Bioinformatics, 2019, 1–7
doi: 10.1093/bioinformatics/btz878
Advance Access Publication Date: 21 November 2019
Original Paper

Sequence analysis

Accurate detection of short and long active ORFs using Ribo-seq data

Saket Choudhary[†], Wenzheng Li[†] and Andrew D. Smith*

Computational Biology and Bioinformatics, University of Southern California, Los Angeles, CA 90089, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Inanc Birol

Received on June 18, 2019; revised on November 4, 2019; editorial decision on November 18, 2019; accepted on November 20, 2019

Abstract

Motivation: Ribo-seq, a technique for deep-sequencing ribosome-protected mRNA fragments, has enabled transcriptome-wide monitoring of translation *in vivo*. It has opened avenues for re-evaluating the coding potential of open reading frames (ORFs), including many short ORFs that were previously presumed to be non-translating. However, the detection of translating ORFs, specifically short ORFs, from Ribo-seq data, remains challenging due to its high heterogeneity and noise.

Results: We present ribotricer, a method for detecting actively translating ORFs by directly leveraging the three-nucleotide periodicity of Ribo-seq data. Ribotricer demonstrates higher accuracy and robustness compared with other methods at detecting actively translating ORFs including short ORFs on multiple published datasets across species inclusive of *Arabidopsis*, *Caenorhabditis elegans*, *Drosophila*, human, mouse, rat, yeast and zebrafish.

Availability and implementation: Ribotricer is available at https://github.com/smithlabcode/ribotricer. All analysis scripts and results are available at https://github.com/smithlabcode/ribotricer-results.

Contact: andrewds@usc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The process of translating messenger RNA into protein is among the greatest investments of energy by cells (Russell and Cook, 1995). Consequently, translation is highly regulated to ensure that each cell synthesizes the right amount of each protein. Our understanding of the mechanisms regulating the translational process remains limited, which has motivated the development of experimental approaches to profile the translation landscape globally. Ribo-seq (Ingolia *et al.*, 2009) is a technology that uses deep-sequencing to identify ribosome-protected fragments, revealing the positions of the entire pool of ribosomes engaged in translation.

Ribo-seq has led to the surprising discovery of prevalent translation through open reading frames (ORFs) which were previously presumed to be non-active (Jackson *et al.*, 2018). Such ORFs include the upstream ORFs (uORFs) located in the 5′ untranslated region, the downstream ORFs located in the 3′ untranslated region, and the ORFs within presumed non-coding genes (Olexiouk *et al.*, 2018).

Transcriptome-wide searches for pairs of in-frame start and stop codons defining potential ORFs in human, and mouse genomes reveal that the sizes of such ORFs are generally 10–20 fold shorter (Calvo *et al.*, 2009) than the known protein-coding sequences (CDS) (Supplementary Fig. S1). Their short size presents challenges in detecting the resulting peptides through proteomic approaches (Fälth *et al.*, 2006). However, there is emerging evidence that these short ORFs, or the products of their translation, serve some function (Andrews and Rothnagel, 2014; Ingolia, 2016). In particular, the role of uORFs in regulating the translation of downstream CDS has been well documented (Barbosa *et al.*, 2013) for individual genes (Hinnebusch *et al.*, 2016), and they are correlated with substantial (30–80%) repression of protein production (Calvo *et al.*, 2009). The same mechanism is also used to encode condition-specific activation: in integrated stress response, where the repressed state is the default, uORF-associated repression is released following the stress stimulus (Andreev *et al.*, 2015).

Ribo-seq has been performed on multiple species ranging from prokaryotes to mammals. Studies over the years have observed that the choice of method of translation inhibition (Gerashchenko and Gladyshev, 2014; Hussmann *et al.*, 2015), the enzyme used for RNA digestion and its concentration (Aeschimann *et al.*, 2015; Gerashchenko and Gladyshev, 2017) and rRNA depletion (Weinberg *et al.*, 2016) can affect the overall signal and reduce its overall reproducibility (Diament and Tuller, 2016). Moreover, the presence of amplification bias, non-ribosomal RNA-protein complexes or other non-ribosomal contamination can often result in apparent ribosome-protected mRNA fragments (RPFs) that do not represent actively translating ribosomes. Some RNAs such as

© The Author(s) 2019. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz878/5637228 by Univ of Southern California, Saket Choudhary on 13 December 2019

telomerase RNA, RNAse P, snRNAs and snoRNAs that are known to be ‘classical’ non-coding RNAs and are predominantly localized in the nucleus have also been reported as origin for RPFs (Guttman *et al.*, 2013). This is an indication that not all RPFs represent actively translating ribosomes. Such fragments could represent non-ribosomal protected regions, such as those protected by RNA binding proteins. When drawing any conclusion about translational regulation from Ribo-seq data it is imperative to focus only on those fragments that represent actively translating ribosomes. However, the presence of noise in the data makes the task of identifying actively translated regions challenging. A shorter translation unit means less total data on average for inference, so detection of short ORFs in Ribo-seq has remained especially difficult.

Several methods exist for analyzing Ribo-seq data to determine the coding potential of the transcribed RNA. FLOSS (Ingolia *et al.*, 2014), one of the earliest methods, identifies actively translating ORF by focusing on the read length distribution. The key assumption is that the distribution of sequenced fragments contains both RPFs and technical noise, and the true RPFs should exhibit a particular length distribution. FLOSS first learns a reference distribution of RPF lengths on a set of protein-coding genes likely to represent active translation, and then compares fragment lengths through the other regions in the transcriptome to this reference distribution. The idea of treating different fragment lengths separately has been adopted in several subsequent methods. Most other methods can be understood broadly through two paradigms. The first hypothesizes that the distribution of number of mapped fragments differs over actively translated regions, and compares this distribution with some selected null model. The other general approach exploits the periodic pattern in the mapped fragment profiles to distinguish actively translating regions.

