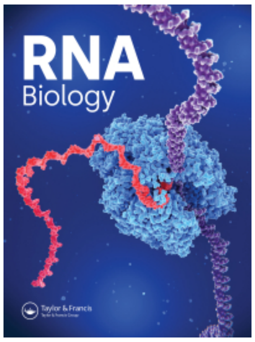

RNA Biology

ISSN: 1547-6286 (Print) 1555-8584 (Online) Journal homepage: www.tandfonline.com/journals/krnb20

Discovering microproteins: making the most of ribosome profiling data

Sonia Chothani, Lena Ho, Sebastian Schafer & Owen Rackham

To cite this article: Sonia Chothani, Lena Ho, Sebastian Schafer & Owen Rackham (2023) Discovering microproteins: making the most of ribosome profiling data, RNA Biology, 20:1, 943-954, DOI: 10.1080/15476286.2023.2279845

To link to this article: https://doi.org/10.1080/15476286.2023.2279845

RNA BIOLOGY
2023, VOL. 20, NO. 1, 943–954
https://doi.org/10.1080/15476286.2023.2279845

Taylor & Francis
Taylor & Francis Group

REVIEW

OPEN ACCESS

Discovering microproteins: making the most of ribosome profiling data

Sonia Chothani[a], Lena Ho[a], Sebastian Schafer[a], and Owen Rackham[a,b,c]

[a]Program in Cardiovascular and Metabolic Disorders, Duke-National University of Singapore, Singapore; [b]School of Biological Sciences, University of Southampton, Southampton, UK; [c]The Alan Turing Institute, The British Library, London, UK

ABSTRACT

Building a reference set of protein-coding open reading frames (ORFs) has revolutionized biological process discovery and understanding. Traditionally, gene models have been confirmed using cDNA sequencing and encoded translated regions inferred using sequence-based detection of start and stop combinations longer than 100 amino-acids to prevent false positives. This has led to small ORFs (smORFs) and their encoded proteins left un-annotated. Ribo-seq allows deciphering translated regions from untranslated irrespective of the length. In this review, we describe the power of Ribo-seq data in detection of smORFs while discussing the major challenge posed by data-quality, -depth and -sparseness in identifying the start and end of smORF translation. In particular, we outline smORF cataloguing efforts in humans and the large differences that have arisen due to variation in data, methods and assumptions. Although current versions of smORF reference sets can already be used as a powerful tool for hypothesis generation, we recommend that future editions should consider these data limitations and adopt unified processing for the community to establish a canonical catalogue of translated smORFs.

ARTICLE HISTORY
Revised 28 August 2023
Accepted 30 October 2023

KEYWORDS
Ribo-seq; ribosome profiling; smorfs; Seps; RNA translation

Small open reading frames: missed key players in biology

To date, there are ~ 22,000 genes annotated in Ensembl [1] that contain an open reading frame (ORF, known ORFs) and which are considered ‘protein coding’. However, the bioinformatic process by which these annotations were made included assumptions about ORF length. This was done in order to account for the fact that start and stop codons appear at random in the genome and as such there are many millions of genomic loci that have ORF-like characteristics [2] (i.e. they are between a start and a stop codon), but only a fraction of these are translated. As such, in the past, to ensure reliable prediction of ORFs a threshold of ~ 100 amino acids has often been used to ensure the validity of predicted ORFs. Not including this assumption would have resulted in the annotation of many millions of possible small ORFs (smORFs), most of which would have been false positives. However, as a consequence we have systematically missed those ORFs that are shorter than 100 amino acids, despite there being no biological reason for their exclusion (see Figure 1).

Evidence is increasingly accumulating that we may have underestimated the prevalence of smORFs. Genes previously annotated as non-coding (e.g. long-non coding RNAs or lncRNA) have been found to be frequently associated with mono- and polysomal complexes [3], suggesting many more may be translated. Indeed, dozens of smORF encoded peptides (SEPs) have been found in lncRNAs playing a role in a diverse range of biological functions (referred to as novel unannotated ORFs or nuORFs). Early studies showed that translation of smORFs located upstream of known ORFs (referred to as upstream ORFs or uORFs) could play a cis-regulatory effect (both negative and positive) [4] on the host ORF, and more recently uORFs have also been shown to encode functional peptides [5–7]. Equally, small open reading frames have been identified downstream of ORFs (known as dORFs). Taken together, these discoveries raise the question of to what extent might a significant subset of the millions of possible small ORFs (smORFs) in the ‘dark matter’ of the genome that were previously disregarded are actually translated?

A step-change in smORF identification came with the introduction of Ribosome profiling (or also called Ribo-seq) [8,9]. This technology can generate a snapshot of ribosome locations across the transcriptome with single nucleotide resolution. For the first time, Ribo-seq provided experimental evidence to estimate global translation levels in vivo [10–12]. Importantly, Ribo-seq was able to demonstrate prevalent binding of ribosomes to RNA outside of the known coding regions. However, there have been disagreements in the field as to what extent of the previously annotated non-coding regions are translated [13,14]. As a result, the challenge that remains is how to accurately distinguish active translation from ribosome occupancy. Specifically, we outline the data, methods and assumptions used in development of reference sets of translated smORFs in humans, the current versions of such sets and the overlap across them. We highlight the relative sparseness of data despite the large number of studies, and following from this, argue for combining multiple Ribo-seq datasets to improve signal-to-noise ratios for smORF identification. We also caution against simply aggregating predictions into consensus smORF

CONTACT Owen Rackham owen.rackham@duke-nus.edu.sg Program in Cardiovascular and Metabolic Disorders, Duke-National University of Singapore, 169857, Singapore

Supplemental data for this article can be accessed online at https://doi.org/10.1080/15476286.2023.2279845

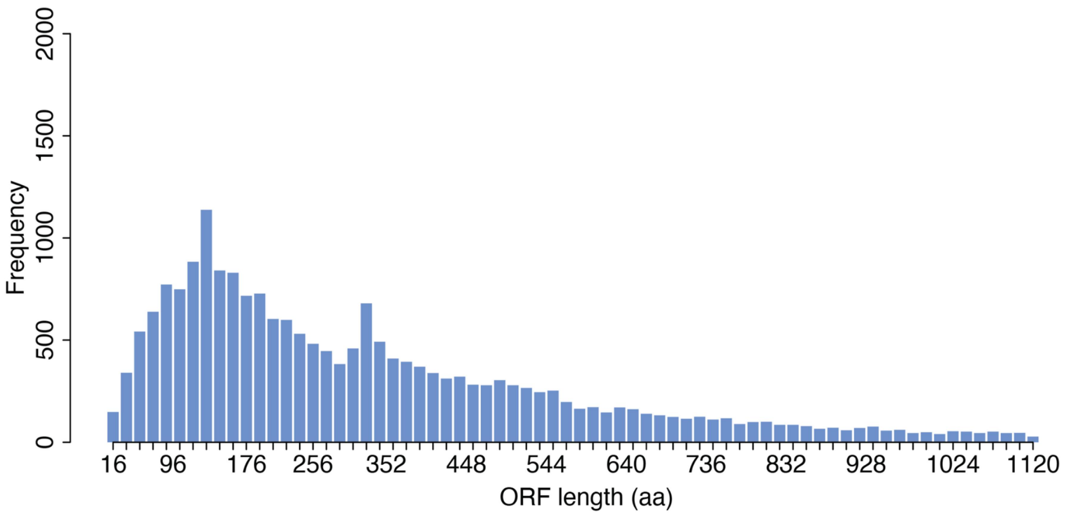

Figure 1. Length distribution of known protein-coding ORFs. Barplot showing frequency of ORFs annotated in Ensembl (hg38) according to their length. Frequency of ORFs longer than 1200 amino-acids are not shown.

reference sets without standardizing the smORF annotation pipeline. Moving forward, we recommend pooling Ribo-seq data and adopting unified processing standards to establish a high-quality, consensus smORF catalogue for the research community. It should be noted that computational tools to detect translated smORFs, classification of smORFs and exemplar SEPs have been reviewed in detail previously [15–17] and as such will not be covered in detail here.

