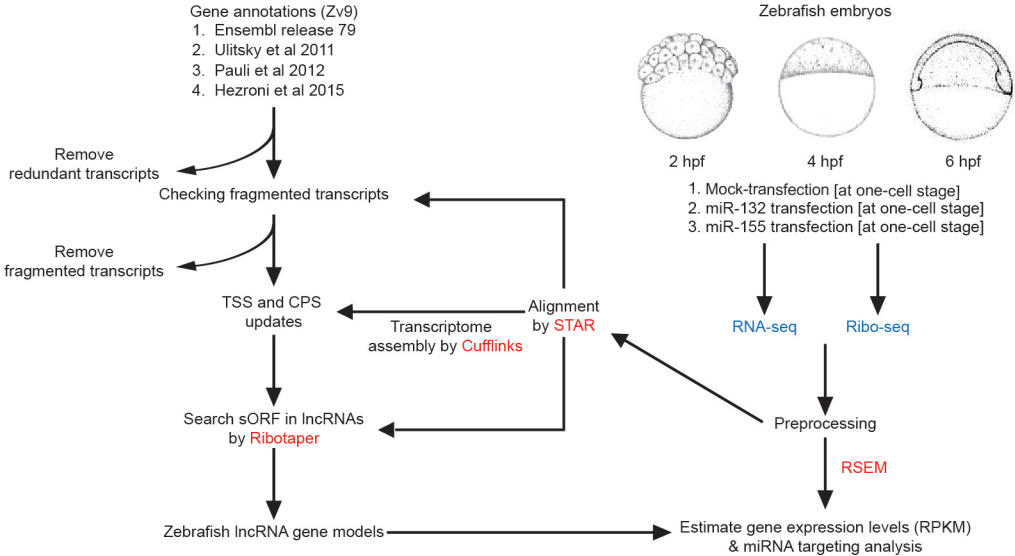
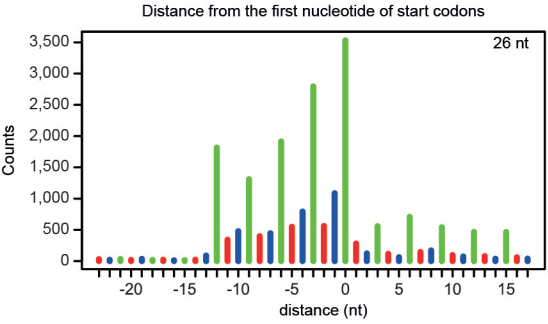
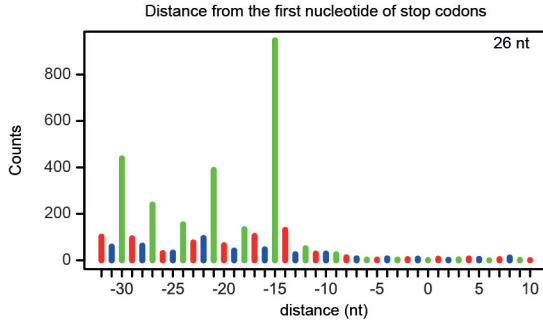
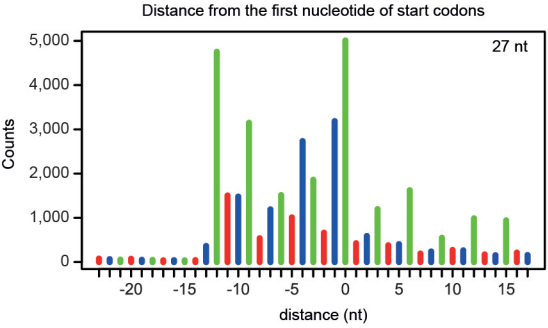
Supplementary Figure S1