In the first paradigm of methods, ORFscore (Bazzini *et al.*, 2014) compares the distribution of reads falling in the three frames to a uniform distribution. ORF-RATER (Fields *et al.*, 2015) uses a combination of regression and random-forest based classification to predict actively translating ORFs. It uses a non-negative least squares fit for regressing Ribo-seq read profile of the transcript against the profile obtained from known protein-coding genes. A random-forest classifier then uses these scores to predict the translational status of the ORF. RiboHMM (Raj *et al.*, 2016), on the other hand, uses a hidden Markov model to detect translating ORFs. It models the contribution of each fragment length separately and then combines them to increase sensitivity. The hidden Markov model learns the distributions of Ribo-seq coverage over the start/stop codons and the translated CDS; the distributions are then used to predict translation status for candidate ORFs. Rp-Bp (Malone *et al.*, 2017) uses probabilistic modeling to estimate if read counts at each position belong to an enriched model or a null uniform model. RiboCode (Xiao *et al.*, 2018) uses a modified Wilcoxon signed-rank test (Wilcoxon, 1945) to assess periodicity by testing for differential enrichment in one of the frames against the other two.

The second paradigm typically leverages spectral approaches to examine the periodic pattern in Ribo-seq data. Mapping RPFs from Ribo-seq onto the mRNA is expected to reveal a ‘high-low-low’ pattern, owing to ribosome’s movement over codons, resulting in a three-nucleotide periodicity. RiboTaper (Calviello *et al.*, 2016) uses multi-tapered windows for calculating a Fourier transform to assess periodicity in the Ribo-seq signal. Based on related principles in signal processing, SPECtre (Chun *et al.*, 2016) makes use of spectral coherence to correlate Ribo-seq signal with the expected ‘high-low-low’ pattern. RiboWave (Xu *et al.*, 2018) uses a wavelet transform based method to denoise the RPF profile by extracting the three-nucleotide periodicity. This denoised RPF profile leads to a better performance when identifying active translation.

Methods within both paradigms have enabled discovery of actively translating ORFs. Each method makes assumptions about the data that are not always satisfied in practice, for different datasets or different data analysis goals. The detection of short ORFs is an example of the latter. However, these methods provide a conceptual foundation that we borrow from to design a simplified method that is more robust to varying statistical features across datasets, and that is capable of detecting both short and long ORFs. Our method, called ribotricer, directly assesses the three-nucleotide periodicity in Ribo-seq data. Ribotricer can account for read length specific P-site offsets and sparsity in Ribo-seq data. Its underlying model emphasizes consistency in the qualitative profile through each codon while down-weighting the influence of the magnitude of the individual values contributing to that profile. This approach helps ribotricer to overcome the challenge of detecting short ORFs in regions of low signal to noise ratio.

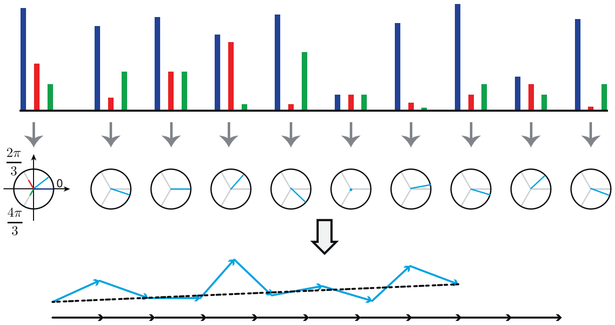

Fig. 1. Methodology design of ribotricer

2 Materials and methods

To detect actively translating ORFs, ribotricer focuses on the characteristic three-nucleotide periodicity in Ribo-seq data. The workflow of ribotricer consists of five major steps. Ribotricer first prepares a candidate set of all potentially translatable ORFs by searching for pairs of start and stop codons genome-wide but inside annotated transcription units. This requires providing gene annotations and the reference genome but is only done once for each genome and gene annotation. Next, ribotricer partitions the mapped reads based on their length. The rationale for processing reads by their length is that each length may be associated with a different P-site offset relative to the $5'$ end of the mapped fragment. For each read length, ribotricer generates a metagene profile using $5'$ ends of the mapped reads (accounting for strand as appropriate). The metagene profiles are used to infer P-site offsets for different read-lengths by choosing the offsets that maximize the cross-correlation of these profiles with the profile for the most abundant read length. The read profiles corresponding to different read-lengths can then be merged using the corresponding inferred P-site offsets, an approach taken previously by Calviello *et al.* for RiboTaper (Calviello *et al.*, 2016) and Xiao *et al.* for RiboCode (Xiao *et al.*, 2018). The previous step produces a single RPF profile for each candidate ORF. In its final step, ribotricer assesses the periodicity of the merged RPF profile using a novel approach to predict its translation status.