The 3-nt periodicity observed using Ribo-seq has revolutionized smORF detection

Many RNA-ribosome interactions are unrelated to translation, and thus polysome profiling is not ideal to discover translated ORFs [18]. Detection of smORFs using Ribo-seq is potentially more accurate and has been a very active field of research since its discovery. Initial studies have shown that ribosome footprints can be found in the known ORFs but also within regions that had previously been annotated as 'non-coding'. However, these initial studies did not take advantage of the single-nucleotide resolution of Ribo-seq data [19]. Active translation of mRNA leads to ribosome footprints with inferred P-sites on the first nucleotide of every codon (or every three base-pairs)-, thus forming a three-nucleotide periodic signal (or periodicity, see Figure 2) observed at a nucleotide resolution. In contrast, a random ribosomal occurrence (i.e. one that is unrelated to mRNA translation) leads to footprints in all nucleotides uniformly. Periodicity observed by combining the ribosome protected fragments (RPFs) around the start and stop of known ORFs has been routinely used to examine the data quality [12,20–23]. Subsequently, scanning the transcriptome for regions with 3-nt periodicity has been at the heart of smORF detection [24].

Current catalogues present a wide-range and inconsistent sets of smORFs

Numerous catalogues listing smORFs using existing Ribo-seq data either to detect or to validate translation have been generated but disparities in data and methods used to identify actively translated smORFs can lead to large differences in the annotation of translated smORFs. Across the seven catalogues [25–32] described (discussed in detail below) in this review, we found that less than 50% of the smORFs could be found in at least two catalogues when using the host gene ID as a reference. This overlap was reduced to ~ 27% for two catalogues and less than 10% for three catalogues when using the exact stop-site position to test for repeated identification (see Figure 3). Although repeated identification in independent catalogues do not necessarily provide unequivocal evidence for the veracity of smORF, and likewise a smORF found only in one catalogue may still be bona fide, the question remains as to how large is the consensus and what should be considered a reference set for future studies. The lack of a consensus reference set such as those for known ORFs (i.e. Uniprot [33], Refseq [34] and Ensembl [1]) creates challenges for researchers trying to understand the function and clinical utility of smORFs.

In an ideal-scenario, translated smORFs would be identified based on continuous 3-nt periodicity from their start-to-end (as in Figure 2) using Ribo-seq as evidence of their translation. In reality, currently available individual sample data are very sparse with uneven coverage across ORFs. The computational methods being developed to detect smORFs have numerous approaches for accounting for this sparseness, with varying levels of stringency. As a result, large differences in smORF detection can occur simply by fluctuations in data quality and smORF detection algorithms, something that is often overlooked. As the field moves towards constructing a consensus set of smORFs these aspects must be considered and accounted for. For this reason, this review will focus on the challenges faced by the community and best practices that the field should consider adopting to reach the goal of a canonical smORF catalogue.

Are the underlying data of high-resolution or not?

Ribo-seq data provide an unprecedented resolution of translation but the available data is very sparse both due to technical

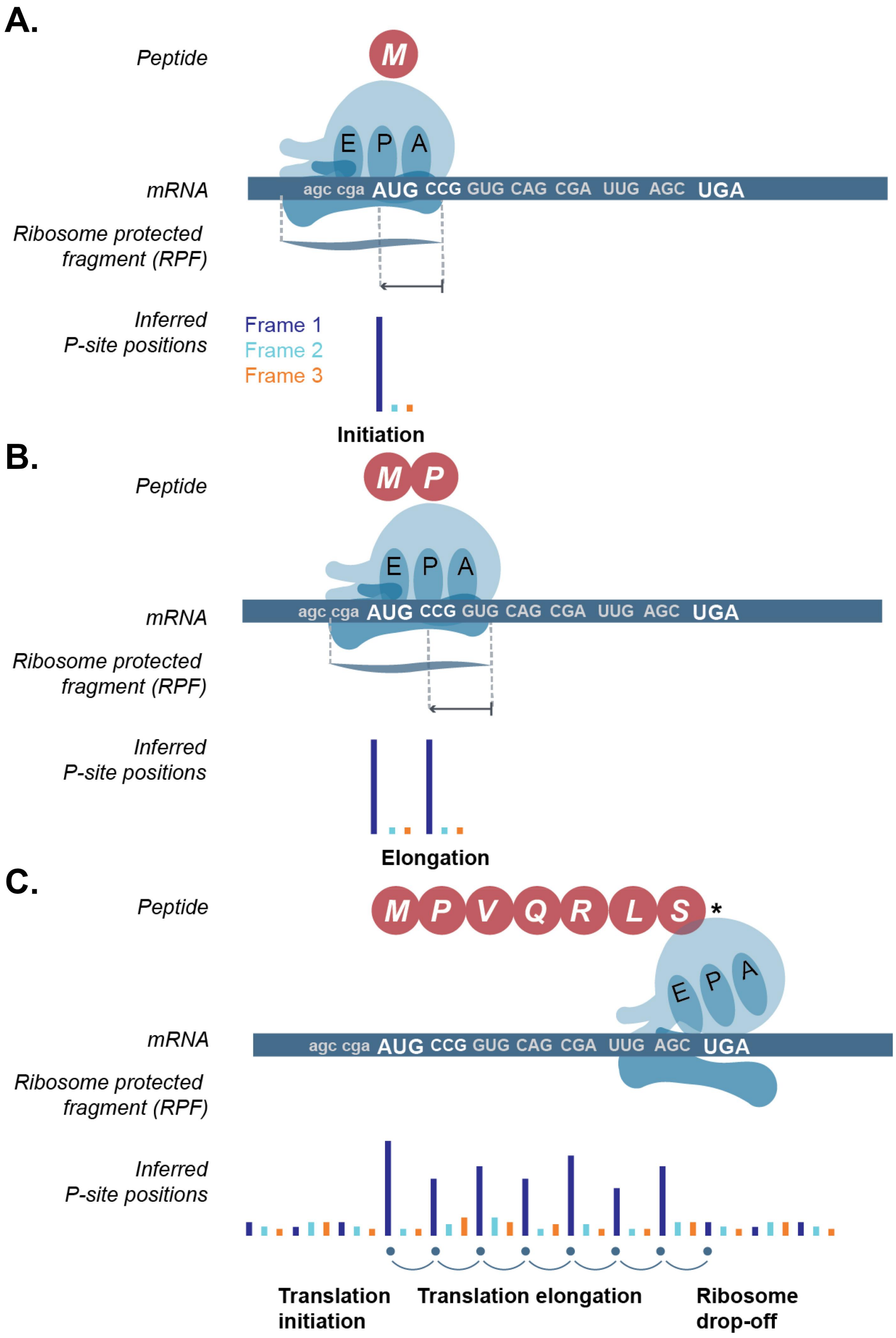
A.
Peptide
M
E P A
mRNA
agc cga AUG CCG GUG CAG CGA UUG AGC UGA
Ribosome protected fragment (RPF)
Inferred P-site positions
Frame 1
Frame 2
Frame 3
Initiation
B.
Peptide
M P
E P A
mRNA
agc cga AUG CCG GUG CAG CGA UUG AGC UGA
Ribosome protected fragment (RPF)
Inferred P-site positions
Elongation
C.
Peptide
M P V Q R L S *
E P A
mRNA
agc cga AUG CCG GUG CAG CGA UUG AGC UGA
Ribosome protected fragment (RPF)
Inferred P-site positions
Translation initiation
Translation elongation
Ribosome drop-off

Figure 2. Three-nucleotide periodicity profile of ribosome footprints during active translation of an open reading frame (ORF). A. Schematic showing translation initiation on the canonical start codon (AUG, methionine) of the ORF, B. translation elongation i.e. translation of the subsequent codons (only one codon shown) of the ORF after initiation, and, C. translation termination with ribosome drop-off or disassembly at the stop-codon (UGA here). E-site: exit site, P-site: peptidyl-site, A-site: aminoacyl site. Sequenced ribosome protected fragment (RPF) can be used to infer the position of the P-site and subsequently the codon that is being translated. Inferred P-site position is coloured based on the ORF frame. Frame 1: dark blue, frame 2: light blue, frame 3: orange.

