eLIFE
elifesciences.org

Figures and figure supplements

Long non-coding RNAs as a source of new peptides

Jorge Ruiz-Orera, et al.

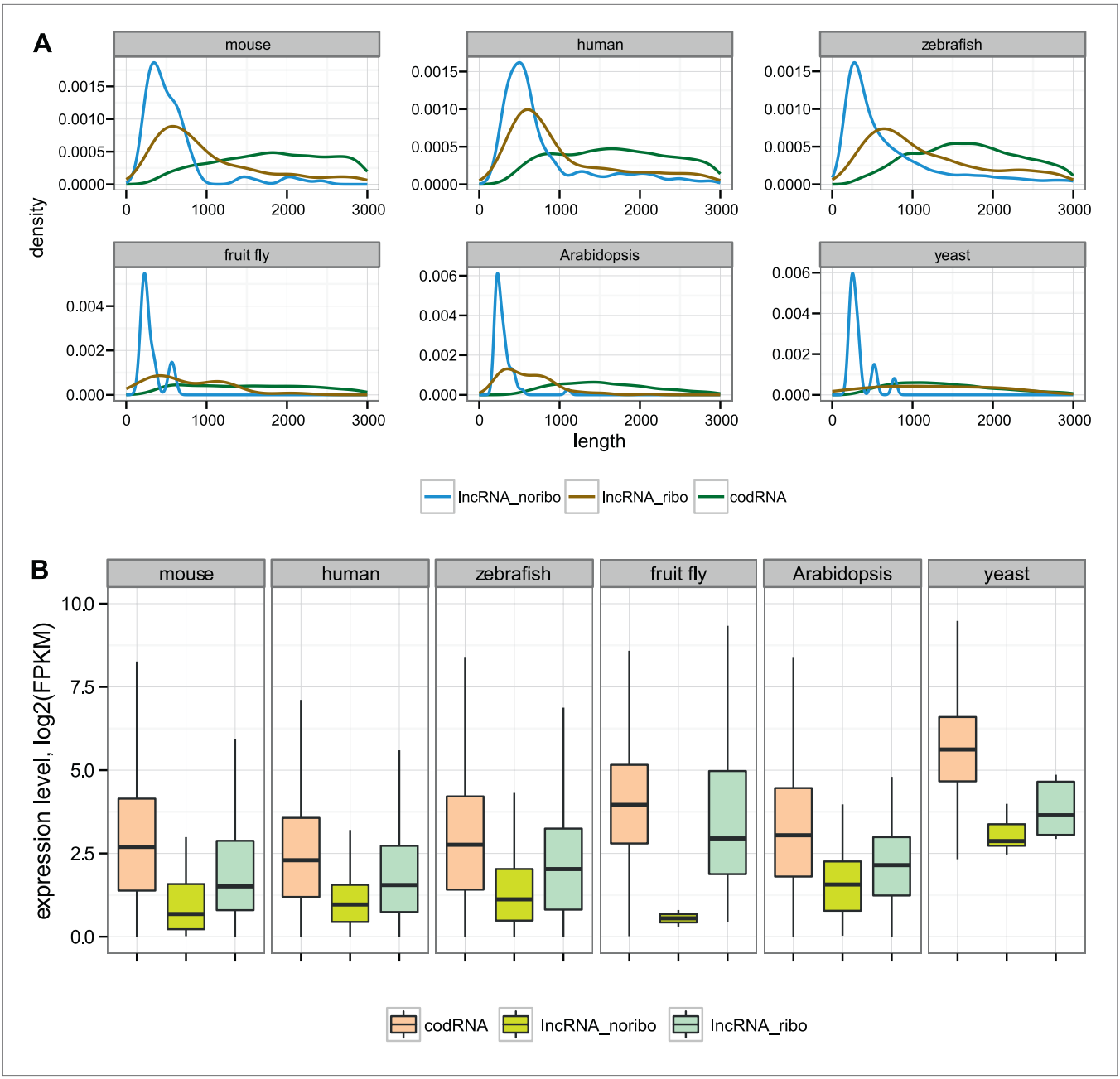

Figure 1. General characteristics of codRNA and lncRNA transcripts. (**A**) Density plots of transcript length. (**B**) Box-plots of transcript expression level in log2(FPKM) units. lncRNA_ribo: lncRNAs associated with ribosomes; lncRNA_noribo: lncRNAs for which association with ribosomes was not detected. codRNA: coding transcripts encoding experimentally validated proteins except for zebrafish in which all transcripts annotated as coding were considered. The area within the box-plot comprises 50% of the data and the line represents the median value. In all studied species, codRNAs were expressed at higher levels than lncRNAs (Wilcoxon test, $p < 10^{-5}$), and lncRNA_ribo at higher levels than lncRNA_noribo (Wilcoxon test, $p < 0.005$).

DOI: 10.7554/eLife.03523.005

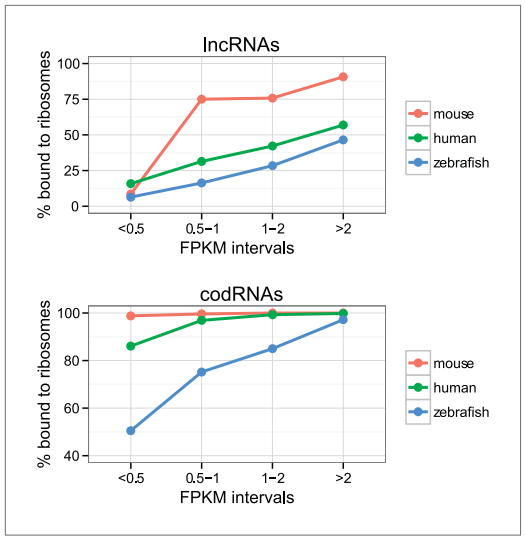

Figure 2. Effect of transcript expression level on the detection of ribosome association. The percentage of transcripts associated with ribosomes is shown for several transcript expression intervals. codRNA: annotated coding transcripts encoding experimentally verified proteins (except in zebrafish for which all coding transcripts were considered). lncRNA: annotated and novel long non-coding RNAs. Only species with at least 20 transcripts in each expression bin were plotted. In the rest of species, the data were consistent with the trends shown.