Our key contribution is a novel method for assessing the three-nucleotide periodicity of RPF profile based on 3D to 2D projection (Fig. 1; Supplementary Fig. S26). Within each codon, we may observe reads with $5'$ ends at each of the three nucleotides, providing three unconstrained count values. These count values can be imagined as vectors in a 3D space with each nucleotide position representing 1D. We hypothesized that using the absolute read count at each nucleotide might obscure the signal of an entire profile when being evaluated for its periodicity. Though genes undergoing translation are expected to accumulate more reads in total, we hypothesized that for many genes an over-emphasis on total counts might amplify the effect of unknown artifacts or noise in the data. Actively translating regions exhibit a distinct ‘high-low-low’ pattern at each codon irrespective of their absolute read count values. Codons in a profile, however, might end up with a high abundance of reads because of the difference in ribosomal decoding speed (Ingolia, 2014), a ribosomal pause (Buskirk and Green, 2017) or presence of non-ribosomal fragments (Andreev *et al.*, 2017). Hence, using absolute read count values at each nucleotide could lead to a non-stationary

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz878/5637228 by Univ of Southern California, Saket Choudhary on 13 December 2019

profile. Applying any spectral method would require that the profiles satisfy conditions to ensure stationarity. Instead, we rely on using the qualitative information at each codon in the form of 'high-low-low' or related pattern. This approach discards much of the quantitative information associated with individual read counts but also simplifies the problem while eliminating the need to explicitly model random variation or systematic trend in total read counts along the RPF profile.

For a given ORF consisting of N codons, let x_{ij} denote the number of P-sites inferred from the reads of Ribo-seq experiment aligning to the i-th codon and j-th frame of the ORF, where $i = 1, 2, \ldots, N$ and $j = 1, 2, 3$. The RPF profile of the ORF can then be denoted as $P = (x_{11}, x_{12}, x_{13}, \ldots, x_{N1}, x_{N2}, x_{N3})$. For each codon profile $x_i = (x_{i1}, x_{i2}, x_{i3})$, a 3D vector, we perform the following transformation to convert it into a 2D unit vector $\phi_i = (a_i, b_i)^{\mathrm{T}}$, more specifically, the angle of the unit vector which is inherently 1D:

$$\phi_i = \frac{w x_i^{\mathrm{T}}}{||w x_i^{\mathrm{T}}||}, \tag{1}$$

where

$$w = \begin{pmatrix} 1 & \cos(-2\pi/3) & \cos(-4\pi/3) \\ 0 & \sin(-2\pi/3) & \sin(-4\pi/3) \end{pmatrix}.$$

With this transformation, the three basis vectors $\{(1,0,0), (0,1,0), (0,0,1)\}$ are mapped as

$$\begin{aligned} (1,0,0) &\to (1,0), \\ (0,1,0) &\to (\cos(-2\pi/3),\ \sin(-2\pi/3)), \\ (0,0,1) &\to (\cos(-4\pi/3),\ \sin(-4\pi/3)). \end{aligned}$$

The three mapped unit vectors lie $2\pi/3$ away from each other to ensure the direction of the transformed vector ϕ_i is equally determined by reads of each frame. These can be replaced by any three unit vectors that are equally spaced on the unit circle, and the results would not change.

For the transformation performed, the direction of the resulting vector is determined by the relative values of x_{i1}, x_{i2}, and x_{i3}. For an actively translating ORF, we expect to see a 'high-low-low' pattern for each codon. This is equivalent to observing x_{i1} as the largest value consistently over all codons. If this holds, we expect the directions of the resulting unit vectors ϕ_i to be consistent across codons. As indicated above, the motivation behind unit normalization of each vector is to help ensure that each codon contributes equally to our assessment of translation status, avoiding bias from the fraction of codons with an over-abundance of reads. This transformation disregards the total read counts at each of the three positions. For example, the two codon profiles (100, 20, 10) and (10, 2, 1) will result in the same unit vectors when applying Equation (1). Another example would be of profiles (100, 99, 99) and (100, 1, 1) which will both result in the same phase score, even though the difference between the first and the rest two frames is much higher in the latter. While this discards quantitative information, it still captures the qualitative 'high-low-low' pattern of the profile. This approach helps ribotricer handle the heterogeneous nature of Ribo-seq data where despite of pervasive active translation, different codons could have completely different coverages either because of the actual difference in ribosome's dwell time or because of usage of drugs like cycloheximide which can alter codon-specific elongation rates (Hussmann *et al.*, 2015).

The l^2-norm of the mean vector of the transformed vectors can be used to assess the periodicity of RPF profile. More consistent directions of the vectors would result in a larger l^2-norm. The mean vector of the transformed vectors is

$$\bar{\phi} = \frac{1}{N} \sum_{i=1}^{N} \phi_i,$$

and its l^2-norm $||\bar{\phi}||$ is

$$||\bar{\phi}|| = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N} a_i\right)^2 + \left(\frac{1}{N}\sum_{i=1}^{N} b_i\right)^2},$$

which falls in [0, 1], with a value of 1 if and only if

$$\begin{aligned} a_1 &= a_2 = \cdots = a_N, \\ b_1 &= b_2 = \cdots = b_N, \end{aligned}$$

in which case the directions for all vectors are the same.

Besides heterogeneity arising from uneven distribution of read counts across codons (O'Connor *et al.*, 2016), another key challenge in Ribo-seq data is sparsity leading to profiles with many empty codons, i.e. codons to which no reads map. We do not use empty codons for phase-score calculation. For a particular dataset with N codons, define the set V of non-empty codons as

$$V = \{i = 1, 2, \ldots, N | x_i \neq (0,0,0)\},$$

and let $N_v = |V|$. If we define $\bar{\phi}^*$ as the mean vector including only non-empty codons, the ratio between $||\bar{\phi}||$ and $||\bar{\phi}^*||$ is

$$\frac{||\bar{\phi}||}{||\bar{\phi}^*||} = \frac{N_v}{N}.$$

With the reasoning outlined above, we use $||\bar{\phi}^*||$ as our measure for assessing the periodicity of the RPF profile of an ORF. This score describes how 'aligned' all the vectors are, and is equivalent to measuring how similar the phases are, i.e. the angles created by the resulting vectors with respect to the abscissa. We will refer to this score as the 'phase score' hereafter. Note that in theory, a high phase score may result from strong consistency of some pattern other than the anticipated 'high-low-low'. In designing our approach, we hypothesized that the only source of consistency in the signal would be an active translation. A consistent 'low-high-low' or 'low-low-high' pattern would most likely result from an inaccurate estimate of the P-site offsets, in which case our assumptions add a layer of robustness.

