Trends in Genetics

CellPress
REVIEWS

Review

Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome

Lorenzo Calviello[1,2] and Uwe Ohler[1,2,3,*]

By mapping the positions of millions of translating ribosomes in the cell, ribosome profiling (Ribo-seq) has established its role as a powerful tool to study gene expression. Several laboratories have introduced modifications to the experimental protocol and expanded the repertoire of biochemical methods to study translation transcriptome-wide. However, the diversity of protocols highlights a need for standardization. At the same time, different computational analysis strategies have used Ribo-seq data to identify the set of translated sequences with high confidence. In this review we present an overview of such methodologies, outlining their assumptions, data requirements, and availability. At the interface between RNA and proteins, Ribo-seq can complement data from multiple omics approaches, zooming in on the central role of translation in the molecular cell.

The Ribo-seq Strategy to Study the Translated Transcriptome

The life of an mRNA molecule is complicated. As it carries all the information needed to synthesize proteins, multiple maturation and modification steps are needed to ensure the correct parsing of this information for correct cellular homeostasis.

RNA-seq [1], which couples RNA isolation with next-generation sequencing, allows us to have a global look at the entire transcriptome, uncovering the staggering diversity of transcripts in a sample. While standard protocols profile steady-state, mature RNA populations, dozens of RNA-seq variants now allow us to zoom in to specific features of the mRNA life cycle: from its synthesis, isolating the nascent pre-mRNA molecule [2,3], to its processing, localization, and decay [4]. Uncovering the dynamics of the mRNA cytosolic life greatly enriches our understanding of gene expression as a dynamic system, in which the turnover of transcripts and proteins shapes their molecular functions [5].

Until a few years ago, a transcriptome-level bridge between transcript detection and translational output, which was mostly inferred from proteomics-based experiments [6], was lacking. To measure the translational output of a transcript, researchers applied polysome profiling, where the degree of association with different polysomal fractions can be used as a proxy to define rates of protein production [7]. Very recently, this technique has been coupled to RNA-seq, obtaining a transcriptome-wide view of polysome association with different RNA species, showing differential translation output across different isoforms per genes [8,9]. From a different angle, Ribo-seq [10] has revolutionized the field of functional transcriptomics by mapping the position of translating ribosomes over the entire transcriptome. Since its inception, the scientific

Trends

Ribo-seq has become an established protocol to identify translated transcript regions via deep sequencing, closing the gap between RNA sequencing and proteomics.

Recently developed Ribo-seq data analysis strategies use different features as hallmarks of translation. Specifically, the ability to monitor the positions of translating ribosomes with single-nucleotide precision has driven the development of computational tools that rely on 'subcodon resolution'. Knowing the concrete assumptions and precise goals of different approaches is crucial.

In addition to addressing translation-focused questions, from defining open reading frames to identifying alternative translation initiation sites and estimating differential translation rates, Ribo-seq data show great promise for integrative efforts combining additional omics approaches.

[1] Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine, Robert-Rössle-Strasse 10, Berlin 13125, Germany
[2] Department of Biology, Humboldt Universität zu Berlin, Unter den Linden 6, Berlin 10117, Germany
[3] Department of Computer Science, Humboldt Universität zu Berlin, Unter den Linden 6, Berlin 10117, Germany

*Correspondence:
uwe.ohler@mdc-berlin.de (U. Ohler).

http://dx.doi.org/10.1016/j.tig.2017.08.003

community has employed Ribo-seq to answer a wide range of questions, ranging from the identification of translated small **open reading frames** (ORFs, see Glossary) [11] to the quantification of translational control [12], while gaining precious mechanistic insights on the translation process itself [13].

The basic idea behind the Ribo-seq protocol is simple: ribosome-protected fragments (RPFs) are protected from RNase digestion and can thus be isolated and sequenced. In brief, the Ribo-seq protocol consists of (i) drug treatment and cell harvesting, (ii) nuclease footprinting and RPF isolation, and (iii) library preparation and sequencing. While the experimental protocol has already been reviewed in detail [14,15], we will focus on interesting variants employed by different laboratories and the implications for downstream data analysis.

Variants of the Original Protocol

In the original protocol cells are preincubated with cycloheximide (CHX), a drug that binds in the ribosome E-site, blocking elongating ribosomes in their pre-translocation step. Such a pre-incubation step is omitted in other protocols [16–18] because cycloheximide is also present in the buffer used for cell lysis. Ribosomes in the post-initiation stage can also be isolated by adding drugs blocking the initiating ribosome, such as harringtonine (HARR) [11] or lactimidomycin (LTM) [19] instead of (or together with) CHX. The QTI-seq protocol [20] introduces a puromycin treatment after adding LTM, triggering the dissociation of elongating ribosomes and thus further enriching for initiating complexes. In eukaryotic cells, organelles such as mitochondria or chloroplasts employ a different set of ribosomes, which resemble prokaryotic ribosomes and are thus different from ribosomes in the cytosol. The use of chloramphenicol (an inhibitor of translation elongation in bacteria) enables actively translating ribosomes to be isolated from mitochondria [21] or chloroplasts [22]. The choice of inhibitor has an impact on the observed kinetics of initiation and elongation, in a concentration-dependent manner [23,24]. Individual profiles over specific codons or small regions must then be considered with care.

During the footprinting step, nucleases are added to digest the mRNA sequences not protected by ribosomes. Given its low sequence bias, RNase I is most commonly used. Despite its more pronounced sequence bias, micrococcal nuclease (MNase) has also been used in different experiments, especially when RNase treatment is inefficient [25]. Different nucleases exhibit different digestion patterns over the isolated monosomes, ultimately resulting in drastically different ribosome profiles [26]. For instance, under specific conditions, Ribo-seq datasets obtained with MNase can show a lower single-nucleotide resolution and higher ribosomal occupancy over 3′ **untranslated regions** (3′-UTRs) [27], in marked difference from profiles resulting from RNase I treatment.

After footprint recovery, rRNA depletion is performed and samples are sequenced. The original protocol suggests a circularization step during the library prep, followed by trimming of the last nucleotide to avoid the presence of untemplated additions [14]. In other protocols linear amplification and the use of randomized oligonucleotides proximal to the ligation site can reduce bias introduced by the ligation and PCR amplification steps [16,28].

Pre-Processing and Quality Control

A characteristic feature of a high-quality Ribo-seq library is its distinct read-length distribution, which usually peaks at ~29 nt in eukaryotic cytosolic ribosomes, reflecting the size of a translating ribosome on the RNA [10]. A broader distribution of reads has been observed in variants of the protocol, depending on the nuclease treatment [26]. An additional shorter footprint of ~20 nt can be detected when performing Ribo-seq in absence of CHX or in

Glossary

Classifier: a machine-learning approach whose objective is to assign datapoints to different classes (two in the case of binary classifiers). In supervised learning, the classifier is trained on known examples, while unsupervised classification methods are used in absence of known (or labeled) data.

Coding sequence (CDS): a sequence that is translated using one (or more) of the three possible reading frames.

Hidden Markov model (HMM): a probabilistic method in which a signal (e.g., a coverage track or a nucleotide sequence) is emitted from a finite succession of unknown (hidden) states. The hidden states can represent different biological concepts (e.g., 5′-UTRs, ORFs, etc. in genomic sequence classification); transitions between them specify possible sequences of the states, and can be defined and trained on available data (e.g., read coverage or nucleotide sequences in annotated genomic regions). Once the model is trained, it can be used to parse a new signal and label it with the optimal sequence of states.

Long non-coding RNAs (lncRNAs): long transcripts (>200 nt) which do not exhibit clear coding potential.

Multitaper: a signal processing method that aims to provide reliable estimates of the spectrum of frequencies present in a signal. In the multitaper method, multiple filters are applied as windows over the same signal, and coefficients for all frequency components are retrieved from each filtered sample (using the Fourier transform). Different types of filters have been proposed; specifically, the use of the so-called Slepian sequences enables the application of a statistical test to each frequency component.

Non-negative least squares: a modified version of the ordinary least squares, in which the regression coefficients cannot be negative values.

Nonsense-mediated decay (NMD): an mRNA surveillance pathway that degrades aberrant transcripts, thus preventing the production of non-functional proteins. One of the proposed mechanisms for NMD involves the recognition of a premature termination codon (PTC), presence of different inhibitors [29]. Distinct read-length distributions can also correspond to other ribosomal conformations [30] or to ribosomes belonging to different subcellular compartments. Mitochondrial ribosomes have been shown to display a bimodal distribution of read lengths, peaking at 27 and 33 nt, thus showing a clear difference when compared to cytosolic-derived RPFs [21] (Figure 1A).

