Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2024

No country for old methods: new tools for studying microproteins

Valdivia-Francia, Fabiola ; Sendoel, Ataman

DOI: https://doi.org/10.1016/j.isci.2024.108972

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-254068
Journal Article
Accepted Version

Originally published at:
Valdivia-Francia, Fabiola; Sendoel, Ataman (2024). No country for old methods: new tools for studying microproteins. iScience, 27(2):108972.
DOI: https://doi.org/10.1016/j.isci.2024.108972

Journal Pre-proof

No country for old methods: new tools for studying microproteins

Fabiola Valdivia-Francia, Ataman Sendoel

PII: S2589-0042(24)00193-7

DOI: https://doi.org/10.1016/j.isci.2024.108972

Reference: ISCI 108972

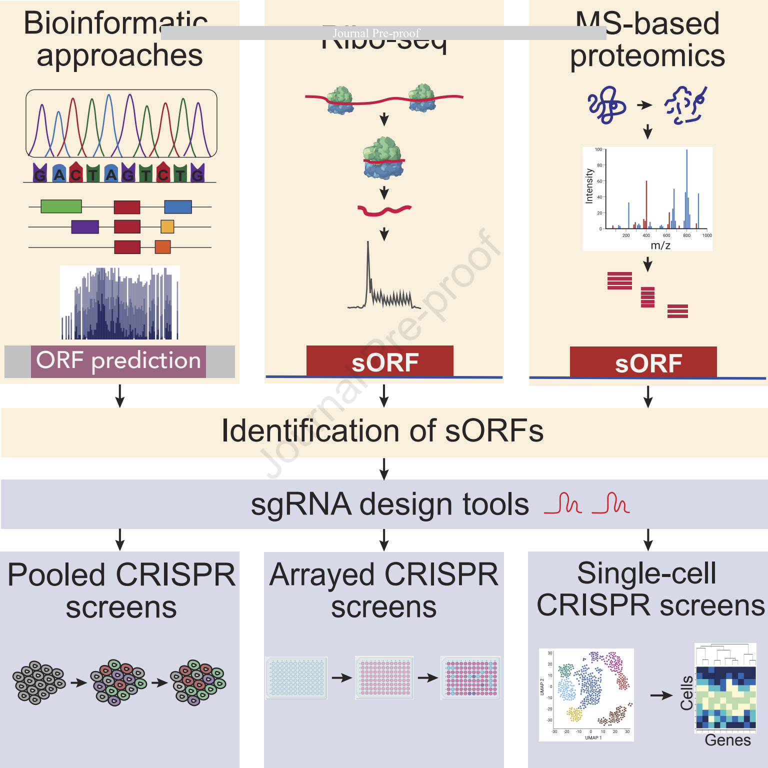
To appear in: *ISCIENCE*

Please cite this article as: Valdivia-Francia, F., Sendoel, A., No country for old methods: new tools for studying microproteins, *ISCIENCE* (2024), doi: https://doi.org/10.1016/j.isci.2024.108972.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2024 The Author(s).

Journal Pre-proof



Bioinformatic approaches

Ribo-seq

MS-based proteomics

ORF prediction

sORF

sORF

Identification of sORFs

sgRNA design tools

Pooled CRISPR screens

Arrayed CRISPR screens

Single-cell CRISPR screens

Journal Pre-proof

Journal Pre-proof

No country for old methods: new tools for studying microproteins

Fabiola Valdivia-Francia[1,2] & Ataman Sendoel[1]

1. University of Zurich
Institute for Regenerative Medicine (IREM)
Wagistrasse 12
CH-8952 Schlieren-Zurich
Switzerland
Phone: +41 (0)44 634 9141

2. Life Science Zurich Graduate School,
Molecular Life Science Program,
University of Zurich/ ETH Zurich, Switzerland

Correspondence: ataman.sendoel@uzh.ch

Journal Pre-proof

Abstract

Microproteins encoded by small open reading frames (sORFs) have emerged as a fascinating frontier in genomics. Traditionally overlooked due to their small size, recent technological advancements such as ribosome profiling, mass spectrometry-based strategies and advanced computational approaches have led to the annotation of more than 7000 sORFs in the human genome. Despite the vast progress, only a tiny portion of these microproteins have been characterized and an important challenge in the field lies in identifying functionally relevant microproteins and understanding their role in different cellular contexts. In this review, we explore the recent advancements in sORF research, focusing on the new methodologies and computational approaches that have facilitated their identification and functional characterization. Leveraging these new tools hold great promise for dissecting the diverse cellular roles of microproteins and will ultimately pave the way for understanding their role in the pathogenesis of diseases and identifying new therapeutic targets.

Journal Pre-proof

Introduction

The mammalian genome is composed of a vast number of uncharacterized and unannotated small open reading frames (sORFs), which are commonly misinterpreted as “junk DNA” with no defined function outside of gene regulation. With the advent of new technologies, thousands of unannotated sORFs – typically located on non-coding RNAs and untranslated regions (UTRs) of protein-coding genes - have been shown to be translated into functional proteins. New technologies, such as proteomics and ribosome profiling, in tandem with advanced bioinformatic methods, have played a critical role in driving forward the sORF field. The combination of these approaches greatly facilitated the genome-wide annotation of sORFs, thereby unraveling their involvement in various cellular functions, including those relevant to human diseases [1–7].

Most of these novel open reading frames (ORF) defy the conventional rules for gene annotation. These rules include a minimal length of 100 codons, an in-frame AUG start and a single ORF per transcript [4,8,9]. Unannotated ORFs smaller than 100 amino acids are classified as sORFs (Figure 1). These include sORFs on non-coding RNAs such as long non-coding RNAs (lncRNAs) and overlapping sequences on annotated ORFs, classified as alternative ORFs (alt-ORFs) [4,10,11]. Moreover, sORFs residing in the 5’ untranslated region (5’UTR) of an mRNA are referred to as upstream open reading frames (uORFs) [6,12,13], while those found in the 3’UTR are known as downstream ORF (dORFs) [12–14].

Recent endeavors to generate standardized sORF catalogs led to the annotation of more than 7000 human sORFs and suggest that sORFs form a substantial part of eukaryotic genomes [4–6,15]. The encoded peptides or microproteins translated from sORFs are involved in a variety of cellular functions in both health and disease [2,12,16]. Microproteins are involved in the downregulation of tumor angiogenesis [17], suppress tumor growth[18], or in cell proliferation [19]. Furthermore, uncharacterized sORFs hold great promise as potential drug targets that drive different cellular processes underlying the pathogenesis of diseases [2,16,20].

The translation of sORFs can either result in a peptide product or it may have a regulatory function, a phenomenon widely observed in the case of most known uORFs

Journal Pre-proof

[12]. About 50% of mammalian genes contain uORFs [21–23], which modulate ribosome access to the downstream ORF and which can reduce translational efficiency by an average of 30-48% [6]. However, under stress conditions, uORF-mediated regulation allows certain genes such as ATF4 to become translationally induced to rapidly mount cellular stress responses [24,25]. Genome-wide uORF translation may be subject to regulation and contribute to the translational program in embryonic stem cells or tumor initiation. Embryonic stem cells (ESCs) decrease their relative uORF translation rates when undergoing differentiation [5]. Similarly, the stemness signature of muscle stem cells has been shown to be partly regulated by uORF-containing mRNAs [26]. Furthermore, tumor-initiating cells increase their relative rate of uORF translation during the early stages of tumorigenesis[27].

Here, we review the current methodologies for studying sORFs in the eukaryotic genome and outline emerging new techniques to study the function of sORFs and their potential involvement in disease. Although we focus on the eukaryotic genome in this review, it is worth noting that substantial progress has been made in the study of bacterial and plant sORFs and that most of the identification techniques described here can also be applied to lower organisms.

Journal Pre-proof

Bioinformatic approaches

New computational approaches and the availability of RNA sequencing data sets have led to better transcriptome annotations and facilitated the classification of microproteins (Figure 2). Historically, ORFs were defined as a sequence of DNA that is delimited by a start codon followed by a downstream in-frame stop codon. However, this approach was biased as it involved an artificial cutoff in annotating only proteins larger than 100 amino acids or 300 nucleotides, mainly because any sequence smaller than this cutoff was considered nonfunctional or derivative artifacts of canonical transcripts and coding sequences [9,28,29]. Additionally, the 100 amino acid cutoff results from the increasing probability of artifactual ORFs and biologically meaningless sequences found in shorter ORFs [9] . The likelihood of a protein-coding ORF increases with its length. Thus the reason why many algorithms had a fixed threshold of 100 amino acids was to avoid dubious non-coding ORFs [9,30–32]. Consequently, current protein catalogs are skewed for larger proteins, which has resulted in a notable underrepresentation of microproteins [5,23]. In light of the mounting evidence supporting the presence of sORFs, new algorithms have been developed to adjust the classification parameters for the annotation of ORFs.

Traditionally, protein annotation has heavily relied on evaluating sequence similarities and conservation across different species, which proves valuable as selection pressure is linked to functional importance [12,33]. Furthermore, examining the similarity between known protein domains has provided valuable insights into the potential function of newly predicted ORFs. However, relying solely on conservation as a criterion for identifying non-canonical ORFs could limit the detection of sORFs that, although not conserved across species, may still encode functional microprotein [12,34]. Short proteins are more challenging to classify than larger ones due to stricter statistical features to distinguish them from non-coding sequences [31]. Sequences shorter than 100 amino acids are less likely to show conservation between species [35,36]. Conserved coding sequences show a higher ratio of synonymous compared to non-synonymous substitutions (dS/dN), which can be exploited to distinguish them from non-coding regions, a difference that is less pronounced in smaller proteins [31].

Journal Pre-proof

To address these challenges, a pipeline called PhyloCSF was developed, which can systematically resolve conservation problems by considering phylogenetic models for shorter sequences. PhyloCSF distinguishes itself from previous tools by using empirical codon models, which can compare alignments of coding regions with alignments of non-coding regions. Moreover, it incorporates genome-wide training data, taking into account codon frequencies and substitution rates to discern protein-coding from non-coding sequences [37]. A recent application of PhyloCSF resulted in the identification of 144 novel coding sequences absent in existing catalogs [38]. Notably, 50 of these newly discovered protein-coding genes encode microproteins containing fewer than 100 amino acids, further advocating for the utilization of PhyloCSF in sORF detection [38].

Annotation of sORFs is not only challenging but also limited by computational algorithms. Many *in silico* approaches in the past restricted in the annotation of ORFs, including the necessity for a single coding sequence per transcript, an AUG start codon, a codon bias, a coding region longer than 100 codons and the sequence conservation [39–42]. With the advent of new technologies and the general increase in sequencing data, computational pipelines have been modified to adjust these criteria. *In silico* approaches can now predict all possible ORF, including overlapping and shorter ORFs from existing annotations, novel predicted isoforms and novel proteins from alternative ORFs [41]. OpenProt is a bioinformatic tool that uses two sources of annotations (NCBI-RefSeq and Ensembl) and publicly available ribosome profiling and mass spectrometry data sets to facilitate the annotation of predicted sORFs [41]. An advantage of the technique is that it allows predicting sORFs using a minimal length of 30 amino acids and, in its latest update, has even removed the restriction for AUG start codons[43]. Other annotation tools such as ORF finder [44], micPDP [4] and uPEPperoni [45] are also used for the detection of sORFs and are reviewed in Table 1. In recent years, sORF databases have become available such as sORFs.org [46] and smProt [47].