The angles made by the resultant vectors when all the codons follow a 'high-low-low' pattern should be concentrated around 0. The distribution we observe for the Ribo-seq data is centered around 0 (Supplementary Figs S6 and S7), which confirms that most codons follow the 'high-low-low' pattern. For the RNA-seq data, the resulting angles follow a multimodal distribution with the highest peaks at $\{-2\pi/3, 0, 2\pi/3\}$ (Supplementary Figs S6 and S7) which corresponds to the three unit vectors. To interpret the multimodal distribution observed in RNA-seq data, we simulated read counts using a Poisson distribution. To account for variation in total data between genes, we simulated means of the Poisson distribution using the per nucleotide coverage from the RNA-seq. The resulting angle distribution of the simulated codon profiles is similar to that obtained from profiles of the RNA-seq data (Supplementary Figs S6 and S7) which explains the observed multimodality.

2.1 Learning cutoff of phase score

The phase score is indicative of how consistent the profile is through a defined region. We require some cutoff to distinguish phase scores that differentiate active from non-active translation, with the latter representing either some form of noise or inactive translation. Our approach is to learn this cutoff empirically using published datasets (Supplementary Table S1) with an assumed ground truth set for regions of active translation and regions lacking active translation. Taking this strategy, we used RPF profiles of expressed consensus coding sequence (CCDS) (Pruitt *et al.*, 2009) exons from Ribo-seq data as the true positives, and mapped read profiles from RNA-seq data for a negative control for human and mouse datasets, as previously described (Calviello *et al.*, 2016; Xiao *et al.*, 2018). In order to choose the best cutoff, we relied on maximizing the $F1$ score statistic. $F1$ score represents the harmonic mean of precision and recall and is considered a more realistic measure of a classifier's performance than precision or recall in isolation. Since the CCDS annotated

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz878/5637228 by Univ of Southern California, Saket Choudhary on 13 December 2019

regions serve as a high confidence ground truth, we first focused on human and mouse datasets for learning the cutoff and benchmarking ribotricer against other methods. The 10 datasets (five in human and five in mouse) are described in Supplementary Tables S1–S5. We envisioned a cutoff that is applicable even if there is no matching RNA-seq sample available. The median phase scores of Ribo-seq samples, however, appear to vary across species (Supplementary Table S7 and Figs S27–S30), and so, we decided to learn the cutoffs in a species-specific manner. Using two arbitrary datasets in human [SRA accession: SRP010679 (Hsieh *et al.*, 2012) and SRP098789 (Lintner *et al.*, 2017)], and two arbitrarily chosen datasets in mouse [SRA accession: SRP003554 (Guo *et al.*, 2010) and SRP115915 (Mariotti *et al.*, 2017)] we determined the human-specific and mouse-specific cutoff as 0.441 and 0.418, respectively (Supplementary Table S6 and Figs S8, S9 and S31). We use these cutoffs for the remaining three datasets in each species to assess ribotricer's performance. One might expect that learning a cutoff within each dataset would yield better performance. We found this not always to be the case (Supplementary Tables S10 and S11 and Figs S35–S39). Here we focus on results using species-specific cutoffs. We also benchmarked ribotricer using species- and dataset-specific cutoffs in *Arabidopsis*, *Caenorhabditis elegans*, *Drosophila*, rat, yeast and zebrafish (Section 3.3).

3 Results

To evaluate the performance of ribotricer and other existing methods, acknowledging the heterogeneity and appreciable noise levels in Ribo-seq data, we first selected five human and five mouse datasets for performance comparison (Supplementary Tables S1–S5 and Figs S2–S5). This includes the human HEK293 cells dataset (SRA accession: SRP063852) (Calviello *et al.*, 2016), which was originally used as a benchmark dataset when RiboTaper was introduced (Calviello *et al.*, 2016) and subsequently used in other studies. The phase scores of Ribo-seq samples show larger variation as compared to RNA-seq samples (Supplementary Figs S27–S30).

We followed the strategy previously established by Calviello *et al.* in assessing RiboTaper (Calviello *et al.*, 2016) and Xiao *et al.* in assessing RiboCode (Xiao *et al.*, 2018). For all the 10 datasets, we obtained the RPF profiles for all the CCDS from the results generated by RiboTaper and used the expressed CCDS profiles from Ribo-seq data as true positives and the corresponding CCDS profiles from RNA-seq data as true negatives. Since RiboTaper was designed and benchmarked for detecting active translation at the exon level, we split the existing methods for active translation detection into two groups; those that support detection at the exon level and those that only allow detection at the transcript level. We compared the performance of ribotricer at both the exon and transcript levels.

3.1 Ribotricer accurately detects translating ORFs at the exon level

We evaluated the performances of methods that support exon-level detection of translation, including ORFscore (Bazzini *et al.*, 2014), RiboTaper (Calviello *et al.*, 2016) and RiboCode (Xiao *et al.*, 2018), and compared their performance with that of ribotricer.