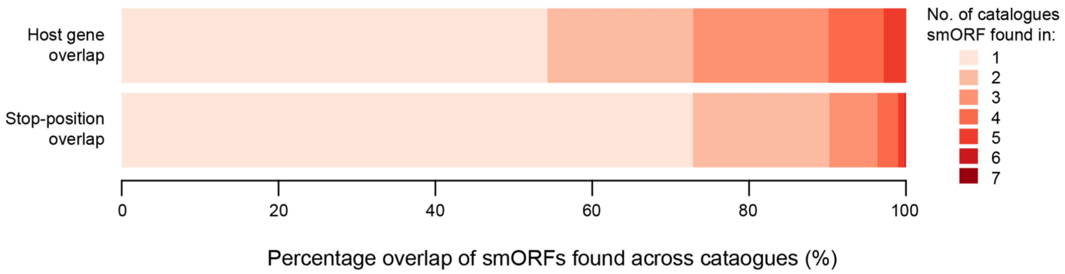

Figure 3. Overlap of smORFs identified across different published catalogues. Stacked barplot showing percentage of smORFs found in various catalogues. A smORF is considered as found in multiple catalogues if it shares the host gene ID (from five catalogues) or shares the same stop-site position (from seven catalogues). Catalogues which did not have host gene symbol information in the downloads file were omitted [28,32] for the host-gene overlap.

and biological limitations. The sample library prep requires high input material and has several optimization steps that can potentially generate distortions as described previously [24]. Thus, sequenced Ribo-seq reads undergo several pre-processing steps including strict quality control and only selected high-quality usable reads are used to infer P-site positions for smORF detection (see Figure 4). Short and low-quality sequenced reads are discarded after trimming sequencing adaptors followed by mapping of remaining reads to a database of contaminant sequences including ribosomal RNA (rRNA) and transfer RNA (tRNA). Such contaminant sequences are very prevalent in Ribo-seq data, making this step very detrimental to the usable read depth. This typically results in as low as only ~ 10% [35] of the sequenced reads to be usable. From the remaining reads about ~ 20–80% are reported as uniquely mapped to the transcriptome in various studies, considering the short read length and varying data quality [36]. Apart from these pre-processing steps, QC such as length distribution and 3-nt periodicity is also tested. The ribosomal footprint during active translation is expected to have a fixed read length range depending on the digestion conditions [37], cellular stress [38] or elongation inhibitor [39]. In a typical Ribo-seq experiment, when cycloheximide is used as the elongation inhibitor, the most prevalent footprints should be ~ 29 nt long in eukaryotic cytosolic ribosomes [8]. Failure to observe these footprints with high 3-nt periodicity in coding sequences may suggest that the data is not reliable for identifying actively translated mRNA. Despite this, many datasets, generated under normal conditions, using cycloheximide, fail to achieve these periodic footprints (see Supp. Figure 1). Overall, a high number of input reads and optimization is required to obtain high-depth and -quality Ribo-seq data that can be used to detect smORF translation accurately and such data are limited.

A low usable read-depth of Ribo-seq leads to low coverage across all codons in ORFs. For instance, in the human genome, there are more than 12 million codons [1] within known proteins. Using a single sample of say 20–30 million sequenced reads typically yields roughly less than, 5–10 million uniquely mapped reads (after QC filters as described above). Therefore, this barely provides one read per codon or P-site location. In reality, this number would be even lower as the coverage is confounded by several factors. First, coverage depends on expression levels and ORFs that are lowly expressed would be more difficult to detect [40]. Second, Ribo-seq typically has non-uniform coverage across codons leading to no-information available for many codons and thus making it impossible to determine if those regions were translated. For instance, the availability of tRNAs can influence the speed of translation by stalled or rapid translation of certain codons, thus leading to higher or lower RPFs mapping to those codons, respectively [41]. Previous studies have shown that there is non-uniform coverage across the length of ORFs [42,43].

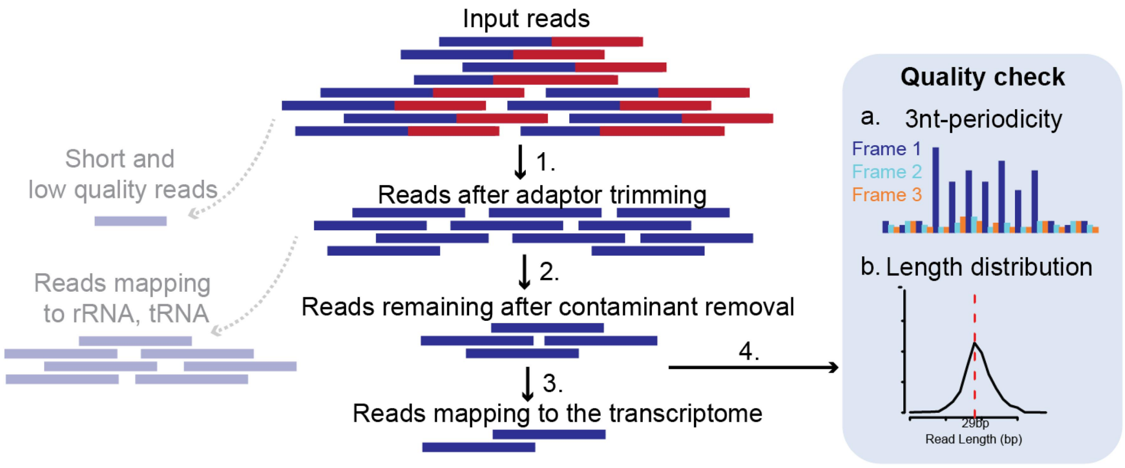
Input reads
Short and low quality reads
1.
Reads after adaptor trimming
Reads mapping to rRNA, tRNA
2.
Reads remaining after contaminant removal
4.
3.
Reads mapping to the transcriptome
Quality check
a. 3nt-periodicity
Frame 1
Frame 2
Frame 3
b. Length distribution
29bp
Read Length (bp)

Figure 4. Pre-processing steps and quality control of sequenced Ribo-seq reads. 1. Sequencing adaptors are trimmed from input reads. Short and low-quality reads are discarded. 2. Contaminant sequences such as reads mapping to ribosomal RNA (rRNA) or transfer RNA (tRNA) are discarded. 3. Remaining reads are mapped to the transcriptome. 4. Quality control steps are carried out such as 3-nt periodicity (a) and length distribution (b).

Theoretically, with sufficient depth and population size, 100% codons would have Ribo-seq read coverage in the translating frame, providing a clear picture of smORF translation (Figure 5A). In reality, various biological and technical biases lead to non-uniform coverage and limited usable read depth. This leads to absence of coverage in several codons. A study showed that in individual Ribo-seq samples, 84% codons are covered only for the top 10 expressed genes while for top 1000 expressed genes, only 36% of the codons are covered [45]. Here, we show an example of a validated uORF encoded protein (SEHBP) [44], using individual sample data only ~ 20% of codons are covered (>1 inferred P-site read) providing no evidence of translation for more than 80% of codons (see Figure 5B). For another example, which is a predicted smORF with no known function, likely a false-positive smORF, also shows only ~ 33% codons covered (Figure 5C). In both the examples, it becomes impossible to confirm the translation of the full length of the smORF to differentiate it from artefacts or false positives as well as accurately define its coordinates. Overall, currently available Ribo-seq samples do not have uniform coverage at the nucleotide-resolution throughout the smORF, thus making it difficult to confirm smORF translation and its coordinates accurately. As such, Ribo-seq can illuminate new translated regions of the genome, but challenges and questions remain as to how best to achieve this.

Is the ORF actively translated?

Several metrics and tools have been developed to detect actively translated smORFs keeping in mind the sparseness of Ribo-seq data. Since there is not enough depth at the nucleotide-resolution, translation of a smORF is often tested by a summarized view of reads mapped to the smORF. For instance, if a smORF is translated, it would be expected to have RPFs of length distribution similar to the distribution within known coding sequences in the given Ribo-seq experiment and the Fragment length organization similarity score (or FLOSS) [14] uses this information to determine a smORFs coding potential. Similarly, if a smORF is translated it is expected to have more reads mapped within the ORF as compared to after its stop codon. The Ribosome release score (or RRS) [13] quantifies the ratio of the reads in the putative smORF and the following 3'UTR normalized to the lengths of the regions and mRNA read depth to test for a smORF's coding potential. TOC classifier [46] uses a random forest classifier that models on scores such as RRS, FLOSS, inside-out (metric to test nucleotides covered by Ribo-seq inside the ORF and outside the ORF) and translational efficiency.