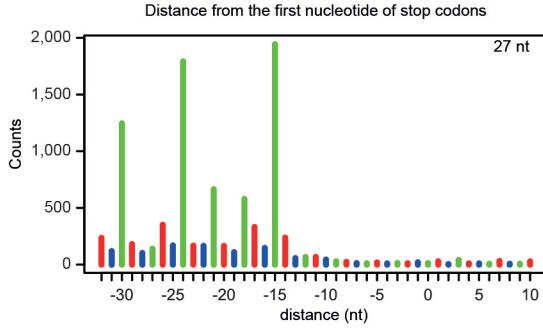


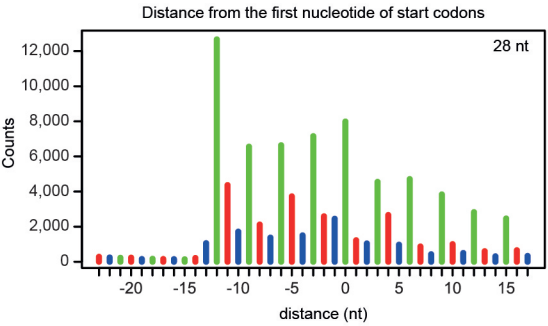
Supplementary Figure S2

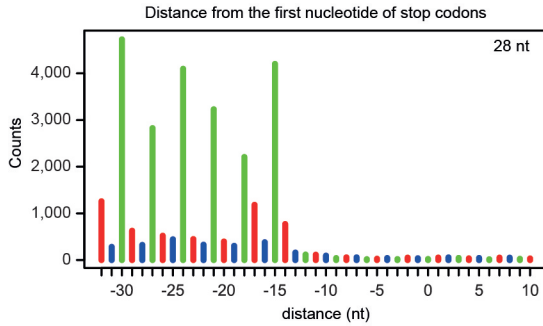


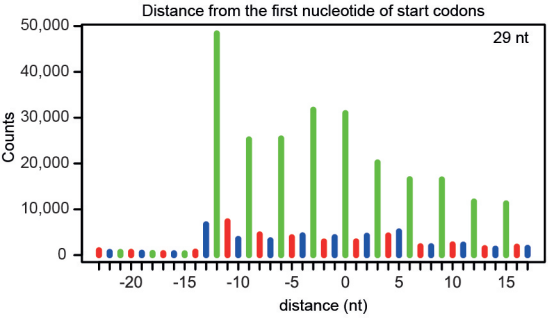


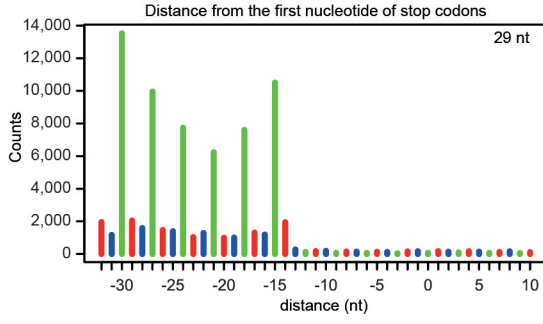


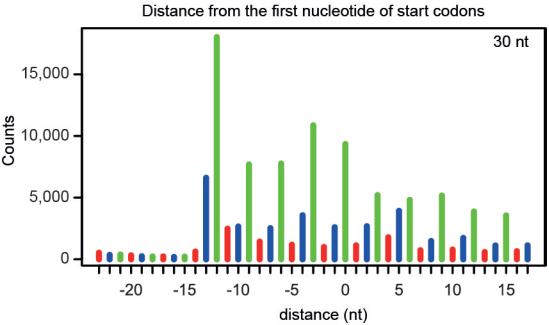


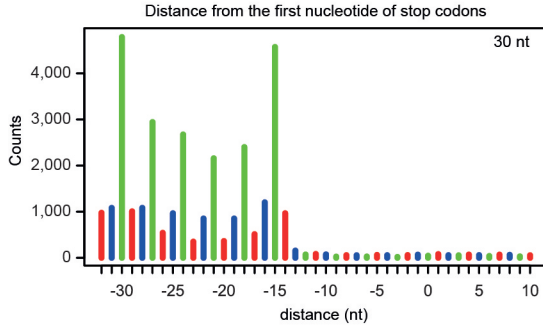






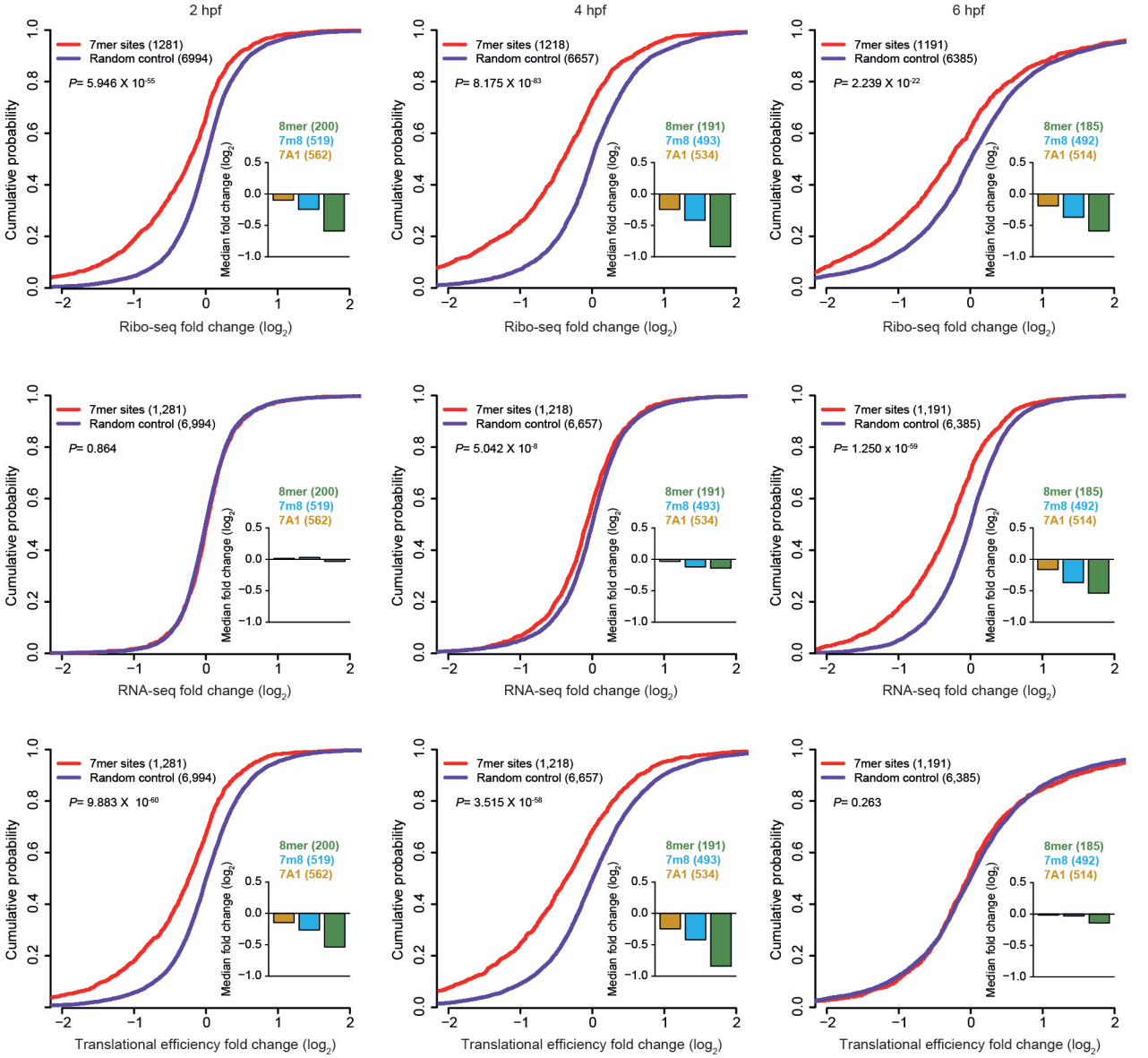




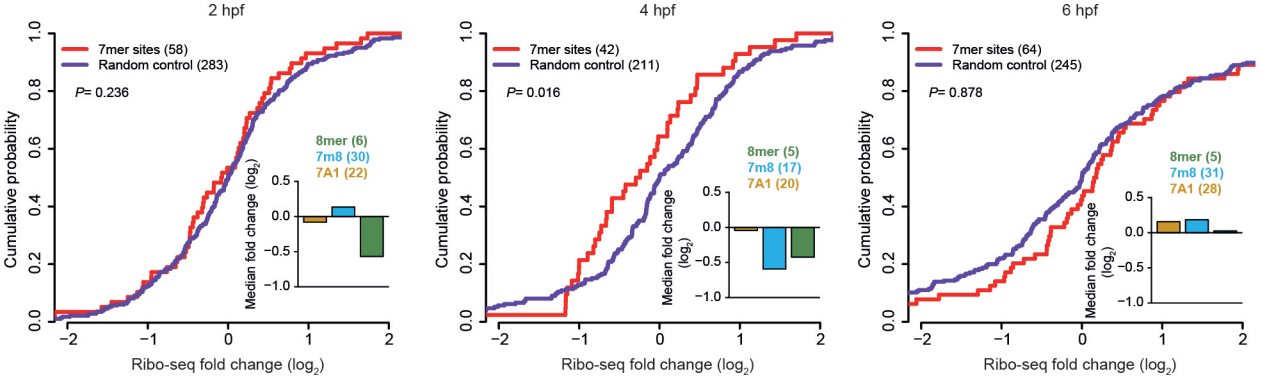


Supplementary Figure S3

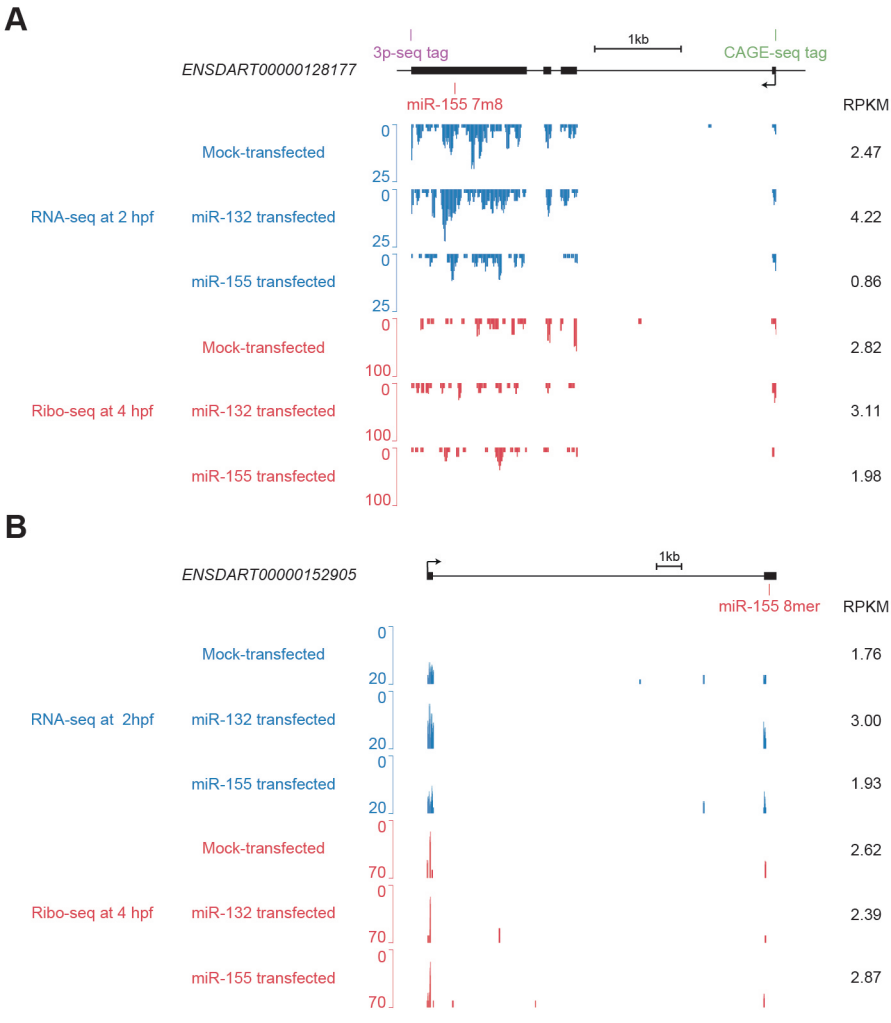
Supplementary Figure S4



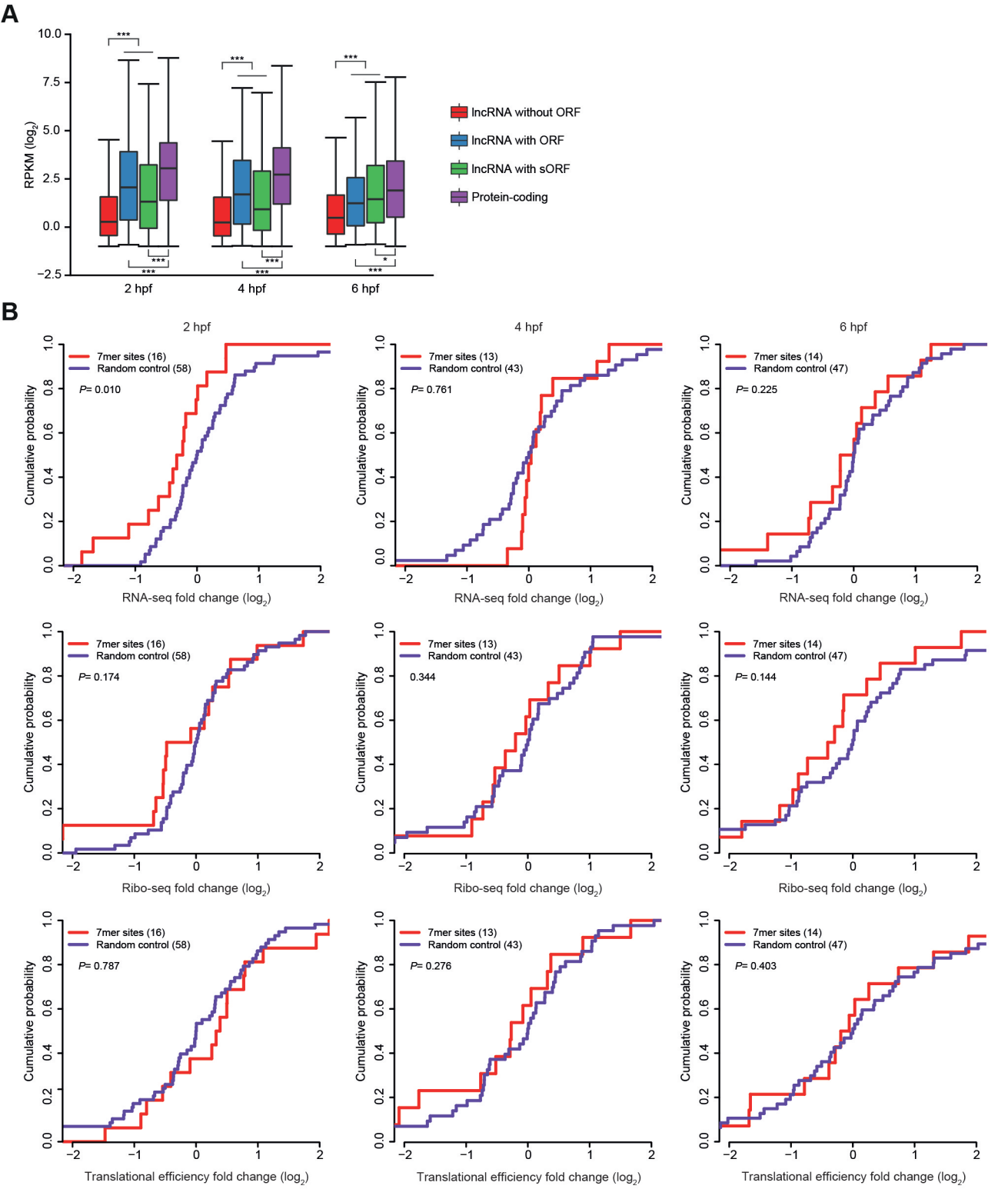
Supplementary Figure S5



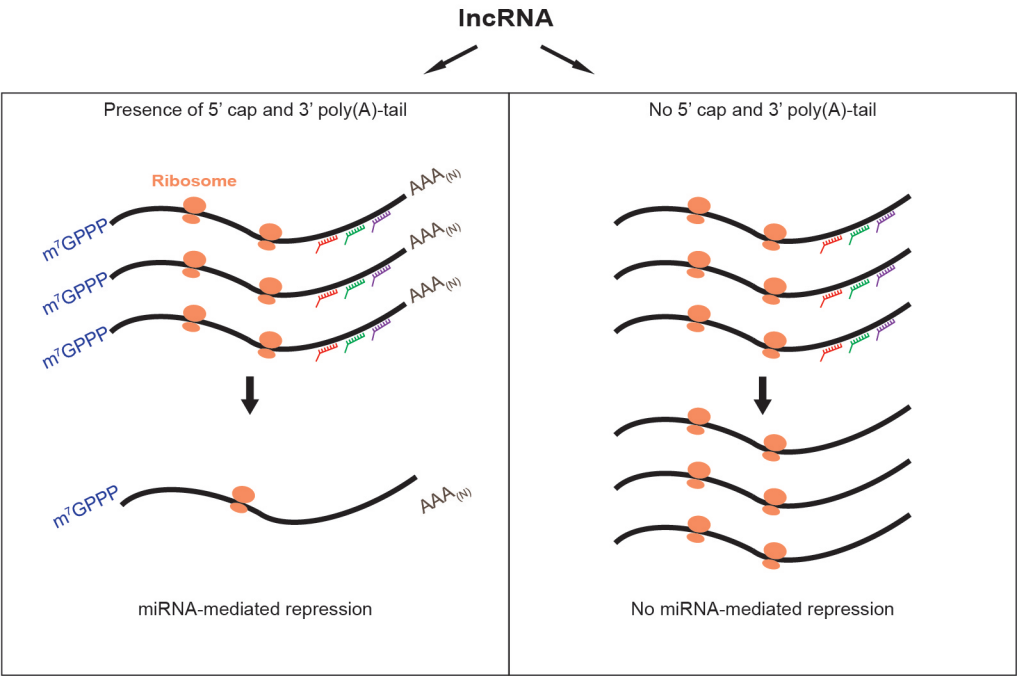
Supplementary Figure S6



Supplementary Figure S7



Supplementary Figure S8



Supplementary Table S1 | Over-represented sequences from each sample of RNA- and Ribo-seq. Reads containing specified sequences in the table at corresponding stage were trimmed off or removed by cutadapt using the options specified in **Supplementary Materials and Methods**.