We first compared the ability of each method to distinguish Ribo-seq profiles from RNA-seq using the area under the receiver operating characteristic (ROC) and precision-recall (PR) curve. For human HEK293 cells dataset (SRA accession: SRP063852) (Calviello *et al.*, 2016), ribotricer achieved an area under the ROC (AUROC) of 0.97. The second best one was achieved by RiboCode with an AUROC of 0.93. RiboTaper and ORFscore achieved an AUROC of 0.88 and 0.87, respectively (Fig. 2A). For the mouse liver tissue dataset (SRA accession: SRP078005) (Fradejas-Villar *et al.*, 2017), ribotricer achieved an AUROC of 0.99 while RiboCode, RiboTaper and ORFscore achieved AUROC of 0.97, 0.92 and 0.92, respectively (Fig. 2A). The difference between AUROC achieved by ribotricer and the next best method is statistically significant ($P < 0.001$, Supplementary Table S8). Ribotricer also outperformed the other three methods consistently under the PR metric (Fig. 2A).

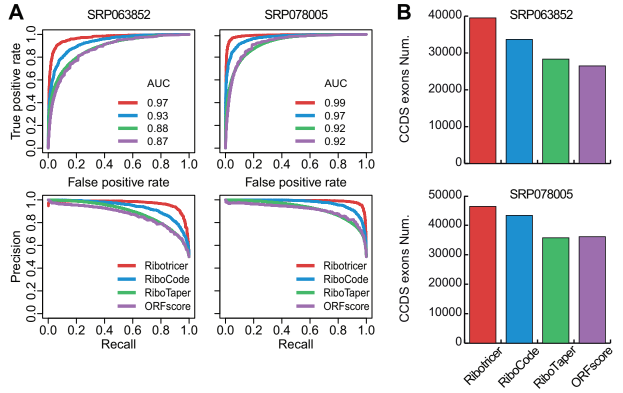

Fig. 2. Comparison of performance on detecting translating exons. The performance of ribotricer is compared with that of RiboCode, RiboTaper and ORFscore. (**A**) The ROC and precision-recall curves summarizing performance of ribotricer, RiboCode, RiboTaper and ORFscore on one human and one mouse dataset. (**B**) The number of translating exons recovered when controlling the false positive rate to be the same

Ribotricer displayed the best performance on almost all the 10 datasets at both ROC and PR metrics (Supplementary Figs S10, S11, S32 and S33 and Table S8).

Next, we compared the performance of ribotricer, ORFscore, RiboTaper and RiboCode by contrasting the number of true positives detected by each method while controlling the false positive rate at 0.1. We calibrated the cutoffs for each method so that the number of false positives reported by each method is 10% of the number of negatives. For human HEK293 cell dataset (SRA accession: SRP063852), ribotricer recovered 39 517 truly translating exons, while RiboCode recovered 33 665. RiboTaper, and ORFscore recovered 28 333 and 26 486 translating exons, respectively (Fig. 2B). For mouse liver tissue dataset (SRA accession: SRP078005), ribotricer recovered 46 380 truly translating exons, RiboCode recovered 43 332, while RiboTaper and ORFscore recovered 35 746 and 36 120 translating exons, respectively (Fig. 2B). We observed a similar performance for the other eight datasets where ribotricer consistently recovered more truly translating exons compared to the other three methods (Supplementary Fig. S12).

Short ORFs (<100 codons) (Basrai *et al.*, 1997) are known to be abundant in mammals, insects, fungi and plants (Frith *et al.*, 2006; Mat-Sharani and Firdaus-Raih, 2019). However, they are often overlooked by proteomic approaches (Fälth *et al.*, 2006). Ribo-seq data provide us with an avenue to bridge this gap. However, the length of shorter ORFs implies less total data on average for inference, making their detection particularly challenging. In order for ribotricer to be capable of detecting both short and long ORFs, the phase scores generated should be minimally dependent on the ORF length. We investigated the effect of ORF length on the scores or the *P*-values generated by each method. The phase score generated by ribotricer is unaffected by the length of ORF while RiboCode, RiboTaper and ORFscore generate a higher score or more significant *P*-value as the ORF gets longer (Supplementary Fig. S13). Ribotricer's phase score remains stable even if the original ORF is truncated to just 10% of its original length, whereas RiboCode and ORFscore show large deviations (Supplementary Figs S24 and S25). Moreover, the difference between ribotricer's phase score of a profile against a 'downsampled' profile with fewer codons is negligible (Supplementary Figs S22 and S23) with as few as 20 codons (see Section 4 and Supplementary Section 5).

Finally, we compared the performance of ribotricer with other methods in terms of *F*1 score using the default cutoff for each method (Supplementary Fig. S33 and Table S9). Since we learned the cutoff for ribotricer from four real datasets, we summarized the performance of ribotricer on the remaining six datasets that were not used to learn the empirical cutoff (Supplementary Figs S14–S17). Notably, for human HeLa cell dataset (SRA accession: SRP029589) (Stumpf *et al.*, 2013), all methods achieved relatively low *F*1 score with the best one to be 0.67 achieved by ribotricer.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz878/5637228 by Univ of Southern California, Saket Choudhary on 13 December 2019

We checked the angle distribution of the 3D to 2D projection described earlier for this dataset (Supplementary Fig. S6), and found that it displays high noise level compared to other datasets analyzed, which indicates low data quality. Consequently, we excluded this dataset from further analysis. For the other two human datasets, ribotricer achieved an average *F*1 score of 0.91, and RiboCode achieved an average *F*1 score of 0.84. RiboTaper and ORFscore achieved an average *F*1 score of 0.73 and 0.12, respectively. For the three mouse datasets, ribotricer achieved an average *F*1 score of 0.93, and RiboCode achieved an average *F*1 score of 0.90. RiboTaper and ORFscore achieved an average *F*1 score of 0.85 and 0.55, respectively.