More recent tools use machine learning for the prediction of sORFs. RNASamba and DeepCPP are two pipelines that predict sORFs based on neuronal networks [48,49]. RNASamba was designed to recognize non-intuitive patterns, such as the Kozak sequence, in a way that it can learn from previous sequence data to distinguish coding from non-coding sequences [48]. On the other hand, DeepCPP also uses the information around the start codon, which the authors term nucleotide bias, to help predict the coding potential of RNA [49]. Considering the critical role of nucleotides surrounding the start codon in translation initiation, DeepCPP evaluates the codon bias at nucleotide positions -3 to +6 from the start codon. It is interesting to point out that this nucleotide bias is not the same for non-coding mRNAs [49]. MiPepid is another machine learning tool that can identify potential microproteins from the DNA sequence [50]. Using a microprotein database for training, MiPepid achieved a 96% accuracy when tested on a blind dataset of high-confidence micropeptides [50]. Finally, a fourth new computational tool is csORF-finder, which emerged as a tool for characterizing the translation potential of sORFs. csORF-finder aims to distinguish coding sORFs (csORFs) from non-coding sORFs in different species and thus facilitate the discovery of new functional microproteins [51].

A pseudo-alignment algorithm, named ORFanage, was recently established for the detection of novel ORFs in the assembled results from RNA-seq [52]. ORFanage can identify ORFs from RNA-seq data based on the similarity to the reference annotation and similarity within genes in the transcripts[52]. The method relies on the assumption that protein-coding genes produced by different transcripts from the same locus should share similarities, which are then exploited to detect new microproteins [52].

With growing data sets and sequencing data available, the new computational tools greatly help the identification of sORFs. While bioinformatic tools have undeniably advanced and facilitated the identification of sORFs, it is important to emphasize that a systematic characterization is essential to validate the existence of the microprotein. An overview of the computation tools can be found in Table 1.

Journal Pre-proof

Ribosome profiling

Ribosome profiling (Ribo-seq) provides real-time snapshots of translation by assessing so-called ribosome-protected fragments (RPF), which indicate the mRNA portions that are being translated into proteins. Ribo-seq is a powerful tool that allows determining ORFs at codon resolution and greatly helps the identification of previously unannotated ORFs (Figure 2) [53,54]. The technique was pioneered by Ingolia et al. and is based on the sequencing of the approximately 30 nucleotide-long fragments that are protected by the ribosomes after nuclease digestion [55]. The reads that are obtained from the sequencing of the RPFs can be aligned to the transcriptome and provide a genome-wide overview of ribosome occupancy. Ribo-seq allows us to determine the position of where translation is taking place and, in recent years, has helped to shed light on unconventional translation and sORF translation in regions previously thought to be non-coding [2,5,56].

With the introduction of this method, Ingolia and colleagues showed that translation could occur in the 5'UTR of an mRNA, first in yeast and later in mouse embryonic stem cells [5,55]. Ribo-seq provided evidence that the mammalian genome undergoes substantially more translation than previously assumed. Ribo-seq studies in Drosophila, zebrafish, mice, and humans led to the discovery of widespread translation on long non-coding RNAs (lncRNAs), upstream and downstream regions and even overlapping coding transcripts [2,6,42,57–59]. Additionally, the translation from non-canonical start sites has helped broaden the repertoire of translated sORFs. Alternative start sites can be mapped by Ribo-seq combined with the treatment of inhibitors such as harringtonine or lactimidomycin, which cause ribosomes to accumulate at sites of translation initiation [5,60,61]. Analysis from harringtonine experiments revealed that 44% of AUG start sites are downstream of annotated proteins and represent a source of alternative ORFs, resulting in truncated proteins [5]. Similarly, protein isoforms can emerge through the utilization of upstream start sites, resulting in N-terminal extensions if the start site is in frame with the main start site and lacks a stop codon in between [5]. Treatment with harringtonine also revealed that translation could be initiated at near-cognate (non-AUG) start sites. Most of the near-cognate initiation sites were mapped to the 5'UTR of transcripts, suggesting that uORFs are most frequently initiated with GUG start codons [5,27] .

Computational tools play an important role in detecting ORFs from ribosome profiling data, enabling the identification of potential microproteins encoded by sORFs. These tools leverage the direct evidence of ribosome-protected fragments captured by ribosome profiling to pinpoint translated regions within the transcriptome in a genome-wide fashion. By analyzing ribosome footprints and especially on the basis of the characteristic three-nucleotide periodicity indicating *bona fide* translation, these algorithms can distinguish between coding and non-coding sequences and therefore map new open reading frames. Some of the commonly used tools for sORF detection from ribosome profiling data include ORF-Rater, RiboTaper/ORF-quant, ORFscore, RiboWave, RiboCode, DeepRibo, ribotricer and riboHMM [4,42,62–69]. Each of these computational tools offers distinct approaches to maximize sensitivity and specificity and together have significantly expanded our ability to annotate novel sORFs from ribosome profiling data.

Since the introduction of Ribo-seq, the technique has allowed us to successfully identify sORFs with the potential of encoding microproteins [2,10,31,56,57,70]. In recent years, numerous researchers in the field have started a joint effort to produce a standardized catalog with more than 7000 human ORFs that were identified based on Ribo-seq. This effort to annotate sORFs in a standardized manner will facilitate future endeavors to dissect the function of these ORFs [15]. Separate studies have identified sORFs over the years, however, thus far, only 3085 ORFs identified by Ribo-seq have been found by more than one research group [15]. Furthermore, despite the wealth of identified sORFs, only a select few have been further characterized, unveiling their

Journal Pre-proof

specific roles in cellular processes. By combining standardization and systematic annotation of sORFs with the development of comprehensive tools to dissect their potential functions, the approaches will likely shed light on which sORFs impact health and disease.

Mass spectrometry-based proteomics

Although Ribo-seq demonstrates the translatability of sORFs, it does not provide direct evidence that these microproteins are present in a cell. Theoretically, RNA-binding proteins could also protect mRNA fragments of similar size to ribosome footprints and would end up in the Ribo-seq library. In addition, it has been argued that some RPF could also result from stalled ribosomes not actively translating RNA. As outlined above, to discern true translation from other types of protected mRNA fragments, the characteristic three-nucleotide periodicity of Ribo-seq datasets can be assessed by computational methods. The three-nucleotide periodicity greatly helps detect *bona fide* translation of longer ORFs. However, the triplet periodicity is challenging to detect on very small ORFs such as certain uORFs.

Mass spectrometry (MS) can detect and quantify proteins and, therefore, verify the presence of the microproteins. In that sense, MS based proteomics is currently the only experimental technique able to provide evidence of the existence of a microprotein. Nevertheless, the identification of microproteins by MS methods can be challenging and is hampered by the fact that sORFs are commonly excluded from protein databases, initiate with near-cognate start sites and produce a few unique tryptic peptides[7]. In recent years, however, MS strategies have been further optimized for the identification of sORF (Figure 2) [4,7,13,71].

Mass spectrometry is the gold standard used to characterize the proteome[72]. Originally based on the observations of four peptides with less than 150 amino acids made by Oyama et al. [73], an MS technique for the detection of microproteins was developed and further optimized. By combining peptidomics with RNA sequencing, Slavoff and colleagues detected a total of 90 sORF-encoded peptides (SEPs), 86 of them being newly discovered [7].

Journal Pre-proof

Due to their small size and abundance, MS-based detection of microproteins requires previous fractionation and enrichment approaches [29]. In MS studies, peptide mapping allows direct identification and quantification of proteins. Proteins are fragmented by tryptic digestion and the molecular weight of the peptides is measured and compared with reference databases. Tryptic digestion can present a problem since the smaller-sized proteins contain very few and sometimes even no tryptic peptide fragments, which biases mapping to more stable and abundant proteins [29,74,75]. Replacing trypsin with different proteases can enhance microprotein detection [74,76,77]. Size exclusion approaches are used to enhance the detection of low molecular weight peptides [78–81].

Mass spectrometry can also be combined with separation techniques such as liquid chromatography to help with the identification of microproteins. In recent years, MS has become important, not only because of its ability to identify proteins (whether small or large), but also for its power to quantify and molecularly characterize them via the identification of posttranslational modifications [79,82].

The challenge of the different MS strategies to detect sORFs is the requirement of reference databases from which the peptides can be identified [7,71]. In most MS experiments, a custom database is generated that contains all potential peptides translated from the transcriptome [7,71,79,83]. The absence of sORF repositories and catalogs encourages the coupling of MS with genomic or transcriptomic data [84]. Many new sORF have been identified by mapping MS data to RNA sequencing and Ribo-seq data [7,85]. Using three- or six-frame translation to generate an expanded reference protein database can improve the detection of sORFs [85–87]. Other challenges include the small size of the encoded peptides and the lack of conservation of sORF sequences between organisms.

To ensure that the peptide identified via MS is not a result of false positive proteomic profiling, it is critical to validate newly discovered microproteins. This is especially true for microproteins for which only 1 peptide has been detected. A widely adopted validation technique involves the use of isotopically labeled standards, which are chemically indistinguishable [79]. The synthetic peptide should show similar MS spectra profiles except for a mass shift introduced from the isotopic label of the synthetic peptide [79,82]. A second method used for the validation of small encoded peptides is

Journal Pre-proof

the siRNA-based silencing of the transcript together with targeted MS and peptide standards, assessed by using RT-qPCR [79].

Mass spectrometry has become a powerful tool to validate the expression of microproteins. Additionally, MS can help infer protein function via interaction partners, as recently shown using the so-called MicroID approach [88]. Proximity biotinylation-based techniques have the potential to systematically map the interaction partners of microproteins and serve as an attractive method for characterizing sORF-encoded microproteins. In the MicroID approach, the authors developed an elegant high-throughput technique by which novel sORFs were identified and mapped to different subcellular localizations such as the nucleus or nucleolus. Furthermore, they also determined functional information based on molecular interactors accessed via transcriptome data [88]. Another powerful example for MS-based identification of microproteins is major human leukocyte antigen class I (HLA-I) peptidomics. Using such a strategy in induced pluripotent stem cells (iPSCs), 240 non-canonical peptides from uORFs but also sORFs on lncRNAs could be identified, indicating that a portion of sORFs can enter the HLA-I presentation pathway to become part of the antigen repertoire [2]. They also employed an elegant, minimally disruptive mNeonGreen split-fluorescent tagging strategy to visualize select microproteins, which is based on a 16 amino acid tag fused to the microprotein and can be detected once it complements with the remainder of the split mNG protein [2,89,90].

Journal Pre-proof

Given the high number of potential microproteins identified by these different methods, it is critical to validate their expression by orthogonal assays. The predicted microproteins can be validated using different tools, including reporter assays, epitope tagging and loss of function assays. In a recent study using comparative proteomics, the authors annotated differentially expressed novel sORFs in leukemia cells and reported their subcellular localization by FLAG-tagged microprotein overexpression and subsequent visualization by immunofluorescence[91]. A potential issue for microproteins is that common fluorescent tags often exceed the size of microproteins, which can affect their biophysical and biochemical properties. To this end, an elegant new technique introduces a single non-canonical amino acid either at the N- or C-terminus of microproteins. This technique, called single-residue terminal labels (STELLA), can be exploited for minimal tagging of microproteins without disturbing the physical or biochemical properties of microproteins[92]. Microproteins can also be validated by chemical labeling coupled with proteomic identification, as previously reviewed[93]. Protein interactions can help elucidate the putative role of a new microprotein. PRISMA, a protein interaction screen on a peptide matrix, was developed to identify the interactome of evolutionary young microproteins via sequence motifs[36].

