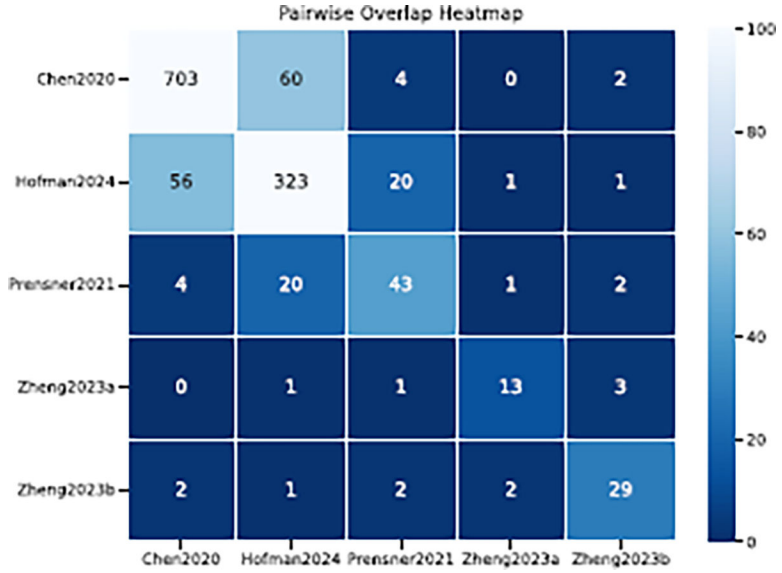

Figure 4: Pairwise overlap of microprotein hits across published CRISPR screens.

The heatmap displays the pairwise overlap of microprotein hits between studies, with each row and column representing a specific study. The heatmap values indicate the number of microprotein hits in one study (row) that overlap with hits in another (column). Overlap between two different ORFs is defined as sharing at least one base pair in the genome. Note that Schlesinger et al. 2024 [16] is excluded due to the absence of genomic coordinate data. Study IDs correspond to those listed in Table 1.

Table 1:

Summary of published CRISPR screens for functional microproteins

Study	Cell type	ncORFs screened	ncORF hits	Microprotein hits
Chen2020 [7]	iPSC	2,353	570 (24%)	703
	K562	2,353	848 (36%)	
Prensner2021 [13]	A375, A549, HA1E, HeLa, HepG2, HT29, MCF7, PC3	553	*57 (10%)	43
Zheng2023a [14]	MCF7	758	28 (3.7%)	13
Zheng2023b [15]	HCT116	1,046	49 (5.4%)	29
Hofman2024 [12]	7 medulloblastoma cell lines	2,019	*387 (21%)	323
Schlesinger2024 [16]	A375	11,776	83 (0.7%)	83

*: Total unique hits from all cell lines screened.