For most of the datasets, the sequenced read length is larger than the RPF length. The adapter sequences used for library preparation must then be removed with tools such as Cutadapt[i] [31]. Depending on the efficiency of the rRNA removal step, a high percentage of reads consists of small structured RNAs (rRNAs, tRNAs, or snoRNAs), which should be removed because their overabundance can skew subsequent quantification. As we are sequencing a pool of RNA fragments, a splice-aware alignment such as STAR[ix] [32] or others [33] can be used. RPFs are short (~29 nt), and many reads will map to multiple locations. To solve this, different 'rescue' strategies are used in popular RNA-seq quantification tools [34,35], but they are typically embedded inside a larger workflow for transcript quantification. Alternatively, the alignments can be filtered either by using specific tools [36] or by extracting one primary alignment per read [16]. In a high-quality Ribo-seq library, reads mostly map to **coding sequence** (CDS) regions (usually >85%) and 5′-UTRs (~5–10%), and very few to 3′-UTRs (Figure 1B). Signals coming from introns and intergenic regions are usually the result of multi-mapping fragments.

The distribution of aligned Ribo-seq reads over the translated ORFs is dependent on the kinetics of the translation process: the assembly of the initiation complex is a relatively slow process, resulting in a pronounced accumulation of signal around the start codon. In most datasets an additional accumulation can be observed at the last codon of the ORF, caused by the slow kinetics of translation termination and peptide release. In aggregate profiles of RPF 5′ ends over annotated start and stop codons, it is possible to appreciate the single-nucleotide resolution of Ribo-seq data (Figure 1C): in most datasets, especially the more recent ones [11,12,16,37], 5′ ends accumulate on one of the possible three frames, thus revealing the translated frame. This level of resolution at the subcodon level is usually accompanied by a distinct offset of the 5′ ends relative to the annotated start codons (~12 nt for many datasets). This distance can be used to shift the positions of Ribo-seq reads and monitor translation at each translated codon, reflecting the positions of the P-site compartment for millions of ribosomal footprints. Aggregate profiles can drastically vary between different read lengths.

Several analysis packages are now available to perform quality control for Ribo-seq data. For instance, riboSeqR[vi] [38] can analyze and visualize read-length distributions and subcodon resolution, while RiboProfiling[v] [39] also provides codon-level statistics and stalling analysis. The Plastid [40] package includes additional utilities to explore profiles over transcripts and genomic regions. In addition, RUST[viiii] performs a more in-depth analysis of the expected P-site profiles over different sequence features [41]. As more datasets became available for multiple organisms, questions about the overall standardization of the Ribo-seq protocol have started to emerge in the field (Box 1).

To facilitate exploration and reanalysis of published Ribo-seq data, RPFdb[xi] [42] contains analyses of dozens of datasets in different organisms, together with few summary statistics. The GWIPS-viz[x] [43] browser hosts publicly available Ribo-seq datasets organized by protocol (initiating vs elongating ribosomes) and organism. Dozens of tracks are available for exploration in a genome browser, and links to other analysis tools are available through an integrated platform, RiboGalaxy[iv] [44]. Additional tools are also present in RiboTools[vii], another Galaxy repository for Ribo-seq analysis [45]. These resources can aid researchers in understanding

aided by the action of proteins that are part of the exon junction complex (EJC).
Open reading frame (ORF): a section of a transcript which contains a start and a stop codon in frame. In eukaryotes, most mRNA transcripts contain one main ORF that is translated into a polypeptide.
Puromycin-associated nascent chain proteomics (PUNCH-P): a technique that isolates nascent protein chains. Ribosome–nascent chain complexes are first isolated, and biotinylated puromycin is incorporated into the complexes. Streptavidin pulldown allows the nascent protein chains to be extracted, and these can by analyzed by LC-MS/MS.
Quantitative proteomics: proteomics techniques aimed at quantifying protein expression. Label-free quantification methods can be used, but techniques such as SILAC that label amino acids can represent superior alternatives for protein quantification.
Random forest classifier: a classification algorithm that combines the classification output of multiple classifiers, called decision trees. Each tree splits the data into different groups ('leaves') and assigns a label to each datapoint in each leaf. Each tree is applied to a subset of the data and features to avoid overfitting. Usually used as a supervised learning method, random forests can also be used for unsupervised learning and for regression tasks.
Regression: this aims to quantify the relationship between a target variable and one (or more) features. To this end, approaches fit a function that minimizes the distance between the predictor and the target variable (e.g., by using the least squares method). The regression coefficient quantifies the relationship between the target variable and the predictor.
Shotgun proteomics: a set of techniques that enable the identification and quantification of protein expression from a mixture of digested peptides, using peptide isolation (usually with liquid chromatography, LC) and tandem mass spectrometry (MS/MS). When they are eluted in the LC step, peptides are ionized, and ions are selected in the first MS step according to their mass-to-charge (*m*/*z*) ratio. Ions are then fragmented, and in the second MS step fragment

their data and also in better appreciating the expected performances of the ORF-finding strategies proposed so far.

ORF Finding with Ribo-seq Data

Ribo-seq, in principle, represents a highly suitable technique for the annotation of CDS regions. However, the identification of high-confidence translated regions from Ribo-seq data is not trivial because heterogeneity and biases in the experimental protocol pose a challenge in the interpretation of the obtained translation profiles. CDS regions only represent a subset of the transcriptome, and the precise identification of translation requires single-nucleotide resolution, which also depends on the overall size of the Ribo-seq library. Starting from early studies [10,46], Ribo-seq has been used to detect the translation of **upstream open reading frames** (uORFs), the use of non-canonical start codons, or translation of presumably non-coding RNAs (ncRNAs). Since then, different algorithms have been developed to identify genuine signal from active translation and to delineate ORF boundaries from transcriptome-wide data (Figure 2, Key Figure).

Translation Efficiency (TE)

TE indicates the amount of ribosomes normalized by transcript abundance, and was one of the first attempts to define actively translated transcripts [47]. Subsequent studies have however pointed out its limitations to define *bona fide* translated regions because it produces a large number of false positives [48].

Ribosome Release Score (RRS)

The RRS[xvii] [48] distinguishes translated from non-translated regions by exploiting the release of translating ribosomes at the stop codon. It is calculated as the ratio (normalized by RNA-seq reads) of RPFs in the CDS with RPFs in the 3′-UTR. At the global scale, the RRS score successfully retains many coding regions and discards known ncRNA regions. However, when RRS is combined with the TE metric it shows a clear separation between CDS and non-coding regions of the transcriptome (e.g., 3′-UTRs). An evaluation of RRS score sensitivity and specificity in detecting translated regions (e.g., over different expression regimes, using simulations, negative data) has never been performed.

Translation ORF Classifier (TOC)

Proposed by Chew *et al.* [49], TOC uses four different features extracted from the Ribo-seq and RNA-seq signal: the TE metric; Inside vs Outside, a metric containing the number of nt covered by Ribo-seq inside and outside the ORF; Fraction Length, representing the size of the ORF over the transcript length; the Disengagement Score, related to the RRS score. A **random forest classifier** is trained on the four different features classifying genes based on their coding behavior. The output of the **classifier** is a label for each locus which distinguishes between coding-like, trailer-like (3′-UTR, no reads), and leader-like (5′-UTR) loci. Most **long non-coding RNAs** (lncRNAs) with a Ribo-seq signal were assigned with a leader-like label, in other words they presented features not resembling *bona fide* protein-coding active translation, but leaving the functional relevance of their translation unanswered. Given the high sequencing depth of the Ribo-seq datasets used (>200 million mapped reads), the classifier also showed good performance on poorly expressed transcripts, but its performance on other datasets is unknown. It has not been released as a software tool for the community.

ORFscore and Related Approaches

In 2014 Bazzini *et al.* [11] produced a deep Ribo-seq dataset with precise subcodon resolution following zebrafish early development. Having precise information about the translated frame, they scored different ORFs based on the number of reads falling on the translated frame compared to a uniform distribution of signal over the three frames. This scoring method, named

ions are again isolated according their *m/z* ratio and quantified. Using a reference protein database, *m/z* values can be mapped to expected values matching peptides from known proteins.