With the improvement in omics technologies, the combination of mass spectrometry, ribosome profiling and bioinformatic tools has enabled a better and more comprehensive discovery of sORFs and their corresponding microproteins[2,56,87,94]. The concurrent identification of sORFs through both Ribo-seq and MS can offer robust evidence for the existence of microproteins, laying the necessary foundation for their subsequent characterization. While integrating these different technologies enabled the annotation of new microproteins, it remains a challenge to characterize and validate these newly discovered sORFs and determine their function in different cellular contexts.

Journal Pre-proof

Translation initiation and regulation of gene expression

Similar to conventional genes, the control of sORF expression involves tight regulation at various stages of the gene expression cascade, including transcriptional and translational control mechanisms. For uORFs, mRNA isoforms can include or exclude uORFs, achieved through alternative transcription start site selection or by alternative splicing, which enables the cell to elegantly regulate uORF-mediated cellular function [95–99]. Long-read sequencing methods, such as those offered by the PacBio and Nanopore platforms, enable comprehensive assessment of mRNA isoforms, including the presence or absence of uORFs, along with the assessment of the full transcript [100–102]. Additional methods geared towards uORF detection and transcription start site selection include the 5’ cap capture methods such as CAGE-seq (Cap Analysis of Gene Expression) to generate snapshots of the 5’ end and the 5’UTR [103,104].

In addition to transcriptional control, translational regulation significantly impacts uORF expression due to the critical role of translation start site recognition by scanning ribosomes [105,106]. Decades of research, particularly focusing on the arguably beststudied uORF-containing gene ATF4, have highlighted how cellular context and changes in the translational machinery influence the recognition of uORFs [24,105].

Translation can be divided into three main steps: initiation, elongation and termination, which includes ribosome recycling for a new round of protein synthesis. Translation initiation is the rate-limiting phase of translation [107–109]. It begins with the formation of the ternary complex (eIF2-GTP-Met-tRNAi), which assembles with the 40S ribosomal subunit to form the preinitiation complex (PIC) to scan the 5' untranslated region (5'UTR) [109,110]. When the PIC reaches the start codon, the GTP in the ternary complex is hydrolyzed and with the release of eIF2-GDP, the 60S ribosomal subunit joins the PIC complex to form the 80S complex and initiate translation [109].

The 5'UTR plays therefore a critical role in ribosome recruitment to mRNA, influencing translation start site selection and initiation [23,105,111,112]. Different structures and elements in the 5'UTR, such as RNA secondary structures, Internal Ribosome Entry Sites (IRES), motifs for RNA binding proteins, single or multiple upstream initiation sites and uORFs can shape the translation of the downstream main ORF [105,111–114]. In the case of the uORFs, translation initiates within the 5'UTR and can represent competition for the PIC to detect the start codon of the main coding sequence, consequently negatively regulating translation of the CDS [61,105,115,116]. Nonetheless, it is important to point out that uORFs do not always repress the translation of the main ORF, as specific conditions such as stress may allow the re-initiation of translation at the main ORF [105,116–119]. Thus, uORFs can add an additional layer of regulation to rapidly boost downstream translation to changes in the cellular environment. Over the last decades, luciferase-based assays in cultured cells have greatly helped in elucidating the regulatory function of uORFs. These approaches could be combined with translation-competent lysates for *in vitro* translation, which enables recapitulation of the key steps of uORF translation and the regulatory role of uORFs with regard to main ORF translation [120].

The sequence context surrounding the initiation codon is important for the translation of uORFs and protein-coding sequences [121]. The PIC recognizes preferentially the correct start codon – usually AUG – in an optimal context known as the Kozak consensus sequence.

Journal Pre-proof

Monitoring the start codon selection is feasible through various techniques, including Global Translation Initiation Sequencing (GTI-seq), which employs a similar principle as the translation start site inhibitor Harringtonine [61,122]. These methods enable genome-wide mapping of translation start sites, which are based on blocking initiating ribosomes while allowing the elongating ribosomes to run off. Coupled with the ribosome profiling protocol, GTI-seq or Harringtonine treatment solely results in translation start site peaks without the reads from elongating ribosomes, therefore providing strong evidence that the start sites of potential sORFs are indeed recognized by the translational machinery. Global translation initiation sequencing (GTI-seq) indicates that approximately 74% of the upstream translation initiation sites (TIS) are non-AUG start codons. CUG is the predominant start codon for uORF, showing a frequency of ~30% compared to ~25% frequency of the conventional AUG TIS [61]. Additionally, conventional ribosome profiling studies similarly suggest that uORFs show a preference for near-cognate start codons, with CUG and GUG being the most frequent TIS [5,27].

Journal Pre-proof

Genome editing using CRISPR

A major challenge in the sORF field is distinguishing functionally relevant sORFs from mere sORF expression. As outlined above, there are multiple approaches to detect and annotate sORFs experimentally and computationally. Consequently, the field will have to progress beyond the essential task of cataloging sORFs and transition toward comprehensive genome-wide analyses to unravel the functional significance of sORFs. For the majority of sORFs, their functions remain untested, necessitating the use of genome editing techniques to explore loss-of-function and gain-of-function effects and analyses of the resulting phenotypes in cell culture studies and *in vivo*. These strategies will be essential in shedding light on the functional roles of sORFs.

The most commonly used and rapidly evolving genome editing technique is the CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) system and their associated Cas endonucleases. These short regularly spaced repeats were found in both bacteria and archaea and are part of their DNA repair system against phages and plasmids [123–125]. Among the different types of immunity, type II can be applied to genome editing [125–128].

The microbial adaptive immune response has been exploited to target any genomic location by using a single guide RNA (sgRNA) in combination with the Cas endonucleases [128]. Endogenous repair mechanisms such as homologous directed repair (HDR) or nonhomologous end joining (NHEJ) are induced by the DSB [127,129,130]. As a result, the DSB can be used to insert, delete, or modify a genomic target. When introducing exogenous DNA fragments as templates for recombination, specific mutations can be introduced. In the absence of a DNA template for HDR, the NHEJ pathway prevails, leading to insertions or deletions introduced to the target locus [131,132].

Altering the sequence of the sgRNA will allow to target any region of interest in the genome. In recent years, optimizing the sequence of the sgRNA has led to improvements in the on-target activity and a reduction of off-target effects [133]. Different tools can be used to design sgRNAs and predict their target activity, however, these tools commonly rely on canonical genes as reference, posing challenges for targeting smaller genes like microproteins. In the human genome, the overall frequency of finding a 'GG' is 5.21%, which means that the 'GG' dinucleotide, critical for the sgRNA design, is found approximately every 42 bases [134]. For smaller genes, such as microproteins, this issue significantly reduces the chances of finding a PAM site and designing good sgRNAs to target them. Considering that the median size of an uORF is 48-78 nucleotides [23,27,35,135,136] and the median size of a translated human sORF from a long non-coding RNA is 72 nucleotides [35], there may be, on average, only 2-4 PAM motifs per microprotein (considering both strands). This notion aligns with our own experience and the design of typically 2-3 sgRNAs per sORF. Furthermore, the design of CRISPR screens can be particularly challenging for sORFs that overlap other genetic elements, such as altORFs. It is important to note that any phenotype attributed to the disruption of an sORF should be complemented by for example reintroducing the sORF.

Over the past decade, numerous studies have demonstrated the remarkable efficiency of the CRISPR/Cas system for genome editing. As interest in the technique grew, different Cas endonucleases have been discovered, modified and used for genome editing. CRISPR interference (CRISPRi) employs a modified version of the Cas9

Journal Pre-proof

protein, the dead Cas9 (dCas9), which is fused with an interference domain and is used to silence the expression of a gene. On the other hand, CRISPR activation (CRISPRa) utilizes a dCas9 that is fused to an activation domain, such as for example the dCas9-SunTag system, to turn on gene expression. Additionally, the CRISPR/Cas12a and Cas12b systems function similarly to CRISPR/Cas9 but exhibit different PAM sequences and editing efficiencies, introduce sticky ends instead of blunt ends, and cut at distinct sites relative to the PAM sequence [137–139]. These variations make Cas12a and Cas12b valuable additions to the CRISPR platform, particularly in targeting sORFs that may have been resistant to Cas9 gene editing due to low editing efficiency or missing NGG PAM sequence in the sORF locus.

Journal Pre-proof

CRISPR screens

CRISPR screens have emerged as a powerful tool to dissect the function of genes. The two main types of CRISPR screens, arrayed and pooled screens, have become instrumental in addressing diverse biological questions (Figure 3). Serving as an unbiased interrogation of gene function, CRISPR screens introduce perturbations into cells, which subsequently reveal cellular phenotypes [140].

In arrayed screens, individual reagents are synthesized and distributed into multi-well plates and are therefore spatially separated. As each well introduces a distinct perturbation, the approach enables the identification of the specific sgRNA responsible for the gene perturbation without the need for sequencing. Arrayed screens can have advantages, such as the possibility to couple them with microscopy-based high-content screenings, but are also laborious and therefore result in lower throughput [133,141,142]. On the other hand, pooled CRISPR screens provide a scalable and powerful platform and allow the targeting of multiple genes using a library of pooled sgRNAs. This library is delivered to cells through lentiviral transduction, resulting in cells harboring single sgRNAs that integrate into the cell's DNA and edit the targeted gene based on the sgRNA sequence. Subsequently, these perturbed cells are subjected to selective pressure or monitored over multiple passages. At the end of the experiment, sequencing and sgRNA identification enable us to infer gene function by calculating the representation of the sgRNAs in the library [141].

In a recent study, 553 non-canonical ORFs were comprehensively analyzed using a CRISPR screen coupled to single-cell RNA sequencing [143]. Among the targeted ORFs, 386 were identified as sORFs with less than 100 amino acids and resided either upstream or downstream of known protein-coding genes and on long non-coding RNAs (lncRNAs). By performing a CRISPR loss-of-function screen in eight different cell lines, the authors observed viability phenotypes in 10% of the targeted ORFs [143]. Subsequent analyses of the 13 top-scoring ORFs provided valuable insights into their functional role in cancer cell survival [143]. This study exemplifies therefore the power of CRISPR screenings in deciphering the function of microproteins as they enable an unbiased approach to simultaneously target them and determine their function in different biological contexts.

The main advantage of CRISPR screens lies in their high throughput and scalability, which enables the simultaneous interrogation of thousands of genes or microproteins. While most of these screens have been conducted *in vitro* using various cell lines, one significant drawback is their inability to account for environmental factors and cellular interactions between different cell types. To address these limitations, different approaches have been taken to design *in vivo* screens in mouse models, which can be challenging to set up but provide the opportunity to assess the consequences of sORF perturbations within a living organism. More recent *in vivo* screens include for example screens to identify modulators for tumor growth or immunotherapy targets [144,145].

Single-cell CRISPR screens

While pooled CRISPR screenings are undeniably powerful, they have the important limitation that they are restricted to simple readouts such as proliferation or the expression of a marker gene. To address this drawback, a suite of new tools has been developed that enable the coupling of single-cell RNA sequencing with pooled CRISPR screening (Table 2), thereby providing a high-throughput functional dissection of genes with single-cell transcriptomic readout.