The above metrics look for features of translation holistically, but they do not confirm the continuous translation of the smORF. The ideal scenario for smORF detection would be observing periodic RPFs over every codon in the smORF. However, because of the shallowness in a typical Ribo-seq library, observing such 3nt-periodicity over every codon of an ORF is rare. For this reason, many of the published methods use strategies to overcome this, for example, combining the observations across codons or alleviating the shallowness. For example, Ribotaper [47] uses a multitaper strategy followed by a Fourier transform as a way to detect

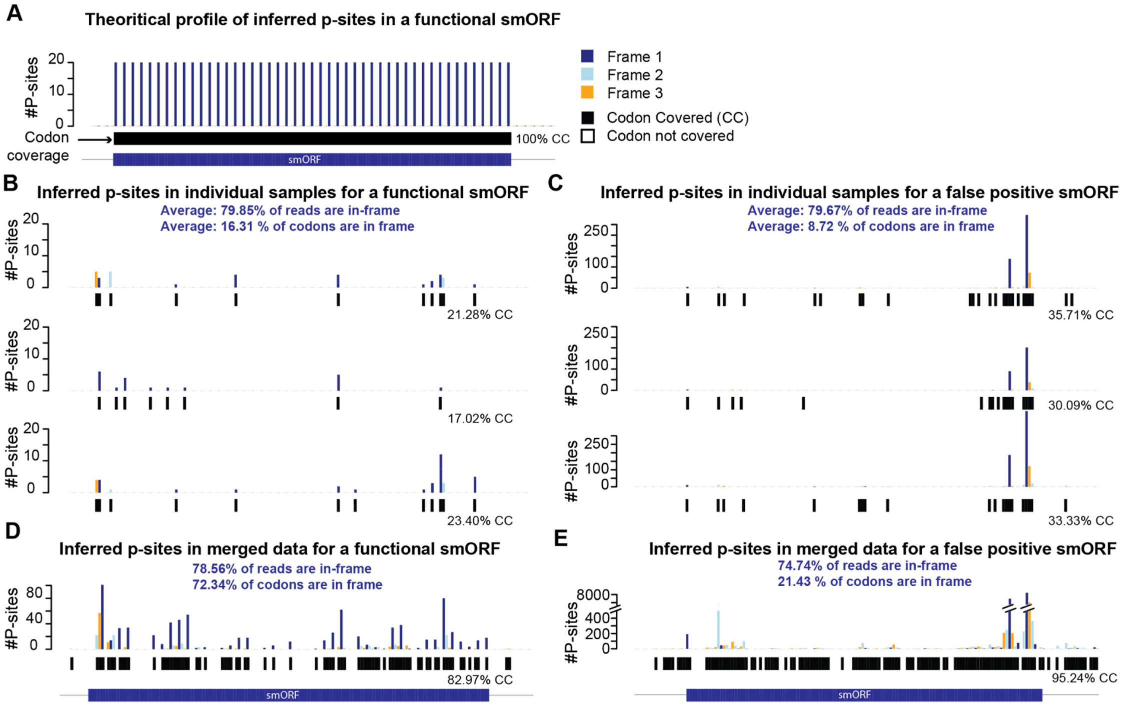

Figure 5. Evidence of translation using Ribo-seq inferred P-sites is sparse across ORFs. A. Barplot showing theoretical distribution of inferred P-site positions (#P-sites) and codon-coverage based on Ribo-seq reads in a given smORF. B. Barplot showing inferred P-site positions and codon coverage across the smORF region using individual samples from hepatocytes [31] for a known and functionally validated smORF SEHBP [44] (B), and a false positive smORF located at chromosome 2: 70,087,581 –70,087,706 (C). D-E. barplot showing inferred P-site positions and codon coverage for smORF shown in B (D), and C (E) using merged high-quality Ribo-seq data [31]. SmORF region is marked in dark blue. P-sites in frame 1: dark blue, frame 2: light blue and frame 3: orange.

periodic signals across all codons. Spectre [48] uses a sliding window approach and RibORF [49] bins the ORF into regions based on the available read coverage within the ORF. RiboWave [50] uses a chi-square test to test for in-frame P-site enrichment in the ORF in comparison to the flanking regions. RiboNT [51] compares data in the translating frame with the two other frames using student t-tests to infer presence of periodicity in smORFs. ORFscore [52] compares the ORF's RPF distribution in each frame with an equally sized uniform distribution using a chi-square test and additionally quantifies % of in-frame positions with reads. Alternatively, in order to alleviate shallow usable data, PRICE [53] attempts to rescue reads such as multi-mappers that would otherwise be discarded. Another tool, RiboHMM [54], assigns modified emission probabilities for base positions with missing data according to the current state of the HMM (e.g. TES, 5'UTS, TIS). Rb-Bp [55] applies LOWESS smoothing for each frame to account for sparse or spiky nature of Ribo-seq data. These tools allow the detection of actively translated smORFs with 3-nt periodicity throughout the length of the smORF, although care must be taken while interpreting the results as they largely depend on data depth and quality.

Where does the smORF start?

The 3-nt periodicity confirms the translating frame and the first encountered stop in this translating frame determines the end of the smORF encoded protein. Determining the start is relatively more difficult. Alternative start-sites and non-canonical start-codons [56] increase the complexity of start-site determination and subsequently can exponentially increase the number of possible isoforms for a given smORF. The sparseness in Ribo-seq data makes it difficult to decipher the most-used start of the smORF and for simplicity, several studies apply prior assumptions such as limiting possible start to only AUG or selecting the most 5' AUG (or longest) as the start-site [49,52,57,58]. Some studies use Ribo-seq data coverage to determine the most used start-site such as SmProt [30], which uses the highest '–framebest' score from RiboTISH [59] tool to select the isoform with the best coverage. Shorter isoforms have a bias towards full coverage, thus another study used the 5' most start-site which maintains uniform coverage of periodicity [31]. Variants of Ribo-seq protocol have been also developed for enrichment of translation initiation (TI) sites using drugs that preferentially inhibit translation initiation only such as harringtonine [9], lactimidomycin [60] and lactimidomycin followed by puromycin [61,62] which is a translation inhibitor that effectively depletes elongating ribosomes. TIS data analysis suggests that the majority of ribosomes initiate translation at cognate AUG codons, followed by near cognate start codons CUG, GUG with ~ 50% initiating at non-AUG start-codons [9,61,62]. Therefore, is critical to consider non-canonical initiation sites when defining smORF start-sites. Computational tools to combine TI-seq and Ribo-seq data from the same biological samples have been developed such as ORF-Rater [63], which uses a regression fit against an expected profile of start- and stop-signals. TISCA [64], which combines translation complex sequencing (TCP-seq) to determine the 40S ribosomal subunit decreasing point along with global TI-seq to more accurately determine initiation sites. RiboTISH [59], detects initiation sites and also quantifies differential initiation site usage across conditions using TI-seq data. Similar to the issue with detecting translation, the start-site determination is also largely dependent on data-depth and quality. Different tools deploy varying prior assumptions and methods to determine the start of a given smORF and care must be taken in interpreting results and combining them.

Pooling data to define a reference set of smORFs

The ultimate goal of smORF detection is to identify potential peptides encoded by genomic regions that could be incorporated into our knowledge base of known proteins. As smORFs were previously excluded only for technical practicality, with new technologies providing high-resolution for translation of smORFs, their reference set development efforts should be treated no different from known ORFs. Historically, gene models were defined using cDNA data which transitioned to using RNA-seq for improved accuracy of 3'UTR boundaries and splice junctions. With the aim to obtain a reference set for a given species, these Ensembl gene models were built based on a pooled dataset with RNA-seq reads merged across tissue-types. Individual tissue dataset gene-models were only used for further refining [65]. In stark contrast, currently, most studies defining smORF sets use single-samples to detect smORF coordinates. This causes two problems: First, considering the sparseness of Ribo-seq data there is not enough evidence to distinguish translation from noise in individual sample data, increasing the false positives (Figure 5B,C). Second, this has led to multiple cell- and tissue-type reference sets instead of a common reference set for a given species. To address the first issue, a study combined three replicates of the same tissue in arabidopsis for smORF detection increasing the codon coverage to 90% [66]. Similarly, technical or biological replicates have been merged in few other studies [30,32]. To address both the issues, in humans, recently a study pooled reads from all published and newly generated high-quality and QC-passed human Ribo-seq from 11 primary human cells and tissues. This led to 1.3 billion inferred P-sites which covered ~ 80% codons across the genome [31]. Trips-Viz [67] also uses aggregated data from multiple studies to improve detection and then uses several features (such as number of codons in regions of interest with higher in-frame reads as compared to out-of-frame, drop-in Ribo-seq density at the stop codon and so on) to rank smORFs for high-confidence of translation. Pooling data in these studies provided increased evidence to define the boundaries of smORFs and allowed testing for their translation more stringently. Specifically, for the two examples described above in Figure 5, by pooling data the codon coverage increased from ~ 20% to ~ 80%. Using the merged data, the difference between a truly translated and false positive is clearer, such as SEHBP (a functionally validated, stable peptide [44], Figure 5D) which is known to be translated has a clear translation signature opposed to a false-positive smORF shown in Figure 5E does not have a translation signature even after merging data. In order to demonstrate the global impact of sequencing depth on ORF detection, we down-sampled published pooled Ribo-seq data [31] to 6 million and up to 1.3 billion inferred P-sites. We then quantified the detection rate of known ORFs and predicted smORFs from the same study across a range of total inferred P-sites (using a codons-in-frame value of 70% as a detection threshold) (see Figure 6A–D). This demonstrates that with ~ 6 million inferred P-sites, the vast majority of ORFs and smORFs have poor codon coverage (Figure 6A) and as a result only ~ 3% (or 431) of expressed known ORFs and ~ 1% (or 103) of smORFs would be detected (Figure 6C,D). Pooling samples increase codon coverage and thus enable us to have stricter quality control by testing each codon within the smORF for translation. Thus, allowing higher resolution to identify the translation as well as the accurate start of the smORF. As smORF reference sets are in early stages, similar to what has been done historically for gene models, smORFs initial set can be defined using pooled data which can be further refined in future versions.