3.2 Ribotricer accurately detects translating ORFs at the transcript level

ORF-RATER (Fields *et al.*, 2015), RibORF (Ji *et al.*, 2015), Rp-Bp (Malone *et al.*, 2017) and RiboWave (Xu *et al.*, 2018) only detect translating ORFs at the full transcript level. To evaluate ribotricer against these methods we use a similar to the comparison strategy as used for exon-level benchmarking. For transcript level comparison, we first used the area under ROC/PR curves to assess the ability of different methods to distinguish Ribo-seq profiles from those from RNA-seq data. For human HEK293 cell dataset (SRA accession: SRP063852), ribotricer correctly distinguished Ribo-seq profiles from the simulated RNA-seq profiles with an AUROC of 1.0, while both Rp-Bp and RibORF achieved an AUROC of 0.96. RiboWave achieved an AUROC of 0.90 (Fig. 3A). For human HeLa cell dataset (SRA accession: SRP098789) (Lintner *et al.*, 2017), ribotricer again perfectly distinguished Ribo-seq profiles from the simulated RNA-seq ones with an AUROC of 1.0, and Rp-Bp achieved an AUROC of 0.91. RibORF and RiboWave achieved an AUROC of 0.96 and 0.83, respectively (Fig. 3A). Ribotricer also consistently outperformed other methods under the PR metric (Fig. 3A). The complete results for all human and mouse samples can be found in Supplementary Figures S18 and S19. It is worth mentioning that RibORF (Ji *et al.*, 2015) uses a classification based method which trains its model by selecting one-third of the CDS profiles as true positives which might give it an extra advantage in this comparison. Notably, here we excluded ORF-RATER from the comparison because it always reports around half the number of detected ORFs compared with other methods, as noticed by Xiao *et al.* previously (Xiao *et al.*, 2018). The difference between ribotricer's AUROC and the second best method in 8 of the 10 human and mouse datasets is statistically significant (Supplementary Table S8).

Next, we compared the performances of different methods by checking the number of truly translating transcripts recovered when controlling the false positive rate to be the same as 0.1. For the human HEK293 cell dataset (SRA accession: SRP063852), ribotricer recovered 577 truly translating transcripts, while Rp-Bp, RibORF and RiboWave recovered 508, 542 and 459 translating transcripts, respectively (Fig. 3B). For the human HeLa cell dataset (SRA accession: SRP098789), ribotricer recovered 2251 truly translating transcripts, and Rp-Bp recovered 1730. RibORF and RiboWave recovered 2130 and 1308 truly translating transcripts, respectively (Fig. 3B and Supplementary Fig. S20).

Finally, we used the *F*1 score to assess the performance of ribotricer in detecting actively translating transcripts in comparison with other tools. For the two human samples, ribotricer achieved an average *F*1 score of 0.99, and Rp-Bp achieved an average *F*1 score of 0.89. RibORF and RiboWave achieved an average *F*1 score of 0.91 and 0.75, respectively. For the three mouse samples, ribotricer achieved an average *F*1 score of 0.99, and Rp-Bp achieved an average *F*1 score of 0.87. RibORF and RiboWave achieved an average *F*1 score of 0.97 and 0.69, respectively (Supplementary Fig. S21).

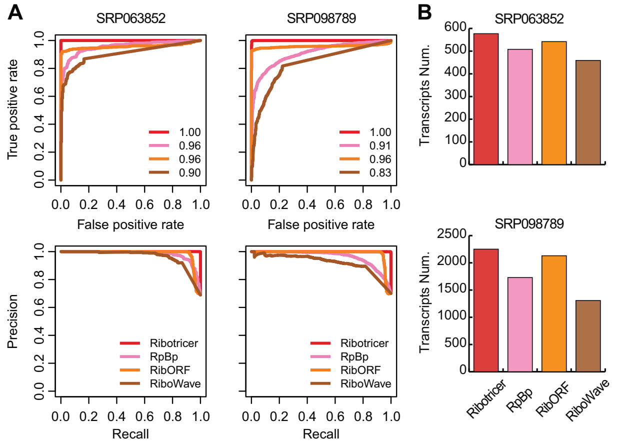

Fig. 3. Comparison of performance on detecting translating transcripts. The performance of ribotricer is compared with that of RibORF, RiboWave and Rp-Bp. (**A**) The ROC and precision-recall curves summarizing performance of ribotricer, RibORF, RiboWave and Rp-Bp on one human and one mouse dataset. (**B**) The number of translating transcripts recovered when controlling the false positive rate to be the same

3.3 Ribotricer achieves high accuracy across species

We further tested the applicability of our method across different species including *Arabidopsis*, *C.elegans*, *Drosophila*, rat, yeast and zebrafish. Though the median scores of RNA-seq samples do not exhibit high levels of variation in the same species, the corresponding Ribo-seq samples appear to have highly varied phase scores (Supplementary Figs S27–S30). Following our previous strategy of learning cutoffs from two datasets, we learned the cutoffs for each species separately. The species-specific cutoffs (Supplementary Table S6) were then used to determine the translation status of Ribo- and RNA-seq profiles.

