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

Saket Choudhary* Wenzheng Li* Andrew D. Smith
{skchoudh, wenzhenl, andrewds}@usc.edu

1 Data description

To evaluate the performance of different methods in a comprehensive way, we selected multiple public datasets from *Arabidopsis*, *C. elegans*, *Drosophila*, human, mouse, rat, yeast, and zebrafish. These datasets span several tissues and cell lines. The treatment employed for inhibiting translation includes, flash freezing, cycloheximide, streptomycin, emetine, tunicamycin, and blasticidin (Supplementary Table S1).

The five *Arabidopsis* datasets include one dataset from inflorescence (SRA accession: SRP108862, unpublished), two datasets from leaf tissue (SRA accession: SRP087624 [50] and SRP059391 [29]). The other two datasets include a whole seedling (SRA accession: SRP029587 [24]) and an etiolated seedling (SRA accession: SRP018118 [28]).

In C. elegans, all the four datasets are from the n2 strain (SRA accession: SRP056647 [35], SRP026198 [45], SRP014427 [43], and SRP010374 [44]).

For *Drosophila*, we have four datasets spanning body wall muscle (SRA accession: SRP108999 [8]), embryo (SRA accession: SRP028243 [12]), oocytes (SRP076919 [14]), and S2 cells (SRA accession: SRP045475 ([3])).

The five human datasets include a prostate cancer cell line (PC3; SRA accession: SRP010679 [20]), two samples from HeLa cells (SRA accession: SRP029589 [46] and SRP098789 [27]), one sample from HEK293 (SRA accession: SRP063852 [7]), and one from H1933 cancer cell line (SRA accession: SRP102021 [42]).

The five mouse datasets include neutrophils cultured from mouse bone marrow (SRA accession: SRP003554 [17]), cultured hippocampal neurons (SRA accession: SRP062407 [9]) , two samples from liver tissue (SRA accession: SRP078005 [15] and SRP115915 [30]), and embryonic stem cells (SRA accession: SRP091889 [47]).

For rat, we have three datasets including one in PC12 cell line (SRA accession: SRP056012 [1]) , one in pheochromocytoma cells (SRA accession: SRP045777 [36]) and one from the BN/SHR strain (SRA accession: ERP007231 [39]).

For yeast, there are five datasets spanning strains by4743 (SRA accession: SRP075766[6]), by4176 (SRA accession: SRP028614 [2]), by4741 (SRA accession: SRP033499 [18], and SRP000637 [21]) and s288 (SRA accession: SRP028552 [32]).

In Zebrafish, all the three datasets from the tuab strain (SRA accession: SRP034750 [4], SRP010040 [5], and SRP023492 [25]).

*equal contribution

For additional benchmarking on completely independent datasets, we also used datasets from *C. albicans* treated with blasticidin (SRA accession: SRP032814 [34]), one from *S. pombe* treated with tunicamycin (SRA accession: SRP107240 [19]) and one each from chimpanzee and macaque lymphoblastoid cell lines involving flash freezing for inhibiting translation (SRA accession: SRP062129 [49]). To the best of our knowledge the datasets we selected for *C. albicans*, chimpanzee and macaque are the only public datasets available for these species.

2 Obtaining and pre-processing data

We downloaded the raw data (Supplementary Table S1) from NCBI's Sequence Read Archive (SRA) using `pysradb` [10]. We used `cutadapt` [31] to perform adapter trimming. The specific adapters for each dataset are either obtained from the corresponding papers or were automatically inferred by checking for over-represented $k-$mers at the 3'end. Sequences of the adapters for each dataset is documented in Supplementary Table S2. All the Ribo-seq and RNA-seq data were mapped using `STAR` [11] by allowing at most two mismatches (`--outFilterMismatchNmax 2`) and forcing end-to-end (`--alignEndsType EndToEnd`) read alignment. Only uniquely mapping reads were retained (`-outFilterMultimapNmax 1`). For human and mouse, we relied on the GENCODE [16] GTF for annotation. For all other species except *C. albicans*, we used ENSEMBL [38]. For *C. albicans*, both the FASTA and the GTF were obtained from the Candida Genomes database [41]. The assembly and GTF information is summarized in Supplementary Table S3. FASTA is handled using the pyfaidx package [40].

The strand-specific protocol, either forward stranded, reverse stranded or unstranded, is inferred by checking the first $20,000$ reads from the mapping results. Since most tools we compared with can only deal with forward stranded protocol, our ten datasets are all forward stranded for both RNA-seq and Ribo-seq samples. BAM files are processed using pysam, a python interface to samtools [26].

To create fragment length specific metagene profile, we counted the number of 5' end of reads at each nucleotide per fragment length. Supplementary Figures S2 and S3 show the distribution of fragment lengths for Ribo-seq and RNA-seq samples across different datasets in human and mouse, respectively. Metagene plots for individual fragment lengths which were retained for downstream analysis for different datasets are shown in Supplementary Figures S4 and S5.

The specific Ribo-seq and RNA-seq samples used from each dataset for the benchmarking along with the read lengths and the corresponding P-site offsets used for the Ribo-seq samples can be found in Supplementary Table S4.

AUROC, F1 scores, and p-values for AUROC difference were calculated using the `pROC` [37] package in R. For calculating p-values, we used the `bootstrap` method and set `alternative='greater'`.

3 Learning species-specific cutoffs

Ribo-seq's protocol was initially developed to profile the translational landscape in yeast [22], but it has been widely used to profile the translational status of ORFs in multiple species [33, 48]. We benchmarked ribotricer first using human and mouse datasets where we have access to CCDS annotated regions as a high confidence ground truth for known protein coding status (Supplementary Figures S12-S21). In order to further benchmark ribotricer against other methods, we used additional public Ribo-seq datasets from *Arabidopsis*, *C. elegans*, *Drosophila*, rat, yeast, and zebrafish (Supplementary Table S1). Unlike human and mouse, CCDS annotations are not available for these species. Hence, for such species, we considered the Ribo-seq profile of annotated CDS regions as the true positive and the corresponding RNA-seq profile

as the true negative. In order to establish if we needed to re-adjust our phase score cutoff for each species separately, we summarized the phase scores for both Ribo-seq and RNA-seq samples from multiple public datasets (Supplementary Figure S30). We observed that phase scores of both RNA-seq and Ribo-seq samples vary across species (Supplementary Figures S27, S28, and S29) with higher variation arising from the Ribo-seq samples. The variation in phase scores for RNA-seq samples in the same species is limited, though it also exhibits a species related trend (Supplementary Figure S29). Ribo-seq samples on the other hand exhibit higher intra-species and across-species heterogeneity. Hence, in order to capture this species-specific differences in RNA-seq and Ribo-seq scores, we learned cutoffs for each species separately (Supplementary Table S5 and S6; Supplementary Figure S31). It is worth noting that, human and mouse samples that we previously used for our benchmark exhibit similar variation in RNA-seq and Ribo-seq phase scores besides having higher Ribo-seq phase scores as compared to all other species. On the other hand, the difference between Ribo- and RNA-seq phase scores appears to be particularly low in *Drosophila* datasets (Supplementary Figure S29).

4 Learning dataset-specific cutoffs

In studies where both Ribo-seq and RNA-seq experiment are available, it is possible to fine-tune the phase-score cutoff to be dataset-specific. The Ribo-seq and RNA-seq samples within the same species can show variation in terms of their phase score (Supplementary Figure S29) and hence, it is possible that learning dataset-specific cutoffs leads to an overall better performance (Supplementary Figures S35-S39). To learn the dataset-specific cutoffs, we calculated the median difference between phase scores of Ribo-seq and RNA-seq profiles for each dataset over only protein-coding regions. Using a sampling strategy where a one-third fraction of protein-coding profiles were used to determine the median difference between Ribo-seq and RNA-seq profiles with replacement ($n_{\text{bootstraps}} = 10000$) [13], the dataset-specific cutoff was assigned to be the median of these differences. It is worth mentioning that this approach is only viable for studies where both Ribo-seq and RNA-seq samples are available. The dataset-specific cutoffs result in ribotricer achieving higher F1 scores in some but not all datasets (Supplementary Tables S9-S11; Supplementary Figures S35-S39). In all our datasets, a median difference of 0.25 or more between Ribo-seq and RNA-seq protein-coding profiles results in an F1 score greater than 0.73 (Supplementary Figure S38). Given a set of Ribo-seq and RNA-seq mapped files (BAM), the dataset-specific cutoffs can be determined by using `ribotricer learn-cutoff` (Section 8.5).

5 Ribotricer's phase score remains stable on truncated ORFs

In order to test the ability of ribotricer to correctly predict the translation status of an ORF whose length has been shortened due to truncation we performed a simulation where for all candidate ORFs which have atleast 50% of non-empty codons, *i.e.* codons with non-zero reads, we truncated it from 3' end such that the truncated length was $10 - 100\%$ of the original length. For each such truncated ORF, we calculated ribotricer's phase score and compared it with the corresponding RiboCode generated p-value. It is worth mentioning, that among the tools of capable of performing exon level classification, we were able to benchmark ribotricer against only RiboCode and ORFscore as RiboTaper requires bam files of both RNA-seq and Ribo-seq samples.

Ribotricer's score for the truncated ORF is negligibly different from the original ORF with a maximum difference of ± 0.05 (Supplementary Figure S24 and S25) as demonstrated using a human (SRA accession: SRP063852) and a mouse dataset (SRA accession: SRP003554). On the other hand, the RiboCode generated

p-values show a clear dependence on the ORF length with the deviation from original score being as high as $\pm$ 100. It is worth mentioning that the differences between truncated and original profile for RiboCode are calculated on a $\log_{10}$ scale as it outputs p-values, while for both ribotricer and ORFscore, the differences are calculated on the same scale as the scores.

6 Ribotricer can detect ORFs as short as 20 codons

In order to determine the minimum length of ORF that can be detected by ribotricer we performed a simulation using the Ribo-seq profiles of genes with total codons > 100 and with at least 50% non-empty codons. We then randomly sampled $10 - 100$ codons, without maintaining their order explicitly, 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 Figures S22 and S23).

7 Running ribotricer on a new species

We provide a list of recommended phase score cutoffs (Supplementary Table S6) for most species where there are at least three or more public Ribo-seq datasets (Supplementary Table S1). The cutoffs for each species were learned empirically by using Ribo-seq and RNA-seq samples from two datasets and maximizing the F1 score by treating the Ribo-seq profiles of CCDS/CDS regions as ground true positive and the corresponding RNA-seq profiles as true negatives (Supplementary Figure S31; Supplementary Table S5). However, this approach is only best suited for species where there are multiple datasets available. For a new species where there are only few or none datasets available and hence the cutoff cannot be learned empirically, we recommend using the median score difference between the profiles of annotated CDS regions of a Ribo-seq and the corresponding RNA-seq sample. This strategy is also used by RibORF [23] which tunes the parameters of its model by selecting one-third of the CDS profiles as true positives. We followed this strategy of using the median phase score difference as the phase score cutoff for each of the four species: *C. albicans*, chimpanzee, macaque and *S. pombe*. Except for *S. pombe*, all other species have only one public dataset available to the best of our knowledge (Supplementary Table S1).

We first generated candidate ORF list for each species using ribotricer over transcripts with annotated CDS regions. Phase scores were then calculated for each RNA-seq and Ribo-seq sample over these CDS annotated candidate ORFs (Supplementary Figure S42). The median differences in Ribo-seq and RNA-seq phase scores for *C. albicans*, chimpanzee, macaque and *S. pombe* is summarized at the end of Supplementary Table S7. We used these differences as species-specific cutoffs for benchmarking ribotricer against other methods.

Ribotricer results in the best AUROC for all the four species with the difference between ribotricer and the second best method statistically significant in all the cases (Supplementary Figure S40; Supplementary Table S8). It is worth mentioning that the AUROC metric is not dependent on the choice of the learned cutoff. Furthermore, ribotricer is also the best method using the F1 score metric (Supplementary Figure S41; Supplementary Table S9).

We recommend using the species-specific cutoffs for all the species as listed in Supplementary Table S6. For any new species, we recommend using median phase score differences on ribotricer generated candidate ORFs over CDS annotated transcripts between Ribo-seq and RNA-seq samples (Supplementary Figure S42). This can be determined by `ribotricer` itself, using the `learn-cutoff` subcommand. (See Section 8.5).

8 Using ribotricer

In order to use `ribotricer`, the following three files are required:

- **GTF:** genome annotation file in GTF format (ENSEMBL/Gencode/others)
- **FASTA:** reference genome file in FASTA format
- **BAM:** alignment file in BAM format

Henceforth, we use the boldface acronyms above to refer to these files as such.

8.1 Preparing candidate ORFs list

ribotricer prepares a candidate list of ORFs given a GTF and FASTA file. For any species, given a reference and a fixed version of GTF, this step only needs to be done once. Ribotricer by default searches for ORFs defined by an 'AUG' start and an in-frame stop codon ('UAG', 'UAA', and 'UGA') and are a minimum of 60 nucleotides long. It is possible to expand the definition of ORF by supplying a list of all start codons using the `--start_codons` parameter. It is also possible to change the minimum length of an ORF by using the `--min_orf_length` option. If multiple potential in-frame start codons exist upstream of a stop codon, we always choose AUG if it exists, otherwise, we take the most upstream one as the start codon.

```
ribotricer prepare-orfs --gtf {GTF} \
                        --fasta {FASTA} \
                        --prefix {RIBOTRICER_INDEX}
```

The command above will create a list of candidate ORFs at the `RIBOTRICER_INDEX` location.

For this study, we used a total of ten codons with a maximum of one nucleotide difference from "ATG" as potential start codons including ATA, ATC, ATT, AAG, ACG, AGG, ATG, CTG, GTG, TTG. Note that we use 'T' as a nucleotide here instead of 'U' as the reference FASTA almost always contains DNA sequences.

8.2 Detecting actively translating ORFs using ribotricer

Ribotricer's ORF list as created above can then be used along with the BAM to define the translation status of these ORFs:

```
ribotricer detect-orfs --bam {BAM} \
                       --ribotricer_index {RIBOTRICER_INDEX}_candidate_ORFs.tsv \
                       --prefix {OUT_PREFIX}
```

For each ORF in the candidate ORFs list, ribotricer calculates the phase score on the read profiles after performing read length appropriate offset shifts. These offsets are determined by maximizing the cross-correlation of these profiles with the profile for the most abundant read length. Additionally, ribotricer automatically infers the sequencing protocol (forward/reverse) and only uses unique mapping reads that conform to the strand orientation in the GTF. For example, a read uniquely mapping to a gene defined on the negative strand for a forward stranded protocol, will be discarded.

In order to assign 'non-translating' or 'translating' status, ribotricer, by default, uses a cutoff threshold of 0.428. ORFs with phase score above 0.428 are marked as translating as long as they have at least five

codons with non-zero read count. Ribotricer does not take coverage into account for predicting an ORF to be translating or not-translating. Apart from these two criteria, there is no other requirement for an ORF to be active. Though, a region with higher overall coverage as defined by number of reads per unit codon might be a more confident 'hit' for active translation, our method is designed to find evidence of active translation based on the qualitative pattern of "high-low-low" and hence our rankings are purely based on phase scores.

The default cutoff (0.428) was learned using public human and mouse Ribo-seq datasets, where the gap between Ribo- and RNA-seq phase scores is the highest amongst other species (Supplementary Table S7) and hence, it is a conservative cutoff for detecting active translation. We provide a list of species-specific recommended cutoffs (Supplementary Table S6), optimized for F1 score based performance.

The main output of the above command is a tab separated file consisting for each candidate ORF, its translation status, the corresponding transcript and gene and the ORF type. Different ORF types defined by ribotricer are described below:

- **annotated:** CDS annotated in the provided GTF file
- **super_uORF:** upstream ORF of the annotated CDS, not overlapping with any CDS of the same gene
- **super_dORF:** downstream ORF of the annotated CDS, not overlapping with any CDS of the same gene
- **uORF:** upstream ORF of the annotated CDS, not overlapping with the main CDS
- **dORF:** downstream ORF of the annotated CDS, not overlapping with the main CDS
- **overlap_uORF:** upstream ORF of the annotated CDS, overlapping with the main CDS
- **overlap_dORF:** downstream ORF of the annotated CDS, overlapping with the main CDS
- **novel:** ORF in non-coding genes or in non-coding transcripts of coding genes

8.3 Filtering actively translating ORFs using multiple criteria

In order to assign 'non-translating' or 'translating' status, ribotricer by default uses a cutoff threshold of '0.428'. ORFs with phase score above '0.428' are marked as translating as long as they have at least five codons with non-zero read count. By default, ribotricer does not take coverage or count information explicitly into account for predicting an ORF to be translating or not-translating. However, this behavior can be changed by following filters:

- `--min_valid_codons` (default=5): Minimum number of codons with non-zero reads for determining active translation
- `--min_valid_codons_ratio` (default=0): Minimum ratio of codons with non-zero reads to total codons for determining active translation
- `--min_reads_per_codon` (default=0): Minimum number of reads per codon for determining active translation
- `--min_read_density` (default=0.0): Minimum read density (total reads/length) over an ORF total codons for determining active translation

For each of the above filters, an ORF failing **any** of the filters is marked as 'non-translating'.

For example, to ensure that each ORF has at least 3/4 of its codons non-empty, we can specify `--min_valid_codons_ratio` to be 0.75:

```
ribotricer detect-orfs --bam {BAM} \
                       --ribotricer_index {RIBOTRICER_INDEX}_candidate_ORFs.tsv \
                       --prefix {OUTPUT_PREFIX}
                       --min_valid_codons_ratio 0.75
```

It might also often be desired to have some minimum density of reads over an ORF. The read density here is defined as the ratio of total number of reads over an ORF to its length. For example to ensure that each 'translating' ORF has at least a read density of 10, we will specify `--min_read_density` to be 10.

```
ribotricer detect-orfs --bam {BAM} \
                       --ribotricer_index {RIBOTRICER_INDEX}_candidate_ORFs.tsv \
                       --prefix {OUTPUT_PREFIX}
                       --min_read_density 10.0
```

The above filters can be combined to give ORFs that have high read density as well as have have reads present over most of the codons in the profile. Note that increasing the value of any of the four filters will usually result in a smaller list of ORFs marked 'translating'.

8.4 Downstream ranking and filtering

It is also possible to filter actively-translating ORFs after running ribotricer. Ribotricer produces a tab separated file with columns that include read-density, number and ratio of valid codons to total codons in the ORF besides the phase score. As such, filtering can be performed downstream using `awk` or any other programming language. Here we provide an example of filtering and sorting the output of a ribotricer run using Python using the `pandas` library:

Listing 1: **Filtering ORFs using python**. The function returns a filtered list of translating ORFs which have a read density of at least 2.5; a total read count of atleast 50; and the ratio of non-empty codons to total codons atleast 0.75.

```
import pandas as pd
def filtered_df(df):
        df_filtered = df.loc[df.status=='translating']
        df_filtered = df.loc[(df['read_density']>=2.5) & \
                             (df['read_count']>=50) & \
                          (df['valid_codons_ratio']>=0.75)]
    df_sorted = df_filtered.sort_values(by=['phase_score',
                                            'read_density'],
                                        ascending=[False,
                                                   False])
    return df_sorted
```

```
# read ribotricer output
ribotricer_output_df = pd.read_csv('/path/to/translating_ORFs.tsv', sep='\t')
# filter and sort ribotricer output
ribotricer_filtered_df = filtered_df(ribotricer_output_df)
```

8.5 Learning cutoff empirically from data

Ribotricer can learn cutoff empirically from the data. Given at least one Ribo-seq and one RNA-seq BAM file, ribotricer learns the cutoff by running one iteration of the algorithm on the provided files with a pre-specified cutoff (`--phase_score_cutoff`, default: 0.428) and then uses the generated output to find the median difference between Ribo-seq and RNA-seq phase scores of only candidate ORFs with `transcript_type` annotated as `protein_coding`:

```
ribotricer learn-cutoff --ribo_bams ribo_bam1.bam,ribo_bam2.bam \
--rna_bams rna_1.bam \
--prefix ribo_rna_prefix \
--ribotricer_index {RIBOTRICER_ANNOTATION}
```

References

[1] Dmitry E Andreev, Patrick B F O'Connor, Alexander V Zhdanov, Ruslan I Dmitriev, Ivan N Shatsky, Dmitri B Papkovsky, and Pavel V Baranov. Oxygen and glucose deprivation induces widespread alterations in mRNA translation within 20 minutes. *Genome Biol.*, 16:90, May 2015.

[2] Carlo G Artieri and Hunter B Fraser. Evolution at two levels of gene expression in yeast. *Genome Res.*, 24(3):411–21, Mar 2014.

[3] Julie L Aspden, Ying Chen Eyre-Walker, Rose J Phillips, Unum Amin, Muhammad Ali S Mumtaz, Michele Brocard, and Juan-Pablo Couso. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *Elife*, 3:e03528, Aug 2014.

[4] Ariel A Bazzini, Timothy G Johnstone, Romain Christiano, Sebastian D Mackowiak, Benedikt Obermayer, Elizabeth S Fleming, Charles E Vejnar, Miler T Lee, Nikolaus Rajewsky, Tobias C Walther, and Antonio J Giraldez. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.*, 33(9):981–93, May 2014.

[5] Ariel A Bazzini, Miler T Lee, and Antonio J Giraldez. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science*, 336(6078):233–7, Apr 2012.

[6] Heidi M Blank, Ricardo Perez, Chong He, Nairita Maitra, Richard Metz, Joshua Hill, Yuhong Lin, Charles D Johnson, Vytas A Bankaitis, Brian K Kennedy, Rodolfo Aramayo, and Michael Polymenis. Translational control of lipogenic enzymes in the cell cycle of synchronous, growing yeast cells. *EMBO J.*, 36(4):487–502, 02 2017.

[7] Lorenzo Calviello, Neelanjan Mukherjee, Emanuel Wyler, Henrik Zauber, Antje Hirsekorn, Matthias Selbach, Markus Landthaler, Benedikt Obermayer, and Uwe Ohler. Detecting actively translated open reading frames in ribosome profiling data. *Nature Methods*, 13(2):165, 2015.

[8] Xun Chen and Dion Dickman. Development of a tissue-specific ribosome profiling approach in Drosophila enables genome-wide evaluation of translational adaptations. *PLoS Genetics*, 13(12):e1007117, Dec 2017.

[9] Jun Cho, Nam-Kyung Yu, Jun-Hyeok Choi, Su-Eon Sim, SukJae Joshua Kang, Chuljung Kwak, Seung-Woo Lee, Ji-il Kim, Dong Il Choi, V Narry Kim, et al. Multiple repressive mechanisms in the hippocampus during memory formation. *Science*, 350(6256):82–87, 2015.

[10] Saket Choudhary. pysradb: A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive. *F1000Research*, 8, 2019.

[11] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[12] Joshua G Dunn, Catherine K Foo, Nicolette G Belletier, Elizabeth R Gavis, and Jonathan S Weissman. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *Elife*, 2:e01179, Dec 2013.

[13] Bradley Efron. Bootstrap Methods: Another Look at the Jackknife. In *Breakthroughs in Statistics*, pages 569–593. Springer, 1992.

[14] Stephen W Eichhorn, Alexander O Subtelny, Iva Kronja, Jamie C Kwasnieski, Terry L Orr-Weaver, and David P Bartel. mRNA poly(A)-tail changes specified by deadenylation broadly reshape translation in Drosophila oocytes and early embryos. *Elife*, 5, 07 2016.

[15] Noelia Fradejas-Villar, Sandra Seeher, Christine B Anderson, Michael Doengi, Bradley A Carlson, Dolph L Hatfield, Ulrich Schweizer, and Michael T Howard. The RNA-binding protein Secisbp2 differentially modulates UGA codon reassignment and RNA decay. *Nucleic Acids Research*, 45(7):4094–4107, 2016.

[16] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, 2018.