In a conventional pooled CRISPR screen, it is not possible to identify which sgRNA is expressed in each cell. The main issue is that the sgRNA, being processed by the human U6 RNA Polymerase III (RNAP III), will not undergo posttranscriptional

Journal Pre-proof

modifications, including the polyadenylation, making it incompatible with the RNA-sequencing techniques [146]. To solve this issue, the sgRNA in each cell can be identified via a Polymerase II transcribed barcode used in Perturb-seq, CRISP-seq and Mosaic-seq methods [147–149], or by detecting the sgRNA within the Pol II transcript used in the CROP-seq method [150]. In the past six years, many more techniques have been developed, allowing for the integration of CRISPR with next-generation sequencing. The advent of single-cell CRISPR screens offers the opportunity to investigate gene function within the context of regulatory pathways and holds great promise for investigating microproteins at single-cell transcriptomic resolution *in vitro* and *in vivo* [147–151].

Only a few large-scale functional characterization screens of microproteins have been performed by using single-cell CRISPR screenings. In a recent study, Chen and colleagues combined Ribo-seq and mass spectrometry techniques to annotate and generate a library of non-canonical ORFs. Initially, a pooled CRISPR screen was performed, leading to the identification of over 500 potential targets exhibiting a significant proliferation phenotype. To gain deeper insights into the role of non-canonical ORFs, a second step involved a Perturb-seq screen, focusing on 83 uORFs and 80 lncRNAs, to assess the transcriptomic changes resulting from the loss-of-function of the sORF [2]. These analyses revealed sORF functions in different cellular pathways suggesting that sORFs play diverse cellular roles and highlighting the power of single-cell CRISPR screenings to analyze the function of microproteins [2]. Another elegant example of such a screening strategy was a study aimed at identifying regulators of zygote genome activation (ZGA)-like transcription in mouse embryonic stem cells, which exploited a modified CROP-seq vector used for a CRISPRa library coupled with single-cell transcriptomics [152]. Out of the 230 genes that were assessed, 24 were identified to have a ZGA-like signature and 9 of those genes were independently validated as ZGA-like transcription regulators [152].

Single-cell CRISPR screens are especially powerful when carried out *in vivo* as they have the potential to interrogate gene function simultaneously in different tissue cell types. To date, only a few *in vivo* single-cell CRISPR screens have been performed. In 2020, the Perturb-seq vector was used for a CRISPR screen *in vivo* targeting 35 risk genes for autism spectrum disorder and developmental delay (ASD/ND)[153]. The

Journal Pre-proof

in vivo Perturb-seq method was able to target the five main different cell types in the brain and uncover common pathways targeted by multiple perturbations [153]. More recently, an *in vivo* immune screen was performed using the CROP-seq vector to elucidate tumor immune evasion mechanisms[154]. These studies expand the power of genetic screens into biological and disease models in mammals, facilitating the understanding of tissue-wide gene function. These advanced single-cell CRISPR technologies provide therefore an attractive set of tools for targeting sORFs and hold great promise in understanding sORF function in different disease contexts.

Conclusion

The rapid advancement in the field of sORFs has been driven by a variety of innovative technologies and computational approaches that emerged rather recently, offering opportunities for functional characterization and understanding of the role of microproteins. The development of various *in silico* tools and pipelines has revolutionized the annotation and identification of sORFs, leading to the expansion of the standardized catalog of these elusive coding sequences in eukaryotic genomes. Ribosome profiling and mass spectrometry-based approaches helped tremendously to identify and validate sORFs by providing direct experimental evidence that they are translated into stable microproteins. As demonstrated in recent studies, the utilization of CRISPR-based systems, especially when coupled with single-cell RNA sequencing, enables comprehensive and systematic analyses of sORF expression and function. We expect that large-scale CRISPR screens will continue to expand the functional repertoire of microproteins. These tools hold great promise for dissecting sORF function in different cellular contexts and unraveling their role in the pathogenesis of disease.

Nevertheless, the field faces challenges in identifying functionally relevant sORFs and sifting through the catalog of thousands of potential sORFs in the mammalian genome to tease apart relevant sORFs from those without a clear cellular function. Further investigations utilizing loss-of-function and gain-of-function approaches in relevant cellular contexts will be crucial for elucidating the functional significance of these microproteins. As sORFs continue to attract growing interest, it is evident that the continued integration of new methodologies will pave the way for new discoveries in this fascinating field of microproteins. Ultimately, unraveling the diverse cellular roles

Journal Pre-proof

of sORFs will have profound implications, shedding light on regulatory mechanisms and uncovering new therapeutic targets for a wide range of diseases.

Bibliography

1. Ruiz-Orera, J., and Albà, M.M. (2019). Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation. Trends Genet. *35*, 186–198. 10.1016/j.tig.2018.12.003.
2. Chen, J., Brunner, A.-D., Cogan, J.Z., Nuñez, J.K., Fields, A.P., Adamson, B., Itzhak, D.N., Li, J.Y., Mann, M., Leonetti, M.D., et al. (2020). Pervasive functional translation of noncanonical human open reading frames. Science (80-.). *367*, 140–146. 10.1126/science.aav5912.
3. McGillivray, P., Ault, R., Pawashe, M., Kitchen, R., Balasubramanian, S., and Gerstein, M. (2018). A comprehensive catalog of predicted functional upstream open reading frames in humans. Nucleic Acids Res. *46*, 3326–3338. 10.1093/nar/gky188.
4. Bazzini, A.A., Johnstone, T.G., Christiano, R., MacKowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C., et al. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. *33*, 981–993. 10.1002/embj.201488411.
5. 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. 10.1016/j.cell.2011.10.002.
6. Chew, G.L., Pauli, A., and Schier, A.F. (2016). Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. Nat. Commun. *7*, 1–10. 10.1038/ncomms11663.
7. Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L., and Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat. Chem. Biol. *9*, 59–64. 10.1038/nchembio.1120.
8. Sieber, P., Platzer, M., and Schuster, S. (2018). The Definition of Open Reading Frame Revisited. Trends Genet. *34*, 167–170. 10.1016/j.tig.2017.12.009.
9. Basrai, M.A., Hieter, P., and Boeke, J.D. (1997). Small Open Reading Frames: Beautiful Needles in the Haystack. Genome Res. *7*, 768–771. 10.1101/GR.7.8.768.
10. Mackowiak, S.D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L.,

Journal Pre-proof

Mastrobuoni, G., Rajewsky, N., Kempa, S., Selbach, M., et al. (2015). Extensive identification and analysis of conserved small ORFs in animals. Genome Biol. *16*, 1–21. 10.1186/s13059-015-0742-x.

11. Wu, Q., Medina, S.G., Kushawah, G., Devore, M.L., Castellano, L.A., Hand, J.M., Wright, M., and Bazzini, A.A. (2019). Translation affects mRNA stability in a codon-dependent manner in human cells. Elife *8*, 1–22. 10.7554/eLife.45396.
12. Couso, J.P., and Patraquim, P. (2017). Classification and function of small open reading frames. Nat. Rev. Mol. Cell Biol. *18*, 575–589. 10.1038/nrm.2017.58.
13. Pueyo, J.I., Magny, E.G., and Couso, J.P. (2016). New Peptides Under the s(ORF)ace of the Genome. Trends Biochem. Sci. *41*, 665–678. 10.1016/j.tibs.2016.05.003.
14. Khitun, A., Ness, T.J., and Slavoff, S.A. (2019). Small open reading frames and cellular stress responses. Mol. Omi. *15*, 108–116. 10.1039/c8mo00283e.
15. Mudge, J.M., Ruiz-Orera, J., Prensner, J.R., Brunet, M.A., Calvet, F., Jungreis, I., Gonzalez, J.M., Magrane, M., Martinez, T.F., Schulz, J.F., et al. (2022). Standardized annotation of translated open reading frames. Nat. Biotechnol. 2022 407 *40*, 994–999. 10.1038/s41587-022-01369-0.
16. Silva, J., Fernandes, R., and Romão, L. (2019). Translational Regulation by Upstream Open Reading Frames and Human Diseases. In The mRNA Metabolism in Human Disease, L. Romão, ed., pp. 99–116. 10.1007/978-3-030-19966-1.
17. 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. J. Exp. Med. *217*. 10.1084/jem_20190950.
18. Huang, J.Z., Chen, M., Chen, D., Gao, X.C., Zhu, S., Huang, H., Hu, M., Zhu, H., and Yan, G.R. (2017). A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. Mol. Cell *68*, 171-184.e6. 10.1016/j.molcel.2017.09.015.
19. Polycarpou-Schwarz, M., Groß, M., Mestdagh, P., Schott, J., Grund, S.E., Hildenbrand, C., Rom, J., Aulmann, S., Sinn, H.P., Vandesompele, J., et al. (2018). The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. Oncogene *37*, 4750–4768. 10.1038/s41388-018-0281-5.
20. Sriram, A., Bohlen, J., and Teleman, A.A. (2018). Translation acrobatics: how cancer cells exploit alternate modes of translational initiation. EMBO Rep. *19*. 10.15252/embr.201845947.
21. Somers, J., Pöyry, T., and Willis, A.E. (2013). A perspective on mammalian upstream open reading frame function. Int. J. Biochem. Cell Biol. *45*, 1690–1700. 10.1016/j.biocel.2013.04.020.

Journal Pre-proof

22. Young, S.K., and Wek, R.C. (2016). Upstream open reading frames differentially regulate genespecific translation in the integrated stress response. J. Biol. Chem. *291*, 16927–16935. 10.1074/jbc.R116.733899.

23. Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc. Natl. Acad. Sci. U. S. A. *106*, 7507–7512. 10.1073/pnas.0810916106.

24. Vattem, K.M., and Wek, R.C. (2004). Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. Proc. Natl. Acad. Sci. U. S. A. *101*, 11269–11274. 10.1073/pnas.0400541101.

25. Pakos-Zebrucka, K., Koryga, I., Mnich, K., Ljujic, M., Samali, A., and Gorman, A.M. (2016). The integrated stress response. EMBO Rep. *17*, 1374–1395. 10.15252/embr.201642195.

26. Zismanov, V., Chichkov, V., Colangelo, V., Jamet, S., Wang, S., Syme, A., Koromilas, A.E., and Crist, C. (2016). Phosphorylation of eIF2α is a Translational Control Mechanism Regulating Muscle Stem Cell Quiescence and Self-Renewal. Cell Stem Cell *18*, 79–90. 10.1016/j.stem.2015.09.020.

27. Sendoel, A., Dunn, J.G., Rodriguez, E.H., Naik, S., Gomez, N.C., Hurwitz, B., Levorse, J., Dill, B.D., Schramek, D., Molina, H., et al. (2017). Translation from unconventional 5′ start sites drives tumour initiation. Nature *541*, 494–499. 10.1038/nature21036.

28. Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. Science (80-.). *309*, 1559–1563. 10.1126/science.1112014.

29. Leong, A.Z.X., Lee, P.Y., Mohtar, M.A., Syafruddin, S.E., Pung, Y.F., and Low, T.Y. (2022). Short open reading frames (sORFs) and microproteins: an update on their identification and validation measures. J. Biomed. Sci. 2022 291 *29*, 1–15. 10.1186/S12929-022-00802-5.

30. Makarewich, C.A., and Olson, E.N. (2017). Mining for Micropeptides. Trends Cell Biol. *27*, 685–696. 10.1016/j.tcb.2017.04.006.

31. Frith, M.C., Forrest, A.R., Nourbakhsh, E., Pang, K.C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T.L., and Grimmond, S.M. (2006). The abundance of short proteins in the mammalian proteome. PLoS Genet. *2*, 515–528. 10.1371/journal.pgen.0020052.