Ribotricer consistently gives the best AUROC and *F*1 score for all samples in *Arabidopsis*, yeast and zebrafish at the exon level (Supplementary Figs S32 and S33 and Tables S8 and S9). Similarly, for *C.elegans*, ribotricer's *F*1 scores are the highest in all the four datasets. In *Drosophila*, where the difference between Ribo-seq and RNA-seq phase scores is low (Supplementary Figs S27–S30), ribotricer consistently results in the best *F*1 scores (Supplementary Figs S33 and S34 and Table S9).

In more challenging datasets, where the AUROC achieved by the best method is not close to one, ribotricer is able to perform well at both AUROC and *F*1 score metrics. Particularly, in *Arabidopsis* dataset SRP087624 (Xu *et al.*, 2017), ribotricer achieves an AURC of 0.690 whereas the second best method, RiboTaper, achieves an AUROC of 0.523 (Supplementary Fig. S32 and Table S8) with the difference between them being statistically significant ($P < 0.001$). It is worth noting that in *Drosophila*, three datasets have AUROC in the range of 0.64–0.73, however ribotricer's AUROC is not the best amongst other methods (Supplementary Table S8). The failure of ribotricer in this case can be attributed to the diminished difference between Ribo-seq and RNA-seq phase scores in these samples (Supplementary Figs S27–S30 and S34). However, ribotricer still results in the highest *F*1 scores for all the datasets (Supplementary Figs S33 and S34 and Table S9).

4 Discussion

Ribo-seq has enabled transcriptome-wide monitoring of translation and has provided avenues for discovering tissue- or condition-specific ORFs. It has expanded the spectrum of translation beyond the annotated coding regions with the discovery of thousands of ORFs that were presumed to be non-active. The presence of amplification bias, non-ribosomal RNA-protein complexes or other contamination can often result in fragments that do not represent active translation. This has made the detection of actively translating ORFs from Ribo-seq data a challenging problem. The correct interpretation of Ribo-seq data requires that only actively translating regions be considered for drawing any conclusion. It is particularly important to do this separation for accurately identifying actively translating short ORFs, since their short length increases the impact

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz878/5637228 by Univ of Southern California, Saket Choudhary on 13 December 2019

of noise. Multiple tools have been developed for detecting actively translating ORFs using Ribo-seq data. However, little focus has been placed on detection of short ORFs. Though the textbook definition of an ORF is a sequence having a multiplicity of three with its ends marked with a start (AUG) and stop codon, a more appropriate definition suggests that such a sequence just be bounded by stop codons (Sieber *et al.*, 2018). As such, Ribo-seq based tools for determining active translation benefit from the capacity to identify translation in all potentially translatable ORFs rather than just known protein-coding regions. Moreover, the detection of true translating ORFs can be used to filter out reads not associated with translation events, which would benefit downstream read count based analysis, such as differential translation efficiency modeling using methods, such as Riborex (Li *et al.*, 2017) and Xtail (Xiao *et al.*, 2016).

Ribotricer assesses the periodicity of RPF profile by projecting the 3D read count vector of each codon to a 2D unit vector. There are several advantages of our method. First, by unit normalizing the projected vector, ribotricer performs a trend correction, allowing for non-uniform coverage across the profile. In particular, this avoids the bias caused by codons with a high number of mapped reads. Second, ribotricer checks the consistency of the pattern across the three frames of each codon but does not assume the exact translating frame which makes it unaffected by any P-site shift. Lastly, as we have demonstrated, the scores generated by ribotricer are not dependent on the length of the ORF.

A key challenge in detecting short ORFs lies in the limited length of the signal. Fourier transform based methods, such as RiboTaper are subject to the uncertainty principle (Donoho and Stark, 1989), which decreases frequency resolution when the signal length is short. Methods that utilize the absolute magnitudes of the count of the profile vector will tend to have a higher error rate in short regions due to the high variance associated with limited observations. Our method, on the other hand, relies on using the qualitative information at each codon in the form of ‘high-low-low’ pattern. This gives it the highest resolution and protects it from bias that might arise from codons with an over-abundance of reads. This explains ribotricer’s higher accuracy even at shorter regions (CCDS exons) as compared to other methods. Species-specific phase-score cutoffs result in good performance across all the datasets that we tested. However, depending on the availability and quality of data, dataset-specific cutoffs can also result in improved performance (Supplementary Tables S10 and S11 and Figs S35–S39).

The strength of ribotricer is derived from its simplicity: we make fewer assumptions about quantitative aspects of the data, and in the face of technically heterogeneous data, this is a positive. However, eventually technical characteristics of Ribo-seq data will converge. When that happens, we expect that by directly modeling those technical characteristics, more intricate methods will be able to more effectively leverage quantitative aspects of RPF profiles. The phase score specifically avoids modeling the distribution of absolute RPF counts along transcripts. If technical characteristics of Ribo-seq data stabilize in the near future, and can be modeled accurately, our approach can be adapted to weigh the contributions of codons based on their total number of reads.

By default, ribotricer searches for ORFs that are at least 60 nt or 20 codons long to build the candidate list but this minimum length can be set to a user-defined value. We arrived at the default value of 20 codons by performing a simulation using the Ribo-seq profiles of genes with total codons >100 and with at least 50% non-empty codons. In the simulation, we randomly sampled 10–100 codons and generated a ‘downsampled’ profile. The mean absolute difference between the original phase score calculated using the full length profile versus the ‘downsampled’ profile with 20 or more codons is smaller than 0.05 and does not change after increasing the number of codons (Supplementary Figs S22 and S23).