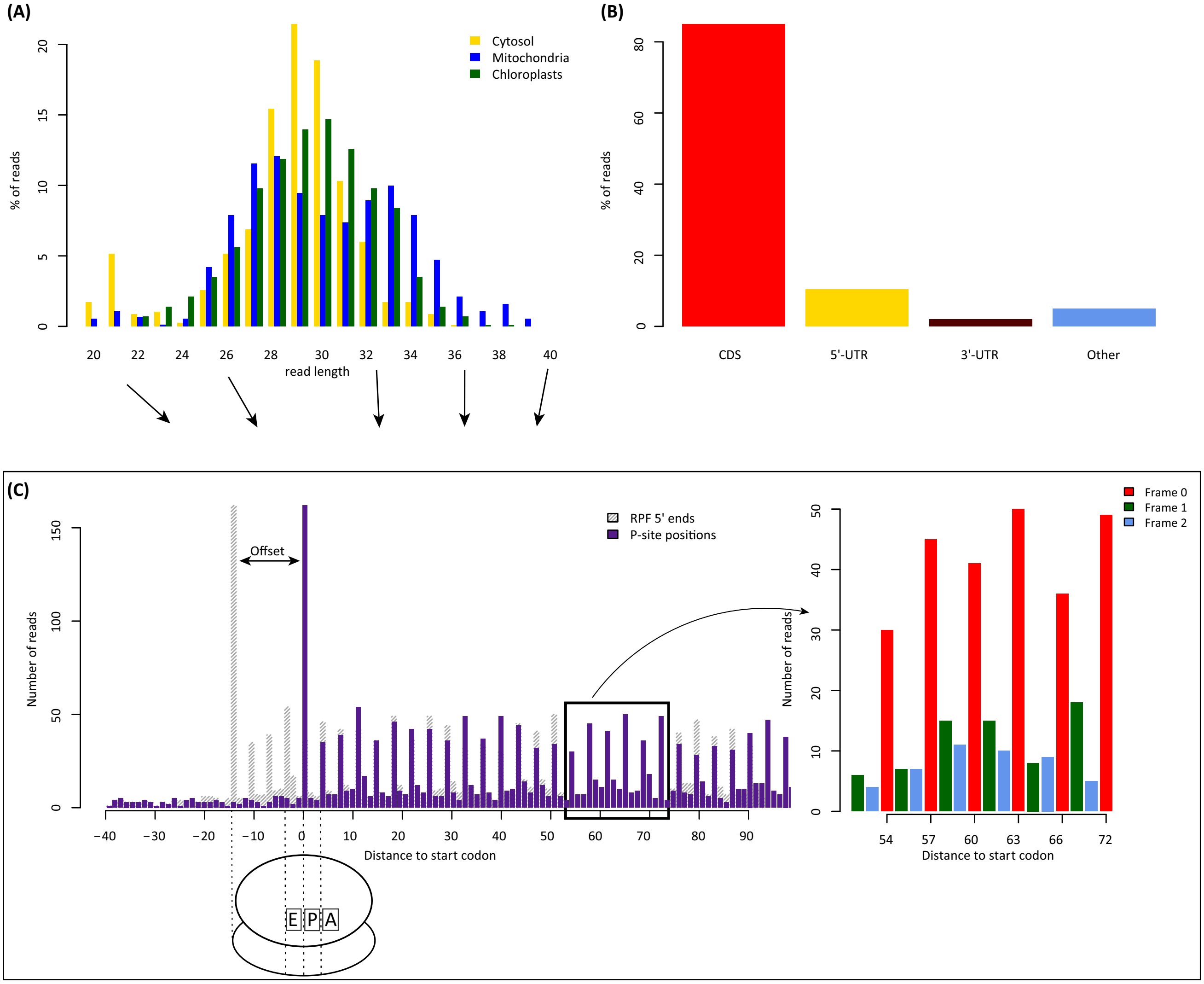
Spectral coherence: a measure of correlation between two frequency spectra. Signals exhibiting a similar set of frequency components will have high coherence.

Stable isotope labeling with amino acids in cell culture (SILAC), and pSILAC: pSILAC is a variant of SILAC in which labeled amino acids are added to the cell culture for short periods of time, thus allowing the kinetics of *de novo* protein synthesis to be monitored.

Support vector machine (SVM) classifier: a binary classification algorithm. SVMs are supervised learning methods and therefore need to be trained on known examples. In the training stage, SVMs aim to define a separating line maximizing the distance between the two sets of data. When a linear separation of the two sets is not effective, SVMs can compute the distance between datapoints in a higher-dimensional space by means of different kernel functions in which a linear separation between the samples is possible. This strategy (the ‘kernel trick’) enables non-linear classification, and has contributed to the popularity of SVMs in the machine-learning community.

Untranslated region (UTR): the section of a coding mature mRNA that does not code for protein. The 5′-UTR is located upstream of the start codon, while the 3′-UTR is downstream of the stop codon.

Upstream open reading frame (uORF): a small (usually <100 aa) ORF whose start codon is located in the 5′-UTR upstream of the main ORF of a transcript. Many uORFs have been shown to regulate the translation of the main ORF. It is generally assumed that uORFs do not encode stable polypeptides.



Trends in Genetics

Figure 1. Quality Control of Ribo-seq Data. (A) The read-length distribution of ribosomal footprints, which may vary between different ribosomal complexes. (B) Read-mapping statistics, where most of the footprints are expected to map to coding sequences (CDS), followed by the 5′-untranslated region (5′-UTR) and other regions. (C) Analysis of subcodon resolution: for each read length, an aggregate profile of the 5′ ends of ribosome-protected fragments (RPFs) is built over open reading frame (ORF) boundaries, here shown for start codons (left). 5′ Ends can be offset to the annotated AUG, allowing precise identification of the ribosomal P-site position over the translated ORFs and reveal translation at single-frame resolution (right).

ORFscore, allowed them to identify a set of translated small ORFs (<100 aa) which were overlooked by automatic annotation pipelines. Despite its high sensitivity and specificity on a deep Ribo-seq dataset in zebrafish (~200 million reads), the performance of ORFscore on different datasets is unknown.

Along similar lines, two other studies in yeast used the subcodon resolution of Ribo-seq reads to identify translated ORFs [50,51]. In one of the two studies, a false discovery rate in ORF identification was calculated using a randomized distribution of P-sites over the three frames [51], in other words following the same assumption as ORFscore. None of these approaches were made available as documented software for the community (ORFscore calculation is now included in the SPECtre[xviii] package).

Box 1. Reproducibility of the Ribo-seq Protocol

Given that different variants of the protocol might influence the obtained profiles, the reproducibility of Ribo-seq experiments has been investigated at different levels. Reproducibility has been assessed by comparing total counts at the gene/transcript level across replicates, often exhibiting excellent correlation. However, Diament and Tuller [97] showed that individual single-nucleotide profiles across transcripts were poorly correlated across biological replicates, using data from different organisms and laboratories. This is of crucial importance, especially considering the importance of subcodon resolution for the many available ORF-finding tools. Given the known, sometimes concentration-dependent, effects of translation inhibitors and other variables [24], O'Connor *et al.* [41] showed how such biases (together with additional features) are dataset-specific and can be used to predict the coverage profiles across single transcripts with good accuracy. The impact that such biases might have when assessing biological variability using Ribo-seq remains to be investigated.

Fragment Length Organization Similarity Score (FLOSS)

The average length of an RPF is ~29 nt. In addition, different contaminants such as rRNA, snoRNAs, and other structured RNA fragments can survive purification steps and thus be sequenced. FLOSS aims at distinguishing between 80S ribosomal footprints and signal from contaminant sources [52]. The idea behind the score is to learn a reference distribution of Ribo-seq fragment lengths on protein-coding regions, which represents actively translating ribosomes. Fragment length distribution over each region in the transcriptome is then compared to the reference distribution to derive a similarity score indicative of its coding-like validity, taking into account the total Ribo-seq coverage. As expected, the FLOSS scores globally distinguish coding from non-coding genes. However, even for some predominantly nuclear lncRNAs such as MALAT1 [53], short elements along the transcript might exhibit a coding-like behavior, thus being masked by the total signal over the transcript. An in-depth analysis of FLOSS performance, to evaluate sensitivity and specificity at different data depths, was not tackled. The method is available as a set of annotated scripts in a supplementary file (now part of the SPECtre package) and was applied to a very deep Ribo-seq dataset in a mouse cell-line (>250 million mapped reads).

ORF-Rater

In the ORF-Rater[xii] strategy [54], aggregate profiles over start and stop codons are used to identify translated regions. Because these profiles become prominent in HARR- or LTM-treated Ribo-seq datasets, the application of multiple Ribo-seq protocols to the same biological samples produces distinct profiles for many translated ORFs. The core of the ORF-Rater method is a **regression** fit (using **non-negative least squares**) of Ribo-seq coverage along the transcript (coming from the multiple Ribo-seq protocols) against its expected coverage given the presence of one (or multiple) translated ORFs. The presence of ORF translation is indicated by a positive regression coefficient of the fit. To evaluate the statistical confidence of ORF translation, a random forest classifier is trained on regression results from known ORFs and used to score the regression fit for the ORFs candidates, yielding a set of ~13 000 translated ORFs that fall into known and novel genomic regions.