Data type	Stage	Condition	Over-represented sequences
RNA-seq	2hpf	Wild-type	GGCATTAACGCGAACTCGGCCTACAATAGTGA
		miR-132 injection	GGCATTAACGCGAACTCGGCCTACAATAGTGA
		miR-155 injection	GGCATTAACGCGAACTCGGCCTACAATAGTGA
	4hpf	Wild-type	GGCATTAACGCGAACTCGGCCTACAATAGT
		miR-132 injection	TTAACGCGAACTCGGCCTACAATAGTGA
		miR-155 injection	GGCATTAACGCGAACTCGGCCTACAATAGT
	6hpf	Wild-type	GGCATTAACGCGAACTCGGCCTACAATAGT
		miR-132 injection	GGCATTAACGCGAACTCGGCCTACAATAGT
		miR-155 injection	GGCATTAACGCGAACTCGGCCTACAATAGT
RPF	2hpf	Wild-type	ACCCGGGGACGCGTGCATTTATCAGAT
		miR-132 injection	ACCCGGGGACGCGTGCATTTATCAGA
		miR-155 injection	TACCCGGGGACGCGTGCATTTATCAGAT
	4hpf	Wild-type	CCCGGGGACGCGTGCATTTATCAGATTCG
		miR-132 injection	ACCCGGGGACGCGTGCATTTATCAGA