[17] Huili Guo, Nicholas T Ingolia, Jonathan S Weissman, and David P Bartel. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835, 2010.

[18] Nicholas R Guydosh and Rachel Green. Dom34 rescues ribosomes in 3' untranslated regions. *Cell*, 156(5):950–62, Feb 2014.

[19] Nicholas R Guydosh, Philipp Kimmig, Peter Walter, and Rachel Green. Regulated Ire1-dependent mRNA decay requires no-go mRNA degradation to maintain endoplasmic reticulum homeostasis in S. pombe. *Elife*, 6, 09 2017.

[20] Andrew C Hsieh, Yi Liu, Merritt P Edlind, Nicholas T Ingolia, Matthew R Janes, Annie Sher, Evan Y Shi, Craig R Stumpf, Carly Christensen, Michael J Bonham, et al. The translational landscape of mTOR signalling steers cancer initiation and metastasis. *Nature*, 485(7396):55, 2012.

[21] Nicholas T Ingolia, Sina Ghaemmaghami, John R S Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–23, Apr 2009.

[22] Nicholas T Ingolia, Sina Ghaemmaghami, John RS Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223, 2009.

[23] Zhe Ji, Ruisheng Song, Aviv Regev, and Kevin Struhl. Many lncRNAs, 5UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife*, 4:e08890, 2015.

[24] Piyada Juntawong, Thomas Girke, Jérémie Bazin, and Julia Bailey-Serres. Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.*, 111(1):E203–12, Jan 2014.

[25] Miler T Lee, Ashley R Bonneau, Carter M Takacs, Ariel A Bazzini, Kate R DiVito, Elizabeth S Fleming, and Antonio J Giraldez. Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. *Nature*, 503(7476):360–4, Nov 2013.

[26] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[27] Nathanael G Lintner, Kim F McClure, Donna Petersen, Allyn T Londregan, David W Piotrowski, Liuqing Wei, Jun Xiao, Michael Bolt, Paula M Loria, Bruce Maguire, et al. Selective stalling of human translation through small-molecule engagement of the ribosome nascent chain. *PLoS Biology*, 15(3):e2001882, 2017.

[28] Ming-Jung Liu, Szu-Hsien Wu, Jing-Fen Wu, Wen-Dar Lin, Yi-Chen Wu, Tsung-Ying Tsai, Huang-Lung Tsai, and Shu-Hsing Wu. Translational landscape of photomorphogenic Arabidopsis. *Plant Cell*, 25(10):3699–710, Oct 2013.

[29] Radoslaw Lukoszek, Peter Feist, and Zoya Ignatova. Insights into the adaptive response of Arabidopsis thaliana to prolonged thermal stress by ribosomal profiling and RNA-Seq. *BMC Plant Biol.*, 16(1):221, 10 2016.

[30] Marco Mariotti, Sumangala Shetty, Lisa Baird, Sen Wu, Gary Loughran, Paul R Copeland, John F Atkins, and Michael T Howard. Multiple RNA structures affect translation initiation and UGA redefinition efficiency during synthesis of selenoprotein P. *Nucleic Acids Research*, 45(22):13004–13015, 2017.

[31] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, 17(1):pp–10, 2011.

[32] C Joel McManus, Gemma E May, Pieter Spealman, and Alan Shteyman. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.*, 24(3):422–30, Mar 2014.

[33] Audrey M Michel, Stephen J Kiniry, Patrick B F OConnor, James P Mullan, and Pavel V Baranov. GWIPS-viz: 2018 update. *Nucleic acids research*, 46(D1):D823–D830, 2017.

[34] Dale Muzzey, Gavin Sherlock, and Jonathan S Weissman. Extensive and coordinated control of allele-specific expression by both transcription and translation in Candida albicans. *Genome Res.*, 24(6):963–73, Jun 2014.

[35] Danny D Nedialkova and Sebastian A Leidel. Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity. *Cell*, 161(7):1606–18, Jun 2015.

[36] Alessandro Ori, Brandon H Toyama, Michael S Harris, Thomas Bock, Murat Iskar, Peer Bork, Nicholas T Ingolia, Martin W Hetzer, and Martin Beck. Integrated Transcriptome and Proteome Analyses Reveal Organ-Specific Proteome Deterioration in Old Rats. *Cell Syst*, 1(3):224–37, Sep 2015.

[37] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, 12(1):77, 2011.

[38] Magali Ruffier, Andreas Kähäri, Monika Komorowska, Stephen Keenan, Matthew Laird, Ian Longden, Glenn Proctor, Steve Searle, Daniel Staines, Kieron Taylor, et al. Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation. *Database*, 2017, 2017.

[39] Sebastian Schafer, Eleonora Adami, Matthias Heinig, Katharina E Costa Rodrigues, Franziska Kreuchwig, Jan Silhavy, Sebastiaan van Heesch, Deimante Simaite, Nikolaus Rajewsky, Edwin Cuppen, Michal Pravenec, Martin Vingron, Stuart A Cook, and Norbert Hubner. Translational regulation shapes the molecular landscape of complex disease phenotypes. *Nat Commun*, 6:7200, May 2015.

[40] Matthew D Shirley, Zhaorong Ma, Brent S Pedersen, and Sarah J Wheelan. Efficient" pythonic" access to fasta files using pyfaidx. Technical report, PeerJ PrePrints, 2015.

[41] Marek S Skrzypek, Jonathan Binkley, Gail Binkley, Stuart R Miyasato, Matt Simison, and Gavin Sherlock. The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Research*, page gkw924, 2016.

[42] Boris Slobodin, Ruiqi Han, Vittorio Calderone, Joachim AF Oude Vrielink, Fabricio Loayza-Puch, Ran Elkon, and Reuven Agami. Transcription impacts the efficiency of mRNA translation via co-transcriptional N6-adenosine methylation. *Cell*, 169(2):326–337, 2017.

[43] Michael Stadler, Karen Artiles, Julia Pak, and Andrew Fire. Contributions of mRNA abundance, ribosome loading, and post- or peri-translational effects to temporal repression of C. elegans heterochronic miRNA targets. *Genome Res.*, 22(12):2418–26, Dec 2012.

[44] Michael Stadler and Andrew Fire. Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA*, 17(12):2063–73, Dec 2011.

[45] Michael Stadler and Andrew Fire. Conserved translatome remodeling in nematode species executing a shared developmental transition. *PLoS Genetics*, 9(10):e1003739, 2013.

[46] Craig R Stumpf, Melissa V Moreno, Adam B Olshen, Barry S Taylor, and Davide Ruggero. The translational landscape of the mammalian cell cycle. *Molecular Cell*, 52(4):574–582, 2013.

[47] Hayami Sugiyama, Kazutoshi Takahashi, Takuya Yamamoto, Mio Iwasaki, Megumi Narita, Masahiro Nakamura, Tim A Rand, Masato Nakagawa, Akira Watanabe, and Shinya Yamanaka. Nat1 promotes

translation of specific proteins that induce differentiation of mouse embryonic stem cells. *Proceedings of the National Academy of Sciences*, 114(2):340–345, 2017.

[48] Hongwei Wang, Ludong Yang, Yan Wang, Leshi Chen, Huihui Li, and Zhi Xie. Rpfdb v2. 0: an updated database for genome-wide information of translated mrna generated from ribosome profiling. *Nucleic acids research*, 47(D1):D230–D234, 2018.

[49] Sidney H Wang, Chiaowen Joyce Hsiao, Zia Khan, and Jonathan K Pritchard. Post-translational buffering leads to convergent protein expression levels between primates. *Genome Biol.*, 19(1):83, 06 2018.

[50] Guoyong Xu, George H Greene, Heejin Yoo, Lijing Liu, Jorge Marqués, Jonathan Motley, and Xinnian Dong. Global translational reprogramming is a fundamental layer of immune regulation in plants. *Nature*, 545(7655):487–490, 05 2017.

List of Tables

List of Figures

9 Supplementary tables

Table S1: List of datasets.

SRA Accession	Species	Cell type	Treatment	Citation
SRP010679	Human	PC3	100 μg/ml cycloheximide	[20]
SRP029589	Human	HeLa	cycloheximide	[46]
SRP063852	Human	HEK293	100 μg/ml cycloheximide	[7]
SRP098789	Human	HeLa	100 μg/ml cycloheximide	[27]
SRP102021	Human	H1933	100 μg/ml cycloheximide	[42]
SRP003554	Mouse	neutrophils cultured from bone marrow	100 μg/ml cycloheximide	[17]
SRP062407	Mouse	hippocampal neurons	100 μg/ml cycloheximide	[9]
SRP078005	Mouse	liver	200 μg/ml cycloheximide	[15]
SRP091889	Mouse	ESC	cycloheximide	[47]
SRP115915	Mouse	liver	200 μg/ml cycloheximide	[30]
SRP108862	Arabidopsis	inflorescences	unavailable	unpublished
SRP087624	Arabidopsis	leaf tissue	50 μg/ml cycloheximide	[50]
SRP029587	Arabidopsis	whole seedlings	50 μg/ml cycloheximide	[24]
SRP059391	Arabidopsis	leaf tissue	100 μg/ml cycloheximide	[29]
SRP018118	Arabidopsis	etiolated seedling	100 μg/ml cycloheximide	[28]
SRP075766	Baker's Yeast	strain by4743	100 μg/ml cycloheximide	[6]
SRP033499	Baker's Yeast	strain: by4741	0.1 mg/ml cycloheximide	[18]
SRP028614	Baker's Yeast	strain: by4176	cycloheximide	[2]
SRP028552	Baker's Yeast	strain: s288	cycloheximide	[32]
SRP000637	Baker's Yeast	strain: by4741	100 μg/ml cycloheximide	[21]
SRP056647	*C. elegans*	strain: n2	100 μg/ml cycloheximide	[35]
SRP026198	*C. elegans*	strain: n2	100 μg/ml cycloheximide	[45]
SRP014427	*C. elegans*	strain: n2	cycloheximide	[43]
SRP010374	*C. elegans*	strain: n2	cycloheximide	[44]
SRP108999	Drosophila	body wall muscle	100 μg/ml cycloheximide	[8]
SRP028243	Drosophila	embryo	20 μg/ml emetine	[12]
SRP076919	Drosophila	oocytes	100 μg/ml cycloheximide	[14]
SRP045475	Drosophila	S2 cell	100 μg/ml cycloheximide	[3]
SRP056012	Rat	PC12 Cells	100 μg/ml streptomycin	[1]
SRP045777	Rat	Pheochromocytoma cells	streptomycin	[36]
ERP007231	Rat	strain: bn/shr	0.1 mg/ml cycloheximide	[39]
SRP034750	Zebrafish	strain: tuab	100 μg/ml cycloheximide	[4]
SRP010040	Zebrafish	strain: tuab	100 μg/ml cycloheximide	[5]
SRP023492	Zebrafish	strain: tuab	50 μg/ml cycloheximide	[25]
SRP032814	*C. albicans*	strain: sc5314	10 μg/mL Blasticidin S	[34]
SRP107240	*S. pombe*	strain: WT	0.15 μg/ml tunicamycin	[19]
SRP062129	Chimpanzee	Lymphoblastoid cell line	flash freezing	[49]
SRP062129	Macaque	Lymphoblastoid cell line	flash freezing	[49]

Table S2: Adapters trimmed from Ribo-seq and RNA-seq samples for each dataset.

SRA Accession	Ribo-seq adapter	RNA-seq adapter
SRP010679	CTGTAGGCAC	CTGTAGGCAC
SRP029589	CTGTAGGCACCATCAAT	CTGTAGGCACCATCAAT
SRP063852	None	None
SRP098789	CTGTAGGCACCATCAAT	CTGTAGGCACCATCAAT
SRP102021	TCGTATGCCGTCTTCTGCTTG	None
SRP003554	TCGTATG	TCGTATG
SRP062407	TGGAATTCTCGGGTGCCAAGG	TGGAATTCTCGGGTGCCAAGG
SRP078005	TGGAATTCTCGGGTGCCAAGG	TGGAATTCTCGGGTGCCAAGG
SRP091889	AGATCGGAAGAGCACACGTCT	AGATCGGAAGAGCACACGTCT
SRP115915	TGGAATTCTCGGGTGCCAAGG	TGGAATTCTCGGGTGCCAAGG
SRP108862	TGGAATTCTCGG	AGATCGGAAGAGC
SRP087624	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP029587	TCGTATGCCGTCTTCTGCTTG	TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC
SRP059391	TGGAATTCTCGG	TGGAATTCTCGG
SRP018118	TGGAATTCTCGG	TGGAATTCTCGG
SRP075766	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP033499	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP028614	AAAAAAAAAAA_AGATCGGAAGAGC	AAAAAAAAAAA_AGATCGGAAGAGC
SRP028552	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP000637	AAAAAAAA_AGATCGGAAGAGC	AAAAAAAA_AGATCGGAAGAGC
SRP056647	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP026198	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP014427	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP010374	AAAAAAA_AGATCGGAAGAGC	AGATCGGAAGAGC
SRP108999	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP028243	CTGTAGGCACCATCAAT	AGATCGGAAGAGC
SRP076919	AGATCGGAAGAGC	TGGAATTCTCGG
SRP045475	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP056012	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP045777	AGATCGGAAGAGC	AGATCGGAAGAGC
ERP007231	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP034750	AGATCGGAAGAGCACACGTCTGAACTCCAGTCA	AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
SRP010040	ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAA	ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAA
SRP023492	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP032814	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP107240	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP062129	TGGAATTCTCGG	AGATCGGAAGAGC

Table S3: Reference assemblies and GTF for each species

Species	Reference assembly	GTF
Human	GRCh38	Gencode (v94)
Mouse	GRCm38	Gencode (v94)
Arabidopsis	TAIR10	ENSEMBL (v96)
C.elegans	WBcel235	ENSEMBL (v96)
Drosophila	BDGP6	ENSEMBL (v96)
Rat	Rnor6.0	ENSEMBL (v96)
Zebrafish	GRCz11	ENSEMBL (v96)
C. albicans	SC5314	Candida Genomes Database (r27)
S. pombe	ASM294v2	ENSEMBL (v96)
Chimpanzee	Pantro3	ENSEMBL (v96)
Macaque	Mmul8	ENSEMBL (v96)

Table S4: Ribo- and RNA-seq samples used for the benchmarking along with the read lengths and P-site offsets used for Ribo-seq samples.

SRA Accession	Ribo-seq sample	Read lengths (nt)	P-site offsets (nt)	RNA-seq sample	Species
SRP010679	SRX118286	28,29,30	12,13,13	SRX118285	Human
SRP029589	SRX345309	29,30,32	12,12,13	SRX345311	Human
SRP063852	SRX1254413	28,29,30	12,12,12	SRX426378	Human
SRP098789	SRX2536421	28,30	12,13	SRX2536426	Human
SRP102021	SRX2647167	28,29,30,31	12,12,12,12	SRX2647164	Human
SRP003554	SRX026871	28,29,30	12,12,12	SRX026872	Mouse
SRP062407	SRX1149649	28,29,30,31	12,12,12,12	SRX1149668	Mouse
SRP078005	SRX1900396	26,27,28,29,30	12,12,12,12,12	SRX1900402	Mouse
SRP091889	SRX2255510	26,27,28,29,30	12,12,12,12,12	SRX2255511	Mouse
SRP115915	SRX3110803	29,30,31,32,33,34	12,12,12,13,13,13	SRX3110807	Mouse
SRP108862	SRX2896566	23	12	SRX2896570	Arabidopsis
SRP087624	SRX2148419	28,29,30,31,32	12,12,12,12,12	SRX2148418	Arabidopsis
SRP029587	SRX345240	26,27	12,12	SRX345251	Arabidopsis
SRP059391	SRX1056790	27,30	12,12	SRX1056791	Arabidopsis
SRP018118	SRX219170	28,29,30,31	11,12,13,13	SRX347226	Arabidopsis
SRP075766	SRX1801603	26,27,28	11,12,13	SRX1801650	Baker's Yeast
SRP033499	SRX386988	29,30,31	12,12,12	SRX386983	Baker's Yeast
SRP028614	SRX333052	28,29,30	12,13,13	SRX334053	Baker's Yeast
SRP028552	SRX332185	28,29,30	11,12,12	SRX332188	Baker's Yeast
SRP000637	SRX003187	28,29,30,31	12,12,12,12	SRX003191	Baker's Yeast
SRP056647	SRX971770	28,29,30,31,32	12,12,12,12,12	SRX971774	*C. elegans*
SRP026198	SRX311784	29,30,31,32	12,12,12,12	SRX311777	*C. elegans*
SRP014427	SRX160518	28,29,30,31,32	12,12,12,12,12	SRX160149	*C. elegans*
SRP010374	SRX118118	28,29,30,31,32	12,12,12,12,12	SRX118116	*C. elegans*
SRP108999	SRX2902857	29,30,31,32	12,13,10,12	SRX2902867	Drosophila
SRP028243	SRX327686	28,29,30,32,33,34	12,12,12,12,12,13	SRX327688	Drosophila
SRP076919	SRX1870218	34	12	SRX1870191	Drosophila
SRP045475	SRX679371	28,29,30,31,32	12,12,12,12,12	SRX679372	Drosophila
SRP056012	SRX915217	29,30,31,32	12,12,13,13	SRX915210	Rat
SRP045777	SRX686499	28,29,30,31	12,12,12,13	SRX686500	Rat
ERP007231	ERX609893	28,29,30,31,32	12,12,12,12,12	ERX609898	Rat
SRP034750	SRX399800	28,29,30,31	12,12,12,12	SRX399817	Zebrafish
SRP010040	SRX113357	27,28,30,31,33,34	12,12,12,12,12,12	SRX113344	Zebrafish
SRP023492	SRX288475	28,29,30	12,12,12	SRX288474	Zebrafish
SRP032814	SRX375317	28,29,30	12,12,12	SRX375318	*C. albicans*
SRP107240	SRX2825796	28,29,30	12,13,13	SRX2825805	*S. pombe*
SRP062129	SRX1135820	28,29,30	12,12,12	SRX333018 (SRP028612)	Chimpanzee
SRP062129	SRX1135825	28,29,30	12,12,12	SRX333023 (SRP028612)	Macaque

Table S5: Datasets used to learn ribotricer phase score cutoffs.

SRA Accession	Species	Used to learn cutoff
SRP010679	Human	Yes
SRP029589	Human	No
SRP063852	Human	No
SRP098789	Human	Yes
SRP102021	Human	No
SRP003554	Mouse	Yes
SRP062407	Mouse	No
SRP078005	Mouse	No
SRP091889	Mouse	No
SRP115915	Mouse	Yes
SRP108862	Arabidopsis	No
SRP087624	Arabidopsis	No
SRP029587	Arabidopsis	No
SRP059391	Arabidopsis	Yes
SRP018118	Arabidopsis	Yes
SRP075766	Baker's Yeast	Yes
SRP033499	Baker's Yeast	No
SRP028614	Baker's Yeast	No
SRP028552	Baker's Yeast	Yes
SRP000637	Baker's Yeast	No
SRP056647	*C. elegans*	No
SRP026198	*C. elegans*	Yes
SRP014427	*C. elegans*	No
SRP010374	*C. elegans*	Yes
SRP108999	Drosophila	Yes
SRP028243	Drosophila	Yes
SRP076919	Drosophila	No
SRP045475	Drosophila	No
SRP056012	Rat	Yes
SRP045777	Rat	No
ERP007231	Rat	Yes
SRP034750	Zebrafish	Yes
SRP010040	Zebrafish	Yes
SRP023492	Zebrafish	No

Table S6: Species specific recommended phase score cutoffs for ribotricer. A "#" indicates the cutoff for the species is taken to be the median phase score difference between CDS annotated Ribo-seq and RNA-seq profiles since they only had one dataset each.

Species	Cutoff
Arabidopsis	0.330
Baker's Yeast	0.318
C. elegans	0.249
Drosophila	0.181
Human	0.440
Mouse	0.418
Rat	0.453
Zebrafish	0.249
C. albicans#	0.228
S. pombe#	0.409
Chimpanzee#	0.334
Macaque#	0.321

Table S7: Species wise mean, median and standard deviation of difference of Ribo-seq and RNA-seq phase scores. SD = Standard Deviation. A "#" indicates that the median phase score difference for these species is also considered as cutoff for ribotricer, since they only had one dataset each.

species	number of samples	mean difference phase score	median difference phase score	SD
Arabidopsis	5	0.308	0.365	0.252
Baker's Yeast	5	0.309	0.287	0.225
C.elegans	4	0.232	0.273	0.235
Drosophila	4	0.048	0.054	0.221
Human	5	0.385	0.428	0.240
Mouse	5	0.468	0.528	0.230
Rat	3	0.260	0.303	0.253
Zebrafish	3	0.325	0.388	0.309
C. albicans#	1	0.228	0.225	0.151
S. pombe#	1	0.380	0.409	0.176
Chimpanzee#	1	0.328	0.334	0.233
Macaque#	1	0.285	0.321	0.218

Table S8: **Best and second to best performing methods at AUROC metric for each dataset.** The p-values were calculated using `pROC` [37] package using bootstrap method and `alternative='greater'`. AUROC (B) and AUROC (SB) denotes area under ROC for the best and the second to best methods respectively. A * indicates the dataset was later used to learn the ribotricer cutoffs by maximizing the F1 score. The AUROC values however do not depend on any cutoff.

SRP	Species	Best (B)	Second Best (SB)	AUROC (B)	AUROC (SB)	p-value
SRP018118*	Arabidopsis	ribotricer	RiboCode	0.982	0.923	$< 2.2 \times 10^{-16}$
SRP029587	Arabidopsis	ribotricer	RiboCode	0.897	0.594	$< 2.2 \times 10^{-16}$
SRP059391*	Arabidopsis	ribotricer	ORFscore	0.690	0.632	$< 2.2 \times 10^{-16}$
SRP087624	Arabidopsis	ribotricer	RiboTaper	0.697	0.523	$< 2.2 \times 10^{-16}$
SRP108862	Arabidopsis	ribotricer	RiboCode	0.732	0.607	$< 2.2 \times 10^{-16}$
SRP000637	Baker's Yeast	ribotricer	RiboCode	0.921	0.837	$< 2.2 \times 10^{-16}$
SRP028552*	Baker's Yeast	ribotricer	RiboCode	0.986	0.951	$< 2.2 \times 10^{-16}$
SRP028614	Baker's Yeast	ribotricer	RiboCode	0.966	0.846	$< 2.2 \times 10^{-16}$
SRP033499	Baker's Yeast	ribotricer	RiboCode	0.947	0.783	$< 2.2 \times 10^{-16}$
SRP075766*	Baker's Yeast	ribotricer	RiboCode	0.996	0.962	$< 2.2 \times 10^{-16}$
SRP010374*	*C. elegans*	ribotricer	RiboCode	0.867	0.776	$< 2.2 \times 10^{-16}$
SRP014427	*C. elegans*	ORFscore	ribotricer	0.927	0.920	3.774×10^{-14}
SRP026198*	*C. elegans*	ORFscore	ribotricer	0.956	0.908	$< 2.2 \times 10^{-16}$
SRP056647	*C. elegans*	RiboCode	RiboTaper	0.745	0.745	0.247
SRP028243*	Drosophila	ribotricer	RiboCode	0.725	0.587	$< 2.2 \times 10^{-16}$
SRP045475	Drosophila	ORFscore	RiboTaper	0.633	0.522	$< 2.2 \times 10^{-16}$
SRP076919	Drosophila	ORFscore	ribotricer	0.638	0.465	0.317
SRP108999*	Drosophila	ribotricer	RiboTaper	0.884	0.727	0.068
SRP010679*	Human	ribotricer	RiboCode	0.944	0.849	$< 2.2 \times 10^{-16}$
SRP029589	Human	ribotricer	RiboCode	0.846	0.701	$< 2.2 \times 10^{-16}$
SRP063852	Human	ribotricer	RiboCode	0.969	0.930	$< 2.2 \times 10^{-16}$
SRP098789*	Human	ribotricer	RiboCode	0.975	0.908	$< 2.2 \times 10^{-16}$
SRP102021	Human	ribotricer	RiboCode	0.961	0.927	$< 2.2 \times 10^{-16}$
SRP003554*	Mouse	RiboCode	ribotricer	0.974	0.972	2.045×10^{-6}
SRP062407	Mouse	RiboCode	ORFscore	0.986	0.981	$< 2.2 \times 10^{-16}$
SRP078005	Mouse	ribotricer	RiboCode	0.989	0.968	$< 2.2 \times 10^{-16}$
SRP091889	Mouse	ribotricer	RiboCode	0.981	0.966	$< 2.2 \times 10^{-16}$
SRP115915*	Mouse	ribotricer	RiboCode	0.926	0.923	1.095×10^{-11}
ERP007231*	Rat	RiboTaper	RiboCode	0.955	0.953	3.321×10^{-9}
SRP045777	Rat	ribotricer	RiboCode	0.793	0.746	$< 2.2 \times 10^{-16}$
SRP056012*	Rat	ORFscore	RiboCode	0.971	0.872	$< 2.2 \times 10^{-16}$
SRP010040*	Zebrafish	ribotricer	ORFscore	0.658	0.562	$< 2.2 \times 10^{-16}$
SRP023492	Zebrafish	ribotricer	ORFscore	0.970	0.958	$< 2.2 \times 10^{-16}$
SRP034750*	Zebrafish	ribotricer	RiboCode	0.995	0.977	$< 2.2 \times 10^{-16}$
SRP032814	*C. albicans*	ribotricer	RiboCode	0.953	0.842	$< 2.2 \times 10^{-16}$
SRP062129	Chimp	ribotricer	ORFscore	0.918	0.883	$< 2.2 \times 10^{-16}$
SRP107240	*S. pombe*	ribotricer	RiboCode	0.972	0.939	$< 2.2 \times 10^{-16}$
SRP062129	Macaque	ribotricer	ORFscore	0.904	0.854	$< 2.2 \times 10^{-16}$

Table S9: **Best and second to best performing methods at F1 score metric for each dataset using dataset-specific cutoff.** F1 (B) and F1 (SB) denotes the F1 scores for the best and the second to best methods respectively. An asterisk (*) indicates that the dataset was used to learn the cutoffs by maximizing the F1 score. A # indicates the ribotricer phase score cutoff for the dataset is taken to be the median phase score difference between CDS annotated Ribo-seq and RNA-seq profiles.