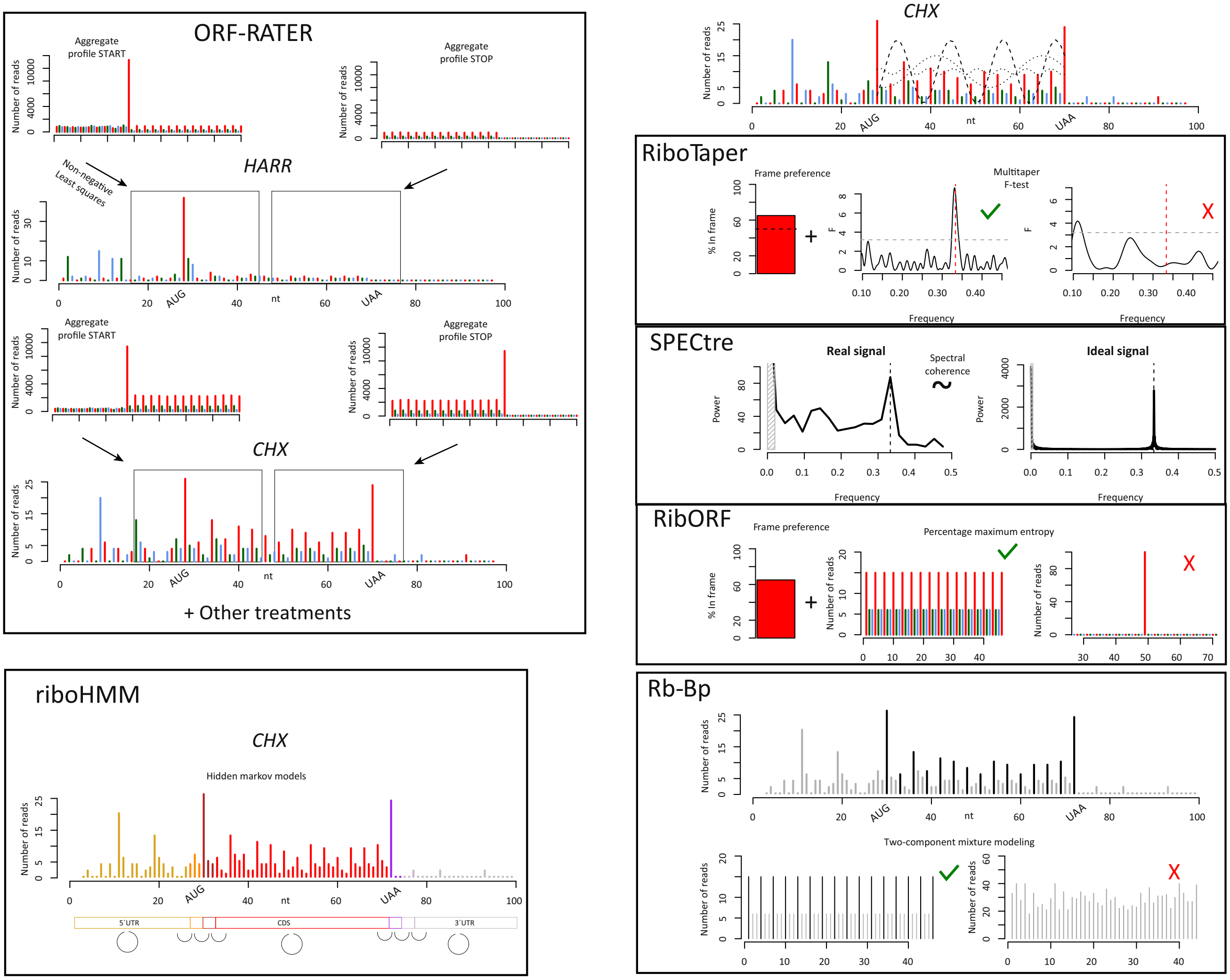
Leveraging on expected profiles at start and stop codons, the ORF-Rater method can identify ORF truncation/extension, out-of-frame ORFs, and small ORFs. The method is implemented as a software available at a Github repository, with well-documented scripts and detailed instructions.

RibORF

RibORF[xv] [55] uses the subcodon resolution of Ribo-seq reads to identify translation. In addition to the amount of in-frame Ribo-seq reads, the method uses percentage maximum entropy (PME) to model a uniform coverage of reads along the ORF. The percentage of in-frame reads and the PME metric are calculated for each ORF in the transcriptome, and a is a method which**support vector machine (SVM) classifier** is used to separate good ORF predictions from unreliable results. Around 10 000 translated genes were detected when applying RibORF

Key Figure

Open Reading Frame (ORF)-Finding Strategies Using Ribo-seq



Trends in Genetics

Figure 2. On the left, ORF-Rater and riboHMM which use features of translation initiation, elongation, and termination. On the right, RiboTaper, SPECtre, RibORF, and Rb-Bp which use the elongating frame to detect translation. Using a toy example, a graphical representation of data requirements is provided for each tool, together with a sketch of the main analytical strategies employed.

to two average-sized Ribo-seq datasets in human cell lines (~40 million reads). Performance metrics were shown as a function of expression levels. The tool is implemented as software available for the community, including essential usage instructions.

RiboTaper

The RiboTaper[xvi] method [16] also exploits subcodon resolution as key feature to identify translation. The method identifies ORFs where Ribo-seq reads display a 3 nt periodic behavior consistent with active translation elongation. To do so, RiboTaper uses the **multitaper** strategy that was introduced by Thomson in 1982 [56]. In this manner, ∼12 000 translated genes are detected on an average depth Ribo-seq dataset from HEK293 cells (∼30 million reads), together with hundreds of genes harboring novel translated ORFs (e.g., uORFs, ORFs in ncRNAs). Measures of sensitivity and specificity are derived from simulations, benchmark application on RNA-seq data, and the use of additional datasets as well as proteomics data from the same cell line. The algorithm has been applied to Ribo-seq datasets from different organisms, including *Arabidopsis thaliana* [37]. An implementation in software, documentation, and guidelines for use, as well as integration in Galaxy, are available.

SPECtre

In the SPECtre [57] method, **spectral coherence** (which measures the correlation between two different frequency spectra) is used to indicate whether the periodic components in the P-sites profile match an ideal profile where reads map only to the translated frame. After normalization of the P-site tracks, the algorithm classifies individual transcripts into translated or not translated. Sensitivity and specificity are addressed at different amounts of input data. The software is well-documented and is available on a public repository.

RiboHMM

Despite a different modeling framework, the riboHMM[xiv] [58] strategy to detect translated ORFs uses a similar idea to ORF-Rater. A **hidden Markov model** (HMM) is trained to recognize distinct Ribo-seq profiles over different ORF positions, leveraging the distinct pattern of Ribo-seq over start and stop codons as well as inside the translated CDS. RiboHMM also explicitly models the contribution of each Ribo-seq read length and sums them to increase sensitivity. The trained HMM is used to parse the Ribo-seq signal transcriptome-wide, yielding predictions for ORF translation. RiboHMM identified ∼36 000 translated transcripts, covering ORF annotations for 7801 annotated protein-coding genes and thousands of novel candidate ORFs, with a large fraction of the latter falling in 5′-UTRs (i.e., uORF candidates). At lower library depths, the algorithm was shown to be robust in terms of its false positive rate despite a marked decrease in sensitivity. The method is available as well-documented software.

Rb-Bp

The Rb-Bp[xiii] [59] strategy uses a probabilistic graphical model to predict translation from P-site profiles. The model is trained on patterns for which a clear enrichment over the translated frame is observed, and it scores ORFs for whether they resemble such patterns or a null uniform model. On an average-sized HEK293 dataset, the algorithm identified ∼17 000 ORFs, including >2000 ORFs in ncRNAs. As with RiboTaper, the predictions of the algorithm were validated with proteomics support and QTI-seq data. The algorithm was run on different datasets of mode depth. An evaluation of the specificity or sensitivity of the method was not extensively presented. The software is well-documented and available on Github.

Figure 2 summarizes the different strategies and their approaches. The diversity of the methods is also reflected by different assumptions and data requirements, outlined below.