		miR-155 injection	ACCCGGGGACGCGTGCATTTATCAGAT
	6hpf	Wild-type	ACCCGGGGACGCGTGCATTTATCAGAT
		miR-132 injection	ACCCGGGGACGCGTGCATTTATCAGAT
		miR-155 injection	ACCCGGGGACGCGTGCATTTATCAGAT

Supplementary Table S2 | Read statistics for RNA-seq and Ribo-seq in each sample

Data type	Stage	Condition	Reads before pre-processing	Reads after pre-processing	Uniquely aligned reads
RNA-seq	2hpf	Wild-type	30,823,253	23,827,257	14,158,189
		miR-132 injection	27,220,096	24,669,611	12,266,788
		miR-155 injection	24,962,449	22,595,872	10,712,927
	4hpf	Wild-type	22,412,745	20,906,430	9,349,676
		miR-132 injection	26,785,344	25,139,562	7,847,362
		miR-155 injection	28,212,007	26,407,805	9,470,116
	6hpf	Wild-type	28,199,307	26,242,998	12,684,872
		miR-132 injection	21,473,374	20,065,657	9,051,650
		miR-155 injection	22,670,965	21,174,663	9,699,067
RPF	2hpf	Wild-type	26,822,908	7,385,388	10,787,524
		miR-132 injection	23,209,679	983,509	8,199,867
		miR-155 injection	19,612,100	7,280,351	17,473,921
	4hpf	Wild-type	9,081,046	23,833,252	3,718,364
		miR-132 injection	9,849,906	8,206,873	4,232,584
		miR-155 injection	10,283,561	8,589,655	4,268,065
	6hpf	Wild-type	11,071,446	9,240,543	4,516,276
		miR-132 injection	11,338,173	9,522,801	4,878,388
		miR-155 injection	10,956,871	9,192,245	4,931,732

Supplementary Table S3 | Statistics for the TSS and CPS update Canonical TSS and CPS supported transcripts are transcripts having TSS and CPS in the annotated start and end position, respectively. CPS updated transcripts are the transcript having TSS in the annotated start position and an alternative CPS, either 5kbp downstream of the annotated end position, or in the exonic, or intronic region, and supported by transcriptome assembly by Cufflinks.