SRP	Species	Best (B)	Second Best (SB)	F1 (B)	F1 (SB)
SRP018118*	Arabidopsis	ribotricer	RiboCode	0.937	0.848
SRP029587	Arabidopsis	ribotricer	RiboCode	0.645	0.176
SRP059391*	Arabidopsis	ribotricer	RiboCode	0.562	0.361
SRP087624	Arabidopsis	ribotricer	ORFscore	0.675	0.338
SRP108862	Arabidopsis	ribotricer	RiboCode	0.628	0.333
SRP000637	Baker's Yeast	RiboCode	ribotricer	0.680	0.503
SRP028552*	Baker's Yeast	ribotricer	RiboCode	0.964	0.859
SRP028614	Baker's Yeast	ribotricer	RiboCode	0.855	0.738
SRP033499	Baker's Yeast	RiboCode	RiboTaper	0.747	0.705
SRP075766*	Baker's Yeast	ribotricer	RiboTaper	0.951	0.877
SRP010374	*C. elegans*	ribotricer	RiboCode	0.799	0.517
SRP014427	*C. elegans*	ribotricer	RiboCode	0.826	0.776
SRP026198*	*C. elegans*	ribotricer	RiboCode	0.828	0.636
SRP056647	*C. elegans*	ribotricer	RiboCode	0.690	0.634
SRP028243*	Drosophila	ribotricer	RiboCode	0.693	0.562
SRP045475	Drosophila	ribotricer	RiboCode	0.561	0.391
SRP076919	Drosophila	ribotricer	RiboCode	0.667	0.125
SRP108999*	Drosophila	ribotricer	RiboCode	0.769	0.400
SRP010679*	Human	ribotricer	RiboCode	0.877	0.773
SRP029589	Human	ribotricer	RiboCode	0.651	0.599
SRP063852	Human	ribotricer	RiboCode	0.919	0.854
SRP098789*	Human	ribotricer	RiboCode	0.932	0.824
SRP102021	Human	ribotricer	RiboCode	0.890	0.835
SRP003554*	Mouse	RiboTaper	ribotricer	0.901	0.899
SRP062407	Mouse	RiboTaper	ribotricer	0.930	0.910
SRP078005	Mouse	ribotricer	RiboCode	0.951	0.901
SRP091889	Mouse	ribotricer	RiboCode	0.938	0.900
SRP115915*	Mouse	ribotricer	RiboCode	0.853	0.842
ERP007231*	Rat	ribotricer	RiboTaper	0.879	0.874
SRP045777	Rat	RiboCode	ribotricer	0.618	0.511
SRP056012*	Rat	ribotricer	RiboCode	0.787	0.786
SRP010040*	Zebrafish	ribotricer	RiboCode	0.670	0.377
SRP023492	Zebrafish	RiboCode	ribotricer	0.838	0.826
SRP034750*	Zebrafish	RiboCode	ribotricer	0.920	0.894
SRP032814#	*C.albicans*	ribotricer	RiboCode	0.883	0.752
SRP062129#	Chimp	ribotricer	RiboCode	0.865	0.436
SRP062129#	Macaque	ribotricer	RiboCode	0.842	0.635
SRP107240#	*S. pombe*	ribotricer	RiboCode	0.913	0.869

Table S10: **Best and second to best performing methods at F1 score metric for each dataset-specific cutoff.** F1 (B) and F1 (SB) denotes the F1 scores for the best and the second to best methods respectively. The cutoff was learned independently for each dataset as the median difference between Ribo-seq and RNA-seq phase scores over protein coding ORFs.

SRP	Species	Best (B)	Second Best (SB)	F1 (B)	F1 (SB)
SRP018118	Arabidopsis	ribotricer	RiboCode	0.920	0.848
SRP029587	Arabidopsis	ribotricer	RiboCode	0.846	0.176
SRP059391	Arabidopsis	ribotricer	RiboCode	0.678	0.361
SRP087624	Arabidopsis	ribotricer	ORFscore	0.671	0.338
SRP108862	Arabidopsis	ribotricer	RiboCode	0.695	0.333
SRP000637	Baker's Yeast	ribotricer	RiboCode	0.850	0.680
SRP028552	Baker's Yeast	ribotricer	RiboCode	0.928	0.859
SRP028614	Baker's Yeast	ribotricer	RiboCode	0.923	0.738
SRP033499	Baker's Yeast	ribotricer	RiboCode	0.904	0.747
SRP075766	Baker's Yeast	ribotricer	RiboTaper	0.935	0.877
SRP010374	*C.elegans*	ribotricer	RiboCode	0.798	0.517
SRP014427	*C.elegans*	ribotricer	RiboCode	0.868	0.776
SRP026198	*C.elegans*	ribotricer	RiboCode	0.846	0.636
SRP056647	*C.elegans*	ribotricer	RiboCode	0.716	0.634
SRP028243	Drosophila	ribotricer	RiboCode	0.679	0.562
SRP045475	Drosophila	ribotricer	RiboCode	0.667	0.391
SRP076919	Drosophila	ribotricer	RiboCode	0.667	0.125
SRP108999	Drosophila	ribotricer	RiboCode	0.818	0.400
SRP010679	Human	ribotricer	RiboCode	0.878	0.773
SRP029589	Human	ribotricer	RiboCode	0.765	0.599
SRP063852	Human	ribotricer	RiboCode	0.919	0.854
SRP098789	Human	ribotricer	RiboCode	0.922	0.824
SRP102021	Human	ribotricer	RiboCode	0.900	0.835
SRP003554	Mouse	ribotricer	RiboTaper	0.919	0.901
SRP062407	Mouse	ribotricer	RiboTaper	0.936	0.930
SRP078005	Mouse	ribotricer	RiboCode	0.944	0.901
SRP091889	Mouse	ribotricer	RiboCode	0.924	0.900
SRP115915	Mouse	ribotricer	RiboCode	0.863	0.842
ERP007231	Rat	RiboTaper	RiboCode	0.874	0.867
SRP045777	Rat	ribotricer	RiboCode	0.722	0.618
SRP056012	Rat	RiboCode	ribotricer	0.786	0.738
SRP010040	Zebrafish	ribotricer	RiboCode	0.668	0.377
SRP023492	Zebrafish	ribotricer	RiboCode	0.918	0.838
SRP034750	Zebrafish	ribotricer	RiboCode	0.937	0.920

Table S11: **Ribotricer's performance at F1 score when considering species-specific or dataset-specific cutoff** F1 (SS) and F1 (DS) denotes the F1 scores for ribotricer when using species-specif and dataset-specific cutoffs respectively. Ribo-RNA indicates the median difference between phase score of protein coding ORFs in Ribo- and RNA-seq samples. 'sampled' indicates the median was calculated using 30% of protein coding ORFs per dataset with resampling ($n_{\text{bootstraps}} = 10000$) while 'all' indicates the median was calculated using the complete list of protein coding ORFs.

SRP	species	F1 (SS)	F1 (DS)	Ribo-RNA (sampled)	Ribo-RNA (all)
SRP018118	Arabidopsis	0.937	0.920	0.455	0.447
SRP029587	Arabidopsis	0.645	0.846	0.206	0.191
SRP059391	Arabidopsis	0.562	0.678	0.109	0.104
SRP087624	Arabidopsis	0.675	0.671	0.233	0.145
SRP108862	Arabidopsis	0.628	0.695	0.181	0.154
SRP000637	Baker's Yeast	0.503	0.850	0.186	0.179
SRP028552	Baker's Yeast	0.964	0.928	0.383	0.382
SRP028614	Baker's Yeast	0.855	0.923	0.267	0.263
SRP033499	Baker's Yeast	0.573	0.904	0.204	0.194
SRP075766	Baker's Yeast	0.951	0.935	0.694	0.671
SRP010374	*C.elegans*	0.799	0.798	0.224	0.222
SRP014427	*C.elegans*	0.826	0.868	0.343	0.334
SRP026198	*C.elegans*	0.828	0.846	0.322	0.316
SRP056647	*C.elegans*	0.690	0.716	0.141	0.135
SRP028243	Drosophila	0.693	0.679	0.109	0.098
SRP045475	Drosophila	0.561	0.667	-0.019	-0.020
SRP076919	Drosophila	0.667	0.667	-0.025	-0.034
SRP108999	Drosophila	0.769	0.818	0.363	0.360
SRP010679	Human	0.878	0.878	0.421	0.404
SRP029589	Human	0.651	0.765	0.234	0.223
SRP063852	Human	0.919	0.919	0.522	0.498
SRP098789	Human	0.932	0.922	0.526	0.514
SRP102021	Human	0.891	0.900	0.427	0.417
SRP003554	Mouse	0.900	0.919	0.542	0.526
SRP062407	Mouse	0.910	0.936	0.588	0.568
SRP078005	Mouse	0.951	0.944	0.603	0.591
SRP091889	Mouse	0.939	0.924	0.509	0.497
SRP115915	Mouse	0.854	0.863	0.372	0.361
ERP007231	Rat	0.879	0.863	0.403	0.388
SRP045777	Rat	0.511	0.722	0.176	0.173
SRP056012	Rat	0.787	0.738	0.264	0.247
SRP010040	Zebrafish	0.670	0.668	0.136	0.108
SRP023942	Zebrafish	0.826	0.918	0.512	0.502
SRP034750	Zebrafish	0.894	0.937	0.660	0.649

10 Supplementary figures

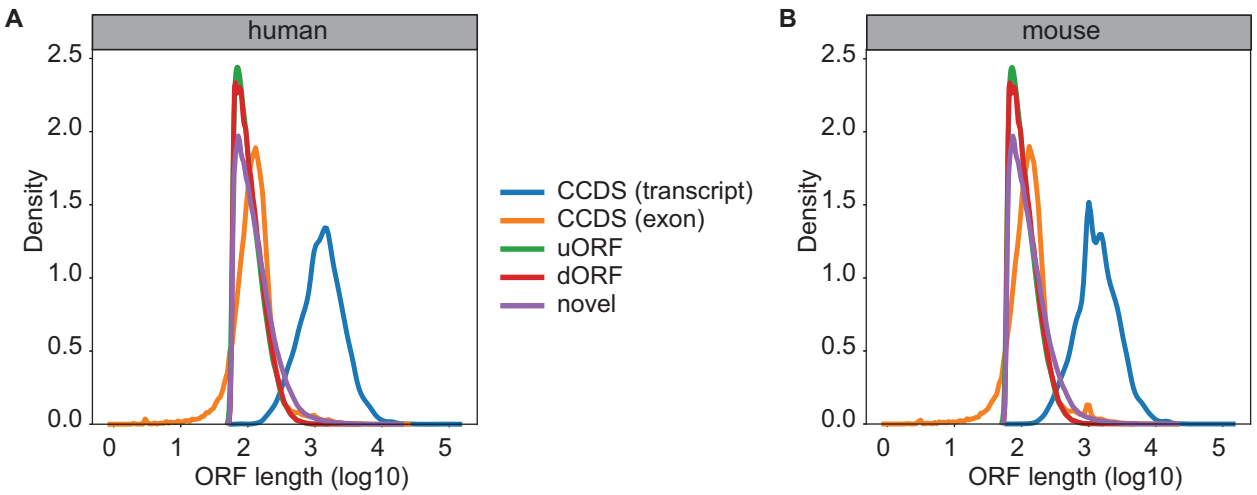

Figure S1: **Length distribution of candidate ORFs for human and mouse.** The length distribution of uORFs, dORFs, and novel ORFs predicted from presumably non-coding genes compared with the CCDS exon and CCDS transcript lengths (CCDS = Canonical Coding Sequence; uORF = upstream ORF in 5' UTR; dORF = downstream ORF in 3' UTR; novel = candidate ORFs in annotated non-coding genes.)

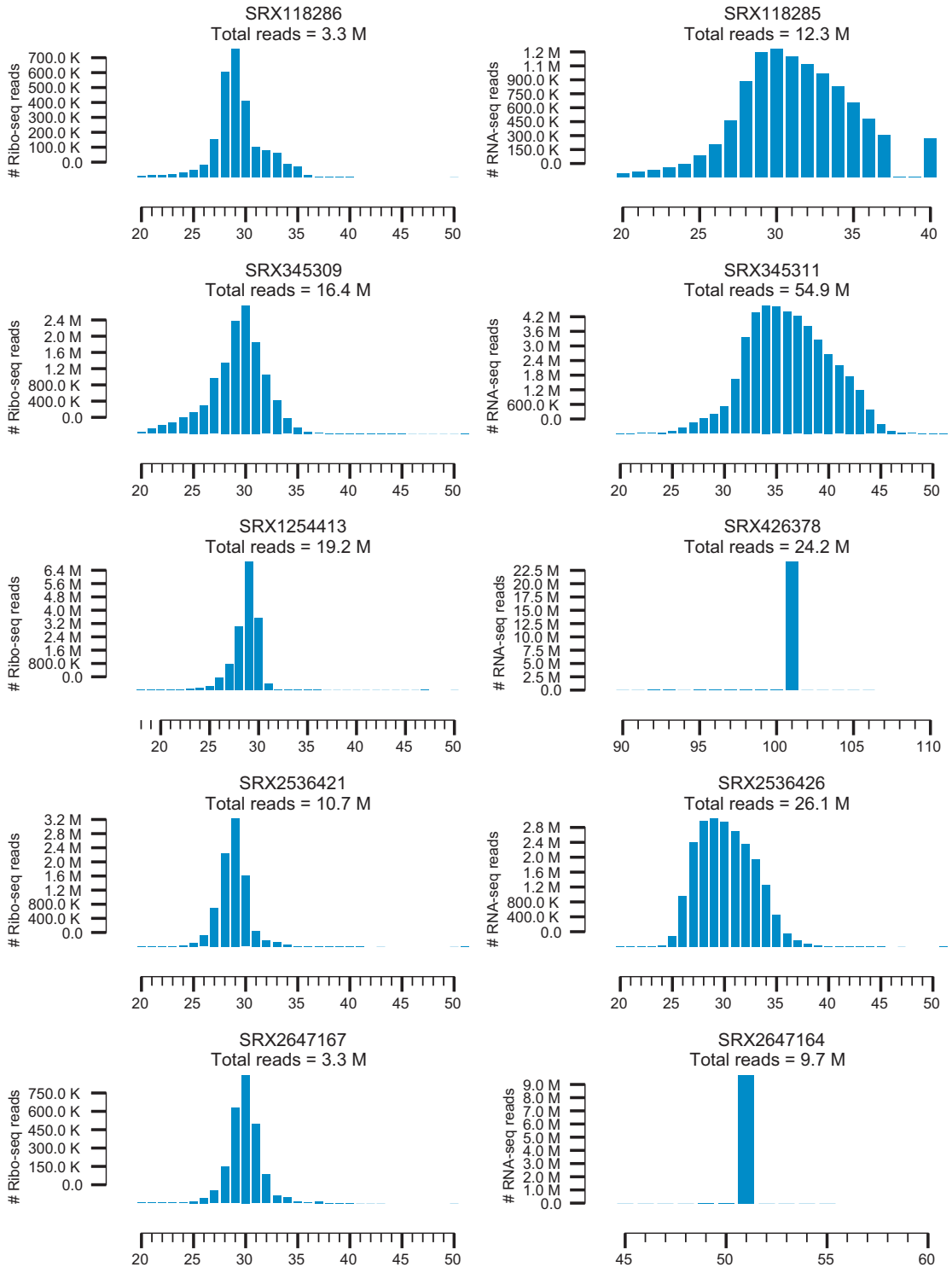

Figure S2: **Read length distribution of Ribo-seq and RNA-seq samples from human datasets.** SRA sample accession and total uniquely mapping reads are shown in individual subplots.

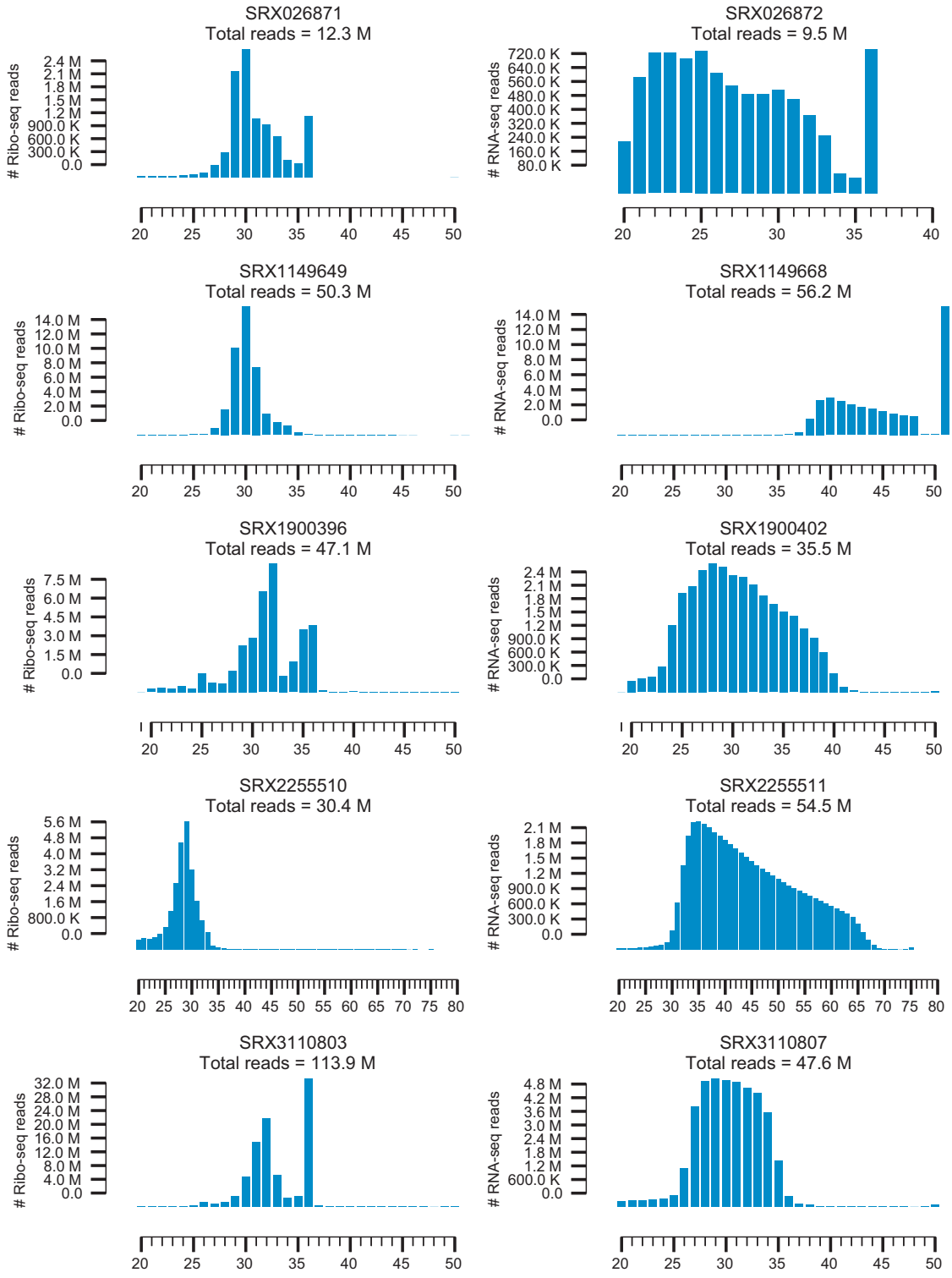

Figure S3: **Read length distribution of Ribo-seq and RNA-seq samples from mouse datasets.** SRA sample accession and total uniquely mapping reads are shown in individual subplots.

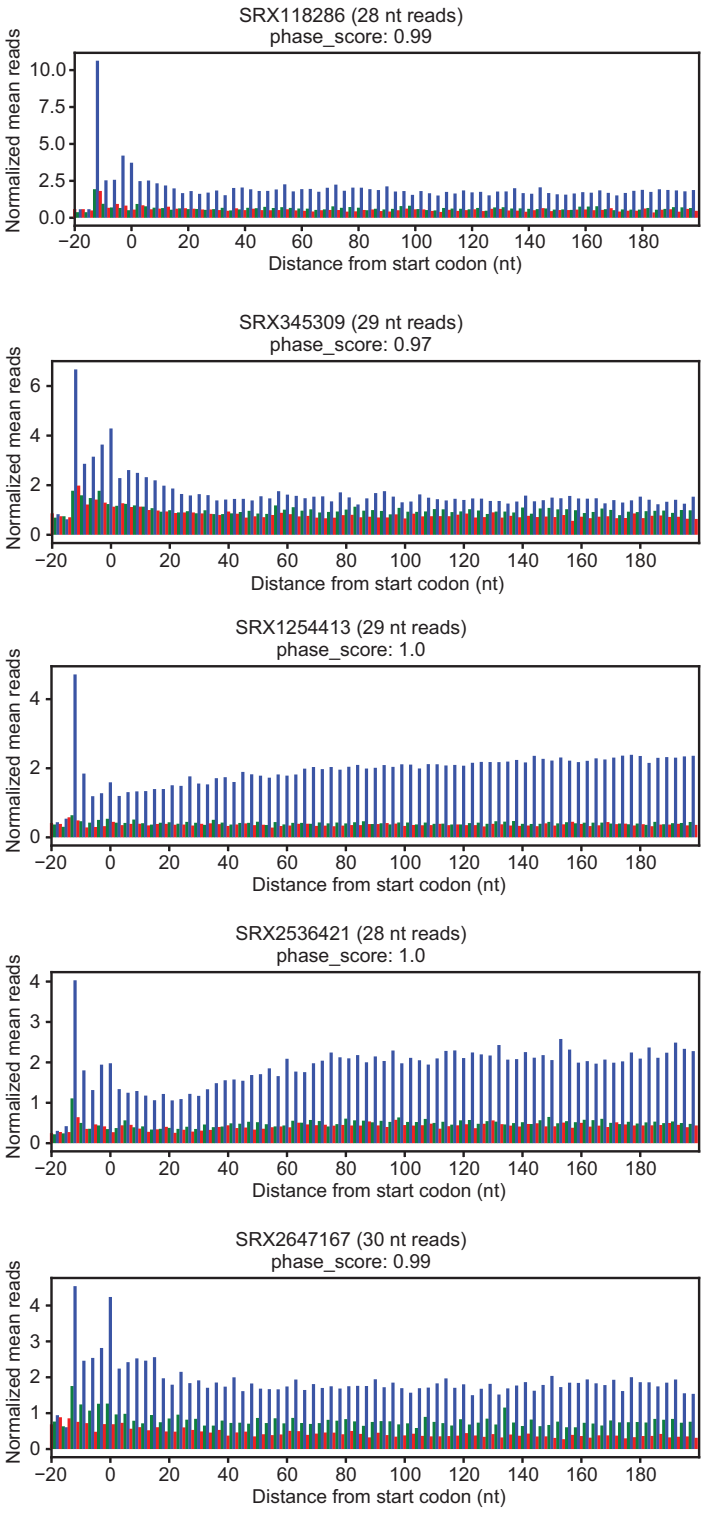

Figure S4: **Metagene plots for representative read lengths for human Ribo-seq samples.** SRA sample accession, read length and phase score are shown in individual subplots.