Current reference sets use different data, methods, and assumptions

Previous cataloguing efforts have allowed researchers access to reference sets of SEPs that can be further tested in a given system of interest (see Table 1). OpenProt [25] and the uORFdb (version 2) [26,27] identify all possible ORFs and uORFs, respectively, using a 3-frame translation of the transcriptome while grouping similar ORFs. OpenProt further annotated several pieces of evidence such as protein conservation based on sequence homology, expression based on mass-spectrometry-based proteomics and translation based on

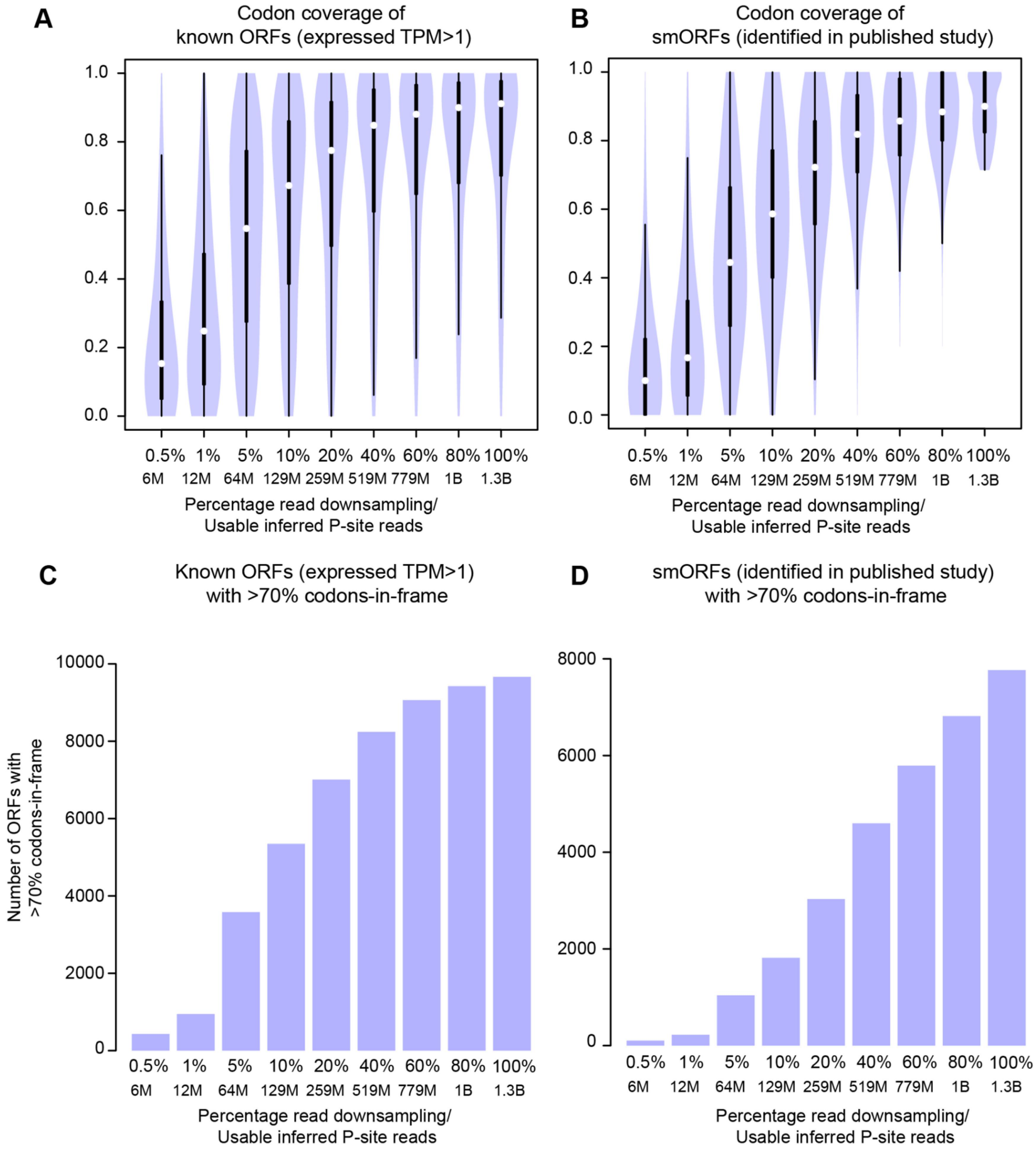

Figure 6. Increase in codon coverage and codons-in-frame within ORFs using pooled Ribo-seq data. A-B. Violin plot showing codon coverage (>1 Ribo-seq read with inferred P-site in the given codon) found with varying usable read depth for known ORFs from Ensembl (A) and smORFs identified in a previous study (B). C-D. Barplots showing number of ORFs with more than 70% codons-in-frame using data with varying usable read depth for known ORFs from Ensembl (C) and smORFs identified in a previous study [31] (D). Usable read depth is the number of inferred P-sites obtained after filtering low-quality sequencing reads, adaptor trimming, removal of contaminant sequences such as rRNA and selecting only uniquely mapped reads for the read lengths that show 3-nt periodicity for annotated ORFs (between 28–30 bp). Known ORFs were filtered for expression (TPM > 1 in at least one sample).

Ribo-seq. Specifically, they used published Ribo-seq data to detect smORFs using PRICE [53] and found only 5,696 of the 461,462 reported on OpenProt with evidence of translation. Not all ORF sequences may be translated and have just occurred by chance, thus such comprehensive approaches have the risk of increased type I error (more false positives). SmProt (version 1) [68], uORFdb [26,27], MetamORF [28] curate smORFs detected using Ribo-seq from the literature and collate them on a database. Such "union sets" have allowed an overview of smORFs found in any published study but led to millions of smORFs catalogued together that were identified from dataset with varying quality and 3-nt periodicity leading to many false positives. In order to reduce the false positives, a study limited their data-source to only seven high-quality Ribo-seq dataset and combined the set of smORFs published in these individual studies [29]. This study presents the Phase I smORFs as the ones found in at least two studies, or an 'intersection set' as high-confidence smORFs based on repeated identification. Although replicability can show repeated evidence of a Ribo-seq signal, it does not indicate that the ones that are not replicated are not translated. Both the union set and intersection set approaches, combined smORFs identified using varying data quality, different computational methods and detection criterion leading to discrepancies in what is considered as a translated smORF.

In order to have a consistent definition of smORF translation, several studies detect smORFs in a unified manner by uniformly processing data and applying common tools and presets [30,32,69]. SmProt [30] (version 2) re-processes 96 human Ribo-seq samples and detects translated smORFs using RiboTISH [59] on individual samples. Sorfs.org uses 34 human Ribo-seq datasets and uses an in-built pipeline to detect smORFs present in individual samples [32]. To account for low-depth in datasets, sorfs.org employs a lenient threshold of 10% in-frame coverage and tests for recurrence of smORFs in multiple datasets. These studies have allowed standardized identification of smORFs but due to sparseness in Ribo-seq samples makes it difficult to distinguish between actual translation and noise as described previously. In humans, a recent study merged high-quality Ribo-seq data to mitigate the sparseness in the data [31] and detected smORFs with a unified pipeline. Pooled data allowed higher codon coverage, thus allowing to test each codon within the smORF and stricter QC, only selecting smORFs which have a high 3-nt periodicity (>75%), percentage of codons-in-frame (>71%) and drop-off score (>92%). Uniform processing and standard definition allows for a more straightforward interpretation of catalogued smORFs as opposed to combining smORF lists from different dataset which were called using a variety of data, tools and assumptions.

What have current smORF sets told us about nuORFs, uORFs, and dORFs?

The reference sets for smORFs, while still evolving, have already enhanced our knowledge on their global properties such as their overall abundance, expression, start-site usage and evolutionary conservation. In humans, uORFs (upstream ORFs) have been found to be most abundant followed by nuORFs (novel unannotated ORFs), and dORFs (downstream ORFs) being the fewest in number across the human genome [29,31]. Human smORFs have also been shown to bemore recently evolved, especially the ones in lncRNAs [19,31]. The generation of a reference set has also allowed quantification of expression levels and translation efficiency for each smORF and several studies have found that uORFs are generally comparable to known ORFs in their translation levels and TE whereas dORFs have been found to have lower levels as compared to known ORFs [31,49]. NuORFs have been found to have low translation levels but the translation-efficiency for nuORFs is nearly comparable to known ORFs [49]. With regard to start-site usage, generally more than half of the translated smORFs have been found to use non-AUG start-codons [9,61,62]. Specifically, uORFs are more enriched for non-canonical start-codons compared to other ORFs [9,49]. Translated uORFs have been found more often in genes encoding transcription factors [31], oncogenes and cellular receptors [70]. These global properties have allowed us to

Table 1. Comparison of data, methods and presets used to catalogue human smORFs.