	Canonical TSS and CPS supported transcripts		CPS updated transcripts	
Chromosome	Protein-coding genes	lncRNAs	Protein-coding genes	lncRNAs
1	541	66	22	9
2	583	72	17	9
3	672	91	24	10
4	354	91	7	3
5	605	148	36	7
6	540	85	22	5
7	589	103	30	2
8	497	70	15	0
9	427	86	16	4
10	429	65	12	3
11	376	63	13	5
12	396	65	12	2
13	455	52	25	2
14	396	81	20	7
15	401	56	12	0
16	535	91	17	4
17	511	68	18	5
18	367	80	16	1
19	540	80	22	3
20	534	79	14	5
21	481	45	21	3
22	377	38	18	3
23	419	87	23	6
24	366	56	14	4
25	410	38	6	1
MT*	0	0	0	0

*MT is mitochondrial chromosome

Supplementary Material

Post-transcriptional and translational regulation of mRNA-like long non-coding RNAs by microRNAs in early developmental stages of zebrafish embryos

Kyung-Tae Lee[1] and Jin-Wu Nam[1,2,3,*]

[1]Department of Life Science, College of Natural Sciences, Hanyang University, Seoul 133791, Republic of Korea

[2]Research Institute for Natural Sciences, Hanyang University, Seoul 133791, Republic of Korea

[3]Research Institute for Convergence of Basic Sciences, Hanyang University, Seoul 133791, Republic of Korea

*Corresponding author

Tel: +82-2-2220-2428, Fax: +82-2-2298-0319, Email: jwnam@hanyang.ac.kr

Keywords: miRNA, lncRNA, 5’ cap, 3′ poly(A)-tail, sORF

Running title: miRNA-mediated silencing of mRNA-like lncRNAs

Supplementary Materials and Method

Dataset The sequence for the zebrafish reference genome assembly (Zv9) and gene annotations used throughout this study were downloaded from the Ensembl database (release 79) (1). Additional lncRNA annotations were derived from the previous studies (2-4). CAGE-seq data for TSS annotations were downloaded from a public website (http://zeprome.genereg.net/downloads/) (5), and 3P-seq data for CPS annotations were downloaded from NCBI Gene Expression Omnibus (GEO, GSE37453) (6). Poly(A)-selected RNA-seq and cycloheximide-treated Ribo-seq data prepared from zebrafish embryos at 2, 4, and 6 hpf from mock- (control) and miRNA-transfected (miR-155 or miR-132) fertilized eggs were obtained from the NCBI GEO (GSE52809) (7).

Preprocessing and mapping of sequencing reads Random barcode sequences at the 5′ end of reads were first removed using seqtk 1.0-r31 (8). The over-represented *k*-mers within reads were examined using FastQC v0.10.1 (9) and were trimmed from the reads using cutadapt 1.9.dev1 (10) with the parameter, “overlap = 6” (**Supplementary Table S1**). Reads were then mapped to the reference genome using Bowtie 1.0.0 (11) and the mismatch rate was examined across read positions. If the mismatch rate was greater than 10% at a certain position, the corresponding sequences from the position to nearby 3’ or 5’ end were trimmed using seqtk. Finally, the remainders were analyzed by the window adaptive trimming tool, Sickle 1.200 (12), using default settings. Preprocessed reads from RNA-seq and Ribo-seq were mapped to the zebrafish reference assembly (Zv9) using STAR 2.4.1a (13) with mapping parameters similar to that for the ENCODE STAR-RSEM pipeline (14): “outFilterType, BySJout; outFilterMultimapNmax, 20; outFilterMismatchNmax, 999; outFilterMismatchNoverLmax, 0.04; alignIntronMin, 20; alignIntronMax, 1000000; alignMatesGapMax, 1000000; alignSJoverhangMin, 8; alignSJDBoverhangMin, 1; sjdbScore, 1; runThreadN, 8; genomeLoad, NoSharedMemory; outSAMtype BAM, Unsorted; quantMode, TranscriptomeSAM”. Statistics of sequencing reads before and after preprocessing, and uniquely aligned reads are shown in **Supplementary Table S2**.

Integration of lncRNA annotations and isoform selection Ensembl release 79 includes 41,703 transcripts classified as ‘protein_coding’ and 5,665 lncRNAs: 1,038 ‘long intergenic non-coding RNAs (lincRNAs)’; 58 ‘sense_intronics’; 9 ‘sense_overlappings’; 3,849 ‘processed_transcripts’; and 711

‘antisense’ transcripts. In addition, 691 lncRNAs from early embryogenesis stages (2), 1133 from late developmental stages (3), and 3,391 from comparative analysis (4) were added to the zebrafish lncRNA catalog. After filtering lncRNAs derived from contigs or scaffolds, 9,858 unique lncRNAs were searched for the presence of TSSs and CPSs. Briefly, 1,959 lncRNAs embedded both TSS and CPS, 3,959 had either TSS or CPS, and 3,940 had no TSS or CPS. For lncRNAs with no TSS and/or CPS, the neighboring lncRNA transcripts within the range of 5 to 95 percentiles of intron length distribution were inspected to see if a putative splice junction exists within the interval using additional RNA-seq junction reads and maximum entropy modeling (15). For junction reads, all RNA-seq bam files at 2 hpf, 4 hpf, and 6 hpf of mock-transfected embryos were used. The maximum entropy model was trained with human and mouse exon-junction data. If there were at least two exon junction reads spanning two neighboring transcripts or there was a predicted splice site, and the entropy score was greater than the cutoff (4.655), the two transcripts were regarded as fragments. In total, 23 fragmented transcripts were removed from the lncRNA catalog.