Ribotricer enables discovery of both short and long ORFs that will deepen our understanding of translational regulation across various biological contexts. We envision ribotricer’s phase score to become a commonly used quality control metric for assessing the quality of Ribo-seq datasets, especially for new datasets in species where no prior datasets exist (Supplementary Figs S40–S42).

Acknowledgements

The authors thank Amal Thomas, Guilherme de Sena Brandine, Hoda Mirsafian, Masaru Nakajima, Meng Zhou, Rish Prabakar, Xiaojing Ji and Qifan Yang for helpful input and discussions on the method.

Funding

This work was supported by National Institutes of Health grant [R01 HG006015].

Conflict of Interest: none declared.

References

Aeschimann,F. *et al.* (2015) Transcriptome-wide measurement of ribosomal occupancy by ribosome profiling. *Methods*, **85**, 75–89.

Andreev,D.E. *et al.* (2015) Translation of 5’ leaders is pervasive in genes resistant to eIF2 repression. *Elife*, **4**, e03971.

Andreev,D.E. *et al.* (2017) Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Res.*, **45**, 513–526.

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

Barbosa,C. *et al.* (2013) Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet.*, **9**, e1003529.

Basrai,M.A. *et al.* (1997) Small open reading frames: beautiful needles in the haystack. *Genome Res.*, **7**, 768–771.

Bazzini,A.A. *et al.* (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.*, **33**, 981–993.

Buskirk,A.R. and Green,R. (2017) Ribosome pausing, arrest and rescue in bacteria and eukaryotes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **372**, 20160183.

Calviello,L. *et al.* (2016) Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods*, **13**, 165.

Calvo,S.E. *et al.* (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. USA*, **106**, 7507–7512.

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.

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

Donoho,D.L. and Stark,P.B. (1989) Uncertainty principles and signal recovery. *SIAM J. Appl. Math.*, **49**, 906–931.

Fälth,M. *et al.* (2006) SwePep, a database designed for endogenous peptides and mass spectrometry. *Mol. Cell. Proteomics*, **5**, 998–1005.

Fields,A.P. *et al.* (2015) A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol. Cell*, **60**, 816–827.

Fradejas-Villar,N. *et al.* (2017) The RNA-binding protein Secisbp2 differentially modulates UGA codon reassignment and RNA decay. *Nucleic Acids Res.*, **45**, 4094–4107.

Frith,M.C. *et al.* (2006) The abundance of short proteins in the mammalian proteome. *PLoS Genet.*, **2**, e52.

Gerashchenko,M.V. and Gladyshev,V.N. (2014) Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res.*, **42**, e134.

Gerashchenko,M.V. and Gladyshev,V.N. (2017) Ribonuclease selection for ribosome profiling. *Nucleic Acids Res.*, **45**, e6.

Guo,H. *et al.* (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, **466**, 835.

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

Hinnebusch,A.G. *et al.* (2016) Translational control by 5‘-untranslated regions of eukaryotic mRNAs. *Science*, **352**, 1413–1416.

Hsieh,A.C. *et al.* (2012) The translational landscape of mTOR signalling steers cancer initiation and metastasis. *Nature*, **485**, 55.

Hussmann,J. A. *et al.* (2015) Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. *PLoS Genet.*, **11**, e1005732.

Ingolia,N.T. (2014) Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.*, **15**, 205.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz878/5637228 by Univ of Southern California, Saket Choudhary on 13 December 2019

Ingolia,N.T. (2016) Ribosome footprint profiling of translation throughout the genome. *Cell*, **165**, 22–33.

Ingolia,N.T. *et al.* (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.

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

Jackson,R. *et al.* (2018) The translation of non-canonical open reading frames controls mucosal immunity. *Nature*, **564**, 434–438.

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

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

Lintner,N.G. *et al.* (2017) Selective stalling of human translation through small-molecule engagement of the ribosome nascent chain. *PLoS Biol.*, **15**, e2001882.

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

Mariotti,M. *et al.* (2017) Multiple RNA structures affect translation initiation and UGA redefinition efficiency during synthesis of selenoprotein P. *Nucleic Acids Res.*, **45**, 13004–13015.

Mat-Sharani,S. and Firdaus-Raih,M. (2019) Computational discovery and annotation of conserved small open reading frames in fungal genomes. *BMC Bioinformatics*, **19**, 551.

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

Olexiouk,V. *et al.* (2018) An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.*, **46**, D497–D502.

Pruitt,K.D. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.

Raj,A. *et al.* (2016) Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife*, **5**, e13328.

Russell,J.B. and Cook,G.M. (1995) Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiol. Rev.*, **59**, 48–62.

Sieber,P. *et al.* (2018) The definition of open reading frame revisited. *Trends Genet.*, **34**, 167–170.

Stumpf,C.R. *et al.* (2013) The translational landscape of the mammalian cell cycle. *Mol. Cell*, **52**, 574–582.

Weinberg,D.E. *et al.* (2016) Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell Rep.*, **14**, 1787–1799.

Wilcoxon,F. (1945) Individual comparisons by ranking methods. *Biometr. Bull.*, **1**, 80–83.

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

Xiao,Z. *et al.* (2018) De novo annotation and characterization of the translatome with ribosome profiling data. *Nucleic Acids Res.*, **46**, e61.

Xu,G. *et al.* (2017) Global translational reprogramming is a fundamental layer of immune regulation in plants. *Nature*, **545**, 487–490.

Xu,Z. *et al.* (2018) Ribosome elongating footprints denoised by wavelet transform comprehensively characterize dynamic cellular translation events. *Nucleic Acids Res.*, **46**, e109.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz878/5637228 by Univ of Southern California, Saket Choudhary on 13 December 2019