32. Skovgaard, M., Jensen, L.J., Brunak, S., Ussery, D., and Krogh, A. (2001). On the total number of genes and their length distribution in complete microbial genomes. Trends Genet. *17*, 425–428. 0.1016/s0168-9525(01)02372-1.

33. Couso, J.P. (2015). Finding smORFs: Getting closer. Genome Biol. *16*, 15–17.

Journal Pre-proof

10.1186/s13059-015-0765-3.

34. Wright, B.W., Yi, Z., Weissman, J.S., and Chen, J. (2022). The dark proteome: translation from noncanonical open reading frames. Trends Cell Biol. *32*, 243–258. 0.1016/J.TCB.2021.10.010.

35. Couso, J.P., and Patraquim, P. (2017). Classification and function of small open reading frames. Nat. Rev. Mol. Cell Biol. 2017 189 *18*, 575–589. 10.1038/nrm.2017.58.

36. Sandmann, C.-L., Schulz, J.F., Ruiz-Orera, J., Kirchner, M., Ziehm, M., Adami, E., Marczenke, M., Christ, A., Liebe, N., Greiner, J., et al. (2023). Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. Mol. Cell *0*. 10.1016/j.molcel.2023.01.023.

37. Lin, M.F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics *27*. 10.1093/bioinformatics/btr209.

38. Mudge, J.M., Jungreis, I., Hunt, T., Gonzalez, J.M., Wright, J.C., 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 PhyloCSF helps elucidate 118 GWAS loci. Genome Res. *29*. 10.1101/gr.246462.118.

39. Cheng, H., Chan, W.S., Li, Z., Wang, D., Liu, S., and Zhou, Y. (2011). Small Open Reading Frames: Current Prediction Techniques and Future Prospect. Curr. Protein Pept. Sci. *12*, 503. 10.2174/138920311796957667.

40. Kute, P.M., Soukarieh, O., Tjeldnes, H., Trégouët, D.A., and Valen, E. Small Open Reading Frames, How to Find Them and Determine Their Function. Front. Genet. *12*, 2903. 10.3389/fgene.2021.796060.

41. Brunet, M.A., Brunelle, M., Lucier, J.F., Delcourt, V., Levesque, M., Grenier, F., Samandi, S., Leblanc, S., Aguilar, J.D., Dufour, P., et al. (2019). OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. Nucleic Acids Res. *47*, D403–D410. 10.1093/nar/gky936.

42. Fields, A.P., Rodriguez, E.H., Jovanovic, M., Stern-Ginossar, N., Haas, B.J., Mertins, P., Raychowdhury, R., Hacohen, N., Carr, S.A., Ingolia, N.T., et al. (2015). A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. Mol. Cell *60*, 816–827. 10.1016/j.molcel.2015.11.013.

43. Brunet, M.A., Lucier, J.F., Levesque, M., Leblanc, S., Jacques, J.F., Al-Saedi, H.R.H., Guilloy, N., Grenier, F., Avino, M., Fournier, I., et al. (2021). OpenProt 2021: Deeper functional annotation of the coding potential of eukaryotic genomes. Nucleic Acids Res. *49*, D380–D388. 10.1093/nar/gkaa1036.

Journal Pre-proof

44. Hanada, K., Akiyama, K., Sakurai, T., Toyoda, T., Shinozaki, K., and Shiu, S.-H. (2010). sORF finder: a program package to identify small open reading frames with high coding potential. Bioinformatics *26*, 399–400. 10.1093/bioinformatics/btp688.

45. Skarshewski, A., Stanton-Cook, M., Huber, T., Al Mansoori, S., Smith, R., Beatson, S.A., and Rothnagel, J.A. (2014). UPEPperoni: An online tool for upstream open reading frame location and analysis of transcript conservation. BMC Bioinformatics *15*, 1–6. 10.1186/1471-2105-15-36.

46. Olexiouk, V., Crappé, J., Verbruggen, S., Verhegen, K., Martens, L., and Menschaert, G. (2016). SORFs.org: A repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. *44*, D324–D329. 10.1093/nar/gkv1175.

47. Li, Y., Zhou, H., Chen, X., Zheng, Y., Kang, Q., Hao, D., Zhang, L., Song, T., Luo, H., Hao, 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.

48. Camargo, A.P., Sourkov, V., Pereira, G.A.G., and Carazzolle, M.F. (2020). RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences. *2*. 10.1093/nargab/lqz024.

49. Zhang, Y., Jia, C., Fullwood, M.J., and Kwoh, C.K. (2021). DeepCPP: a deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for RNA coding potential prediction. Brief. Bioinform. *22*, 2073–2084. 10.1093/bib/bbaa039.

50. Zhu, M., and Gribskov, M. (2019). MiPepid: MicroPeptide identification tool using machine learning. BMC Bioinformatics *20*, 1–11. 10.1186/s12859-019-3033-9.

51. Zhang, M., Zhao, J., Li, C., Ge, F., Wu, J., Jiang, B., Song, J., and Song, X. (2022). csORF-finder: an effective ensemble learning framework for accurate identification of multi-species coding short open reading frames. Brief. Bioinform. *23*. 10.1093/bib/bbac392.

52. Varabyou, A., Erdogdu, B., Salzberg, S.L., and Pertea, M. (2023). Investigating open reading frames in known and novel transcripts using ORFanage. Nat. Comput. Sci. 2023, 1–9. 10.1038/s43588-023-00496-1.

53. Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., and Weissman, J.S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc. *7*, 1534–1550. 10.1038/nprot.2012.086.

54. Ingolia, N.T. (2016). Ribosome Footprint Profiling of Translation throughout the Genome. Cell *165*, 22–33. 10.1016/j.cell.2016.02.066.

55. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009).

Journal Pre-proof

Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science (80-.). *324*, 218–223. 10.1126/science.1168978.

56. Martinez, T.F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M.N., and Saghatelian, A. (2020). Accurate annotation of human protein-coding small open reading frames. Nat. Chem. Biol. *16*, 458–468. 10.1038/S41589-019-0425-0.

57. Aspden, J.L., Eyre-Walker, Y.C., Phillips, R.J., Amin, U., Mumtaz, M.A.S., Brocard, M., and Couso, J.P. (2014). Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. Elife *3*, 1–19. 10.7554/eLife.03528.

58. Ji, Z., Song, R., Regev, A., and Struhl, K. (2015). Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife *4*, 1–21. 10.7554/eLife.08890.

59. Chothani, S.P., Adami, E., Widjaja, A.A., Langley, S.R., Viswanathan, S., Pua, C.J., Zhihao, N.T., Harmston, N., D'Agostino, G., Whiffin, N., et al. (2022). A high-resolution map of human RNA translation. Mol. Cell *82*, 2885-2899.e8. 10.1016/j.molcel.2022.06.023.

60. Michel, A.M., and Baranov, P. V. (2013). Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. Wiley Interdiscip. Rev. RNA *4*, 473–490. 10.1002/wrna.1172.

61. Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B., and Qian, S.-B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc. Natl. Acad. Sci. U. S. A. *109*, E2424–E2432. 10.1073/pnas.1207846109.

62. Dunn, J.G., and Weissman, J.S. (2016). Plastid: Nucleotide-resolution analysis of next-generation sequencing and genomics data. BMC Genomics. 10.1186/s12864-016-3278-x.

63. Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., and Ohler, U. (2016). Detecting actively translated open reading frames in ribosome profiling data. Nat. Methods. 10.1038/nmeth.3688.

64. Xu, Z., Hu, L., Shi, B., Geng, S., Xu, L., Wang, D., and Lu, Z.J. (2018). Ribosome elongating footprints denoised by wavelet transform comprehensively characterize dynamic cellular translation events. Nucleic Acids Res. *46*. 10.1093/nar/gky533.

65. Raj, A., Wang, S.H., Shim, H., Harpak, A., Li, Y.I., Engelmann, B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2016). Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. Elife *5*. 10.7554/eLife.13328.

66. Xiao, Z., Huang, R., Xing, X., Chen, Y., Deng, H., and Yang, X. (2018). De novo annotation and characterization of the translatome with ribosome profiling data.

Journal Pre-proof

Nucleic Acids Res. *46*, e61. 10.1093/nar/gky179.

67. Clauwaert, J., Menschaert, G., and Waegeman, W. (2019). DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns. Nucleic Acids Res. *47*, e36. 10.1093/nar/gkz061.

68. Calviello, L., Hirsekorn, A., and Ohler, U. (2020). Quantification of translation uncovers the functions of the alternative transcriptome. Nat. Struct. Mol. Biol. 2020 278 *27*, 717–725. 10.1038/s41594-020-0450-4.

69. Choudhary, S., Li, W., and Smith, A.D. (2020). Accurate detection of short and long active ORFs using Ribo-seq data. Bioinformatics *36*, 2053–2059. 10.1093/bioinformatics/btz878.

70. Ruiz-Orera, J., Messeguer, X., Subirana, J.A., and Alba, M.M. (2014). Long non-coding RNAs as a source of new peptides. Elife *3*. 10.7554/ELIFE.03523.

71. Ma, J., Ward, C.C., Jungreis, I., Slavoff, S.A., Schwaid, A.G., Neveu, J., Budnik, B.A., Kellis, M., and Saghatelian, A. (2014). Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue. J. Proteome Res. *13*, 1757–1765. 10.1021/pr401280w.

72. Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nat. 2014 5097502 *509*, 575–581. 10.1038/nature13302.

73. Oyama, M., Kozuka-Hata, H., Suzuki, Y., Semba, K., Yamamoto, T., and Sugano, S. (2007). Diversity of transplantation start sites may define increased complexity of the human short ORFeome. Mol. Cell. Proteomics *6*, 1000–1006. 10.1074/mcp.M600297-MCP200.

74. Bartel, J., Varadarajan, A.R., Sura, T., Ahrens, C.H., Maaß, S., and Becher, D. (2020). Optimized proteomics workflow for the detection of small proteins. J. Proteome Res. *19*, 4004–4018. 10.1021/acs.jproteome.0c00286.

75. Müller, S.A., Kohajda, T., Findeiß, S., Stadler, P.F., Washietl, S., Kellis, M., von Bergen, M., and Kalkhof, S. (2010). Optimization of parameters for coverage of low molecular weight proteins. Anal. Bioanal. Chem. *398*, 2867–2881. 10.1007/s00216-010-4093-x.

76. Kaulich, P.T., Cassidy, L., Bartel, J., Schmitz, R.A., and Tholey, A. (2021). Multi-protease Approach for the Improved Identification and Molecular Characterization of Small Proteins and Short Open Reading Frame-Encoded Peptides. J. Proteome Res. *20*, 2895–2903. 10.1021/acs.jproteome.1c00115.

77. Cassidy, L., Kaulich, P.T., Maaß, S., Bartel, J., Becher, D., and Tholey, A. (2021). Bottom-up and top-down proteomic approaches for the identification, characterization, and quantification of the low molecular weight proteome with focus on short open

Journal Pre-proof

reading frame-encoded peptides. Proteomics *21*. 10.1002/pmic.202100008.

78. Ma, J., Diedrich, J.K., Jungreis, I., Donaldson, C., Vaughan, J., Kellis, M., Yates, John R., I., and Saghatelian, A. (2016). Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. Anal. Chem. *88*, 3967–3975. 10.1021/acs.analchem.6b00191.

79. Khitun, A., and Slavoff, S.A. (2019). Proteomic Detection and Validation of Translated Small Open Reading Frames. Curr. Protoc. Chem. Biol. *11*, e77. 10.1002/cpch.77.