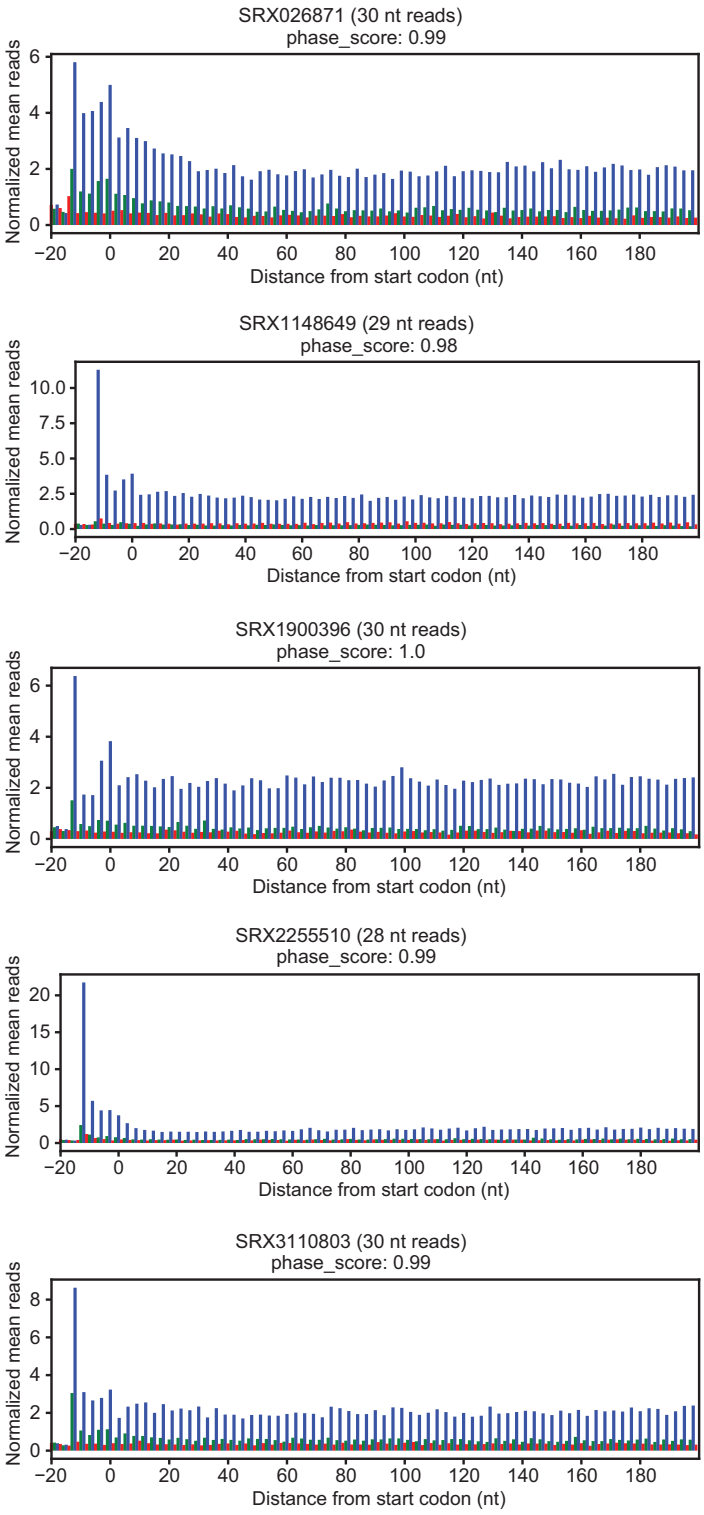

Figure S5: **Metagene plots for representative read lengths for mouse Ribo-seq samples.** SRA sample accession, read length and phase score are shown in individual subplots.

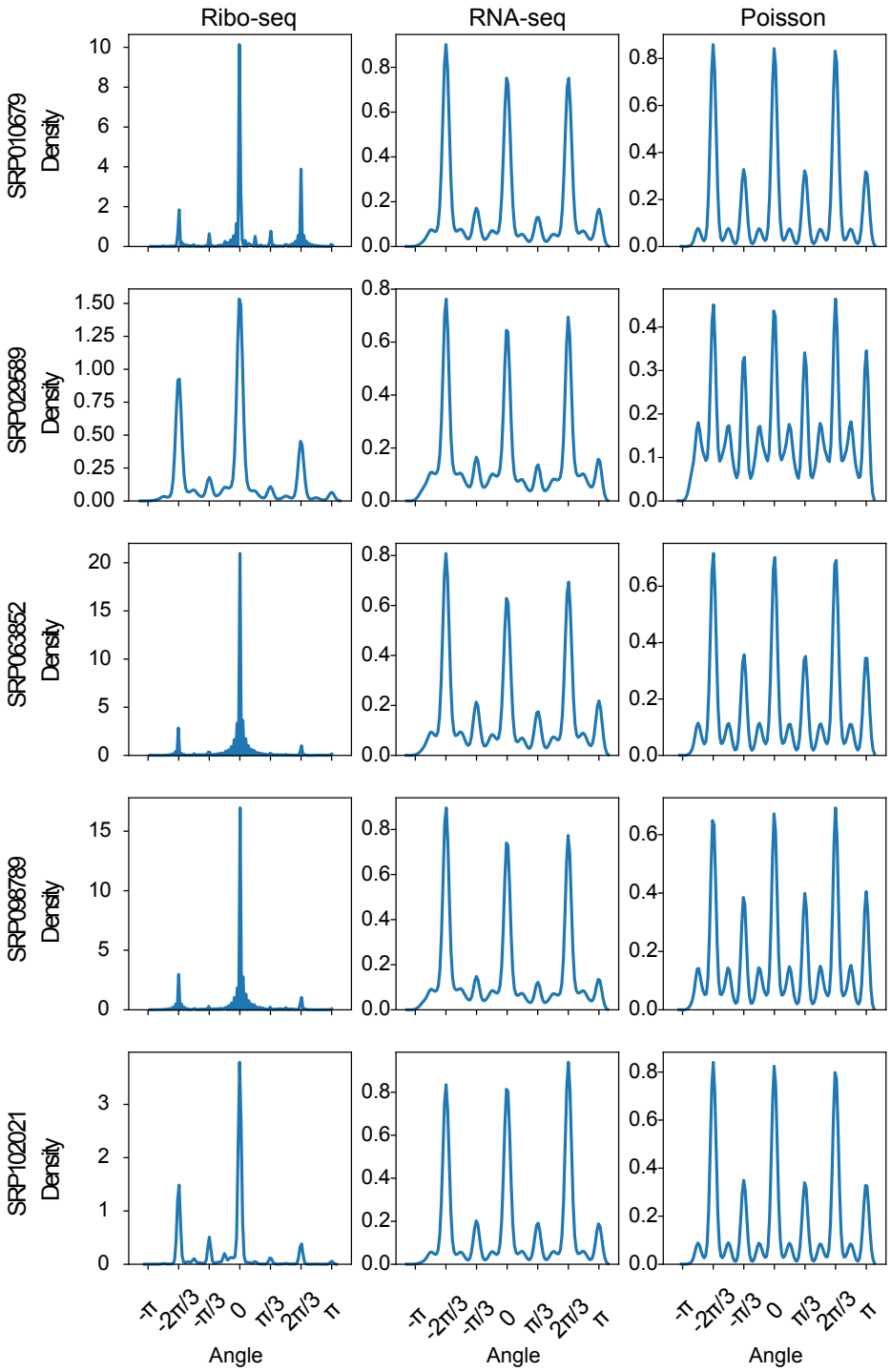

Figure S6: **Distribution of the resulting vector angles for datasets in human.** Angles are formed by projecting the CCDS 3D codon profiles to 2D unit vectors. The left sub-panel indicates the distribution for Ribo-seq sample; the center sub-panel shows the distribution for its corresponding RNA-seq sample; the right sub-panel shows the distribution of angles resulting from a RNA-seq profile simulated from a Poisson distribution with the mean parameter estimated from the RNA-seq data.

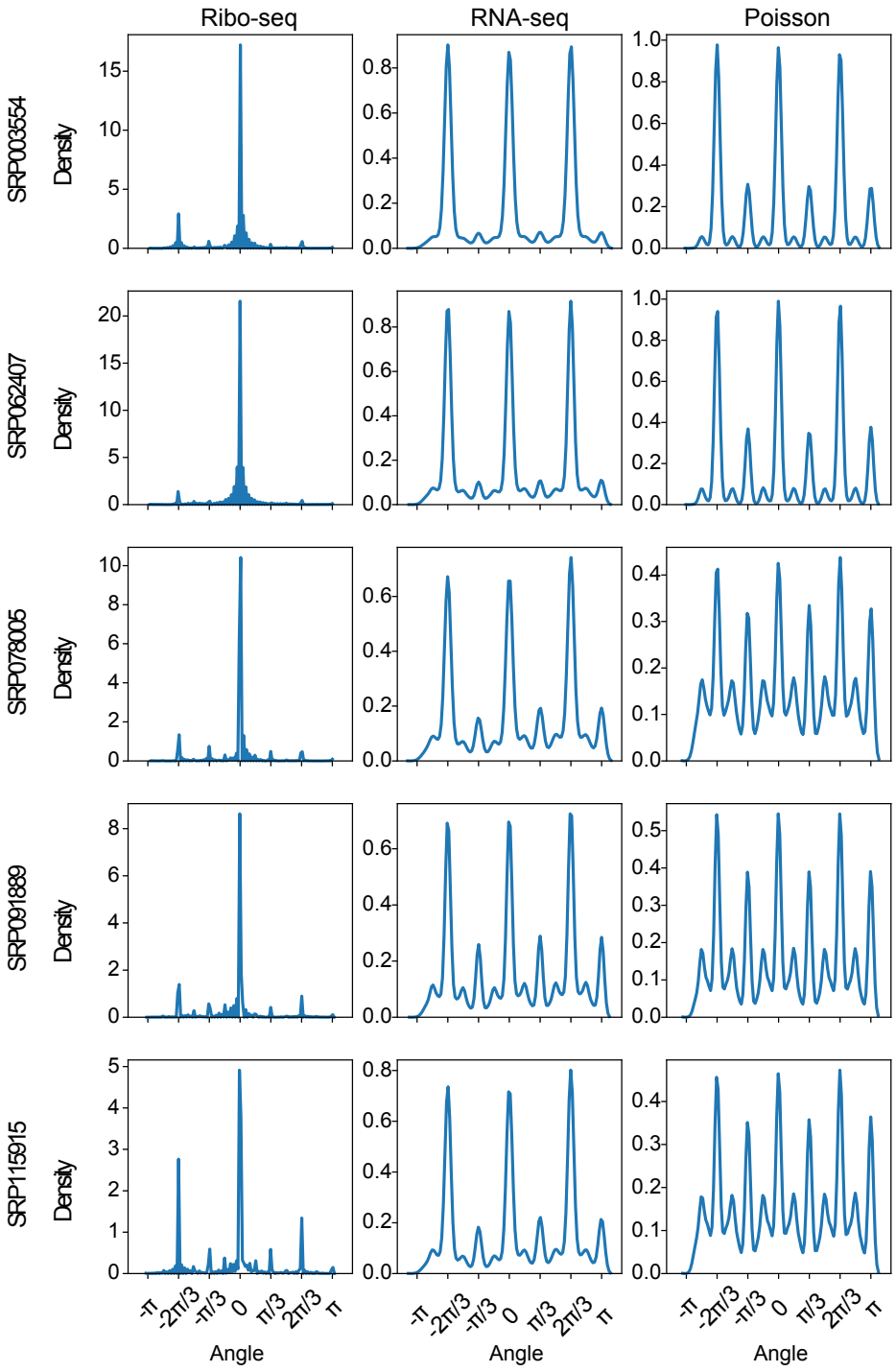

Figure S7: **Distribution of the resulting vector angles for datasets in mouse.** Angles are formed by projecting the CCDS 3D codon profiles to 2D unit vectors. The left sub-panel indicates the distribution for Ribo-seq sample; the center sub-panel shows the distribution for its corresponding RNA-seq sample; the right sub-panel shows the distribution of angles resulting from a RNA-seq profile simulated from a Poisson distribution with the mean parameter estimated from the RNA-seq data.

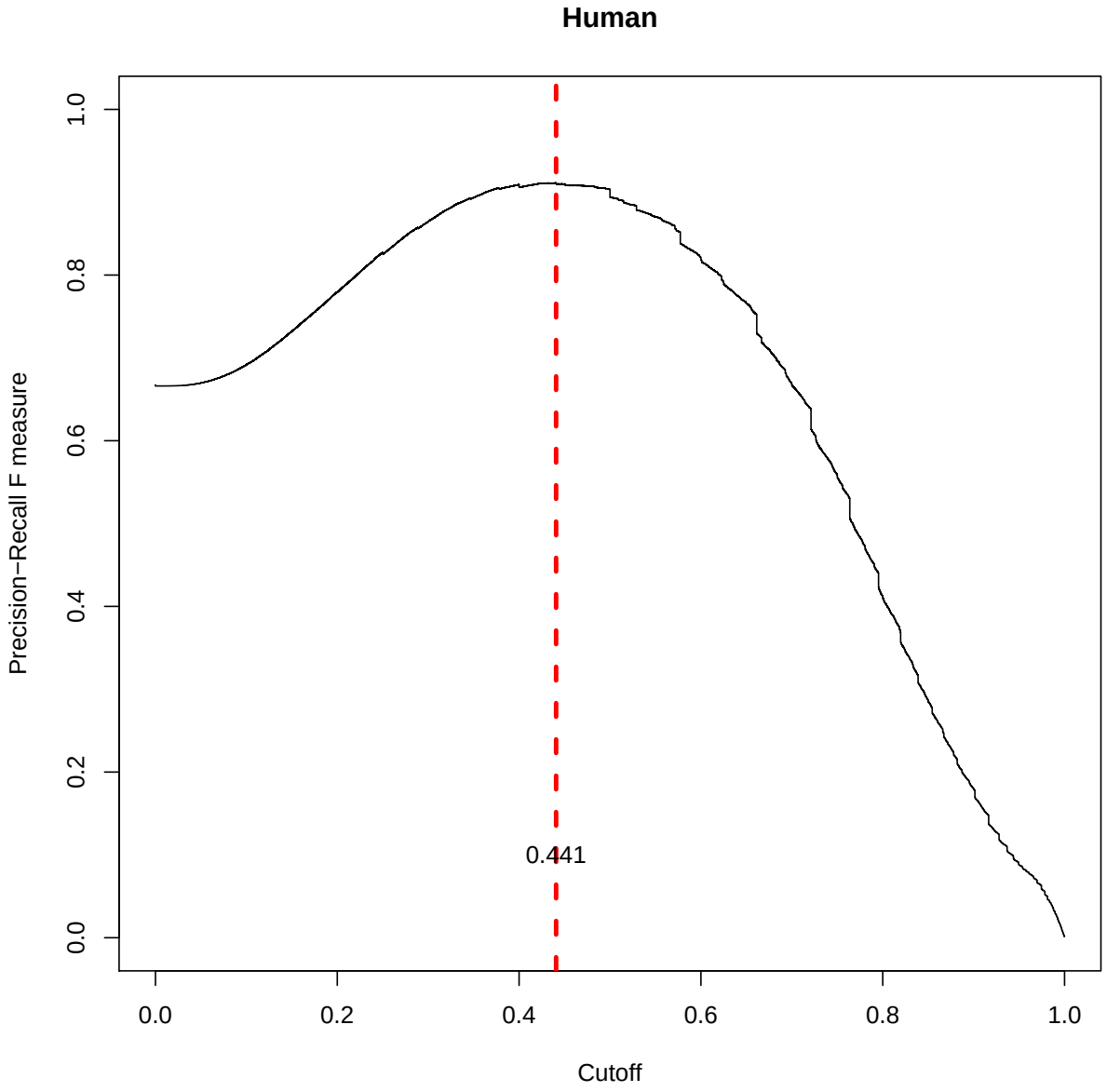

Figure S8: **Learning the cutoff for phase scores for human datasets.** The optimum cutoff for distinguishing actively translating regions from non-active translation was learned by maximizing the F1 score. The profiles from expressed CCDS exons in Ribo-seq data were treated as positives and corresponding profiles from RNA-seq were treated as negatives. Two datasets in human (SRA accession: SRP010679, SRP098789) were used for learning this cutoff.

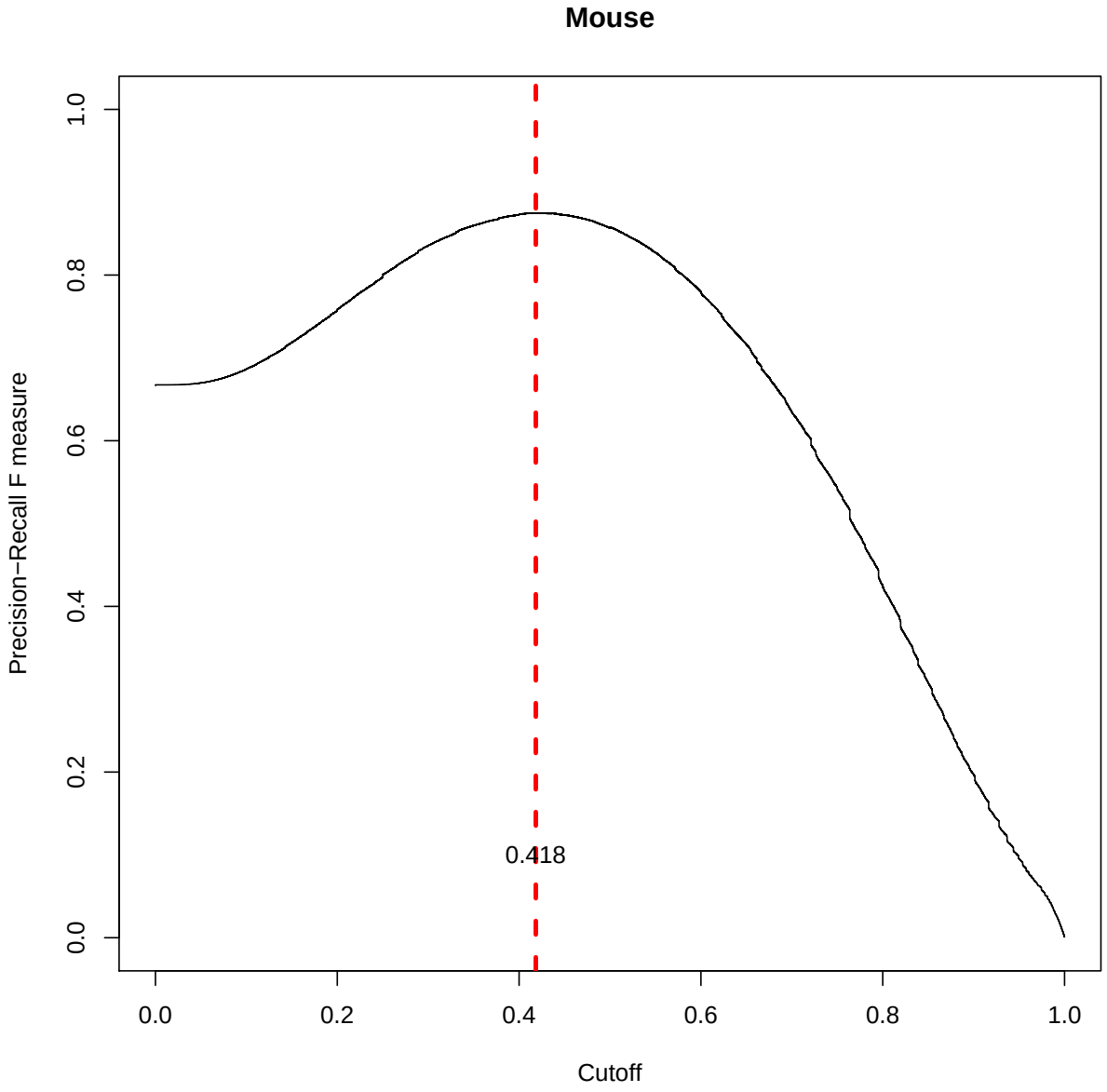

Figure S9: **Learning the cutoff for phase scores for mouse datasets.** The optimum cutoff for distinguishing actively translating regions from non-active translation was learned by maximizing the F1 score. The profiles from expressed CCDS exons in Ribo-seq data were treated as positives and corresponding profiles from RNA-seq were treated as negatives. Two datasets in mouse (SRA accession: SRP003554, and SRP115915)were used for learning this cutoff.