For miRNA targeting analysis in protein-coding mRNAs and lncRNAs, isoform selection was used. If a gene had multiple isoforms, the isoform with the longest coding DNA sequence (CDS) was selected. If there were multiple isoforms with the longest CDS, then the isoform with the longest 3′ UTR was selected. For the lncRNAs that did not have an annotated ORF, the isoform with the longest exonic length was selected. In this manner, 16,297 protein-coding and 6,488 lncRNA isoforms were identified.

Periodicity analysis of Ribo-seq To check the validity and periodicity of the Ribo-seq data used in this study, multiple metaplots for each read length were created using Ribotaper v1.3’s metagene analysis (16). The plots were manually inspected for reads of 26, 27, 28, 29, and 30 nt long where the periodicity was presented and the 12^{th} nucleotide was a P-site. These reads were regarded as high confidence RPFs and subjected to downstream analysis using Ribotaper. The selected lengths were re-analyzed to search for sORFs, which encode peptides shorter than 100 amino acids, in lncRNA. As a result, 701 unique isoforms out of 812 lncRNAs were found to contain at least one predicted ORF that translates a peptide less than 100 amino acids in length, and were regarded as lncRNAs with sORFs. In addition, 448 unique isoforms out of 512 lncRNAs with ORFs that translate a peptide greater than 100 amino acids were regarded as lncRNAs with putative ORFs, and were analyzed separately.

Coding potential To exclude lncRNAs that were not detected by Ribotaper to have sORFs but could still have substantial protein-coding potential, an additional analysis using CPC-0.9-r2 software (17) was performed. As a result, 1,083 lncRNAs, that were determined to have protein-coding potential by CPC, were excluded from the lncRNA set determined to have no ORF.

CPS and TSS annotations For CPS annotations, we adopted the same protocol, described in the 3P-seq paper (18). In addition, the 3P-seq tags that overlapped with the repeat elements, that were identified using RepeatMasker annotation for Zv9 zebrafish from the UCSC table browser (19), were removed. In total, 107,636 CPSs were annotated from different developmental stages (53,669 for the plus strand and 53,967 for the minus strand). The TSS annotations were performed using the same method as described in the original CAGE-seq paper (5), except for a less stringent cutoff of 0.5 tags per million (TPM) that was used to filter out transcript clusters of CAGE-seq tags. In total, 201,916 TSSs were annotated (99,482 for the plus strand and 102,074 for the minus strand).

TSS and CPS update Protein-coding and lncRNA gene annotations were updated with TSSs and CPSs using an in-house script. If TSS was located within 500 bps from the 5′ end of the transcript, the TSS was assigned to that transcript. If CPS was located within 100 bps from the 3′ end of the transcript, the CPS was assigned to that transcript. For transcripts supported by TSSs but not by CPSs, we inspected other CPSs that were 5 kilo base pairs (kb) downstream or in the exonic/intronic region. If there was CPS in the downstream 5 kb region, we examined whether the 3′ end of the transcript could be extended to the CPS via transcriptome assembly using Cufflinks v2.1.1 (20). On the other hand, if there was CPS in the exonic/intronic region, we examined whether the 3′ end of the transcript can be shorted to the CPS, and if a substantial depth of RNA-seq reads exists using transcriptome assembly with Cufflink. Statistics of TSS and CPS update on protein-coding mRNAs and lncRNAs are shown in **Supplementary Table S3**.

Gene expression profiles Gene expression values were estimated using RSEM v1.2.25 (14) with the running parameter “strand-specific”.

miRNA target prediction Canonical target sites of miR-132 and miR-155 were searched for in the 3′ UTR of protein-coding mRNAs and in whole exons of lncRNAs whose expression levels of RNA- and Ribo-seq exceeded 0.5 RPKM. To remove the ribosome shadowing effect in miRNA targeting (21), sites located 15 nucleotides (nt) downstream from the stop codon were excluded. Seeing as miR-132 and miR-155 were transfected in the duplex form, we searched for canonical miRNA binding sites in both strands. RNAs embedding 7mer sites (i.e., 7A1, 7m8, and 8mer, **Figure 3A** and **Supplementary Figure S3**) were considered as miRNA targets. For control targets, 100 dinucleotide-shuffled sequences of miR-132 and miR-155 were generated. Then the 7mer sites in the random sequences were searched for in the protein-coding mRNAs and lncRNAs using the same method described above. The genes embedding at least one of random 7mer sites, but not the original seed sites, were defined as random control.

Meta-analysis of miRNA targeting Differences in the expression and translational efficiency of targets (i.e., mRNAs and lncRNAs) containing 7mer sites between the mock- and miRNA-transfected conditions were compared to those with random 7mer sites using CDF graphs. Due to the relatively low number of lncRNAs with mRNA like features, fold changes ($\log_2$) of miR-132 and -155 7mer targets were combined to test the significance of changes in expression following miRNA transfection. RPKM from RNA-seq and Ribo-seq were used as expression values.

Translational efficiency Translational efficiency was calculated as $\log_2$ fold changes of RPF expression levels normalized by those of RNA to remove possible contribution of changes in RNA levels to changes in RPF levels. All fold changes for RNA-seq, Ribo-seq and the translational efficiency of miRNA target and random control RNAs were normalized to median fold changes in the random control RNAs.

Statistical analysis and in-house script All of the statistical analysis and in-house scripting in this paper were conducted by and written in R 3.2.3 (22) in python 2.7.11 (23), respectively.

Supplementary Figure Legends

Supplementary Figure S1 | The workflow for annotations of lncRNAs with mRNA-like features in zebrafish. Publicly available sequencing data and programs used in this study are indicated in blue and red fonts, respectively.