80. He, C., Jia, C., Zhang, Y., and Xu, P. (2018). Enrichment-Based Proteogenomics Identifies Microproteins, Missing Proteins, and Novel smORFs in Saccharomyces cerevisiae. J. Proteome Res. *17*, 2335–2344. 10.1021/acs.jproteome.8b00032.

81. Ahrens, C.H., Wade, J.T., Champion, M.M., and Langer, J.D. (2022). A Practical Guide to Small Protein Discovery and Characterization Using Mass Spectrometry. J. Bacteriol. *204*. 10.1128/jb.00353-21.

82. Cassidy, L., Kaulich, P.T., and Tholey, A. (2023). Proteoforms expand the world of microproteins and short open reading frame-encoded peptides. iScience *26*, 106069. 10.1016/J.ISCI.2023.106069.

83. Vanderperre, B., Lucier, J.F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., Wisztorski, M., Salzet, M., Boisvert, F.M., and Roucou, X. (2013). Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. PLoS One *8*, e70698. 10.1371/journal.pone.0070698.

84. Schlesinger, D., and Elsässer, S.J. (2021). Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. FEBS J. 10.1111/febs.15769.

85. 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. 10.1093/nar/gkz646.

86. Crappé, 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. 10.1093/NAR/GKU1283.

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

88. Na, Z., Dai, X., Zheng, S.-J., Bryant, C.J., Loh, K.H., Su, H., Luo, Y., Buhagiar, A.F., Cao, X., Baserga, S.J., et al. (2022). Mapping subcellular localizations of unannotated microproteins and alternative proteins with MicroID. Mol. Cell *82*, 2900-2911.e7. 10.1016/j.molcel.2022.06.035.

Journal Pre-proof

89. Kamiyama, D., Sekine, S., Barsi-Rhyne, B., Hu, J., Chen, B., Gilbert, L.A., Ishikawa, H., Leonetti, M.D., Marshall, W.F., Weissman, J.S., et al. (2016). Versatile protein tagging in cells with split fluorescent protein. Nat. Commun. *7*. 10.1038/ncomms11046.

90. Feng, S., Varshney, A., Coto Villa, D., Modavi, C., Kohler, J., Farah, F., Zhou, S., Ali, N., Müller, J.D., Van Hoven, M.K., et al. (2019). Bright split red fluorescent proteins for the visualization of endogenous proteins and synapses. Commun. Biol. 2019 21 *2*, 1–12. 10.1038/s42003-019-0589-x.

91. Cao, X., Khitun, A., Na, Z., Dumitrescu, D.G., Kubica, M., Olatunji, E., and Slavoff, S.A. (2020). Comparative Proteomic Profiling of Unannotated Microproteins and Alternative Proteins in Human Cell Lines. J. Proteome Res. *19*, 3418–3426. 10.1021/acs.jproteome.0c00254.

92. Lafranchi, L., Schlesinger, D., Kimler, K.J., and Elsässer, S.J. (2020). Universal Single-Residue Terminal Labels for Fluorescent Live Cell Imaging of Microproteins. J. Am. Chem. Soc. *142*, 20080–20087. 10.1021/jacs.0c09574.

93. Chen, Y., Cao, X., Loh, K.H., and Slavoff, S.A. (2023). Chemical labeling and proteomics for characterization of unannotated small and alternative open reading frame-encoded polypeptides. Biochem. Soc. Trans. *51*, 1071–1082. 10.1042/BST20221074.

94. van Heesch, S., Witte, F., Schneider-Lunitz, V., Schulz, J.F., Adami, E., Faber, A.B., Kirchner, M., Maatz, H., Blachut, S., Sandmann, C.L., et al. (2019). The Translational Landscape of the Human Heart. Cell *178*, 242–260. 10.1016/j.cell.2019.05.010.

95. Cheng, Z., Otto, G.M., Powers, E.N., Keskin, A., Mertins, P., Carr, S.A., Jovanovic, M., and Brar, G.A. (2018). Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis. Cell *172*, 910–923. 10.1016/j.cell.2018.01.035.

96. Weber, R., Ghoshdastider, U., Spies, D., Duré, C., Valdivia-Francia, F., Forny, M., Ormiston, M., Renz, P.F., Taborsky, D., Yigit, M., et al. (2023). Monitoring the 5’UTR landscape reveals isoform switches to drive translational efficiencies in cancer. Oncogene *42*, 638–650. 10.1038/S41388-022-02578-2.

97. Hollerer, I., Barker, J.C., Jorgensen, V., Tresenrider, A., Dugast-Darzacq, C., Chan, L.Y., Darzacq, X., Tjian, R., Ünal, E., and Brar, G.A. (2019). Evidence for an integrated gene repression mechanism based on mRNA isoform toggling in human cells. G3 Genes, Genomes, Genet. *9*, 1045–1053. 10.1534/g3.118.200802.

98. Floor, S.N., and Doudna, J.A. (2016). Tunable protein synthesis by transcript isoforms in human cells. Elife *5*. 10.7554/eLife.10921.

99. Arribere, J.A., and Gilbert, W. V. (2013). Roles for transcript leaders in translation and

Journal Pre-proof

mRNA decay revealed by transcript leader sequencing. Genome Res. *23*, 977–987. 10.1101/GR.150342.112.

100. Rhoads, A., and Au, K.F. (2015). PacBio Sequencing and Its Applications. Genomics. Proteomics Bioinformatics *13*, 278–289. 10.1016/j.gpb.2015.08.002.

101. Kono, N., and Arakawa, K. (2019). Nanopore sequencing: Review of potential applications in functional genomics. Dev. Growth Differ. *61*, 316–326. 10.1111/DGD.12608.

102. De Coster, W., Weissensteiner, M.H., and Sedlazeck, F.J. (2021). Towards population-scale long-read sequencing. Nat. Rev. Genet. *22*, 572–587. 10.1038/S41576-021-00367-3.

103. Haberle, V., Forrest, A.R.R., Hayashizaki, Y., Carninci, P., and Lenhard, B. (2015). CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. Nucleic Acids Res. *43*, e51–e51. 10.1093/NAR/GKV054.

104. Murata, M., Nishiyori-Sueki, H., Kojima-Ishiyama, M., Carninci, P., Hayashizaki, Y., and Itoh, M. (2014). Detecting expressed genes using CAGE. Methods Mol. Biol. *1164*, 67–85. 10.1007/978-1-4939-0805-9_7.

105. Hinnebusch, A.G., Ivanov, I.P., and Sonenberg, N. (2016). Translational control by 5′-untranslated regions of eukaryotic mRNAs. Science (80-.). *352*, 1413 LP – 1416. 10.1126/science.aad9868.

106. Hinnebusch, A.G. (2017). Structural Insights into the Mechanism of Scanning and Start Codon Recognition in Eukaryotic Translation Initiation. Trends Biochem. Sci. 10.1016/j.tibs.2017.03.004.

107. Andreev, D.E., O'Connor, P.B.F., Loughran, G., Dmitriev, S.E., Baranov, P. V., and Shatsky, I.N. (2017). Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. Nucleic Acids Res. *45*, 513–526. 10.1093/NAR/GKW1190.

108. Hershey, J.W.B., Sonenberg, N., and Mathews, M.B. (2012). Principles of translational control: An overview. Cold Spring Harb. Perspect. Biol. *4*. 10.1101/cshperspect.a009829.

109. Sonenberg, N., and Hinnebusch, A.G. (2009). Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. Cell *136*, 731–745. 10.1016/j.cell.2009.01.042.

110. Hao, P., Yu, J., Ward, R., Liu, Y., Hao, Q., An, S., and Xu, T. (2020). Eukaryotic translation initiation factors as promising targets in cancer therapy. Cell Commun. Signal. *18*, 1–20. 10.1186/s12964-020-00607-9.

111. Zhang, H., Wang, Y., and Lu, J. (2019). Function and Evolution of Upstream ORFs in Eukaryotes. Trends Biochem. Sci., 1–13. 10.1016/j.tibs.2019.03.002.

112. Schuster, S.L., and Hsieh, A.C. (2019). The Untranslated Regions of mRNAs in

Journal Pre-proof

Cancer. Trends in Cancer *5*, 245–262. 10.1016/j.trecan.2019.02.011.

113. Sajjanar, B., Deb, R., Raina, S.K., Pawar, S., Brahmane, M.P., Nirmale, A. V., Kurade, N.P., Manjunathareddy, G.B., Bal, S.K., and Singh, N.P. (2017). Untranslated regions (UTRs) orchestrate translation reprogramming in cellular stress responses. J. Therm. Biol. *65*, 69–75. 10.1016/j.jtherbio.2017.02.006.

114. Lacerda, R., Menezes, J., and Romão, L. (2017). More than just scanning: the importance of cap-independent mRNA translation initiation for cellular stress response and cancer. Cell. Mol. Life Sci. *74*, 1659–1680. 10.1007/s00018-016-2428-2.

115. Morris, D.R., and Geballe, A.P. (2000). Upstream Open Reading Frames as Regulators of mRNA Translation. Mol. Cell. Biol. *20*, 8635–8642. 10.1128/mcb.20.23.8635-8642.2000.

116. Wethmar, K. (2014). The regulatory potential of upstream open reading frames in eukaryotic gene expression. Wiley Interdiscip. Rev. RNA *5*, 765–778. 10.1002/wrna.1245.

117. Silva, J., Fernandes, R., and Romão, L. (2017). Gene expression regulation by upstream open reading frames in rare diseases. J. Rare Dis. Res. Treat. *2*, 33–38. 10.29245/2572-9411/2017/4.1121.

118. Araujo, P.R., Yoon, K., Ko, D., Smith, A.D., Qiao, M., Suresh, U., Burns, S.C., and Penalva, L.O.F. (2012). Before it gets started: Regulating translation at the 5′; UTR. Comp. Funct. Genomics *2012*. 10.1155/2012/475731.

119. Jackson, R.J., Hellen, C.U.T., and Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. Nat. Rev. Mol. Cell Biol. *11*, 113–127. 10.1038/nrm2838.

120. Gurzeler, L.A., Ziegelmüller, J., Mühlemann, O., and Karousis, E.D. (2022). Production of human translation-competent lysates using dual centrifugation. RNA Biol. *19*, 78–88. 10.1080/15476286.2021.2014695.

121. Kozak, M. (2002). Pushing the limits of the scanning mechanism for initiation of translation. Gene *299*, 1–34. 10.1016/S0378-1119(02)01056-9.

122. Fresno, M., Jiménez, A., and Vázquez, D. (1977). Inhibition of Translation in Eukaryotic Systems by Harringtonine. Eur. J. Biochem. 10.1111/j.1432-1033.1977.tb11256.x.

123. Mojica, F.J.M., Díez-Villaseñor, C., Soria, E., and Juez, G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Mol. Microbiol. *36*, 244–246. 10.1046/J.1365-2958.2000.01838.X.

124. Makarova, K.S., Aravind, L., Grishin, N. V., Rogozin, I.B., and Koonin, E. V. (2002). A DNA repair system specific for thermophilic Archaea and bacteria predicted by

Journal Pre-proof

genomic context analysis. Nucleic Acids Res. *30*, 482–496. 10.1093/NAR/30.2.482.

125. Barrangou, R., and Marraffini, L.A. (2014). CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity. Mol. Cell *54*, 234–244. 10.1016/J.MOLCEL.2014.03.011.

126. Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J.M., Wolf, Y.I., Yakunin, A.F., et al. (2011). Evolution and classification of the CRISPR–Cas systems. Nat. Rev. Microbiol. 2011 96 *9*, 467–477. 10.1038/nrmicro2577.