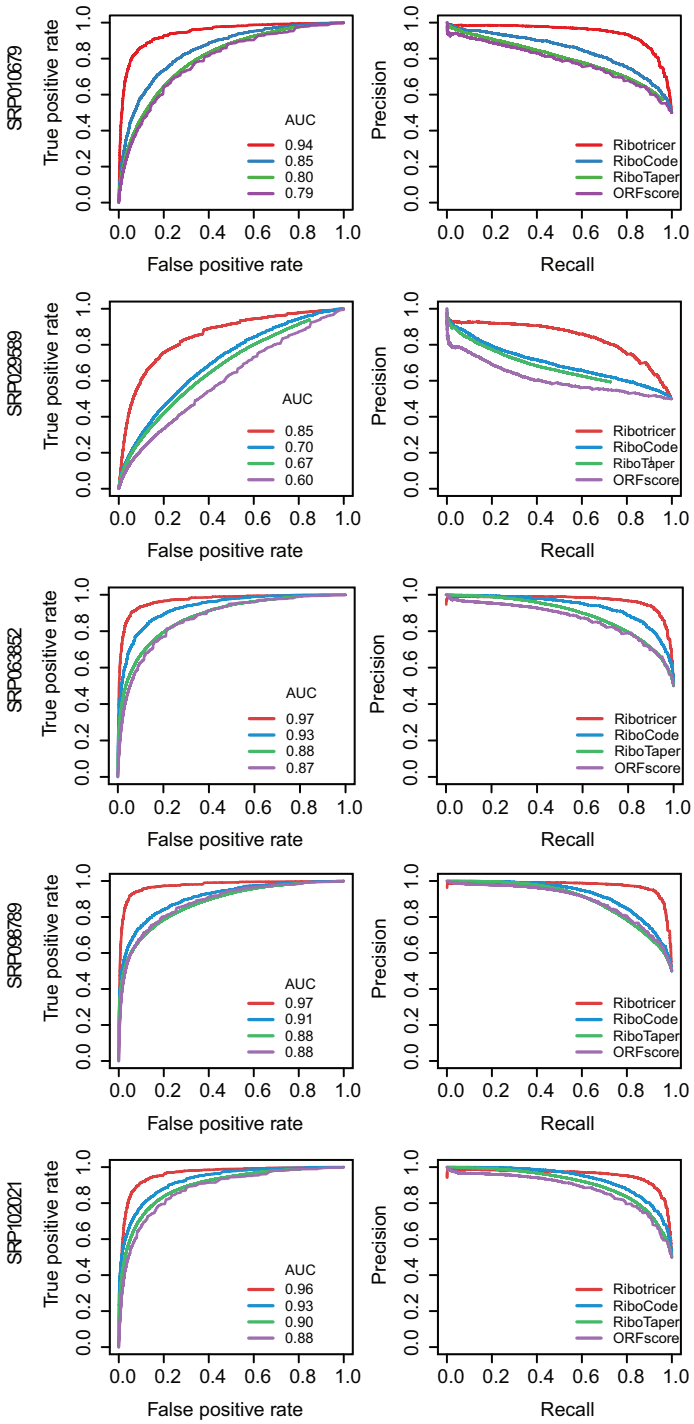

Figure S10: **ROC plots and Precision-Recall plots for human datasets for exon level classification.** Performance of ribotricer for detecting translating ORFs at exon level is compared with RiboCode, RiboTaper and ORFscore. The profiles of expressed CCDS exons in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

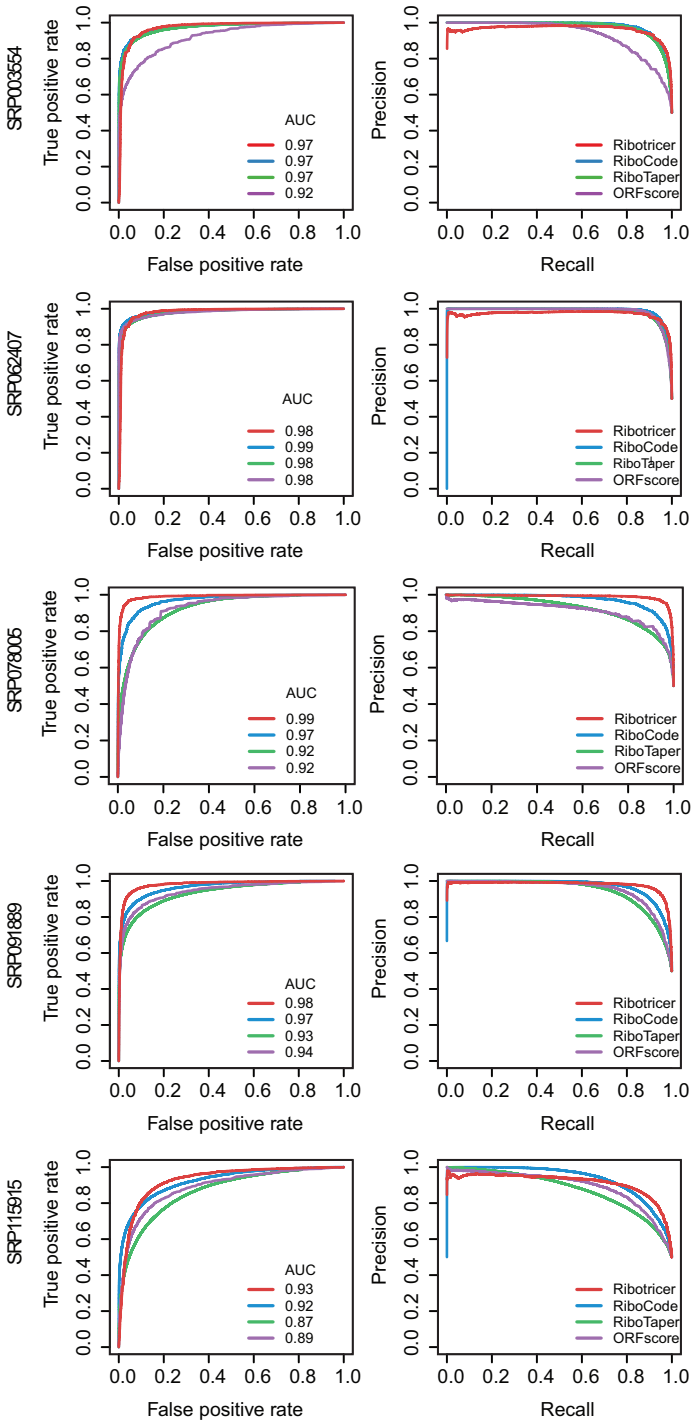

Figure S11: **ROC plots and Precision-Recall plots for mouse datasets for exon level classification.** Performance of ribotricer for detecting translating ORFs at exon level is compared with RiboCode, RiboTaper and ORFscore. The profiles of expressed CCDS exons in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

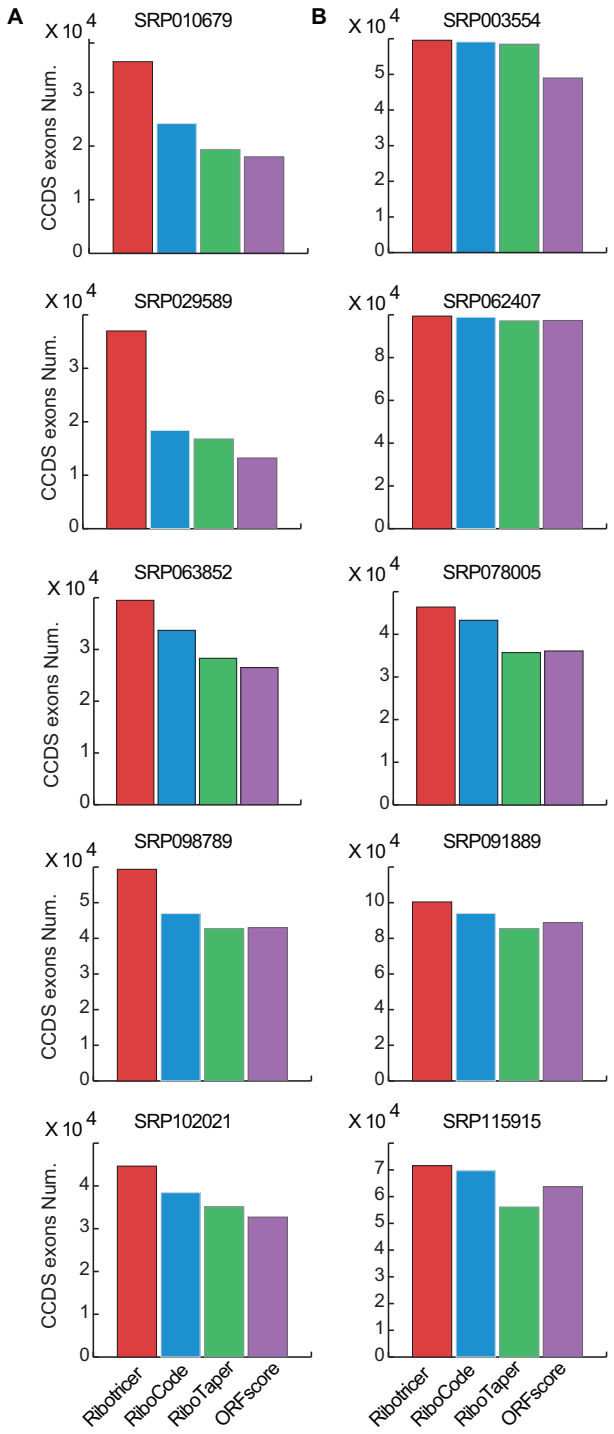

Figure S12: **Number of translating exons recovered when controlling the false positive rate to be the same.** Performance of ribotricer is compared with RiboCode, RiboTaper, and ORFscore when the false positive rate is controlled to be 0.1. The number of truly translating exons are shown for both human (A) and mouse (B) datasets. The profiles of expressed CCDS exons in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

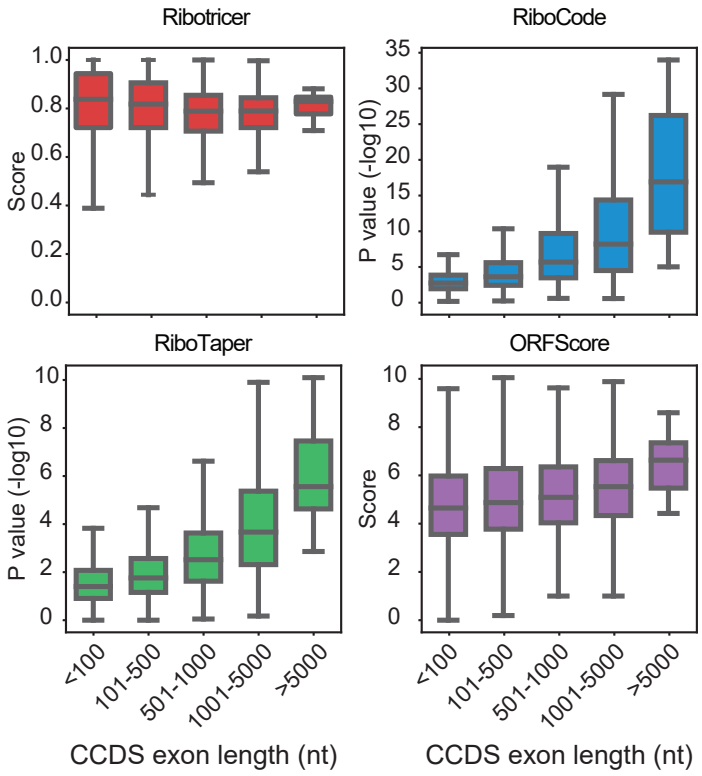

Figure S13: **Effect of ORF length on output scores.** Distribution of scores generated by ribotricer and ORFscore, and the P-values generated by RiboCode and RiboTaper over different CCDS exon lengths.

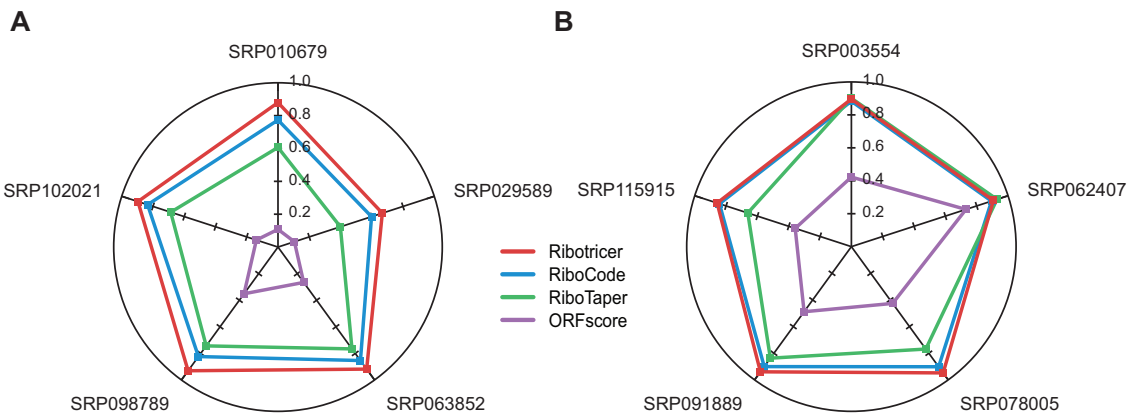

Figure S14: **Comparison of F1 score (exon level) of ribotricer with RiboCode, RiboTaper, and ORFscore.** Performance of ribotricer is compared with RiboCode, RiboTaper, and ORFscore in terms of F1 score when the default threshold score is used for each tool. Results are shown for human (A) and mouse (B) datasets. The profiles of expressed CCDS exons in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

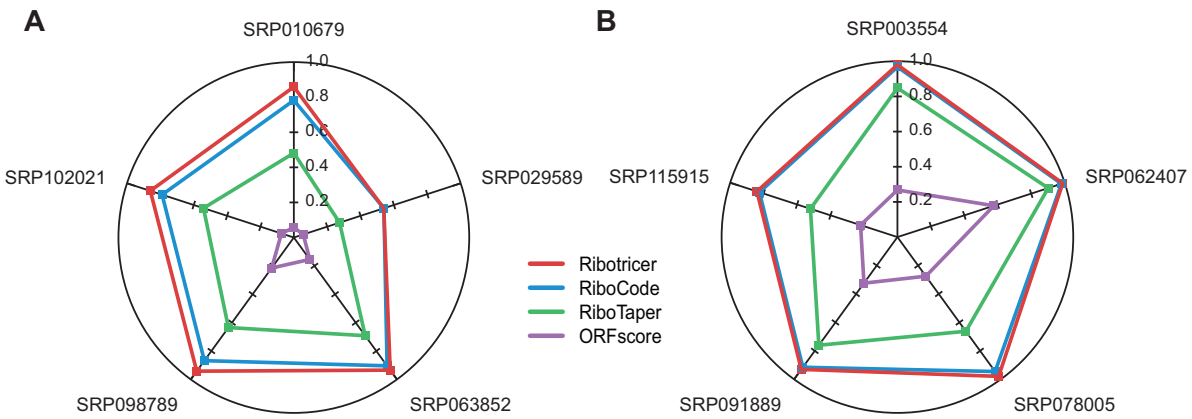

Figure S15: **Comparison of sensitivity (exon level) of ribotricer with RiboCode, RiboTaper, and ORFscore.** Performance of ribotricer is compared with RiboCode, RiboTaper, and ORFscore in terms of sensitivity when the default threshold score is used for each tool. Results are shown for human (A) and mouse (B) datasets. The profiles of expressed CCDS exons in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

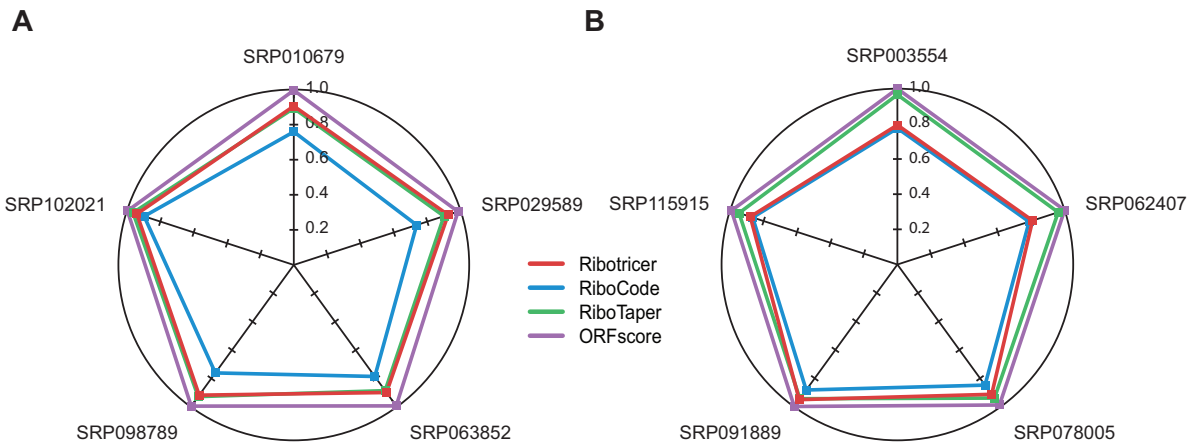

Figure S16: **Comparison of specificity (exon level) of ribotricer with RiboCode, RiboTaper, and ORFscore.** Performance of ribotricer is compared with RiboCode, RiboTaper, and ORFscore in terms of specificity when the default threshold score is used for each tool. Results are shown for human (A) and mouse (B) datasets. The profiles of expressed CCDS exons in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

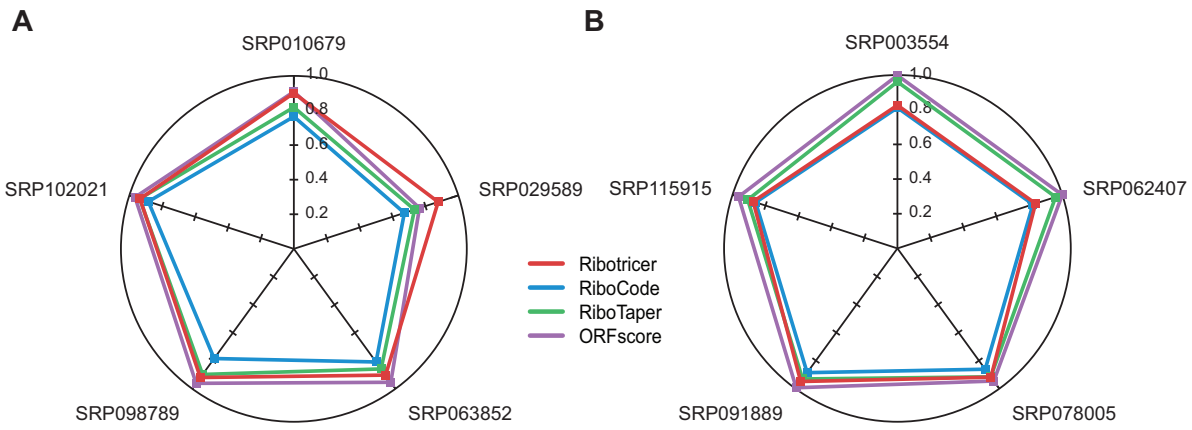

Figure S17: **Comparison of precision (exon level) of ribotricer with RiboCode, RiboTaper, and ORFscore.** Performance of ribotricer is compared with RiboCode, RiboTaper, and ORFscore in terms of precision when the default threshold score is used for each tool. Results are shown for human (A) and mouse (B) datasets. The profiles of expressed CCDS exons in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

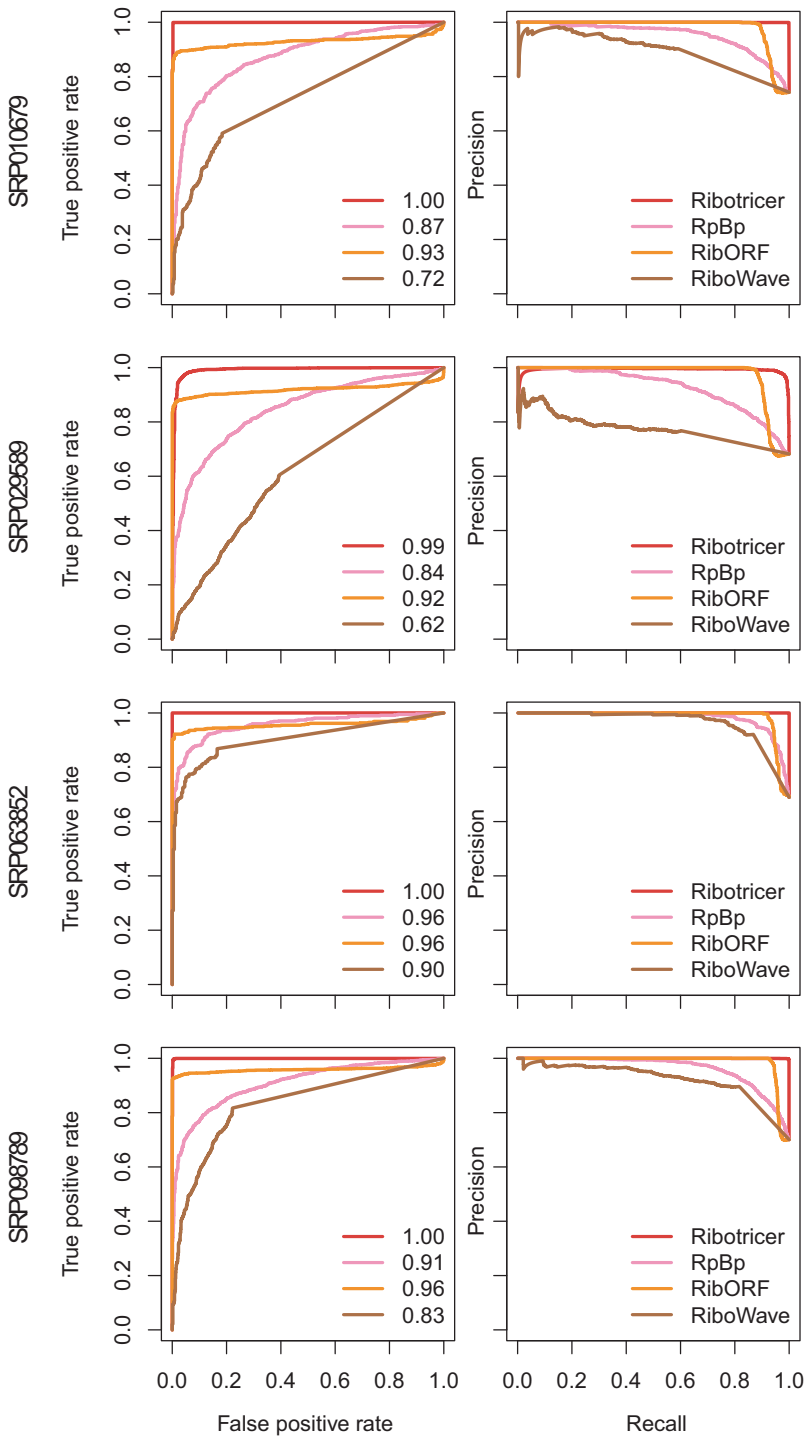

Figure S18: **ROC plots and Precision-Recall plots on transcript level for human datasets.** Performance of ribotricer for detecting translating ORFs at transcript level is compared with RpBp, ribORF and RiboWave. The profiles of expressed CCDS transcripts in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

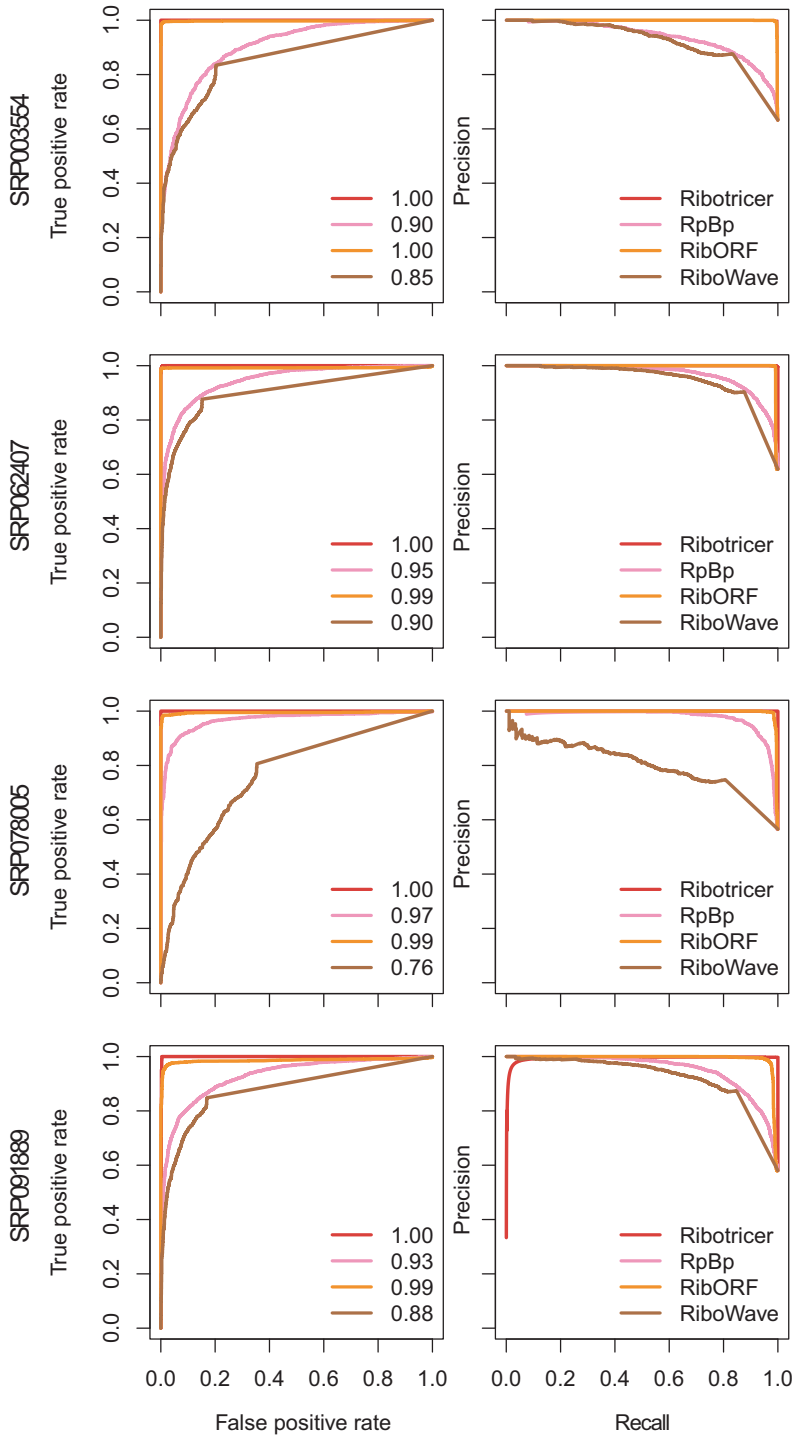

Figure S19: **ROC plots and Precision-Recall plots on transcript level for mouse datasets.** Performance of ribotricer for detecting translating ORFs at transcript level is compared with RpBp, ribORF and RiboWave. The profiles of expressed CCDS transcripts in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