	Uses Ribo-seq to detect smORFs	No. of Ribo-seq samples used	Cell-types/ tissues	Uniform Re-processing of Ribo-seq data	Individual/ merged data used	Method to call smORFs	Method to select most probable isoform	Start-codon	Length threshold (amino acids)	No. of smORFs reported	uORFs: nuORFs: dORFs reported
sorfs.org	Y	34 dataset	Various	Y	Only replicates merged	Unified	N	AUG and near-cognate	10	555,927	-
SmPROT v2	Y	96 samples	Various	Y	Only replicates merged	Unified	Using the –frame best score	AUG and near-cognate	5	327,995	NA
OpenPROT	N	NA	NA	N	NA	NA	N	AUG	30	461,462	NA
MetaMORF	Y	-	Various	N	Various	Various	N	Various	No	664,771	-
uORFdb	N	NA	Various	N	NA	NA	N	AUG and near-cognate	No	>2.4 million	Only uORFs
Mudge et al.	Y	139 samples	Various	N	Various	Various	Longest ORF	AUG	16	7,264	3771: 2208: 565
Chothani et al.	Y	187 samples	Various	Y	Merged	Unified	Longest ORF with best uniformity score	AUG and near-cognate	No	7,767	5280: 1652: 802

view smORFs and known ORFs together to be able to understand the similarities and differences between them, their potential functions and evolutionary history.

Apart from global properties of smORFs, several different possibilities of functions of their translation have been described. NuORFs have been generally linked to generate functional proteins that are important in humans such as for heart development [71], muscle formation [72–74], regulating calcium uptake in muscle [75,76] and play important roles in the mitochondria [77–79]. For smORFs in the untranslated regions of known protein-coding ORFs there are various alternative fates [80]. UORFs have been frequently shown to repress the known ORF on its transcript and few global studies show lower readout of protein levels for known ORFs containing uORFs using proteomics readout [70,81] and translation-efficiency [49,58]. Although, several negatively regulated uORF-mORF pairs exist, translation of uORFs has also been shown to positively affect the translation of the main ORF [4,82,82,83] and protect the translation of the known ORF under stress [84]. Recently, several human genome-wide studies have highlighted that although there do exist several uORF-mORF pairs that are negatively regulated, the predominant trend shows uORF and mORF being regulated in the same direction. This has been shown independently by various studies, such as for cell-identity of fibroblasts, endothelial cells, kidney, brain and heart tissues [31] and in disease conditions such as fibrosis [85] and in glioblastoma [5] and dilated cardiomyopathy patients [10]. Another study found deleting start-codon for peptide-forming uORFs only minimally increased the expression of the main CDS indicating non repressive function of uORFs [6]. uORFs are also increasingly being shown to encode peptides with important functions in disease [5], and form complexes or directly inhibit other proteins [6,7]. Apart from uORFs, dORFs, which are found on the 3'UTR of known ORFs, have been shown to enhance translation of the main ORFs and the number of dORFs rather than the length is shown to further enhance this effect [86]. A study also showed a dORF encoded a protein that is a cancer antigen [87]. While these studies show evidence for the possible function of smORFs and their encoded peptides, with 1000s of smORFs detected in Ribo-seq with translation signatures identical to known proteins, more studies are needed to understand the roles of nuORFs, uORFs and dORFs.

Road ahead and challenges

There is a growing concern in the scientific community that the current reference set of long ORFs may have overlooked a significant number of smORFs. This has led to a community call for the development of a translated smORF reference set that can be integrated into existing annotations [29]. Currently available Ribo-seq data have shallow-depth and are sparse in nature and thus to detect smORFs accurately it requires pooling of data, as has been presented for human smORFs [31]. This study uniformly processed, analysed and pooled high-quality data to obtain >80% codon coverage on the human translatome and thus was able to detect a reference set of smORFs that have undergone a stricter quality control by testing each codon for translation. Moving forward, we recommend pooling existing high-quality data for a given species to account for the data-sparseness and uniformly identify smORFs to obtain a reference set. All subsequent newly generated Ribo-seq data should then be added to the original data release to re-analyse and identify smORFs further improving the annotation. To ensure stability across versions, two primary areas need to be considered. First, the approach to updating and sharing revisions and second, the method for ranking smORFs to provide a confidence level. Learning from the experience of incorporating known ORFs into the genebuild can help in developing a smORF reference set. Ensembl is updated every 3 months and significant updates, such as the most recent genome build was updated after 5 years. Similarly, smORF reference set updates should be released every few years with new Ribo-seq data pooled with previous data builds to identify and update the smORF reference set. Criteria should be established to add or remove smORFs. Instead of removing smORFs with slightly reduced scores in newer versions, they could be assigned a confidence level based on consistent or improved uniformity and periodicity in reference set updates. A translation support level (TrSL) can be assigned to smORFs similar to transcript support level (TSL) scores provided by Ensembl for transcripts. With technological advances from cDNA sequencing to RNA-seq with longer sequencing reads, revisions of gene models have now become fairly consistent. Future revisions of smORF reference sets will also need to be dynamic and aim to follow the same trajectory, and thus, be updated with not only new data but also incorporate advances in Ribo-seq protocols improving data resolution and quality.

A reference set of translated smORFs can be a powerful tool for discoveries of new proteins or regulatory control elements. Previously conducted studies to understand known ORFs can be used as blueprints to discover smORF biology. Here, we provide a few examples of applications in each layer of protein production, i.e. DNA, RNA and protein. At the DNA level, a study showed uORF start-creating and stop-disrupting mutations are under strong negative selection [88]. smORFs can be tested for presence of GWAS and eQTLs and have been reviewed for cardiovascular disorders recently [89]. At the RNA level, the smORFs can be used for differential expression analysis using RNA-sequencing and Ribo-seq to understand their role in a given biological system [12]. Generation of a reference set incorporated along with the known ORFs would allow researchers to use publicly available RNAseq data to infer which smORFs are differentially regulated in a given disease or perturbation of interest without having to perform Ribo-seq and call smORFs in every new system of study. For those interested to investigate lncRNA coding potential, several sequence- and evolutionary conservation-based tools [90,91], and more recently deep learning models, have been developed [92,93] to identify cryptic ORFs using *in silico* prediction. As has recently become evident, lncRNAs tend to encode young proteins [19] and thus traditional ORF prediction methods which rely on length-biased and evolutionary conservation-biased methods would not be able to discern coding potential efficiently [46]. Instead, Ribo-seq provides experimental evidence for translation within lncRNAs and thus smORF reference sets can be used as a direct-measure of lncRNA coding potential. Tools such as 'Is it a smORF?' available in http://smorfs.ddnetbio.com/ can be used to identify lncRNAs encoding high-confidence translated smORFs using ribo-seq evidence. Lastly, protein-level studies can help us delineate whether a smORF makes a stable peptide or is degraded. Thus, several studies have verified the evidence of their presence in mass spectrometry data [11,25,31,32,52,94–96], but due to technical limitations [97,98] for detecting short peptide sequences accurately, best practices for methods to detect smORFs in-vivo are still developing. CRISPR-based screening strategies have also been deployed to identify smORFs essential for cellular growth [6,99] and cancer cell survival [99]. Depending on the number of smORFs that can be directly used for testing, the reference set may need to be filtered to obtain a more feasible number. The activity and role of smORFs can be better understood with such studies and thus have the potential to uncover new insights into cellular processes as well as disease mechanisms. As our understanding of smORFs grows, they can be incorporated into widely used databases such as Uniprot, GENCODE and Ensembl similar to the approach taken with known protein-coding ORFs, further expanding our knowledge of the translatome and proteome.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

S.C. is supported by a Khoo post-doctoral fellowship award and Singapore MOH NMRC under OF-YIRG (MOH-001340). L.H. is supported by NMRC MOH-000960 and Howard Hughes Medical Institute IRSP 55008732.

Data availability statement

Data sharing is not applicable to this article as no new data were generated in this study.

References

[1] Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. Nucleic Acids Res. 2022;50(D1):D988–95.