127. Wright, A. V., Nuñez, J.K., and Doudna, J.A. (2016). Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell *164*, 29–44. 10.1016/J.CELL.2015.12.035.

128. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science (80-.). *337*, 816–821. 10.1126/science.1225829.

129. Doudna, J.A., and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. Science (80-.). *346*. 10.1126/science.1258096.

130. Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. *8*, 2281–2308. 10.1038/NPROT.2013.143.

131. Weterings, E., and Chen, D.J. (2008). The endless tale of non-homologous end-joining. Cell Res. 2008 181 *18*, 114–124. 10.1038/cr.2008.3.

132. Jiang, W., and Marraffini, L.A. (2015). CRISPR-Cas: New Tools for Genetic Manipulations from Bacterial Immunity Systems. https://doi.org/10.1146/annurev-micro-091014-104441 *69*, 209–228. 10.1146/annurev-micro-091014-104441.

133. Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. *34*, 184–191. 10.1038/nbt.3437.

134. Scherer, S. (2008). A Short Guide to the Human Genome (Cold Spring Harbor Laboratory Press).

135. Rodriguez, C.M., Chun, S.Y., Mills, R.E., and Todd, P.K. (2019). Translation of upstream open reading frames in a model of neuronal differentiation. BMC Genomics *20*, 1–18. 10.1186/s12864-019-5775-1.

136. Renz, P.F., Valdivia Francia, F., and Sendoel, A. (2020). Some like it translated: small ORFs in the 5′UTR. Exp. Cell Res. *396*, 112229. 10.1016/j.yexcr.2020.112229.

137. Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., Van Der Oost, J., Regev, A., et al. (2015).

Journal Pre-proof

Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell *163*, 759–771. 10.1016/j.cell.2015.09.038.

138. Strecker, J., Jones, S., Koopal, B., Schmid-Burgk, J., Zetsche, B., Gao, L., Makarova, K.S., Koonin, E. V., and Zhang, F. (2019). Engineering of CRISPR-Cas12b for human genome editing. Nat. Commun. *10*. 10.1038/S41467-018-08224-4.

139. Paul, B., and Montoya, G. (2020). CRISPR-Cas12a: Functional overview and applications. Biomed. J. *43*, 8–17. 10.1016/J.BJ.2019.10.005.

140. Shalem, O., Sanjana, N.E., and Zhang, F. (2015). High-throughput functional genomics using CRISPR–Cas9. Nat. Rev. Genet. 2015 165 *16*, 299–311. 10.1038/nrg3899.

141. Bock, C., Datlinger, P., Chardon, F., Coelho, M.A., Dong, M.B., Lawson, K.A., Lu, T., Maroc, L., Norman, T.M., Song, B., et al. (2022). High-content CRISPR screening. Nat. Rev. Methods Prim. 2022 21 *2*, 1–23. 10.1038/s43586-021-00093-4.

142. Hanna, R.E., and Doench, J.G. (2020). Design and analysis of CRISPR-Cas experiments. Nat. Biotechnol. *38*, 813–823. 10.1038/S41587-020-0490-7.

143. Prensner, J.R., Enache, O.M., Luria, V., Krug, K., Clauser, K.R., Dempster, J.M., Karger, A., Wang, L., Stumbraite, K., Wang, V.M., et al. (2021). Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *39*, 697–704. 10.1038/s41587-020-00806-2.

144. Braun, C.J., Bruno, P.M., Horlbeck, M.A., Gilbert, L.A., Weissman, J.S., and Hemann, M.T. (2016). Versatile in vivo regulation of tumor phenotypes by dCas9-mediated transcriptional perturbation. Proc. Natl. Acad. Sci. U. S. A. *113*, E3892–E3900. 10.1073/pnas.1600582113.

145. Manguso, R.T., Pope, H.W., Zimmer, M.D., Brown, F.D., Yates, K.B., Miller, B.C., Collins, N.B., Bi, K., La Fleur, M.W., Juneja, V.R., et al. (2017). In vivo CRISPR screening identifies Ptpn2 as a cancer immunotherapy target. Nat. 2017 5477664 *547*, 413–418. 10.1038/nature23270.

146. Nowak, C.M., Lawson, S., Zerez, M., and Bleris, L. (2016). Guide RNA engineering for versatile Cas9 functionality. Nucleic Acids Res. *44*, 9555–9564. 10.1093/nar/gkw908.

147. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell *167*, 1853-1866.e17. 10.1016/j.cell.2016.11.038.

148. Jaitin, D.A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T.M., Tanay, A., van Oudenaarden, A., and Amit, I. (2016). Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. Cell *167*,

Journal Pre-proof

1883-1896.e15. 10.1016/J.CELL.2016.11.039.

149. Xie, S., Duan, J., Li, B., Zhou, P., and Hon, G.C. (2017). Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. Mol. Cell *66*, 285-299.e5. 10.1016/J.MOLCEL.2017.03.007.

150. Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. Nat. Methods *14*, 297–301. 10.1038/nmeth.4177.

151. Cheng, J., Lin, G., Wang, T., Wang, Y.Y., Guo, W., Liao, J., Yang, P., Chen, J., Shao, X., Lu, X., et al. (2022). Massively Parallel CRISPR-Based Genetic Perturbation Screening at Single-Cell Resolution. Adv. Sci. *10*, 2204484. 10.1002/ADVS.202204484.

152. Alda-Catalinas, C., Bredikhin, D., Hernando-Herraez, I., Santos, F., Kubinyecz, O., Eckersley-Maslin, M.A., Stegle, O., and Reik, W. (2020). A Single-Cell Transcriptomics CRISPR-Activation Screen Identifies Epigenetic Regulators of the Zygotic Genome Activation Program. Cell Syst. *11*, 25-41.e9. 10.1016/j.cels.2020.06.004.

153. Jin, X., Simmons, S.K., Guo, A., Shetty, A.S., Ko, M., Nguyen, L., Jokhi, V., Robinson, E., Oyler, P., Curry, N., et al. (2020). In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. Science (80-.). *370*.

154. Hou, J., Liang, S., Xu, C., Wei, Y., Wang, Y., Tan, Y., Sahni, N., McGrail, D.J., Bernatchez, C., Davies, M., et al. (2022). Single-cell CRISPR immune screens reveal immunological roles of tumor intrinsic factors. NAR Cancer *4*. 10.1093/narcan/zcac038.

Journal Pre-proof

Figure legends

Figure 1: The classification of sORFs

Schematic representation of different open reading frames (ORFs) and their genomic location. A large fraction of the mammalian genome is composed of small open reading frames (sORFs) in the untranslated regions (red). Canonical ORFs, conventionally more than 100 amino acids long, are depicted at the top, with exons delimited with a known start and stop codon flanked by 5' untranslated region (UTR) and 3'UTR. The mammalian genome encodes transcribed and potentially functional sORFs between 10 and 100 amino acids, which can be classified according to their genomic location. Upstream open reading frames (uORFs) are found in the 5'UTR of conventional ORFs, while downstream ORFs (dORFs) are found in the 3'UTR of conventional ORFs. In some cases, alternative ORFs arise from alternative initiation start sites within canonical ORFs and lead to shorter isoforms of a known ORF. sORFs can also be found in intronic regions of canonical ORFs and in intergenic regions between two canonical ORFs, known as intronic and intergenic sORFs, respectively. Finally, an important source of sORFs are long non-coding ORFs.

Figure 2: Identification of sORFs

Schematic workflow of the different methods used for the identification of small ORFs. Samples from diverse sources, human biopsies, mouse cells and cultured cells can be processed using ribosome profiling (Ribo-seq), mass spectrometry (MS) and/or computational approaches. Ribo-seq captures snapshots of ribosome-protected fragments that are purified and sequenced. Small ORFs showing 3-nucleotide periodicity are most likely to be translated into microproteins. Microproteins can be extracted, digested, fractionated and enriched by size selection followed by proteomics. Data are searched against custom databases containing the potential sORFs. Computational approaches to determine sORFs rely on predictions based on the conservation between species, codon bias and coding potential and transcriptomic and proteomic data analysis. The different algorithms can predict the presence of sORF based on detecting similarity to known proteins or domains, nucleotide composition, codon substitution or machine learning approaches.

Figure 3: Targeting sORFs using CRISPR

Schematic representation of the CRISPR screening workflow. Top panel: For pooled CRISPR screens, the sgRNA library is transduced into Cas9-expressing cells *in vitro*. Cells are harvested at the end of the experiment (e.g. following a certain number of passages or treatment) and submitted to sequencing. The enrichment and depletion of the sgRNAs is then used to infer gene function. Middle panel: Arrayed CRISPR screens are carried out in different

Journal Pre-proof

wells, where one sgRNA is targeted per well. In an arrayed screen, the phenotype can be linked directly to the sgRNA to determine gene function. Lower panel: single-cell CRISPR screens *in vitro* and *in vivo*. Similar to pooled CRISPR screens, cells are transduced with a pooled library. Single cells are then subjected to single-cell RNA-seq to obtain the transcriptomic readout coupled to cell-type specific sgRNA representation. In an *in vivo* single-cell CRISPR screen, the sgRNA library is delivered, for example, directly into mouse embryos or adult mice. At a later time point, the organ of interest is collected, and cells are isolated for single-cell RNA-seq, which can determine proliferative changes and the transcriptomic consequences of the sgRNA in different cell types.

Acknowledgments.

The Sendoel lab was supported by an SNSF Professorship grant (grant number 176825), by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 759006), by the Swiss Cancer Research foundation (KFS-5023-02-2020-R), the National Center of Competence in Research (NCCR) on RNA and Disease funded by the SNSF and the UZH Candoc Grant 2022 from the University of Zurich (Nr. FK-22-039).

Author Contributions:

Conceptualization, FVF and AS; Writing - Original Draft, FVF and AS; Writing - Review and Editing, FVF and AS; Visualization, FVF and AS; All authors contributed to the article and approved the submitted version.

Declaration of Interests.

The authors declare no competing financial interests.

Journal Pre-proof

Table 1: Bioinformatic tools

Methods based on sequence prediction		Reference
PhyloCSF	Codon substitution and conservation elements	36
OpenProt	Uses public available datasets to asses Ribo-seq and MS	41
RNASamba	Neuroal network to predict sORFs, recognizes Kozak sequence	44
DeepCPP	Algorithm using nucleotide bias to predict sORFs	46
MiPepid	Machine learning tool using for identificaton of sORFs based on known sORFs	49
csORF-finder	Uses trinucleotide deviation from expected mean to distinguish between coding sORFs from non-coding sORFs	50
ORFanage	Pseudo alignment algorithm for the detection of sORFs from RNA-seq data	51
sORF finder	Detection of sORFs according to 3-nucleotide composition bias	43
micPDP	Search for sORFs based on codon substitutions observed in whole-genome alignments	4
uPEPperoni	Detection of uORFs based on location and transcript conservation	45
sORFs.org	Database of sORFs based on Ribo-seq data and integrates MS evidence and conservation searches	47
SmPROT	sORFs reported in literature, databases, Ribo-seq and MS data	49
Methods based on ribosome profiling		
ORF-Rater	ORF Regression algorithm using Ribo-seq data to quantify translation regardless of start codon, overlap and length	40
ORFscore	Codon in-frame reads	4
ORFquant	Annotation and quantification of tranlation level of ORFs considering multiple transcript isoforms	71
RiboTaper	Identification of translated regions based on 3-nucleotide periodicity	62
Ribotricer	Identification of translating ORFs based on 3-nucleotide periodicity	73
RiboWave	Uses wavelet transform and 3-nucleotide periodicity to located the P-site	63
RiboCode	*De novo* annotation of the translatome using 3-nucleotide periodicity	65
DeepRibo	Neuronal network using ribo-seq data to determin binding patterns and translation initiation sites	66
riboHMM	Identification of coding sequences based on abundance and codon pediodicity	64

Table 1: Compilation of different bioinformatic tools, divided according to sequence prediction and ribosome profiling methods, used for the predictions and identification of small open reading frames.