Supplementary Figure S2 | Meta-plots of Ribo-seq. The 3 nt periodicity of the Ribo-seq reads are shown in the metaplots drawn using Ribotaper meta-analysis. Plots were drawn using Ribo-seq reads of length showing 3 nt periodicity. The x-axis indicates positions from the start codon (left) and the upstream position from the stop codon (right). The y-axis is the number of aligned reads at specific positions.

Supplementary Figure S3 | Canonical miRNA target sites. (A) miR-132 (left) and miR-132* (right) target sites. Colored letters for each type of target site represent the actual miRNA target sequences in RNAs. **(B)** miRNA target sites of miR-155 (left) and miR-155* (right), with coloring as described in **(A)**.

Supplementary Figure S4 | miRNA targeting on protein-coding mRNAs. The cumulative distribution functions (CDFs) show the changes of RNA expression (top), RPF (middle) and translational efficiency (bottom) for protein-coding mRNAs with miRNA (red) or random target sites (purple) (**see Supplementary Materials and Methods** for more details), at each developmental stage in zebrafish embryos. The number of miRNA and random target sites are shown in parenthesis in top left corner. *P* values were calculated using a one-tailed Wilcoxon rank-sum test and is shown in top left corner. For each CDF, the median $\log_2$ fold change and the number of 7mer sites of miR-132 and miR-155 (right) that are 8mer (green), 7m8 (sky blue), and 7A1 (dark orange), are depicted.

Supplementary Figure S5 | The changes of RPFs for lncRNAs upon miRNA transfection. The CDFs show RPF expression changes between mock- and miRNA-transfected (miR-132 or miR-155) zebrafish embryos for lncRNAs with miRNA (red) or random target sites (purple)**,** otherwise as in **Supplementary Figure S4**

Supplementary Figure S6 | miRNA-mediated repression on lncRNA with canonical miRNA target site, 5' cap and 3' poly(A)-tail. (A-B) Read histogram of RNA-seq and Ribo-seq in exemplified lncRNAs. **(A)** Example of a lncRNA (*ENSDART00000128177*) supported by TSS and CPS and predicted to contain miR-155 7m8 target site. RNA expression levels at 2 hpf (red) and RPF expression levels at 4 hpf (blue) of different conditions are shown. The RPKM estimated from the whole transcripts by RSEM are shown on the right **(B)** Example of a lncRNA (*ENSDART00000152905*) that is not supported by TSS and CPS but predicted to contain miR-155 8mer site, otherwise as in (**A)**.

Supplementary Figure S7 | miRNA targeting for lncRNAs with putative ORFs

(A) Expression levels of the lncRNA sets without ORFs, with sORFs, and with putative ORFs and protein-coding mRNAs, at various zebrafish embryo development stages(2, 4 and 6 hpf), otherwise as shown in **Figure 2A.** (**B**) The CDFs show RNA (top), RPF levels (middle) and translational efficiency (bottom) changes for lncRNAs with the evidence of putative ORFs encoding peptides longer than 100 amino acids and with miRNA (red) or random target sites (purple) at each developmental stage in zebrafish embryos, otherwise as in **Supplementary Figure S4**.

Supplementary Figure S8 | The miRNA and lncRNA interaction model

The hypothetical model of interaction between miRNA and lncRNA. Briefly, miRNA targets lncRNAs but only exerts repression in RNA and RPF expression levels for those having 5′ cap and 3′ poly(A)-tail.

References

1. Aken BL, Ayling S, Barrell D et al (2016) The Ensembl gene annotation system. Database 2016, baw093-baw093
2. Ulitsky I, Shkumatava A, Jan CH, Sive H and Bartel DP (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell 147, 1537-1550
3. Pauli A, Valen E, Lin MF et al (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res 22, 577-591
4. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP and Ulitsky I (2015) Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep 11, 1110-1122
5. Nepal C, Hadzhiev Y, Previti C et al (2013) Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. Genome Res 23, 1938-1950
6. Ulitsky I, Shkumatava A, Jan CH et al (2012) Extensive alternative polyadenylation during zebrafish development. Genome Res 22, 2054-2066
7. Subtelny AO, Eichhorn SW, Chen GR, Sive H and Bartel DP (2014) Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature 508, 66-71
8. Shen W, Le S, Li Y and Hu F (2016) SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE 11, e0163962
9. Andrews S FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
10. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17
11. Langmead B, Trapnell C, Pop M and Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10
12. Joshi NA FJ (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.
13. Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21
14. Li B and Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12
15. Yeo G and Burge CB (2004) Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. Journal of Computational Biology 11, 377-394
16. Calviello L, Mukherjee N, Wyler E et al (2016) Detecting actively translated open reading frames in ribosome profiling data. NATURE METHODS 13, 165-170
17. Kong L, Zhang Y, Ye ZQ et al (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35, W345-349
18. Jan CH, Friedman RC, Ruby JG and Bartel DP (2011) Formation, regulation and evolution of

Caenorhabditis elegans 3'UTRs. Nature 469, 97-101
19. Karolchik D, Hinrichs AS, Furey TS et al (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Research 32, D493-D496
20. Trapnell C, Williams BA, Pertea G et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511-515
21. Pasquinelli AE (2012) MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nat Rev Genet 13, 271-282
22. Team RC (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing
23. Foundation. PS Python Language Reference, version 2.7.