Data Requirements

The outlined approaches differ in terms of data requirements and assumptions because they rely on different statistical methods to identify translation (Figure 2). Moreover, the outlined tools differ in terms of start codon definition, software implementation, and other features, as summarized in Table 1.

Tools such as RiboTaper, Rb-Bp, Spectre, and RibORF rely on features reflecting the dynamics of elongation to find translated ORFs, and they were applied to datasets of average coverage

Table 1. An Overview of the Available ORF-Finding Algorithms Using Ribo-seq

Method	*De novo* ORF finding; start codon	Sensitivity assessment(s)	Specificity assessment(s)	Input data	Specific requirements	Datasets analyzed; depth	Validation strategy	Implemented as software; detailed usage guidelines	Version control; strategy	Computational requirements; expected runtime	Ref.
TE	NA[a]	Retains coding	Discards non-coding	Ribo-seq; RNA-seq	Ribo-seq coverage over RNA abundance	Several Datasets	NA	NA	NA	NA	[10]
RRS	NA	Retains coding	Discards non-coding	Ribo-seq; RNA-seq	Coverage drop after stop codon	1	NA	Yes; NA	NA	NA	[48]
TOC	Yes; AUG/CUG	Retains coding; vs coverage	Discards non-coding; vs coverage	Ribo-seq; RNA-seq	Ribo-seq coverage over RNA abundance, ORF structure	NA	NA	NA	NA	NA	[49]
ORFScore	NA	Retains coding	Discards non-coding	Ribo-seq	Frame preference	1; >100 million reads	Proteomics; conservation	In SPECtre	NA	NA	[11]
FLOSS	NA	Retains coding	Discards non-coding	Ribo-seq	Read length distribution	NA	Ribosome affinity purification; individual validations	In SPECtre; (documented scripts in original publication)	NA	NA	[52]
ORF-Rater	Yes; NUG	Retains coding	Discards non-coding; agreement with annotation	Different Ribo-seq variants	Frame preference, aggregate profiles over start/stop codons	2; >100 million reads	Proteomics; conservation; individual validations	Yes; Yes	Yes; Github repository	HPC cluster; >1 day	[54]
RibORF	Yes; NUG	Retains coding	Discards non-coding	Ribo-seq	Frame preference; signal uniformity	Several datasets	Conservation	Yes; Yes	NA	NA	[55]
RiboTaper	Yes; AUG	Retains coding; vs coverage; proteomics	Discards non-coding; vs simulations; benchmark on RNA-seq	Ribo-seq; RNA-seq	Frame preference; 3 nt periodicity	Several datasets	Initiation mapping; Proteomics; Conservation	Yes; Yes	Yes; updates on website.	HPC cluster; ≥1 day	[16]
SPECtre	NA	Retains coding; vs coverage	Discards non-coding; vs coverage	Ribo-seq, quantification estimates	3 nt periodicity	1	NA	Yes; Yes	Yes; Github repository	NA; <1 day	[57]
RiboHMM	Yes; NUG and others	Retains coding; vs coverage	Discards non-coding; vs coverage	Ribo-seq	Frame preference, aggregate profiles over start/stop codons	1; >100 million reads	Initiation mapping; proteomics; conservation	Yes; Yes	Yes; Github repository	HPC cluster; >1 day	[58]
Rb-Bp	Yes; NUG and others	Retains coding	Discards non-coding; proteomics	Ribo-seq	Frame consistency	Several datasets	Initiation mapping; proteomics	Yes; Yes	Yes; Github repository	HPC cluster; <1 day	[59]

[a] NA, not available.

(<50 million reads) to identify translation for thousands of genes. ORF-Rater and riboHMM use features around translation initiation, elongation, and termination as distinctive features of translation and were applied to high-coverage datasets (ORF-Rater: ~150 million reads per each of the four Ribo-seq variants; riboHMM, >500 million reads). The ORF-Rater pipeline also requires data from multiple Ribo-seq protocols to extract additional profiles around start/stop codons and thus provide a more reliable set of predictions. Precise subcodon resolution is essential for the first group of tools, while high-coverage datasets and the availability of multiple types of experiments might favor the other strategies.

All these analysis strategies aim at detecting translation, regardless of its genomic location (e.g., uORFs or ORFs in lncRNAs). The identification of short ORFs in the non-coding transcriptome might require high-coverage datasets, perhaps in conjunction with multiple Ribo-seq protocol variations. In any case, ORF identification in UTRs or lncRNAs is more challenging, and false identification can dramatically increase. Benchmarking analyses will be necessary to investigate the performance of the outlined strategies in detecting small ORF translation, for instance using simulated Ribo-seq experiments and introducing different levels of resolution, read coverage, and non-canonical translation events, akin to RNA-seq simulations [60].

Moreover, a detailed performance of the outlined methods (when provided) is often available for one dataset only, thus limiting our current understanding of their relative performance. As discussed in the previous sections, Ribo-seq datasets show high variability between laboratories, and this problem may have influenced the choice of datasets used to evaluate method performance. In fact, none of the presented algorithms is designed to intrinsically handle biological replicates, and, although agreements across a few datasets have been reported, a thorough assessment of reproducibility of detected ORFs has not been tackled. Concerted efforts for example from genomics consortia would help to alleviate this bottleneck because only a few studies investigating the effect of protocol variations are currently available.

In addition to these recently published methods, prior studies addressed the identification of translated ORFs with Ribo-seq using custom analysis strategies, as in the comprehensive PROTEOFORMER pipeline [61]. Of particular note is a study in murine myoblast differentiation, for which the entire analysis workflow is freely available [62]. Moreover, the strategies discussed so far deal with ORF finding in eukaryotic systems, and were not originally applied to bacterial data. However, Ribo-seq has also been performed in bacteria, for instance to study ribosomal pausing [63] or to accurately quantify protein synthesis rates [64]. Additional efforts are now exploring the presence of novel translation events in bacteria using Ribo-seq [65].

Alternative Translation Events

Features of Ribo-seq signals along the translated frames can also be used to identify the presence of non-canonical events. Using a change-point algorithm, Zupanic *et al.* [66] detected sharp changes in the Ribo-seq coverage to identify novel initiation sites, premature stop-codon usage, and novel splice junctions. Despite convincing support for individual new events, it is not clear to what extent the detected change-point events all reflect the presence of true alternative translation events, especially considering the high non-uniformity of the Ribo-seq signal.

Leveraging the subcodon resolution of Ribo-seq reads, Michel *et al.* [67] developed a strategy to identify regions where translation occurs in multiple frames. The authors identified ~100 candidates where the ribosomal coverage switches between two different frames along a single transcript, mostly due to overlapping translation of multiple ORFs. Additional attempts were made to identify rare read-through events using Ribo-seq. In *Drosophila*, dozens of targets were identified despite the lack of subcodon resolution in the data [68]. While rare, such events appear to be conserved, with some genes consistently being identified across different studies [105].

As the kinetics of elongation create a distinct periodic pattern over an ORF, regions displaying high occupancy on isolated regions can indicate the presence of non-ribosomal complexes. In the Rfoot[iii] [69] method, the principle underlining the RibORF [55] approach (see above) is employed to find regions exhibiting isolated pile-ups of Ribo-seq signal. Such regions have been shown to map to small RNAs and also to structured regions of lncRNAs and 3′-UTRs.

Integration with Other Omics Data

Translation output of is of course dependent on transcript steady-state abundance in the cytosol. RPFs have been shown to highly correlate with RNA-seq counts: since early studies [10] RNA-seq has been used together with Ribo-seq to account for transcript abundance and enable the calculation of transcript-specific translation rates. This analysis strategy opened the door for the quantification of translational control, with a great increase in statistical tools tackling this problem, many of which use already established statistical methods for the analysis of count data (Box 2).

Coupled with RNA-seq information, Ribo-seq has furthermore been used to understand the functions of many regulators such as miRNAs [70] or to quantify the impact of poly(A) tail length dynamics [71].

Additional techniques such as polysome profiling might complement the shortcomings of Ribo-seq, as they can show the extent of translation regulation over full-length transcripts, retaining information about UTR structure [8,9]. However, a thorough investigation of the quantitative aspects of polysome profiling data has not been carried out. The comparison between the profile over polysomal fractions and protein synthesis rates is not straightforward. Even the monosomal fraction has been shown to contain translated RNAs, mostly containing short ORFs encoding regulatory proteins [72]. Most importantly, information about the translated sequences is lacking in polysome profiling data.