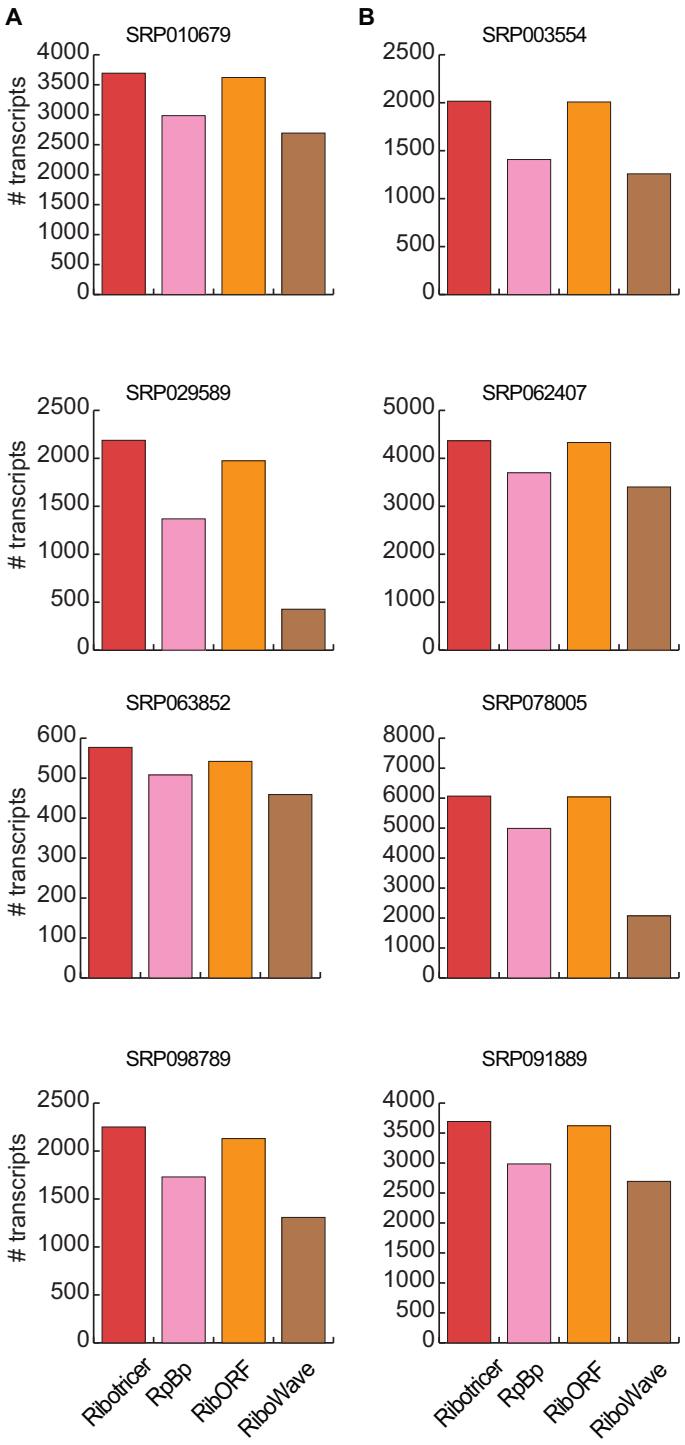

Figure S20: **Number of translating transcripts recovered when controlling the false positive rate to be the same.** Performance of ribotricer is compared with RpBp, ribORF, and RiboWave when the false positive rate is controlled to be 0.1. The number of truly translating transcripts are shown for both human (A) and mouse (B) datasets. The profiles of expressed CCDS transcripts in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

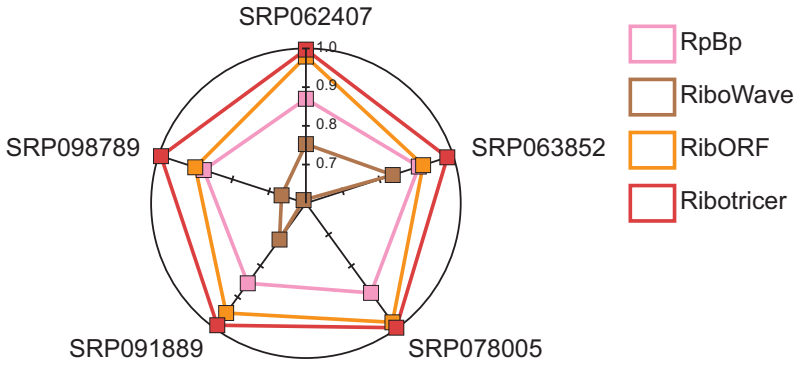

Figure S21: **Performance of different methods on transcript level measured using F1 score.** Performance of ribotricer is compared with RpBp, ribORF, and RiboWave in terms of F1 score when the default threshold score is used for each tool. The profiles of expressed CCDS transcripts in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

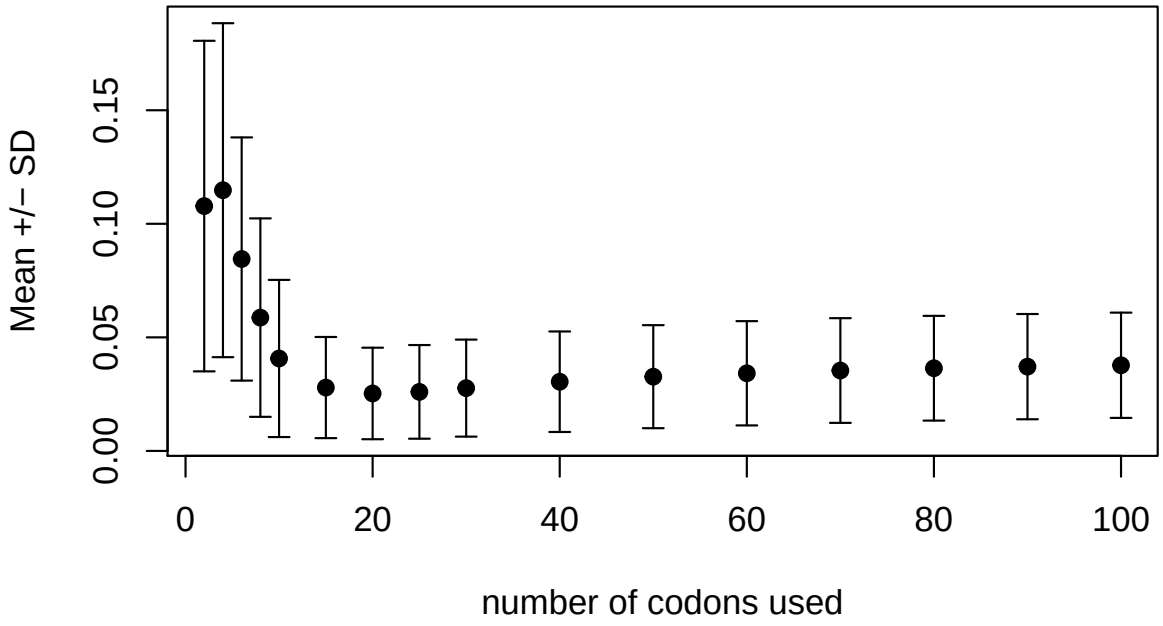

Figure S22: **Effect of number of codons on ribotricer's phase score in human dataset.** Mean absolute difference and standard deviation between original phase score using all codons and the one with down-sampled number of codons. The plot was generated on human dataset (SRA accession: SRP063852) using 5K genes with at least 50% valid codons, the down-sampling is repeated 100 times for each gene. Similar trend is observed for other human datasets.

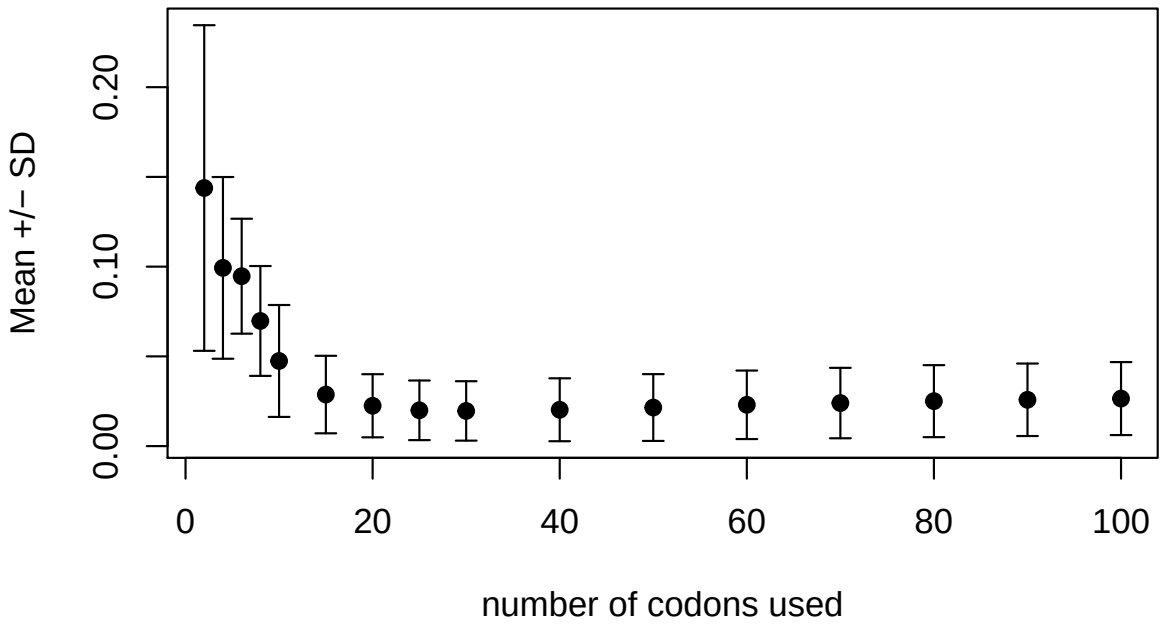

Figure S23: **Effect of number of codons on ribotricer's phase score in mouse dataset.** Mean absolute difference and standard deviation between original phase score using all codons and the one with down-sampled number of codons. The plot was generated on mouse dataset (SRA accession: SRP003554) using 5K genes with at least 50% valid codons, the down-sampling is repeated 100 times for each gene. Similar trend is observed for other mouse datasets.

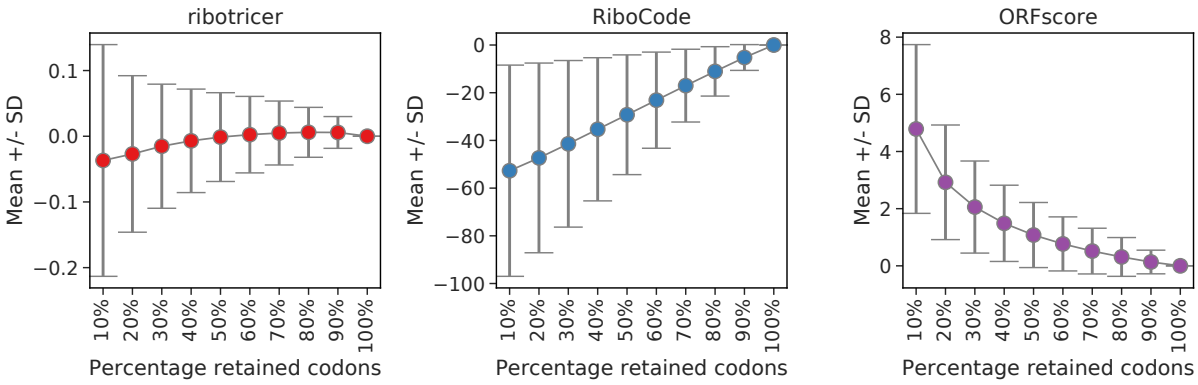

Figure S24: **Effect of truncating an ORF on ribotricer's phase score, RiboCode's p-values and ORF-score in human dataset.** Mean difference and standard deviation between original phase score using full length ORF and the ones after truncating it from the 3' end. The plot was generated on human dataset (SRA accession: SRP063852) using 5K genes with at least 50% valid codons and truncating it to have indicated percentage (X-axis) of codons.The differences between truncated and original profile for RiboCode are calculated on a $\log_{10}$ scale as it outputs p-values, while for both ribotricer and ORFscore, the differences are calculated on the same scale as the scores.

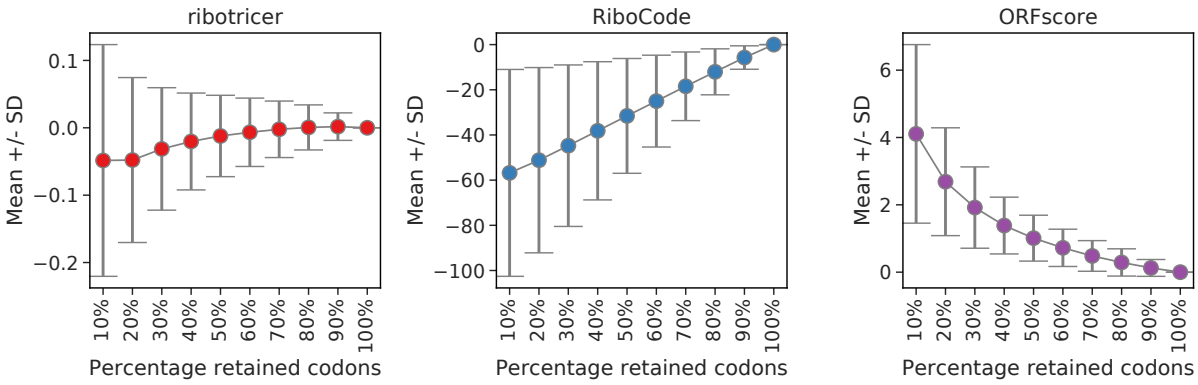

Figure S25: **Effect of truncating an ORF on ribotricer's phase score, RiboCode's p-values and ORF-score in mouse dataset.** Mean difference and standard deviation between original phase score using full length ORF and the ones after truncating it from the 3' end. The plot was generated on mouse dataset (SRA accession: SRP003554) using 5K genes with at least 50% valid codons and truncating it to have indicated percentage (X-axis) of codons. The differences between truncated and original profile for RiboCode are calculated on a $\log_{10}$ scale as it outputs p-values, while for both ribotricer and ORFscore, the differences are calculated on the same scale as the scores.

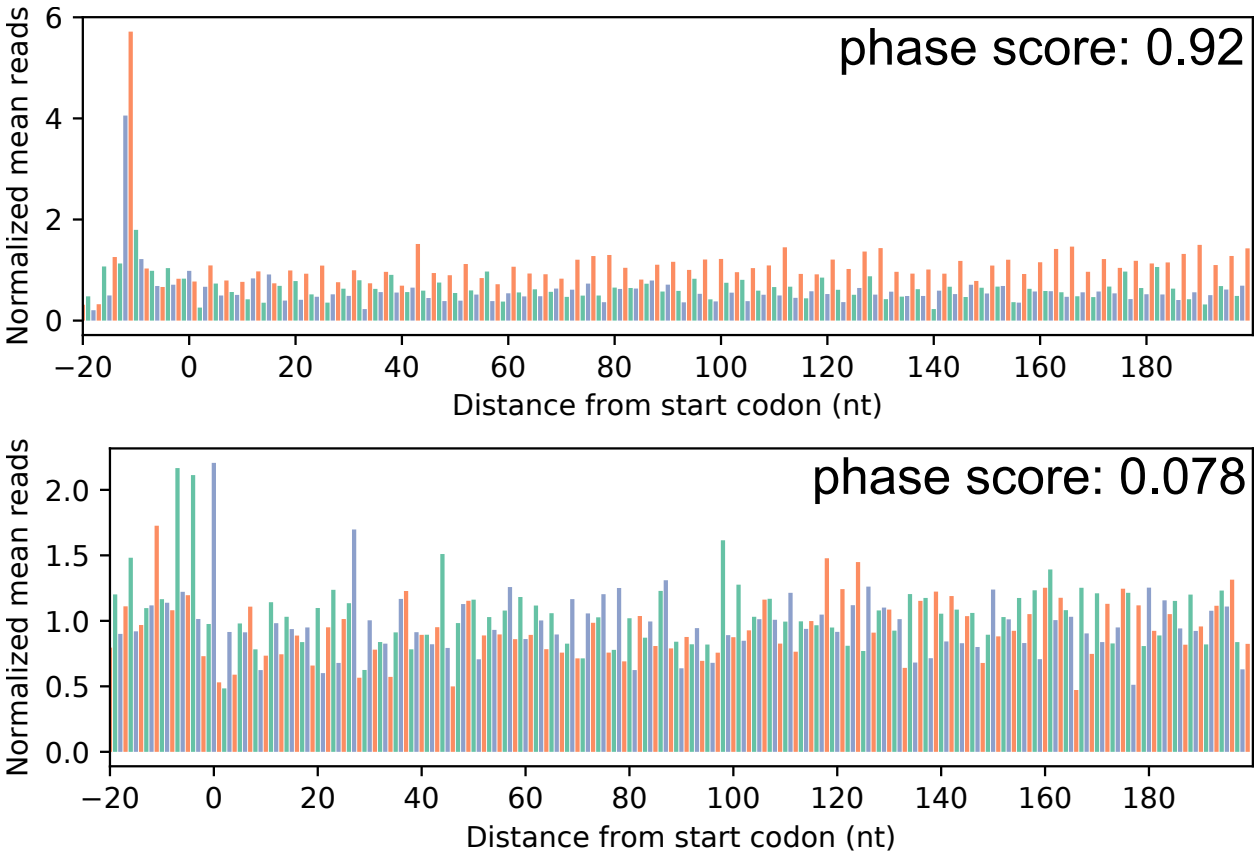

Figure S26: **Example of phase scores for an active and a non-active ORF.** Phase score generated by ribotricer for two different profiles.

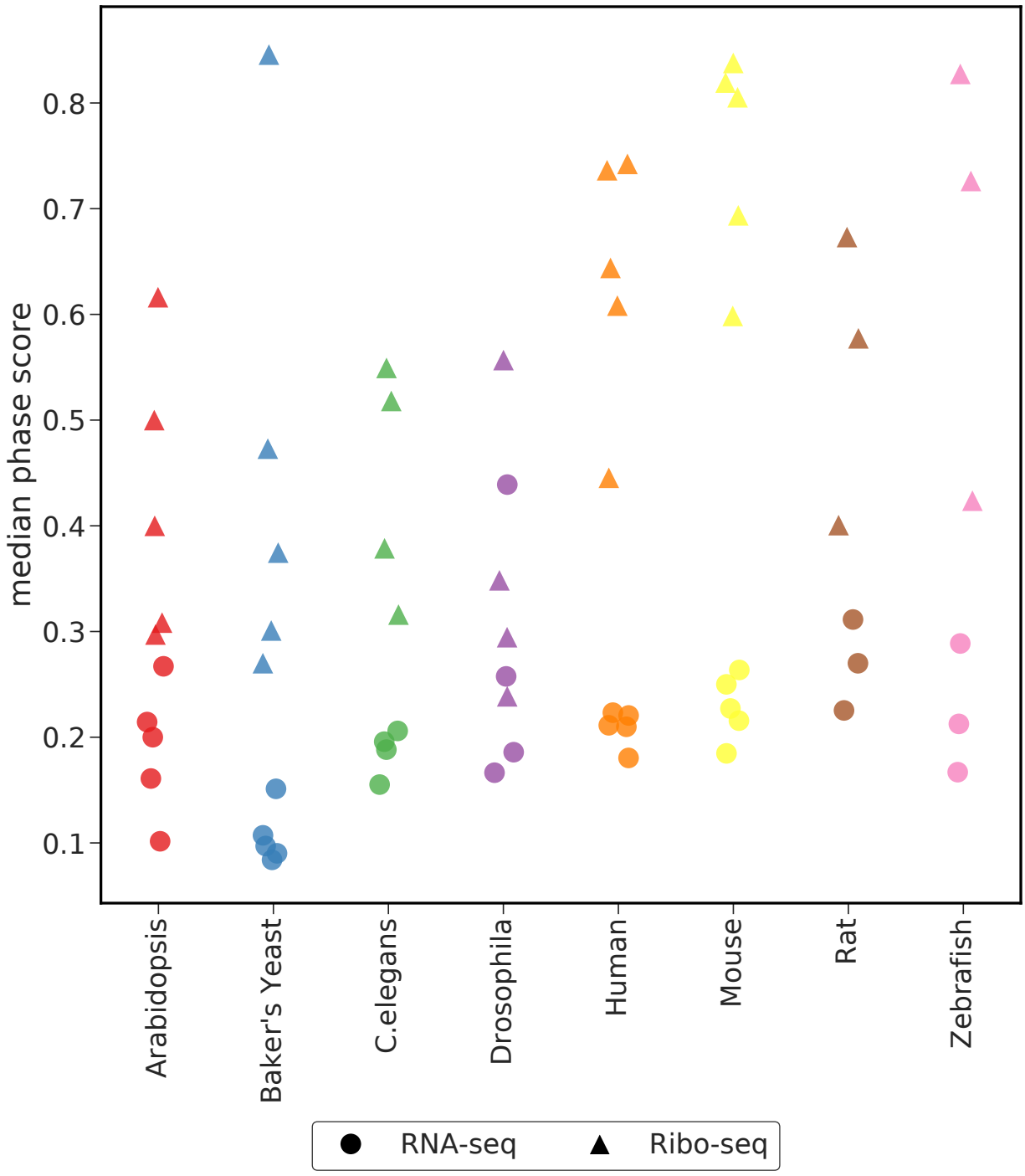

Figure S27: **Summarized median phase score for RNA-seq and Ribo-seq for all datasets.** For each dataset, the median phase score was calculated for all the candidate ORFs for both Ribo-seq and the corresponding RNA-seq sample.