Journal Pre-proof

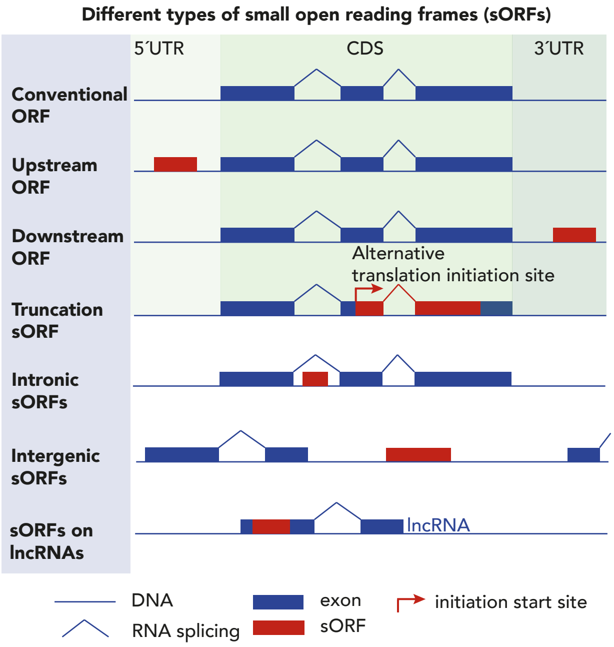
Table 2: Single-cell CRISPR methods

Method		Reference
Perturb-seq	The Perturb-seq method combines a pooled CRISPR screen with single-cell RNA-seq by a guide barcode (GBC) expressed for each perturbation.	143
CRISP-seq:	The CRISP-seq technique allows the identification of the sgRNA that infects each individual cell by using a vector that contains the gRNA sequence and a transcribed polyadenylated unique guide index with a fluorescent selection marker.	142
Mosaic-seq: mosaic single-cell analysis by indexed CRISPR sequencing	Mosaic-seq couples CRISPRi with single-cell RNA sequencing. The library of vectors targeting enhancers also carries a unique barcode that allows the identification of the sgRNA.	144
CROP-seq: CRISPR droplet sequencing	The CROP-seq method, unlike the other methods, does not pair the sgRNA with a barcode. Instead, the CROP-seq method uses the sgRNA as a barcode overlapping the Pol II transcript.	145

Table 2: Summary of the different single-cell CRISPR methods currently available with a small description of the principle of the method.

Journal Pre-proof

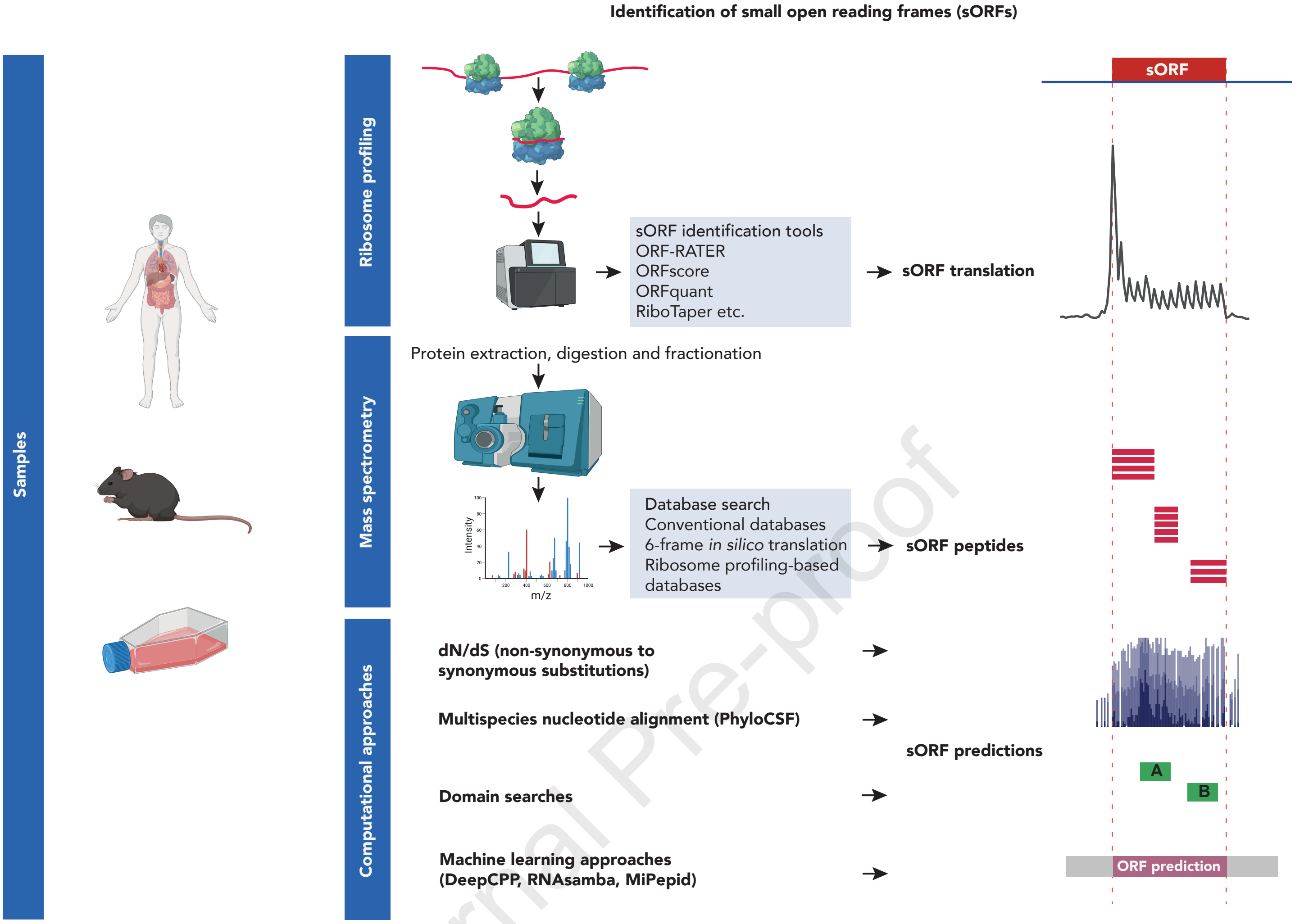
Figure 1:



Journal Pre-proof

Journal Pre-proof

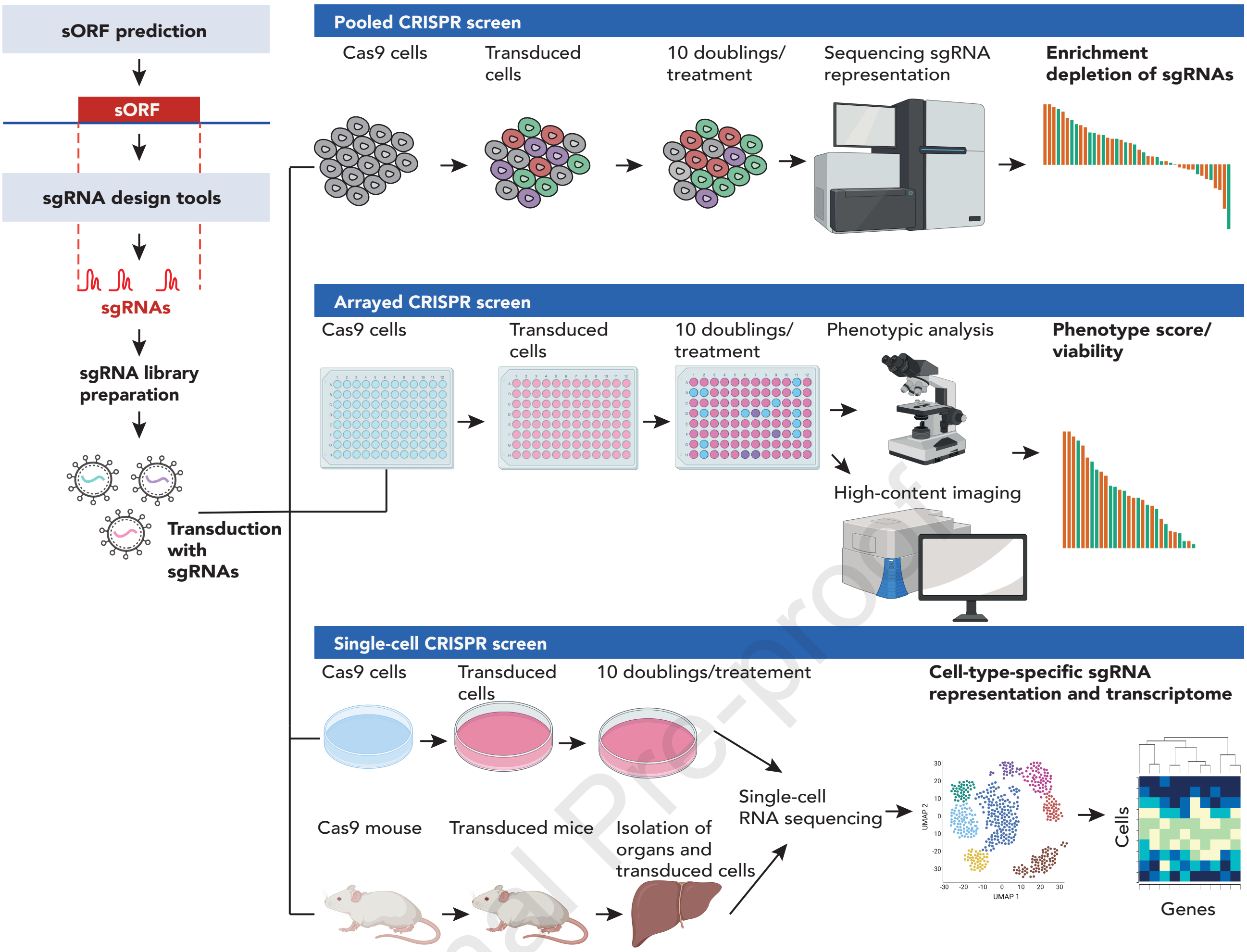
Figure 2:



Identification of small open reading frames (sORFs)
Samples
Ribosome profiling
sORF identification tools
ORF-RATER
ORFscore
ORFquant
RiboTaper etc.
sORF translation
sORF
Mass spectrometry
Protein extraction, digestion and fractionation
Intensity
m/z
Database search
Conventional databases
6-frame *in silico* translation
Ribosome profiling-based databases
sORF peptides
Computational approaches
dN/dS (non-synonymous to synonymous substitutions)
Multispecies nucleotide alignment (PhyloCSF)
Domain searches
Machine learning approaches (DeepCPP, RNAsamba, MiPepid)
sORF predictions
A
B
ORF prediction

Journal Pre-proof

Figure 3:



Journal Pre-proof

Highlights:

- Microproteins are encoded from small open reading frames in noncanonical regions
- Bioinformatics, Ribo-Seq, and proteomics led to the annotation of >7000 sORFs
- CRISPR screens are a powerful tool for identifying the function of microproteins

Journal Pre-proof