Precise information about translated ORF coordinates allows us to zoom in on the functions of translation (Figure 3), and thus consider additional aspects of RNA biology. For instance, different RNA surveillance pathways, such as **nonsense-mediated decay** (NMD), act in a translation-dependent fashion [73]. Ribo-seq data can thus provide links to other aspects of the RNA life cycle [74]. Such integrative efforts provide a more comprehensive view on the transcriptome, including valuable insights on the possible function of lncRNA transcripts, many of which have been shown to undergo some level of translation in most of the ORF-finding studies described above. However, the question remains open of whether such translation

Box 2. Detecting and Quantifying Translation Control with Ribo-seq and RNA-seq

To quantify translation relative to RNA levels, the translation efficiency measure (TE) was introduced into the original Ribo-seq study [10]. The TE metric provides a quantitative measure of translation per transcript by dividing the Ribo-seq by the RNA-seq signal. Despite being used in many publications, the consistency of the TE measure for indicating translation changes has been called into question in follow-up studies [98], and a few alternatives have been proposed to better detect differential translation. After a first wave of analytical strategies such as Babel[xx] [99] and Anota[xix] [100], recent tools including RiboDiff[xxi] [101], Xtail[xxiii] [102], and Riborex[xxii] [103] employ generalized linear model strategies akin to DESeq2 [104] to model the Ribo-seq and RNA-seq read counts and obtain distributions of fold changes between conditions. Modeling the two distributions enables significance testing for genes belonging to a concordant versus discordant mode of regulation (on the translation or expression level) across the assayed conditions. Alternative approaches for the analysis of the distribution of RNA-seq and Ribo-seq fold changes also aim to distinguish between gene regulation at the level of steady-state transcript abundance and the level of translation [12].

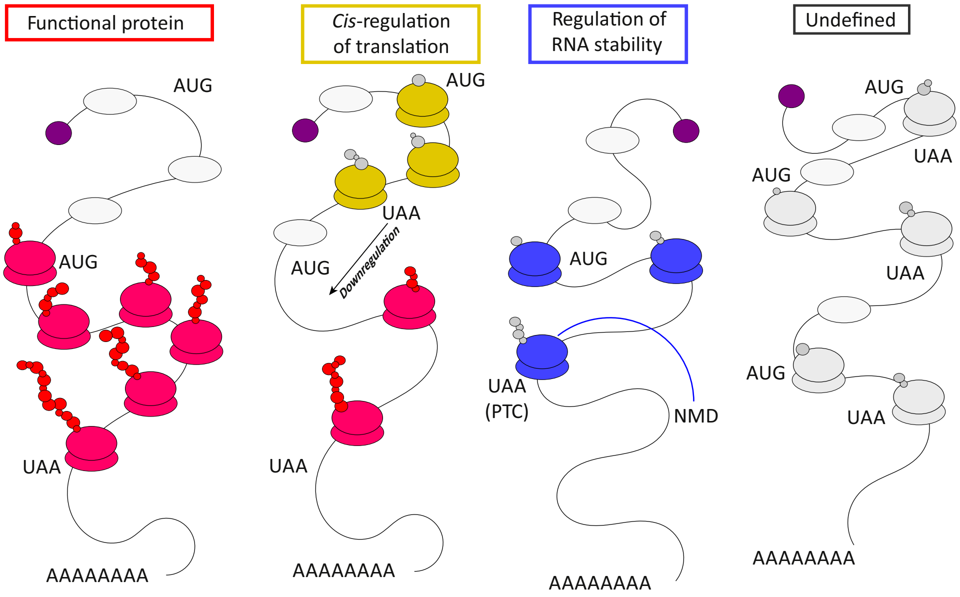


Figure 3. Multiple Roles of mRNA Translation. Together with the synthesis of important proteins (in red, left), translation has been shown to have regulatory roles such as influencing the usage of additional ORFs in *cis* (e.g., uORFs, yellow) or regulating transcript stability (in blue), for instance by triggering NMD via the recognition of a PTC. In other instances, translation might simply occur on a cytosolic transcript without any important function. Multiple scenarios can occur on the same RNA molecule. Abbreviations: NMD, nonsense-mediated decay; ORF, open reading frame; uORF, upstream ORF; PTC, premature stop codon.

events produce important proteins or might have other regulatory functions (as in the case of uORFs [75]).

Evolutionary conservation patterns can help pinpointing such differences by contrasting nucleotide-level conservation to possible constraints at the codon composition level [16,76]. However, the pervasive nature of translation of many lncRNAs and 5′-UTRs challenges our abilities to define the boundaries of translation and thus infer biological function (if any).

The primary function of ribosomes is to synthesize proteins. The Ribo-seq method thus represents a powerful link between transcriptomics and proteomics techniques. Such links allow us to improve gene annotation and protein detection [61]. However, a comparison between the two worlds of RNAs and proteins, depicted by Ribo-seq and **shotgun proteomics**, respectively, must carefully consider fundamental differences between the two approaches.

Despite relatively weak standardization of the experimental protocol, the excellent sensitivity of Ribo-seq enables us to identify translation even for lowly expressed genes and small ORFs, which has generated considerable excitement in the research community, with many publications focusing on the detection and characterization of small peptides [77]. On the other hand, proteomics methods rely on extremely precise detection of fragmented peptide ions. Current shotgun proteomics methods can provide evidence for only a subset (usually <5000) of the synthesized proteins. Such limitations may result from the inefficient detection of the entire spectrum of possible peptides. Another aspect to consider is the presence of nearly ~200 post-translational modifications (PTMs) that can occur in multiple protein residues. At least a dozen PTMs are common, and only a limited number can be considered when trying to reconstitute the original mixture of proteins. It has been recently shown how different PTMs can cause 20–50% of false peptide identification, producing modified spectra which can perfectly match a different peptide sequence [78].

In a similar fashion, the choice of sequence database has a heavy impact on results, a well-documented and discussed phenomenon in the proteomics community [79]. In our own study we could confirm how the set of ORFs identified by RiboTaper alone can achieve excellent coverage of the detectable proteome, underlining the validity of our ORF identification strategy [16]. However, the size of the protein database will have an impact on the quality of peptide identification, and researchers must perform additional analysis to validate the results of the custom database search. To confirm the presence of novel peptides in our work, we merged a database built from RiboTaper ORF candidates with the entire catalog of annotated proteins, and then used the merged database to confirm the presence of novel peptides [16]. Because Ribo-seq data is available for a variety of conditions, it is now possible to apply this strategy to the annotation of multiple tissues and different species for which several mass spectrometry datasets are available for reanalysis [80,81].

To further confirm the presence of alternative proteoforms, techniques such as N-terminal COFRADIC aim at isolating the N-termini of the synthesized proteins, aiding the quantification of alternative start-codon usage [82,83]. However, the low throughput of this technique limits our ability in deriving reliable performance metrics for the identification of alternative translation events together with their protein products.

Many studies observed good proteome-wide correlation between protein steady-state abundance estimates and translation quantification estimates. Again, many factors must be carefully considered when performing such a correlative analysis [84]. Ribo-seq reflects the protein synthesis rates of different ORFs, which would ideally correspond to the estimates given by **quantitative proteomics** techniques such as the pSILAC variant of **stable isotope labeling with amino acids in cell culture** (SILAC) [85] or **puromycin-associated nascent chain proteomics** (PUNCH-P) [86]. Quantification estimates from Ribo-seq and the two mentioned techniques have been shown to highly correlate [87,88]. However, as in many other studies integrating transcriptomics and proteomics data [5], the analysis was limited to a few thousand genes, which represent only a fraction of the endogenous proteome, especially when compared to more sensitive estimates from RNA-seq or Ribo-seq.

Overall, a cautious attitude must be taken before dismissing the potential of proteomics methods; equating ribosomal density to the production of a stable functional protein product would represent an oversimplistic approach, ignoring many additional aspects of as-yet poorly understood phenomena such as cotranslational protein folding and degradation [89].

Concluding Remarks and Future Directions

Ribo-seq provides a very rich and detailed description of a key step in the mRNA life cycle. Thanks to the use of different compounds affecting different aspects of the kinetics of translation, we can extract multiple meaningful features from ribosome profiles over thousands of genes, enabling distinct methods of analysis to study translation in entire transcriptomes.

We have collected and discussed different analysis strategies to identify translation in Ribo-seq data, together with important aspects about their availability, assumptions, and possible limitations. Unfortunately, our incomplete understanding of the impact of different steps in the Ribo-seq protocol might pose a problem for the level of resolution required by most of the presented approaches, underlining the need for replicate information when assessing the quality of the inferred ORF boundaries.

Outstanding Questions

Many variations of the Ribo-seq protocol have been applied by different laboratories over recent years. Despite its widespread usage, has Ribo-seq become a standardized technique?