[2] Basrai MA, Hieter P, Boeke JD. Small open reading frames: beautiful needles in the haystack. Genome Res. 1997;7(8):768–771. doi: 10.1101/gr.7.8.768

[3] van Heesch S, van Iterson M, Jacobi J, et al. Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. Genome Biol. 2014;15(1):R6. doi: 10.1186/gb-2014-15-1-r6

[4] Mueller PP, Hinnebusch AG. Multiple upstream AUG codons mediate translational control of GCN4. Cell. 1986;45(2):201–207. doi: 10.1016/0092-8674(86)90384-3

[5] Huang N, Li F, Zhang M, et al. An upstream open reading frame in phosphatase and tensin homolog encodes a circuit breaker of lactate metabolism. Cell Metab. 2021;33(2):454.

[6] Chen J, Brunner A-D, Cogan JZ, et al. Pervasive functional translation of noncanonical human open reading frames. Science. 2020;367(6482):1140–1146.

[7] Parola AL, Kobilka BK. The peptide product of a 5' leader cistron in the beta 2 adrenergic receptor mRNA inhibits receptor synthesis. J Biol Chem. 1994;269(6):4497–4505. doi: 10.1016/S0021-9258(17)41806-0

[8] Ingolia NT, Ghaemmaghami S, Newman JRS, et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324(5924):218–223.

[9] Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell. 2011;147(4):789–802. doi: 10.1016/j.cell.2011.10.002

[10] Schafer S, Adami E, Heinig M, et al. Translational regulation shapes the molecular landscape of complex disease phenotypes. Nat Commun. 2015;6(1):7200.

[11] van Heesch S, Witte F, Schneider-Lunitz V, et al. The translational landscape of the human heart. Cell. 2019;178:242–60.e29. doi: 10.1016/j.cell.2019.05.010

[12] Chothani S, Adami E, Ouyang JF, et al. deltaTE: detection of translationally regulated genes by Integrative analysis of Ribo-seq and RNA-seq data. Curr Protoc Mol Biol. 2019;129:e108. doi: 10.1002/cpmb.108

[13] Guttman M, Russell P, Ingolia NT, et al. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell. 2013;154(1):240–251.

[14] Ingolia NT, Brar GA, Stern-Ginossar N, et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep. 2014;8(5):1365–1379.

[15] Calviello L, Ohler U. Beyond read-counts: ribo-seq data analysis to understand the functions of the transcriptome. Trends Genet. 2017;33:728–744. doi: 10.1016/j.tig.2017.08.003

[16] Plaza S, Menschaert G, Payre F. In search of lost small peptides. Annu Rev Cell Dev Biol. 2017;33(1):391–416. doi: 10.1146/annurev-cellbio-100616-060516

[17] Couso J-P, Patraquim P. Classification and function of small open reading frames. Nat Rev Mol Cell Biol. 2017;18(9):575–589. doi: 10.1038/nrm.2017.58

[18] Ingolia NT; Ingolia NT. Ribosome profiling: new views of translation, from single codons to genome scale. Nat Rev Genet. 2014;15(3):205–213. doi: 10.1038/nrg3645

[19] Ruiz-Orera J, Messeguer X, Subirana JA, et al. Long non-coding RNAs as a source of new peptides. Elife. 2014;3:e03523. doi: 10.7554/eLife.03523

[20] Carja O, Xing T, Wallace EWJ, et al. Riboviz: analysis and visualization of ribosome profiling datasets. BMC Bioinf. 2017;18(1):461.

[21] Liu Q, Shvarts T, Sliz P, et al. RiboToolkit: an integrated platform for analysis and annotation of ribosome profiling data to decode mRNA translation at codon resolution. Nucleic Acids Res. 2020;48(W1):W218–29.

[22] Lauria F, Tebaldi T, Bernabò P, et al. riboWaltz: Optimization of ribosome P-site positioning in ribosome profiling data. PLoS Comput Biol. 2018;14(8):e1006169.

[23] Chung BY, Hardcastle TJ, Jones JD, et al. The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis. RNA. 2015;21(10):1731–1745.

[24] Brar GA, Weissman JS. Ribosome profiling reveals the what, when, where and how of protein synthesis. Nat Rev Mol Cell Biol. 2015;16(11):651–664. doi: 10.1038/nrm4069

[25] Brunet MA, Brunelle M, Lucier J-F, et al. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. Nucleic Acids Res. 2019;47:D403–10. doi: 10.1093/nar/gky936

[26] Wethmar K, Barbosa-Silva A, Andrade-Navarro MA, et al. uORFdb–a comprehensive literature database on eukaryotic uORF biology. Nucleic Acids Res. 2014;42:D60–7. doi: 10.1093/nar/gkt952

[27] Manske F, Ogoniak L, Jürgens L, et al. The new uOrfdb: integrating literature, sequence, and variation data in a central hub for uORF research. Nucleic Acids Res. 2023;51(D1):D328–36.

[28] Choteau SA, Wagner A, Pierre P, et al., MetamORF: a repository of unique short open reading frames identified by both experimental and computational approaches for gene and metagene analyses. Database [Internet]. 2021;2021. Available from 10.1093/database/baab032.

[29] Mudge JM, Ruiz-Orera J, Prensner JR, et al. Standardized annotation of translated open reading frames. Nat Biotechnol. 2022;40(7):994–999.

[30] Li Y, Zhou H, Chen X, et al. SmProt: a reliable repository with comprehensive annotation of small proteins identified from ribosome profiling. Int J Genomics Proteomics. 2021;19(4):602–610.

[31] Chothani SP, Adami E, Widjaja AA, et al. A high-resolution map of human RNA translation. Mol Cell. 2022;82(15):2885–99.e8.

[32] Olexiouk V, Van Criekinge W, Menschaert G. An update on sOrfs.Org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. 2018;46(D1):D497–502. doi: 10.1093/nar/gkx1130

[33] Consortium U, Martin M-J, Orchard S. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 2023;51: D523–31. doi: 10.1093/nar/gkac1052

[34] O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45.

[35] Gerashchenko MV, Lobanov AV, Gladyshev VN. Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. Proc Natl Acad Sci U S A. 2012;109 (43):17394–17399. doi: 10.1073/pnas.1120799109

[36] Halpin JC, Jangi R, Street TO. Multimapping confounds ribosome profiling analysis: a case-study of the Hsp90 molecular chaperone. Proteins. 2020;88(1):57–68. doi: 10.1002/prot.25766

[37] Ingolia NT, Brar GA, Rouskin S, et al. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc. 2012;7 (8):1534–1550.

[38] CC-C W, Zinshteyn B, Wehner KA, et al. High-resolution ribosome profiling defines discrete ribosome elongation states and translational regulation during cellular stress. Mol Cell. 2019;73:959–70.e5. doi: 10.1016/j.molcel.2018.12.009

[39] Lareau LF, Hite DH, Hogan GJ, et al. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. Elife. 2014;3:e01257. doi: 10.7554/eLife.01257

[40] Martinez TF, Chu Q, Donaldson C, et al. Accurate annotation of human protein-coding small open reading frames. Nat Chem Biol. 2020;16(4):458–468.

[41] Varenne S, Buc J, Lloubes R, et al. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J Mol Biol. 1984;180(3):549–576.

[42] Tuller T, Carmi A, Vestsigian K, et al. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell. 2010;141(2):344–354.

[43] Weinberg DE, Shah P, Eichhorn SW, et al. Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. Cell Rep. 2016;14(7):1787–1799.

[44] Koh M, Ahmad I, Ko Y, et al. A short ORF-encoded transcriptional regulator. Proc Natl Acad Sci U S A. 2021;118(4):118. doi: 10.1073/pnas.2021943118

[45] Michel AM, Choudhury KR, Firth AE, et al. Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res. 2012;22(11):2219–2229.

[46] Chew G-L, Pauli A, Rinn JL, et al. Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. Development. 2013;140:2828–2834. doi: 10.1242/dev.098343

[47] Calviello L, Mukherjee N, Wyler E, et al. Detecting actively translated open reading frames in ribosome profiling data. Nat Methods. 2016;13(2):165–170.

[48] Chun SY, Rodriguez CM, Todd PK, et al. Spectre: a spectral coherence–based classifier of actively translated transcripts from ribosome profiling sequence data. BMC Bioinf. 2016;17:482. doi: 10.1186/s12859-016-1355-4

[49] Ji Z, Song R, Regev A, et al. Many lncRnas, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife. 2015;4:e08890. doi: 10.7554/eLife.08890

[50] Xu Z, Hu L, Shi B, et al. Ribosome elongating footprints denoised by wavelet transform comprehensively characterize dynamic cellular translation events. Nucleic Acids Res. 2018;46(18):e109.