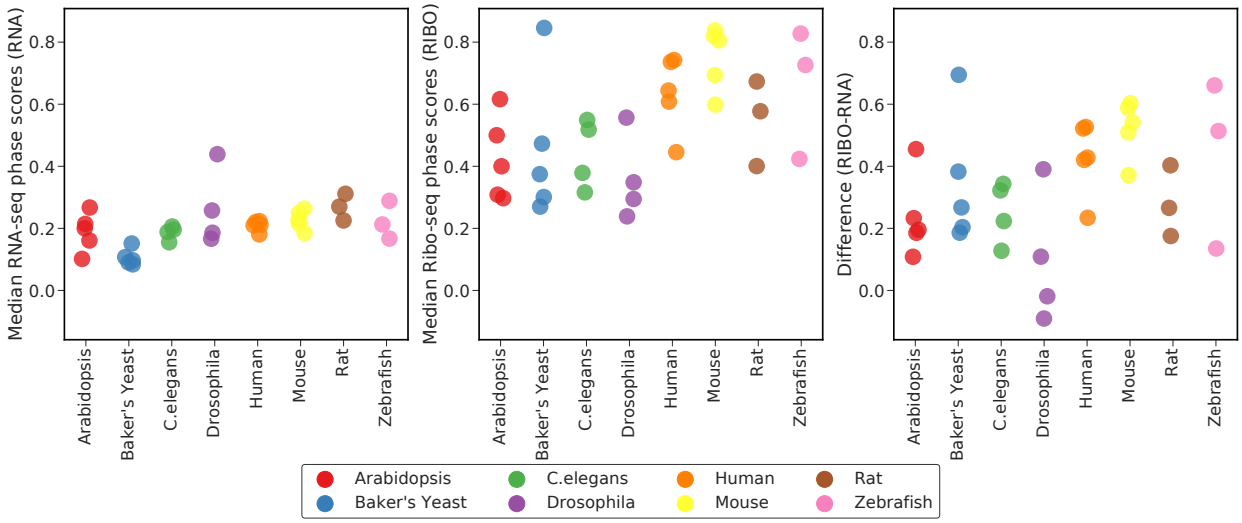

Figure S28: **Median phase score for RNA-seq and Ribo-seq and their differences across multiple species.** For each dataset, the median phase score was calculated for all the candidate ORFs for both Ribo-seq and the corresponding RNA-seq sample. Same as Supplementary Figure S27 except that the RNA- and Ribo-seq samples have been separated into individual panels.

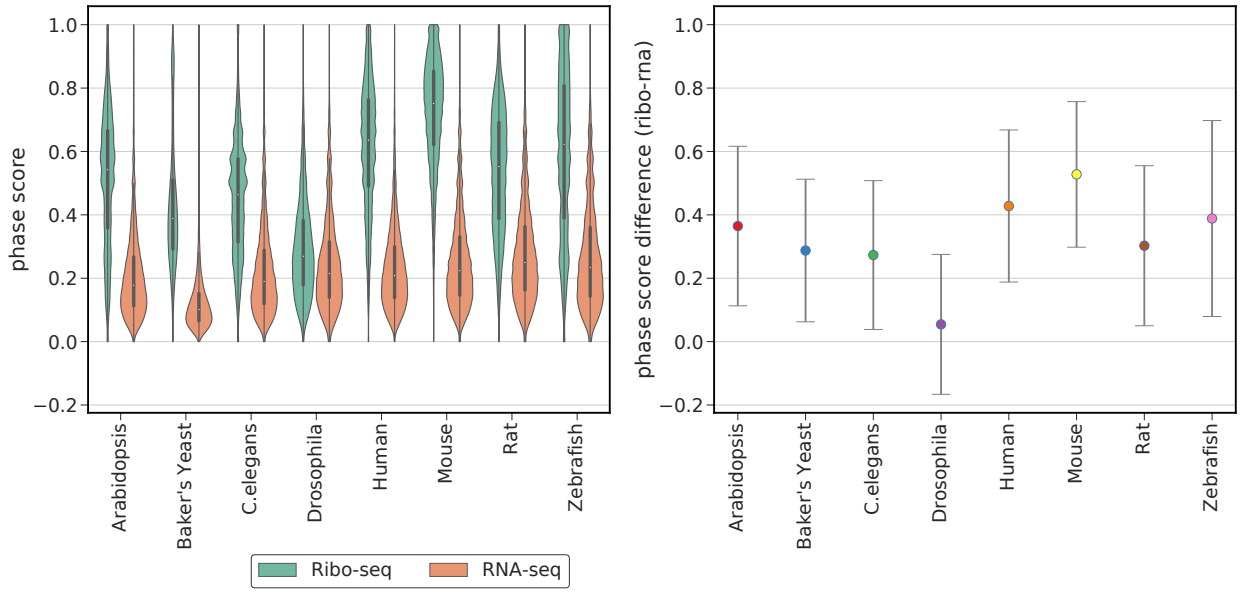

Figure S29: **Distribution of median phase scores for RNA-seq and Ribo-seq samples and their differences across multiple species.** For each species, medians were calculated on the collection of merged datasets for that species.

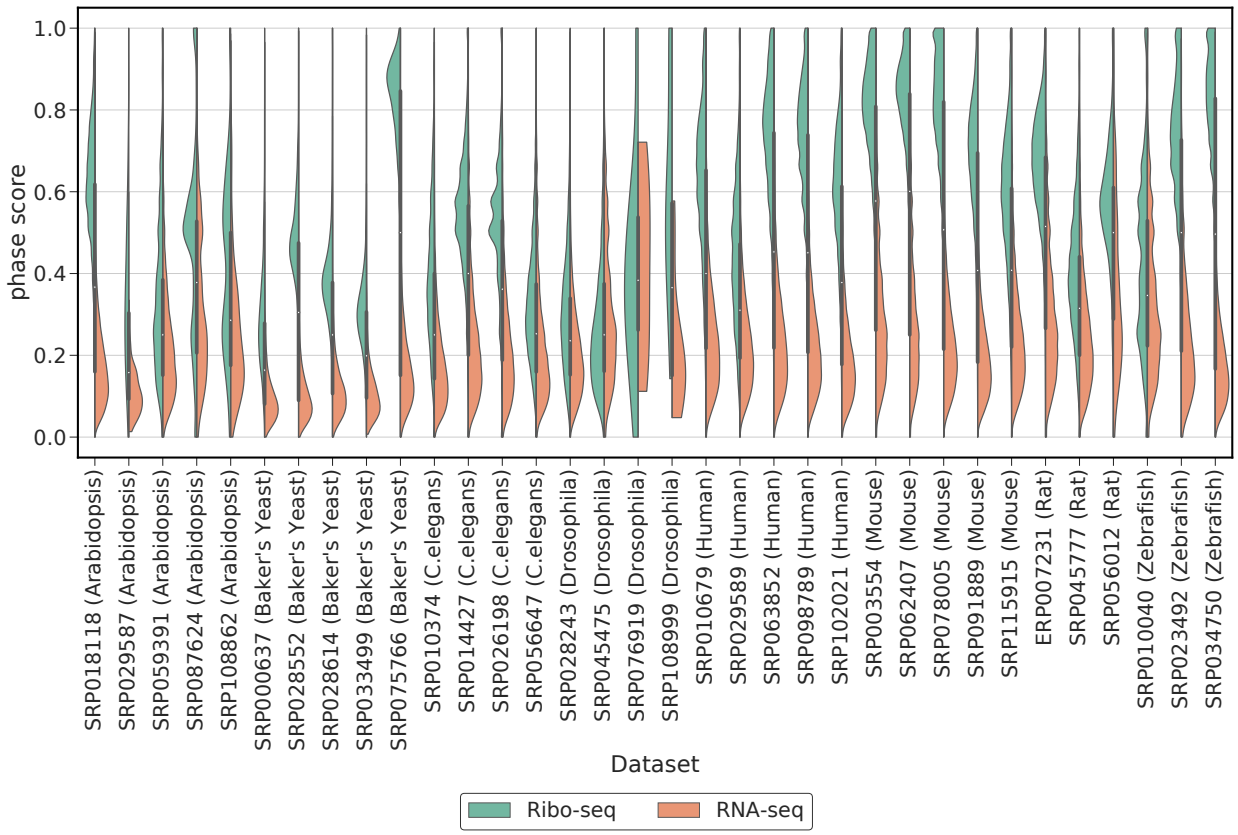

Figure S30: **Distribution of individual RNA-seq and Ribo-seq samples' phase scores across species.** For each dataset phase scores were calculated for all candidate ORFs. For human and mouse, Ribo-seq CCDS profiles were treated as true positive and the corresponding RNA-seq profile was treated as true negative. For all other species Ribo-seq profile of annotated CDS regions were treated as true positive and the corresponding RNA-seq profile treated as true negative.

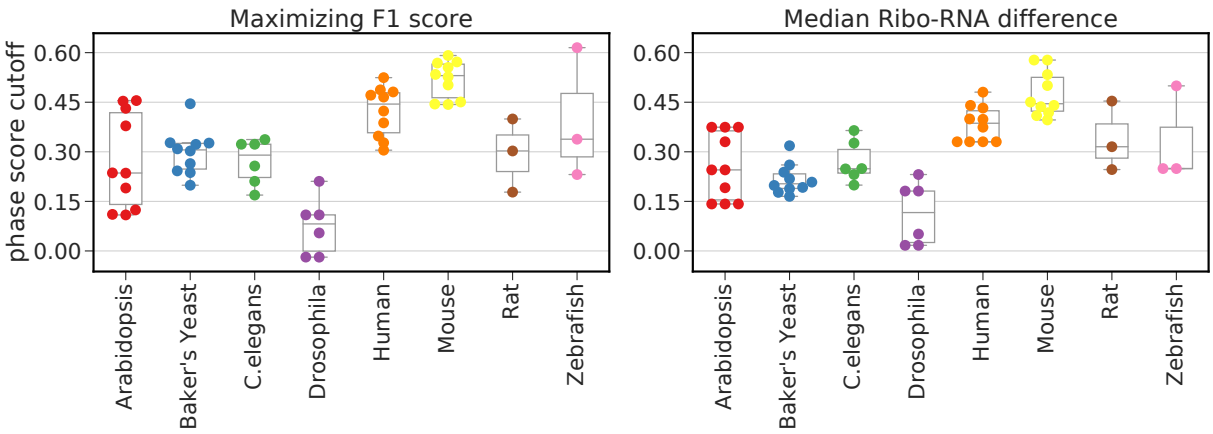

Figure S31: **Distribution of median difference between Ribo-seq and RNA-seq sample as determined using only two datasets per species.** For each species all possible combinations of two datasets were chosen and median difference between phase scores of Ribo-seq and RNA-seq determined.

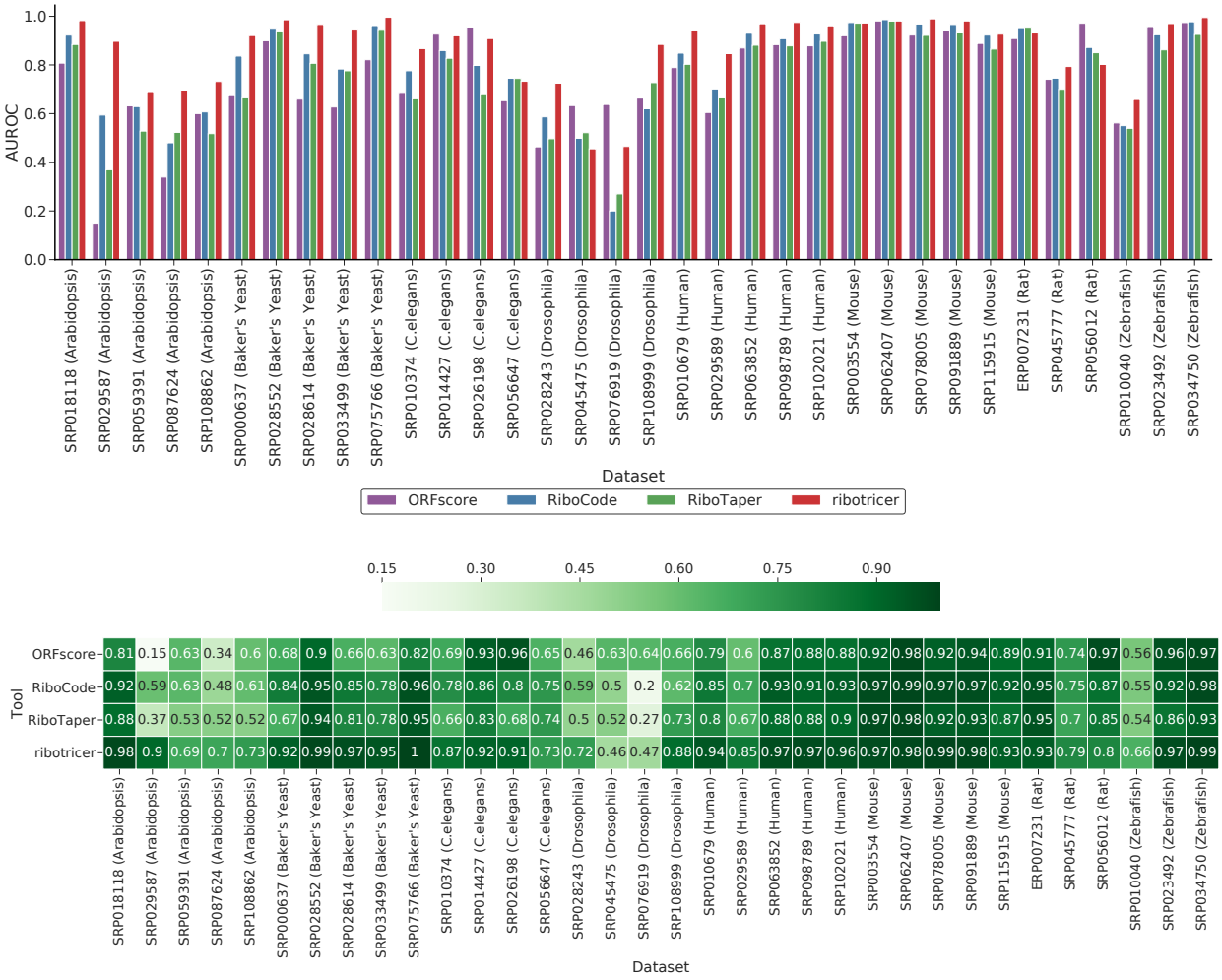

Figure S32: **Distribution of area under ROC (AUROC) across multiple species.** For each Ribo-seq and RNA-seq pair in a dataset, area under ROC was calculated for exon level classification using Ribotricer, Ribotaper, RiboCode and ORFScore.

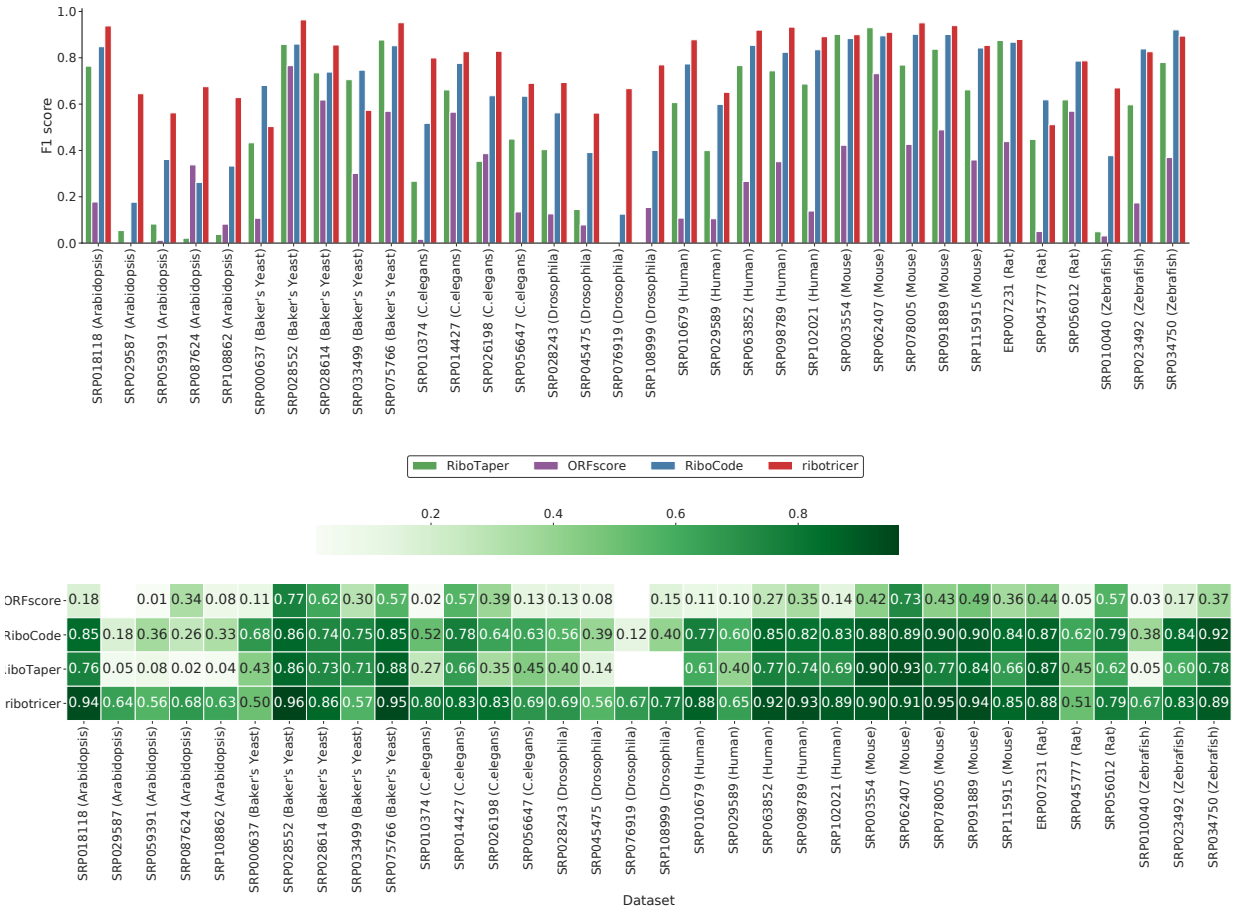

Figure S33: **Distribution of F1 scores across species using species-specific cutoff.** For each species two datasets were used to learn the cutoff score of ribotricer for that species.

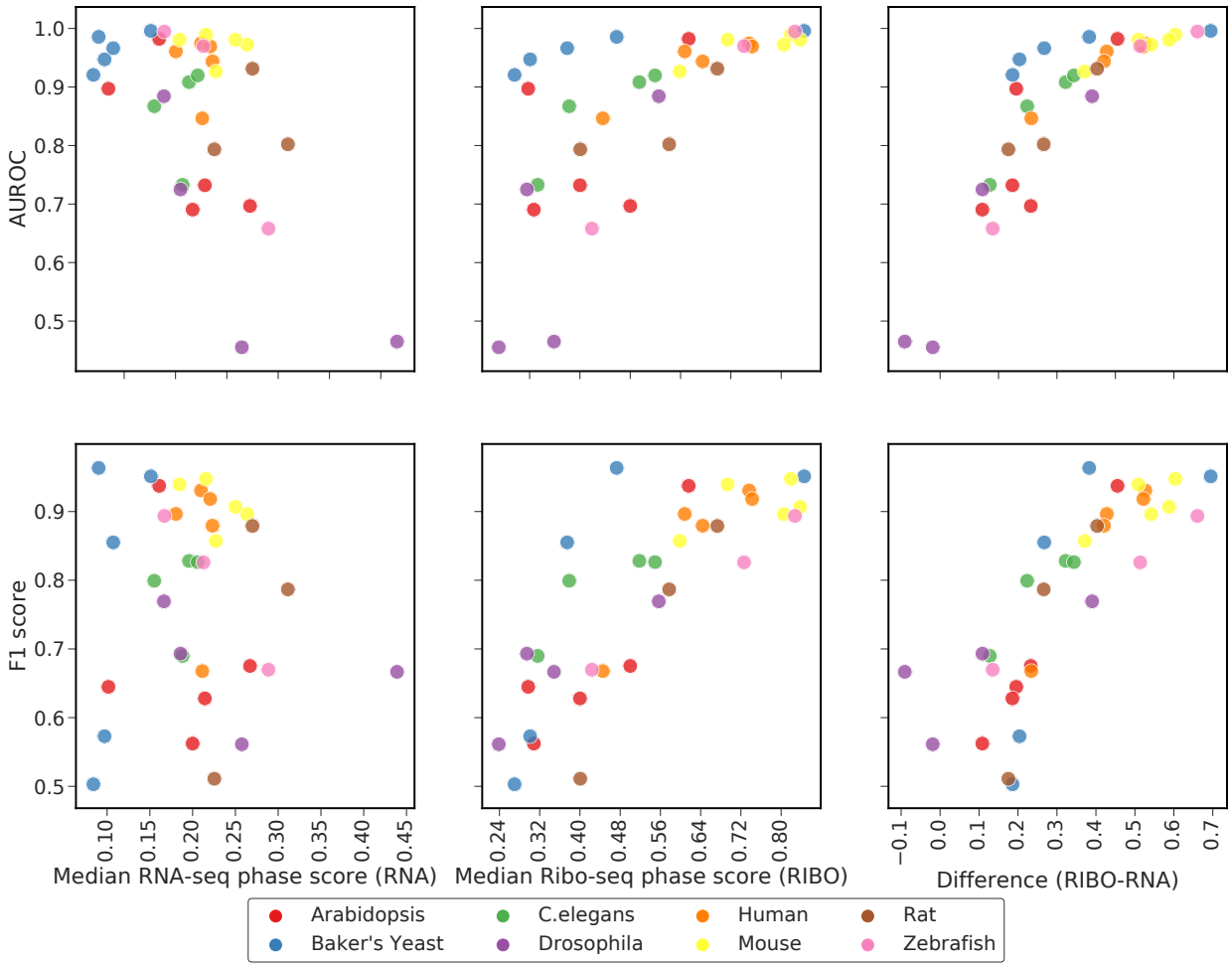

Figure S34: **Performance of ribotricer at AUROC and F1 scores metrics across species at different median phase scores of RNA-seq and Ribo-seq samples using species-specific cutoff.** For each dataset, median phase score was calculated for both RNA-seq and Ribo-seq samples for the same list of candidate ORFs.

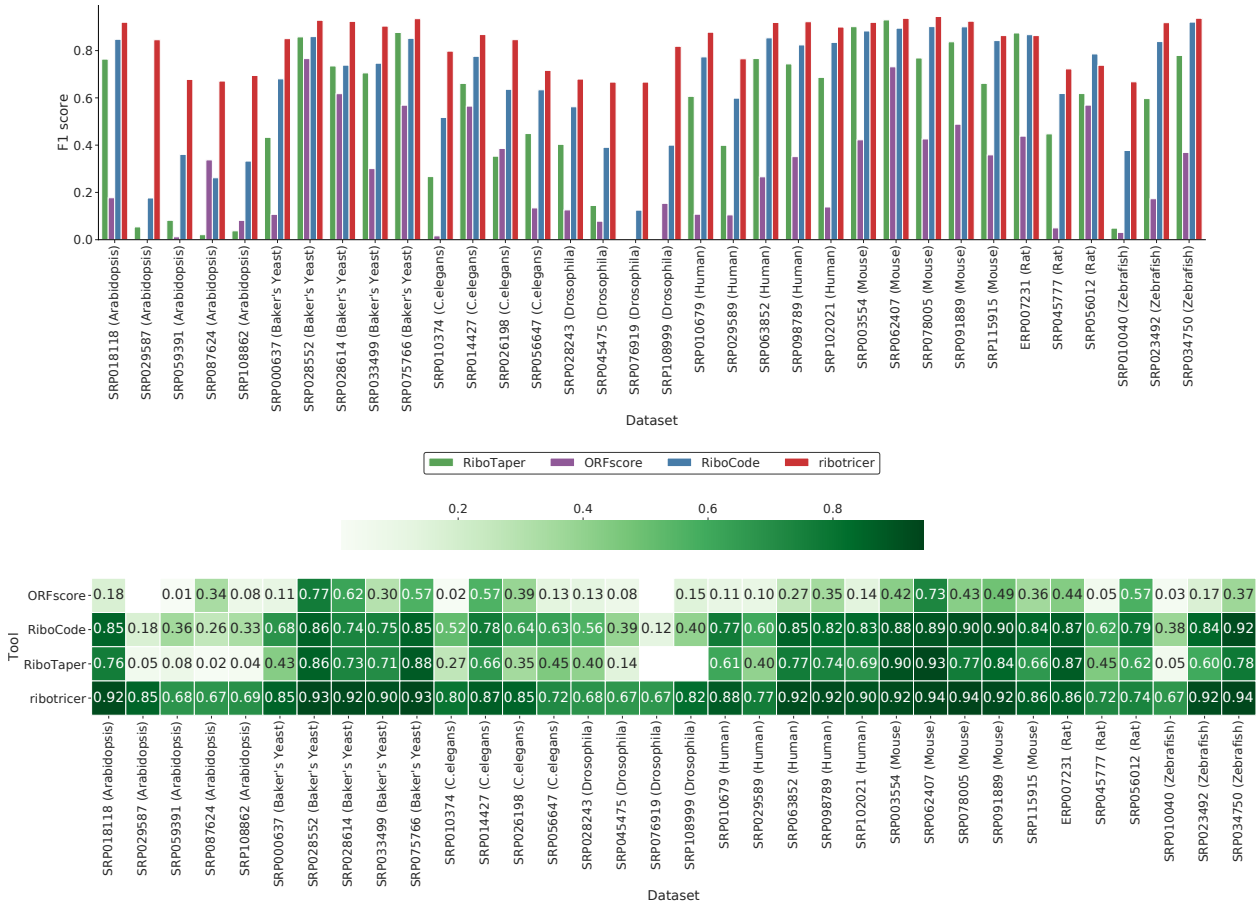

Figure S35: **Distribution of F1 scores across species using dataset-specific cutoff.** For each dataset, the cutoff was learned by determining the median phase score difference between Ribo-seq and RNA-seq profiles by sampling one-third of the total protein-coding transcripts $n_{\text{bootstrap}} = 10000$ times.