Are Ribo-seq data alone a good proxy to detect and quantify the nascent proteome?

How many translated small ORF have important biological functions?

Can we use Ribo-seq data to quantify the functional heterogeneity of the expressed transcriptome?

Together with many other useful approaches to study protein synthesis [90], a new exciting alternative is represented by the TCP-seq method [30], where all stages of translation can be captured in a single experiment, which is albeit slightly more involved than Ribo-seq. Distinct ribosomal states can be inferred by the captured read lengths, providing additional mechanistic insights about the translation process. The additional information coming from the read-length distribution adds a crucial feature and can be used as a hallmark of translation. As for Ribo-seq, the presence of multiple useful features requires a more general analysis approach that is not tailored to one specific feature only (e.g., 3 nt periodicity, slow initiation, etc.).

Deep neural networks might prove a valuable analytical strategy to infer multiple features directly from the data itself, and in turn use those features to classify different genomic regions. The first deep-learning approaches for Ribo-seq data have predicted translation initiation sites [91] and ribosome stalling events [92].

We should not ignore that translation is not only about protein synthesis: Ribo-seq data must be interpreted in the broader context of RNA metabolism. Multiple RNA degradation pathways have been characterized, and several of them occur in a translation-dependent fashion. Quantitative estimates of translation can be integrated with rates of RNA synthesis, degradation, and additional features to better aid the functional characterization of the expressed transcriptome [74].

This avenue is however currently limited by our ability to account for the functional heterogeneity derived from alternative splicing. A very recent study has shown that alternative splicing events can in principle be identified using Ribo-seq [93], confirming the presence of a translated alternative transcriptome [94]. However, the lack of evidence at the protein level challenges the functional interpretation of such events, favoring the ‘one gene – one protein’ view of protein synthesis [95]. Again, the translation machinery might have a primary impact on the fate of RNA molecules rather than solely representing a way to expand the cellular proteome.

Moreover, the ribosome itself represents a central node in post-transcriptional regulation because it can recruit different protein complexes to the mRNA and thus regulate its fate. The profiling of different ribosomal complexes and their interactome will further expand our understanding of the link between translation and other molecular pathways [96]. The correct identification and quantification of ribosomal movement across different transcripts is an excellent starting point to focus on the spectrum of functions of the transcriptome, from protein synthesis to other mechanisms of gene expression control.

Acknowledgments

The authors received funding from the German Federal Ministry of Education and Research (BMBF grant 031 A538A RBC, de.NBI).

Resources

Preprocessing and Quality Control

[i]Cutadapt: http://cutadapt.readthedocs.io/en/stable/index.html

[ii]Plastid: http://plastid.readthedocs.io/en/latest/

[iii]Rfoot: www.broadinstitute.org/~zheji/software/Rfoot.0.1.tar.gz

[iv]RiboGalaxy: http://ribogalaxy.ucc.ie

[v]RiboProfiling: https://bioconductor.org/packages/release/bioc/html/RiboProfiling.html

[vi]RiboSeqR: http://bioconductor.org/packages/release/bioc/html/riboSeqR.html

[vii]RiboTools: https://testtoolshed.g2.bx.psu.edu/repository?repository_id=d5063de8c6e62bf2

[viii]RUST: http://lapti.ucc.ie/rust/

[ix]STAR: https://github.com/alexdobin/STAR

Databases

[x]GWIPS-viz: http://gwips.ucc.ie

[xi]RPFdb: http://sysbio.sysu.edu.cn/rpfdb/index.html

ORF Finding

[xii]ORF-Rater: https://github.com/alexfields/ORF-RATER

[xiii]Rb-Bp: https://github.com/dieterich-lab/rp-bp

[xiv]RiboHMM: https://github.com/rajanil/riboHMM

[xv]RibORF: https://personal.broadinstitute.org/zheji/software/RibORF.html

[xvi]RiboTaper: https://ohlerlab.mdc-berlin.de/software/

[xvii]RRS: www.lncrna-test.caltech.edu/software.php

[xviii]SPECtre: https://github.com/mills-lab/spectre

Differential Analysis

[xix]Anota: www.bioconductor.org/packages/release/bioc/html/anota.html

[xx]Babel: https://cran.r-project.org/web/packages/babel/index.html

[xxi]RiboDiff: https://github.com/ratschlab/RiboDiff

[xxii]Riborex: https://github.com/smithlabcode/riborex

[xxiii]Xtail: https://github.com/xryanglab/xtail

References

1. Mortazavi, A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628
2. Churchman, L.S. and Weissman, J.S. (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* 469, 368–373
3. De Hoon, M. and Hayashizaki, Y. (2008) Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *Biotechniques* 44, 627–632
4. Rabani, M. *et al.* (2011) Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* 29, 436–442
5. Schwanhäusser, B. *et al.* (2011) Global quantification of mammalian gene expression control. *Nature* 473, 337–342
6. Vogel, R.C. *et al.* (2010) Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* 6, 400
7. Arava, Y. *et al.* (2003) Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* 100, 3889–3894
8. Sterne-Weiler, T. *et al.* (2013) Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res.* 23, 1615–1623
9. Floor, S.N. and Doudna, J.A. (2016) Tunable protein synthesis by transcript isoforms in human cells. *Elife* 5, e10921
10. Ingolia, N.T. *et al.* (2009) Genome-wide analysis in vivo of resolution using ribosome profiling. *Science* 324, 218–223
11. Bazzini, A.A. *et al.* (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* 33, 981–993
12. Schafer, S. *et al.* (2015) Translational regulation shapes the molecular landscape of complex disease phenotypes. *Nat. Commun.* 6, 7200
13. Andreev, D.E. *et al.* (2017) Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Res.* 45, 513–526
14. Ingolia, N.T. *et al.* (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* 7, 1534–1535
15. Bartholomäus, A. *et al.* (2016) Mapping the non-standardized biases of ribosome profiling. *Biol. Chem.* 397, 23–35
16. Calviello, L. *et al.* (2015) Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* 13, 1–9
17. Couvillion, M.T. *et al.* (2016) Synchronized mitochondrial and cytosolic translation programs. *Nature* 533, 1–17
18. Andreev, D.E. *et al.* (2015) Translation of 5′ leaders is pervasive in genes resistant to eIF2 repression. *Elife* 4, e03971
19. Lee, S. *et al.* (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U. S. A.* 109, E2424–E2432
20. Gao, X. *et al.* (2015) Quantitative profiling of initiating ribosomes in vivo. *Nat. Methods* 12, 147–153
21. Rooijers, K. *et al.* (2013) Ribosome profiling reveals features of normal and disease-associated mitochondrial translation. *Nat. Commun.* 4, 2886
22. Chotewutmontri, P. *et al.* (2016) Dynamics of chloroplast translation during chloroplast differentiation in maize. *PLoS Genet.* 12, e1006106
23. Gerashchenko, M.V. and Gladyshev, V.N. (2014) Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res.* 42, 1–7
24. Hussmann, J.A. *et al.* (2015) Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. *PLoS Genet.* 11, e1005732
25. Dunn, J.G. *et al.* (2013) Ribosome profiling reveals pervasive and regulated stop codon read through in *Drosophila melanogaster*. *Elife* 2013, e01179
26. Gerashchenko, M.V. and Gladyshev, V.N. (2017) Ribonuclease selection for ribosome profiling. *Nucleic Acids Res.* 45, e6
27. Miettinen, T.P. and Bjorklund, M. (2015) Modified ribosome profiling reveals high abundance of ribosome protected mRNA fragments derived from 3′ untranslated regions. *Nucleic Acids Res.* 43, 1019–1034
28. Kivioja, T. *et al.* (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9, 72–74
29. Lareau, L.F. *et al.* (2014) Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *Elife* 3, e01257
30. Archer, S.K. *et al.* (2016) Dynamics of ribosome scanning and recycling revealed by translation complex profiling. *Nature* 535, 570–574
31. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal* 17, 10–12
32. Dobin, A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21
33. Baruzzo, G. (2016) Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods* 14, 135–139
34. Trapnell, C. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515
35. Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, 323

36. Kahles, A. *et al.* (2015) MMR: a tool for read multi-mapper resolution. *Bioinformatics* 32, 770–772

37. Hsu, P.Y. *et al.* (2016) Super-resolution ribosome profiling reveals novel translation events in *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* 113, E7126–E7135

38. Chung, B.Y. *et al.* (2015) The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis. *RNA* 21, 1731–1745

39. Popa, A. (2016) RiboProfiling: a Bioconductor package for standard Ribo-seq pipeline processing. *F1000Research* 5, 1309