[51] Song B, Jiang M, Gao L. RiboNT: a noise-tolerant predictor of open reading frames from ribosome-protected footprints. Life. 2021;11:701. doi: 10.3390/life11070701.

[52] Bazzini AA, Johnstone TG, Christiano R, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. 2014;33(9):981–993.

[53] Erhard F, Halenius A, Zimmermann C, et al. Improved Ribo-seq enables identification of cryptic translation events. Nat Methods. 2018;15(5):363–366.

[54] Raj A, Wang SH, Shim H, et al. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. eLife [Internet]. 2016;5. doi: 10.7554/eLife.13328

[55] Malone B, Atanassov I, Aeschimann F, et al. Bayesian prediction of RNA translation from ribosome profiling. Nucleic Acids Res. 2017;45:2960–2972. doi: 10.1093/nar/gkw1350

[56] Cao X, Slavoff SA. Non-AUG start codons: expanding and regulating the small and alternative ORFeome. Exp Cell Res. 2020;391 (1):111973. doi: 10.1016/j.yexcr.2020.111973

[57] Mackowiak SD, Zauber H, Bielow C, et al. Extensive identification and analysis of conserved small ORFs in animals. Genome Biol. 2015;16(1):179.

[58] Johnstone TG, Bazzini AA, Giraldez AJ. Upstream ORFs are prevalent translational repressors in vertebrates. EMBO J. 2016;35 (7):706–723. doi: 10.15252/embj.201592759

[59] Zhang P, He D, Xu Y, et al. Genome-wide identification and differential analysis of translational initiation. Nat Commun. 2017;8(1):1749.

[60] Lee S, Liu B, Lee S, et al. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A. 2012;109(37):E2424–32.

[61] Gao X, Wan J, Liu B, et al. Quantitative profiling of initiating ribosomes in vivo. Nat Methods. 2015;12(2):147–153.

[62] Fritsch C, Herrmann A, Nothnagel M, et al. Genome-wide search for novel human uOrfs and N-terminal protein extensions using ribosomal footprinting. Genome Res. 2012;22(11):2208–2218.

[63] Fields AP, Rodriguez EH, Jovanovic M, et al. A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. Mol Cell. 2015;60(5):816–827.

[64] Ichihara K, Matsumoto A, Nishida H, et al. Combinatorial analysis of translation dynamics reveals eIF2 dependence of translation initiation at near-cognate codons. Nucleic Acids Res. 2021;49(13):7298–7317.

[65] Collins JE, White S, Searle SMJ, et al. Incorporating RNA-seq data into the zebrafish Ensembl genebuild. Genome Res. 2012;22(10):2067–2078.

[66] Hsu PY, Calviello L, Wu H-Y, et al. Super-resolution ribosome profiling reveals unannotated translation events in. Proc Natl Acad Sci U S A. 2016;113:E7126–35. doi: 10.1073/pnas.1614788113

[67] Kiniry SJ, Judge CE, Michel AM, et al. Trips-Viz: an environment for the analysis of public and user-generated ribosome profiling data. Nucleic Acids Res. 2021;49(W1):W662–70.

[68] Hao Y, Zhang L, Niu Y, et al. SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. Brief Bioinform. 2018;19:636–643. doi: 10.1093/bib/bbx005

[69] Olexiouk V, Crappé J, Verbruggen S, et al. sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. 2016;44(D1):D324–9.

[70] Ye Y, Liang Y, Yu Q, et al. Analysis of human upstream open reading frames and impact on gene expression. Hum Genet. 2015;134(6):605–612.

[71] Ho L, van Dijk M, Chye STJ, et al. ELABELA deficiency promotes preeclampsia and cardiovascular malformations in mice. Science. 2017;357:707–713. doi: 10.1126/science.aam6607

[72] Bi P, Ramirez-Martinez A, Li H, et al. Control of muscle formation by the fusogenic micropeptide myomixer. Science. 2017;356 (6335):323–327.

[73] Zhang Q, Vashisht AA, O'Rourke J, et al. The microprotein Minion controls cell fusion and muscle formation. Nat Commun. 2017;8 (1):15664.

[74] Quinn ME, Goh Q, Kurosaka M, et al. Myomerger induces fusion of non-fusogenic cells and is required for skeletal muscle development. Nat Commun. 2017;8(1):15665.

[75] Magny EG, Pueyo JI, Pearl FMG, et al. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. Science. 2013;341(6150):1116–1120.

[76] Anderson DM, Anderson KM, Chang C-L, et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell. 2015;160(4):595–606.

[77] Lee CQE, Kerouanton B, Chothani S, et al. Coding and non-coding roles of MOCCI (C15ORF48) coordinate to regulate host inflammation and immunity. Nat Commun. 2021;12(1):2130.

[78] Zhang S, Reljić B, Liang C, et al. Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. Nat Commun. 2020;11(1):1312.

[79] Lee C, Zeng J, Drew BG, et al. The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. Cell Metab. 2015;21(3):443–454.

[80] Morris DR, Geballe AP. Upstream open reading frames as regulators of mRNA translation. Mol Cell Biol. 2000;20(23):8635–8642. doi: 10.1128/MCB.20.23.8635-8642.2000

[81] Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A. 2009;106 (18):7507–7512. doi: 10.1073/pnas.0810916106

[82] Starck SR, Tsai JC, Chen K, et al. Translation from the 5' untranslated region shapes the integrated stress response. Science. 2016;351:aad3867. doi: 10.1126/science.aad3867

[83] Hinnebusch AG. Translational regulation of Yeast GCN4: a window on factors that control initiator-tRNA binding to the ribosome *. J Biol Chem. 1997;272(35):21661–21664. doi: 10.1074/jbc.272.35.21661

[84] Andreev DE, O'Connor PBF, Fahey C, et al. Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. Elife. 2015;4:e03971. doi: 10.7554/eLife.03971

[85] Chothani S, Schäfer S, Adami E, et al. Widespread translational control of fibrosis in the human heart by RNA-Binding proteins. Circulation. 2019;140(11):937–951.

[86] Wu Q, Wright M, Gogol MM, et al. Translation of small downstream ORFs enhances translation of canonical main open reading frames. EMBO J [Internet]. 2020 [[cited 2023 Mar 15]];39. doi: 10.15252/embj.2020104763

[87] Chong C, Müller M, Pak H, et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. Nat Commun. 2020;11(1):1293.

[88] Whiffin N, Karczewski KJ, Zhang X, et al. Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. Nat Commun. 2020;11(1):2523.

[89] Soukarieh O, Meguerditchian C, Proust C, et al. Common and rare 5'UTR variants altering upstream open reading frames in cardiovascular genomics. Front Cardiovasc Med. 2022;9:841032. doi: 10.3389/fcvm.2022.841032

[90] Wang L, Park HJ, Dasari S, et al. CPAT: Coding-potential assessment tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013;41:e74. doi: 10.1093/nar/gkt006

[91] Washietl S, Findeiss S, Müller SA, et al. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA. 2011;17(4):578–594.

[92] Clauwaert J, McVey Z, Gupta R, et al. TIS transformer: remapping the human proteome using deep learning. NAR Genom Bioinform. 2023;5:lqad021. doi: 10.1093/nargab/lqad021

[93] Nabi A, Dilekoglu B, Adebali O, et al. Discovering misannotated lncRnas using deep learning training dynamics. Bioinformatics [Internet]. 2023 [cited 2023 Mar 19];39(1). doi: 10.1093/bioinformatics/btac821.

[94] Zheng EB, Zhao L. Protein evidence of unannotated ORFs in Drosophila reveals diversity in the evolution and properties of young proteins. eLife [Internet]. 2022;11:11. doi: 10.7554/eLife.78772.

[95] Lu S, Zhang J, Lian X, et al. A hidden human proteome encoded by "non-coding" genes. Nucleic Acids Res. 2019;47:8111–8125. doi: 10.1093/nar/gkz646

[96] Duffy EE, Finander B, Choi G, et al. Developmental dynamics of RNA translation in the human brain. Nat Neurosci. 2022;25 (10):1353–1365.

[97] Wacholder A, Carvunis A-R. Rare detection of noncanonical proteins in yeast mass spectrometry studies. Biorxiv. 2023. doi: 10.1101/2023.03.09.531963

[98] Prensner JR, Abelin JG, Kok LW, et al. What can Ribo-seq and proteomics tell us about the non-canonical proteome? Biorxiv. 2023. doi: 10.1101/2023.05.16.541049

[99] Prensner JR, Enache OM, Luria V, et al. Noncanonical open reading frames encode functional proteins essential for cancer cell survival. Nat Biotechnol. 2021;39(6):697–704.