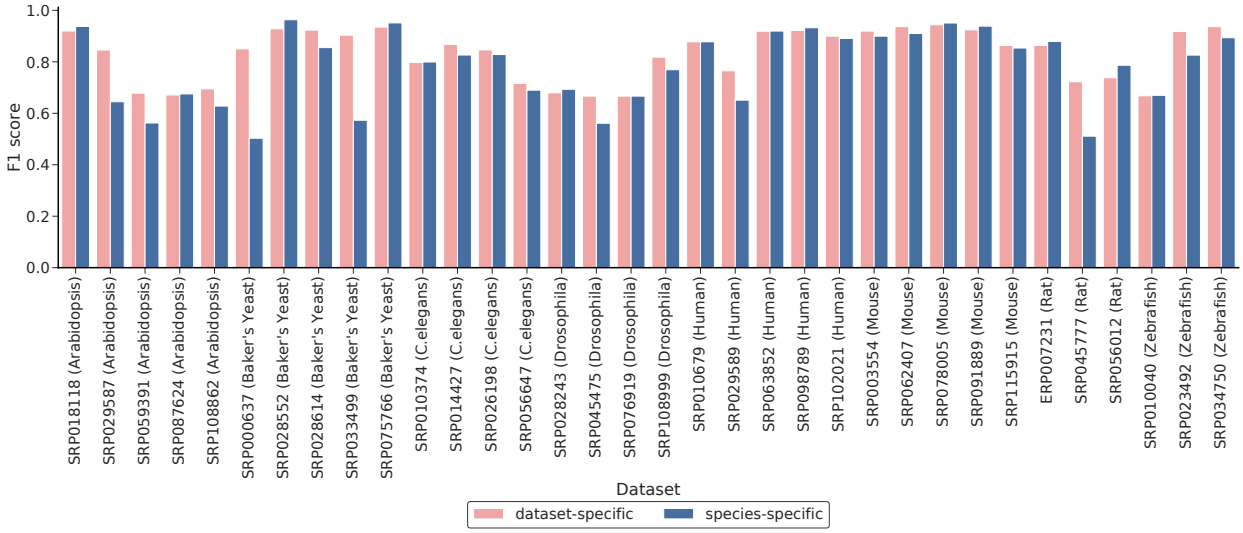

Figure S36: **Difference in performance of ribotricer using species-specific or dataset-specific cutoffs.** Species-specific cutoffs were learned by maximizing the F1 scores for two datasets per species while dataset-specific cutoffs were learned per dataset using the median difference of phase score of Ribo-seq and RNA-seq protein coding profiles.

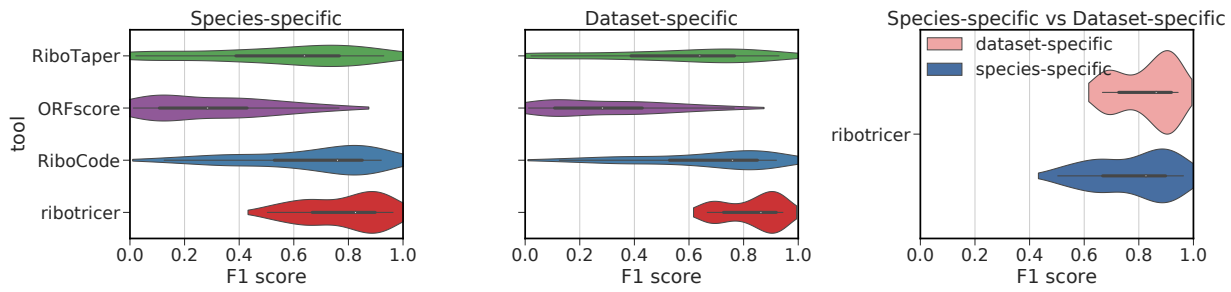

Figure S37: **Summarized performance of ribotricer using species-specific and dataset-specific strategies.** Species-specific cutoffs were learned by maximizing the F1 scores for two datasets per species while dataset-specific cutoffs were learned per dataset using the median difference of phase score of Ribo-seq and RNA-seq protein coding profiles. Species-specific and dataset-specific cutoffs only apply to ribotricer.

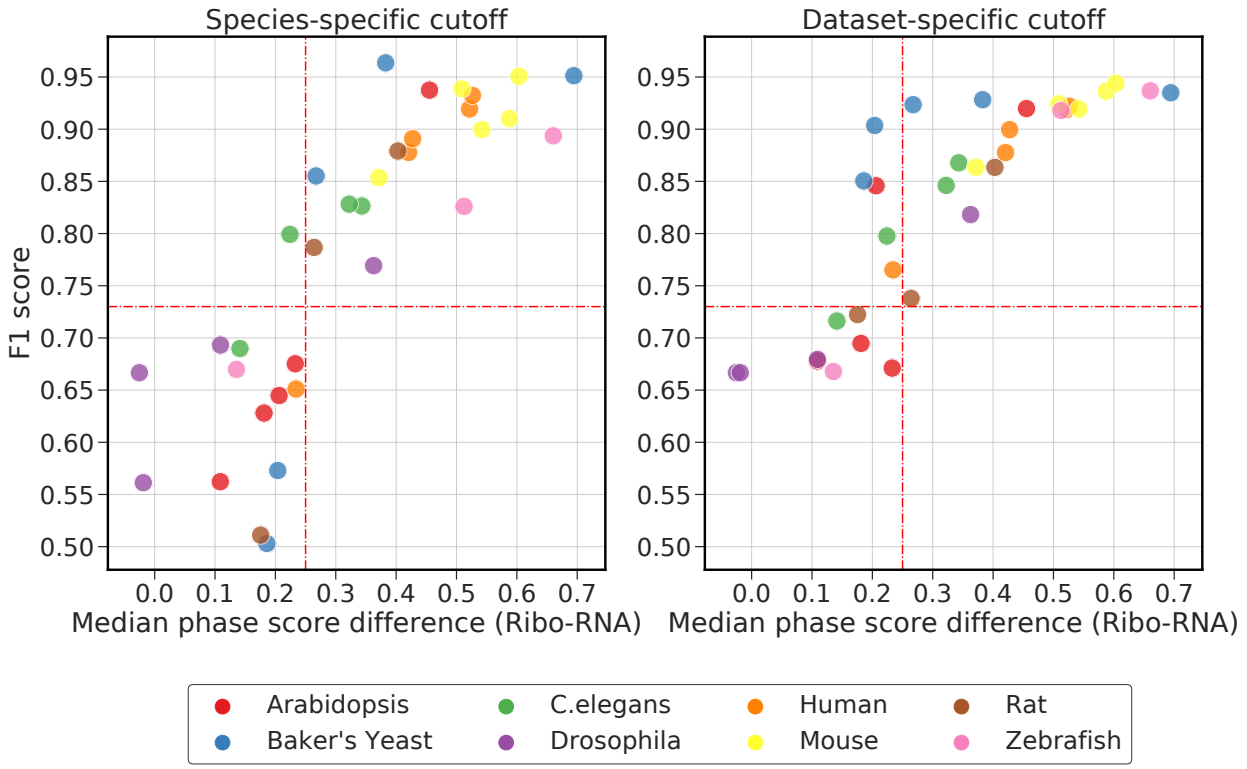

Figure S38: **Distribution of ribotricer's F1 scores with respect to median phase score difference of Ribo-seq and RNA-seq, using species-specific and dataset-specific cuoffs.** Species-specific cutoffs were learned by maximizing the F1 scores for two datasets per species while dataset-specific cutoffs were learned per dataset using the median difference of phase score of Ribo-seq and RNA-seq protein coding profiles. The dashed red lines indicate a median difference of 0.25 between Ribo-seq and RNA-seq phase scores results in a F1 score of 0.73 and above.

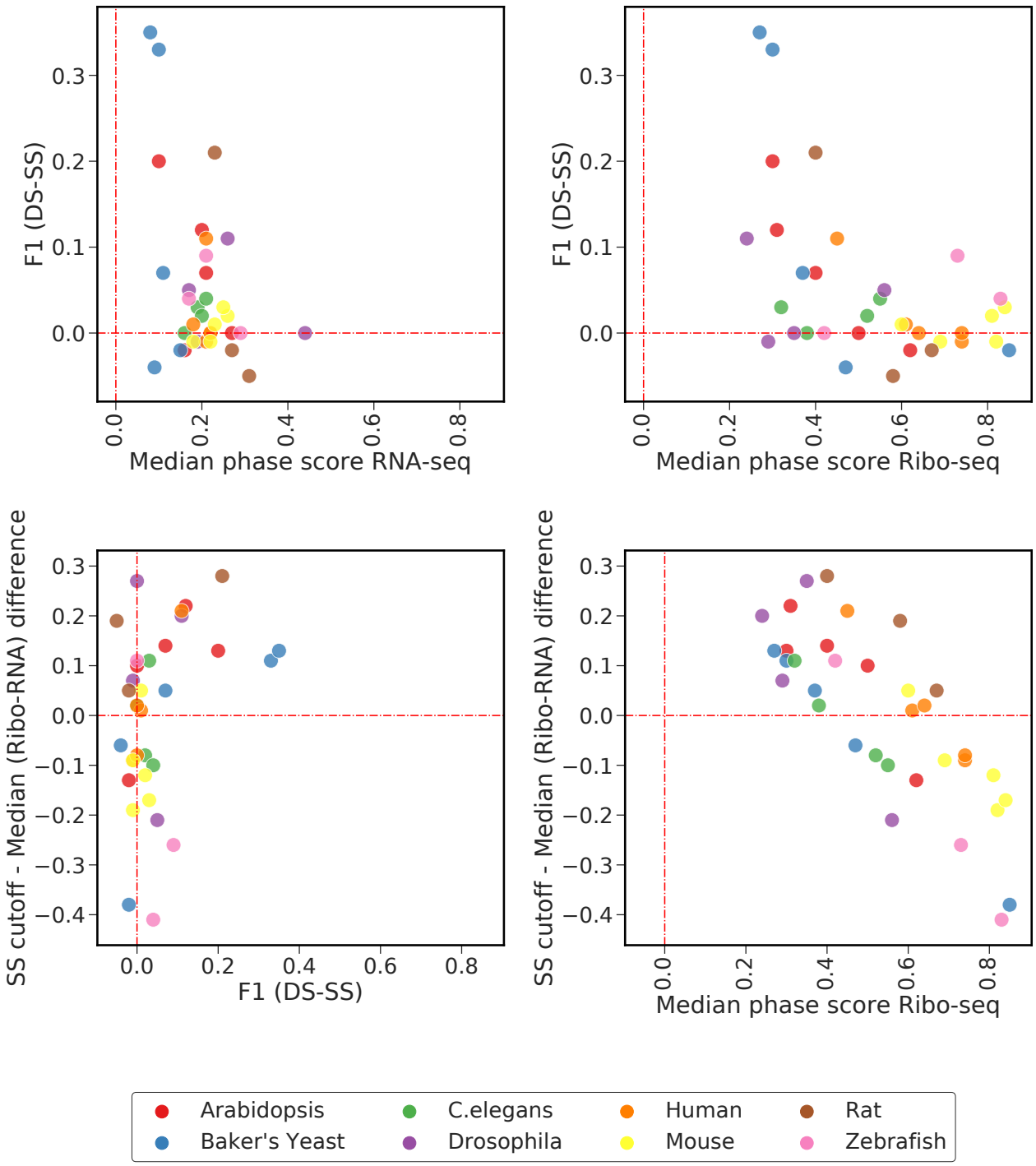

Figure S39: **Effect of Ribo-seq and RNA-seq phase scores on species-specific and dataset-specific based F1 performance.** F1 (DS-SS) indicates difference in F1 scores using species-specific (SS) or dataset-specific (DS) cutoff. Each single data point represents one dataset. Median phase scores were calculated using all the candidate ORF profiler of either RNA-seq or Ribo-seq sample.

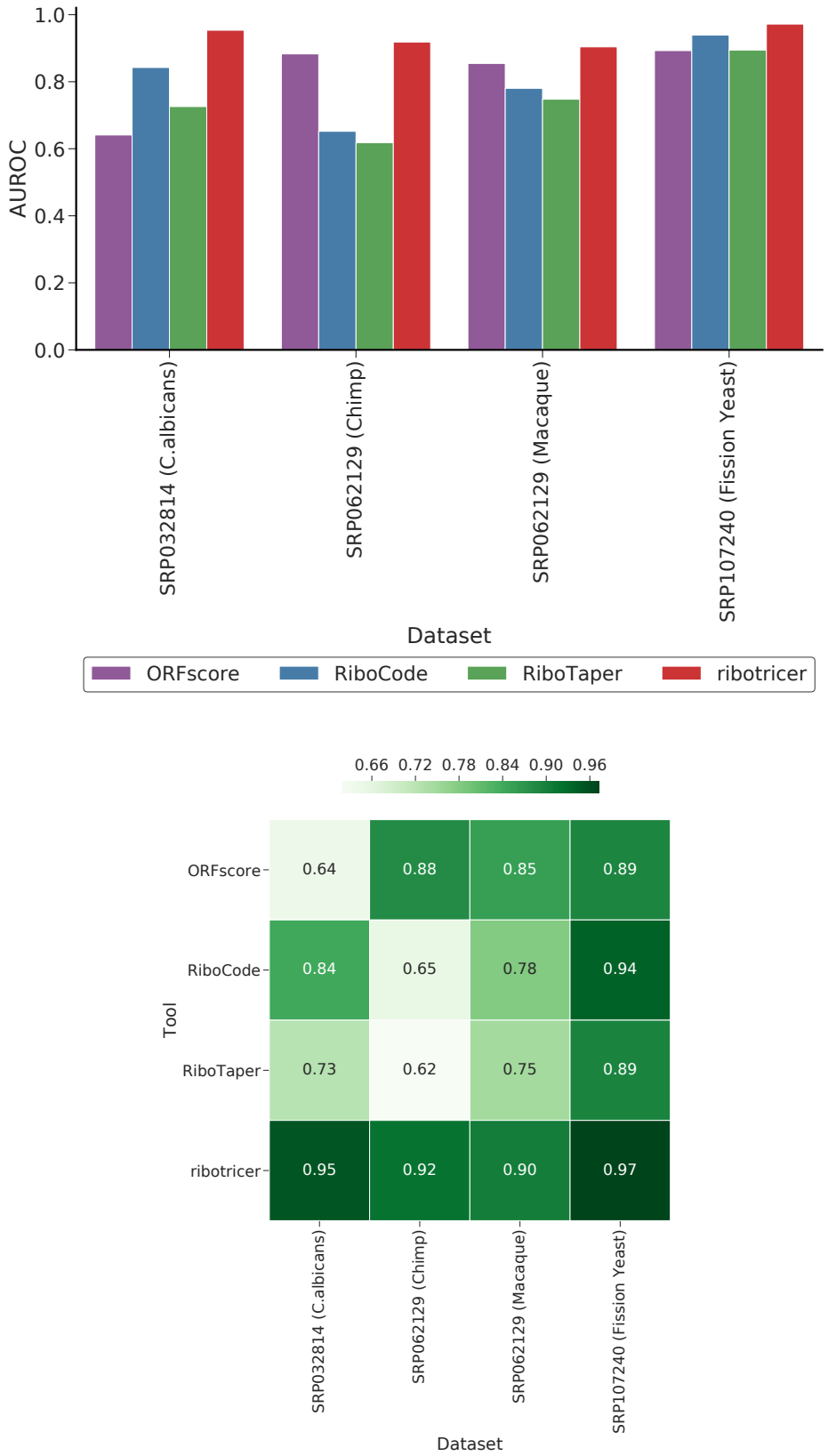

Figure S40: **Distribution of area under ROC in the independent datasets.** For each Ribo-seq and RNA-seq pair in a dataset, area under ROC was calculated for exon level classification using Ribotricer, Ribotaper, RiboCode and ORFScore.

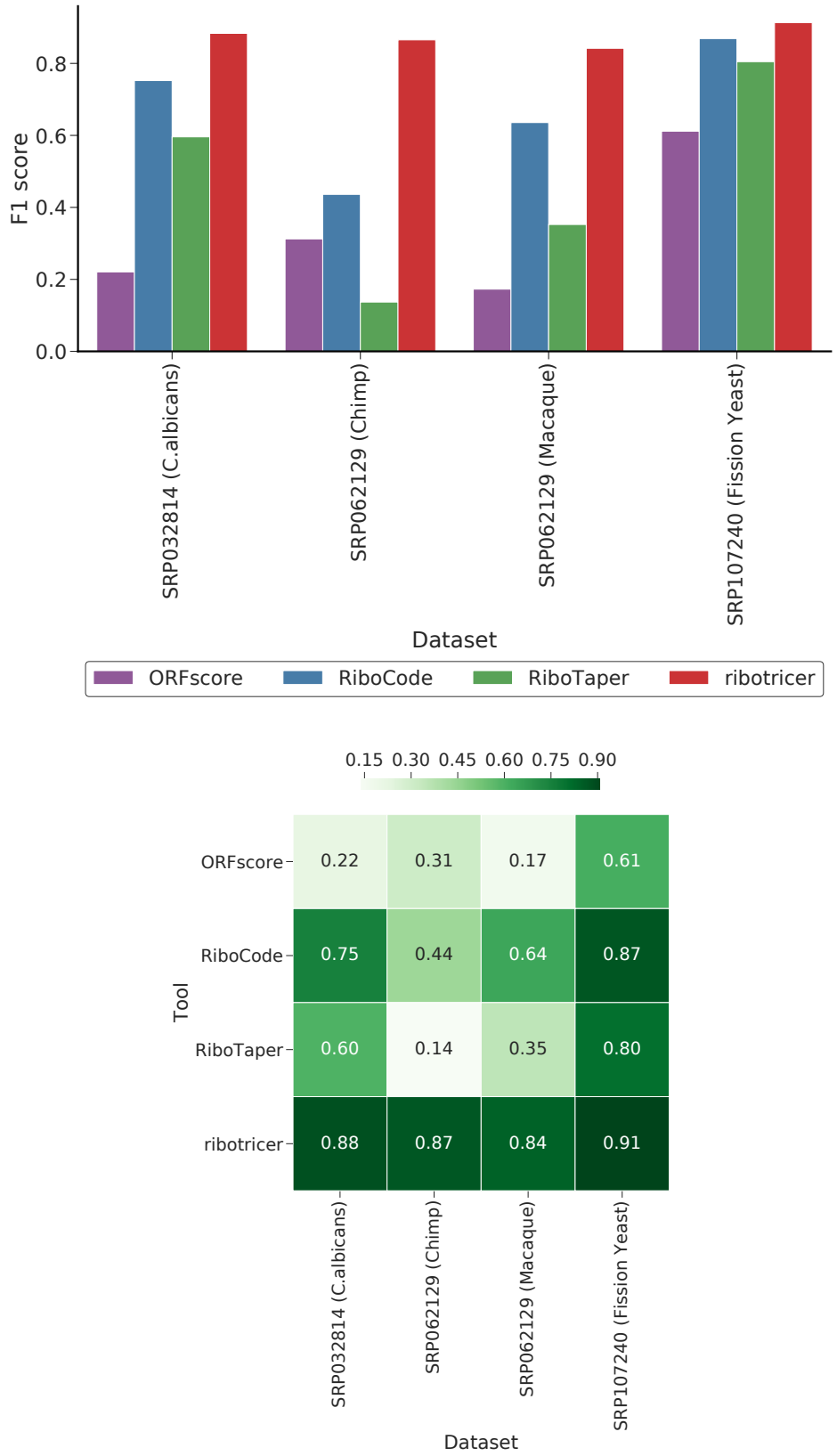

Figure S41: **Distribution of F1 scores in the independent datasets.** For each Ribo-seq and RNA-seq pair in a dataset, F1 score was calculated for exon level classification using Ribotricer, Ribotaper, RiboCode and ORFScore.

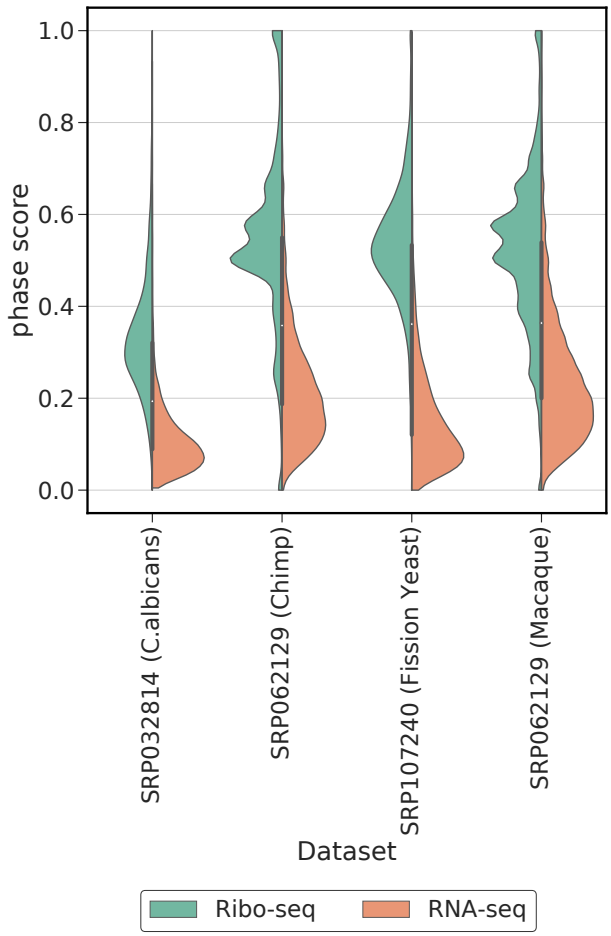

Figure S42: **Distribution of ribotricer's phase scores for RNA-seq and Ribo-seq samples in the independent datasets.** For each dataset phase scores were calculated for all candidate ORFs. For human and mouse, Ribo-seq CCDS profiles were treated as true positive and the corresponding RNA-seq profile was treated as true negative. For all other species Ribo-seq profile of annotated CDS regions were treated as true positive and the corresponding RNA-seq profile treated as true negative.