40. Dunn, J.G. and Weissman, J.S. (2016) Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. *BMC Genomics* 17, 958

41. O'Connor, P.B.F. *et al.* (2016) Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. *Nat. Commun.* 7, 12915

42. Xie, S.-Q. *et al.* (2016) RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res.* 44, D254–D258

43. Michel, A.M. *et al.* (2014) GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res.* 42, D859–D864

44. Michel, A.M. *et al.* (2016) RiboGalaxy: a browser based platform for the alignment, analysis and visualization of ribosome profiling data. *RNA Biol.* 13, 316–319

45. Legendre, R. *et al.* (2015) RiboTools: a galaxy toolbox for qualitative ribosome profiling analysis. *Bioinformatics* 31, 2586–2588

46. Ingolia, N.T. *et al.* (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147, 789–802

47.. Ruiz-Orera, J. *et al.* (2014) Long non-coding RNAs as a source of new peptides. *Elife* 3, e03523

48. Guttman, M. *et al.* (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154, 240–251

49. Chew, G.-L. *et al.* (2013) Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. *Development* 140, 2828–2834

50. Smith, J.E. *et al.* (2014) Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*. *Cell Rep.* 7, 1858–1866

51. Duncan, C.D.S. and Mata, J. (2014) The translational landscape of fission-yeast meiosis and sporulation. *Nat. Struct. Mol. Biol.* 21, 641–647

52. Ingolia, N.T. *et al.* (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* 8, 1365–1379

53. Hutchinson, J.N. *et al.* (2007) A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* 8, 39

54. Fields, A.P. *et al.* (2016) A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol. Cell* 60, 816–827

55. Ji, Z. *et al.* (2015) Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* 4, e08890

56. Thomson, D.J. (1982) Spectrum estimation and harmonic analysis. *Proc. IEEE* 70, 1055–1096

57. Chun, S.Y. *et al.* (2016) SPECtre: a spectral coherence-based classifier of actively translated transcripts from ribosome profiling sequence data. *BMC Bioinformatics* 17, 482

58. Raj, A. *et al.* (2016) Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* 5, 1586–1591

59. Malone, B. *et al.* (2017) Bayesian prediction of RNA translation from ribosome profiling. *Nucleic Acids Res.* 45, 2960–2972

60. Griebel, T. *et al.* (2012) Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.* 40, 10073–10083

61. Crappe, J. *et al.* (2015) PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res.* 43, e29

62. de Klerk, E. *et al.* (2015) Assessing the translational landscape of myogenic differentiation by ribosome profiling. *Nucleic Acids Res.* 43, 4408–4428

63. Mohammad, F. *et al.* (2016) Clarifying the translational pausing landscape in bacteria by ribosome profiling. *Cell Rep.* 14, 686–694

64. Li, G.W. *et al.* (2014) Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* 157, 624–635

65. Ndah, E. *et al.* (2017) REPARATION: ribosome profiling assisted (re-)annotation of bacterial genomes. *bioRxiv* 113530

66. Zupanic, A. *et al.* (2014) Detecting translational regulation by change point analysis of ribosome profiling data sets. *RNA* 20, 1507–1518

67. Michel, A.M. *et al.* (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res.* 22, 2219–2229

68. Dunn, J.G. *et al.* (2013) Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *Elife* 2013, e01179

69. Ji, Z. *et al.* (2016) Transcriptome-scale RNase-footprinting of RNA–protein complexes. *Nat. Biotechnol.* 34, 410–413

70. Bazzini, A.A. *et al.* (2012) Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science* 336, 233–237

71. Park, J.E. *et al.* (2016) Regulation of poly(A) tail and translation during the somatic cell cycle. *Mol. Cell* 62, 462–471

72. Heyer, E.E. *et al.* (2016) Redefining the translational status of 80S monosomes. *Cell* 164, 757–769

73. Lykke-Andersen, S. and Jensen, T.H. (2015) Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat. Rev. Mol. Cell Biol.* 16, 665–677

74. Mukherjee, N. *et al.* (2016) Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nat. Struct. Mol. Biol.* 24, 86–96

75. Johnstone, T.G. *et al.* (2016) Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.* 35, 1–18

76. Mackowiak, S.D. *et al.* (2015) Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.* 16, 179

77. Olexiouk, V. *et al.* (2016) sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* 44, D324–D329

78. Bogdanow, B. *et al.* (2016) Systematic errors in peptide and protein identification and quantification by modified peptides. *Mol. Cell. Proteomics* 15, 2791–2801

79. Knudsen, G.M. *et al.* (2011) The effect of using an inappropriate protein database for proteomic data analysis. *PLoS One* 6, e20873

80. Martens, L. and Vizcaíno, J.A. (2017) A golden age for working with public proteomics data. *Trends Biochem. Sci.* 42, 333–341

81. Vaudel, M. *et al.* (2015) PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* 33, 22–24

82. Gawron, D. *et al.* (2016) Positional proteomics reveals differences in N-terminal proteoform stability. *Mol. Syst. Biol.* 12, 858–858

83. Willems, P. *et al.* (2017) N-terminal proteomics assisted profiling of the unexplored translation initiation landscape in *Arabidopsis thaliana*. *Mol. Cell. Proteomics* 16, 1064–1080

84. Liu, Y. *et al.* (2016) On the dependency of cellular protein levels on mRNA abundance. *Cell* 165, 535–550

85. Schwanhäusser, B. *et al.* (2009) Global analysis of cellular protein translation by pulsed SILAC. *Proteomics* 9, 205–209

86. Aviner, R. *et al.* (2014) Genome-wide identification and quantification of protein synthesis in cultured cells and whole tissues by puromycin-associated nascent chain proteomics (PUNCH-P). *Nat. Protoc.* 9, 751–760

87. Liu, T.-Y. *et al.* (2017) Time-resolved proteomics extends ribosome profiling-based measurements of protein synthesis dynamics. *Cell Syst.* 6, 21635

88. Zur, H. *et al.* (2016) Complementary post transcriptional regulatory information is detected by PUNCH-P and ribosomeprofiling. *Sci. Rep.* 6, 21635

89. Nedialkova, D.D. and Leidel, S.A. (2015) Optimization of codon translation rates via tRNA modifications maintains proteome integrity. *Cell* 161, 1606–1618

90. Iwasaki, S. and Ingolia, N.T. (2017) The growing toolbox for protein synthesis studies. *Trends Biochem. Sci.* 163, 799–810

91. Zhang, S. *et al.* (2017) TITER: predicting translation initiation sites by deep learning. *Bioinformatics* 33, i234–i242

92. Zhang, S. *et al.* (2016) ROSE: a deep learning based framework for predicting ribosome stalling. *bioRxiv* 2016, 067108

93. Weatheritt, R.J. *et al.* (2016) The ribosome-engaged landscape of alternative splicing. *Nat. Struct. Mol. Biol.* 23, 1117–1123

94. Blencowe, B.J. (2017) The relationship between alternative splicing and proteomic complexity. *Trends Biochem. Sci.* 42, 407–408

95. Tress, M.L. *et al.* (2017) Alternative splicing may not be the key to proteome complexity. *Trends Biochem. Sci.* 42, 98–110

96. Simsek, D. *et al.* (2017) The mammalian ribo-interactome reveals ribosome functional diversity and heterogeneity. *Cell* 169, 1051–1065

97. Diament, A. and Tuller, T. (2016) Estimation of ribosome profiling performance and reproducibility at various levels of resolution. *Biol. Direct* 11, 24

98. Larsson, O. *et al.* (2010) Identification of differential translation in genome wide studies. *Proc. Natl. Acad. Sci. U. S. A.* 107, 21487–21492

99. Olshen, A.B. *et al.* (2013) Assessing gene-level translational control from ribosome profiling. *Bioinformatics* 29, 2995–3002

100. Larsson, O. *et al.* (2011) Anota: analysis of differential translation in genome-wide studies. *Bioinformatics* 27, 1440–1441

101. Zhong, Y. *et al.* (2017) RiboDiff: detecting changes of mRNA translation efficiency from ribosome footprints. *Bioinformatics* 33, 139–141

102. Xiao, Z. *et al.* (2016) Genome-wide assessment of differential translations with ribosome profiling data. *Nat. Commun.* 7, 11194

103. Li, W. *et al.* (2017) Riborex: fast and flexible identification of differential translation from Ribo-seq data. *Bioinformatics* 33, 1735–1737

104. Love, M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550

105. Schueren, F. *et al.* (2016) Functional translational readthrough: a systems biology perspective. *PLoS Genet.* 12, e1006196