DOI: 10.7554/eLife.03523.007

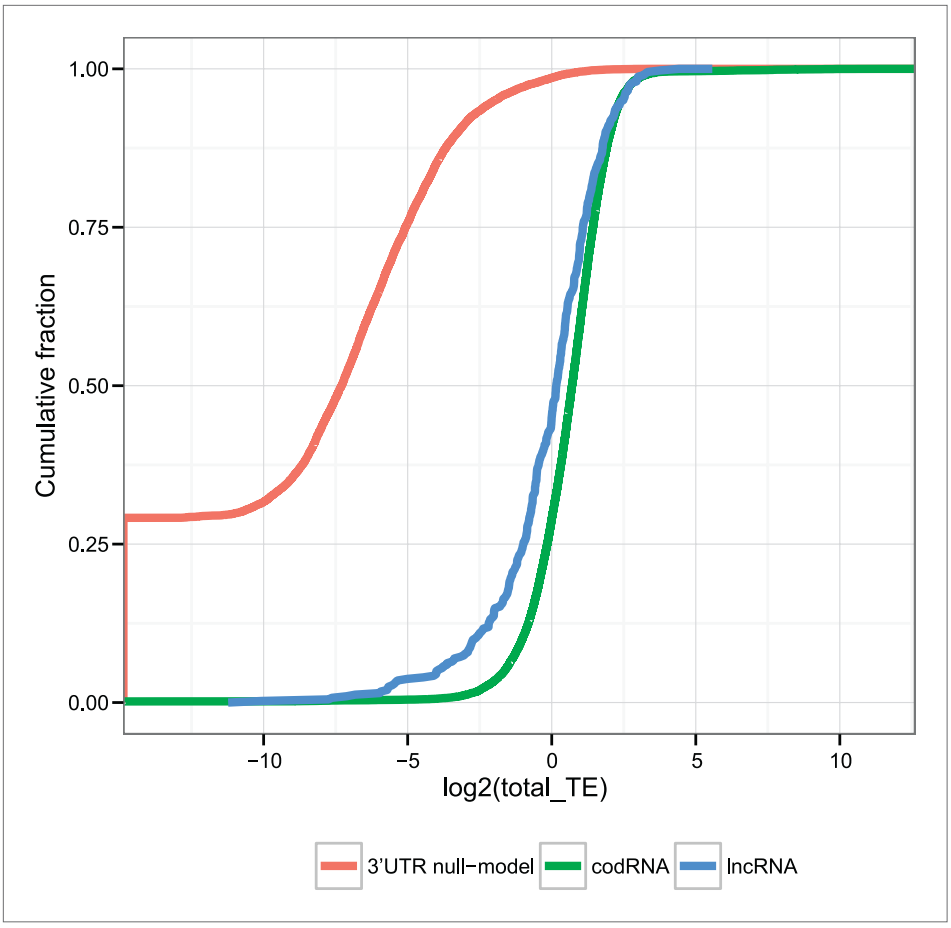

Figure 3. TE distribution in human transcripts and 3'UTRs (null-model). Cumulative distribution of TE values in human codRNAs, lncRNAs, and 3'UTR sequences. We randomly selected 3'UTRs with a minimum length of 30 nucleotides to build a set of 3'UTR sequences with the same size distribution as the complete transcripts.

DOI: 10.7554/eLife.03523.008

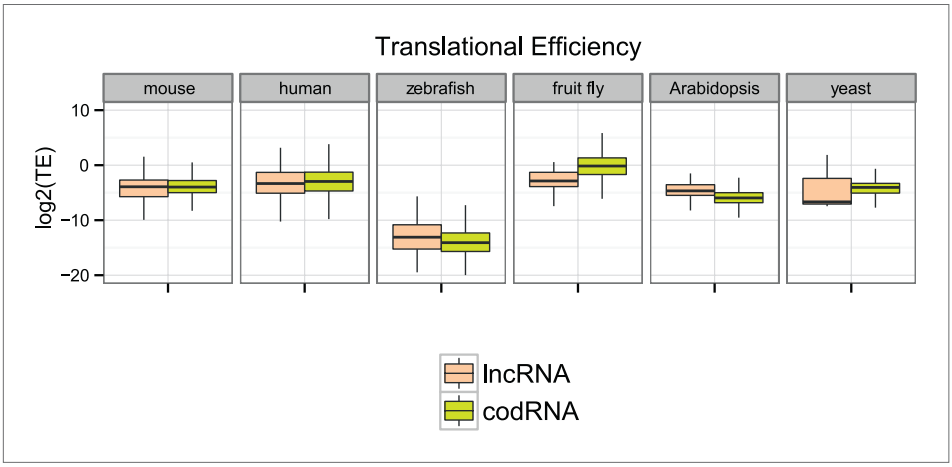

Figure 4. Ribosome association profiles for codRNAs and lncRNAs. Box-plots of transcript translational efficiency (TE) in log2(TE) units. The area within the box-plot comprises 50% of the data, and the line represents the median value. lncRNA: lncRNAs for which association with ribosomes was detected. codRNA: coding RNAs transcripts encoding experimentally validated proteins except for zebrafish in which all transcripts annotated as coding were considered.

DOI: 10.7554/eLife.03523.009

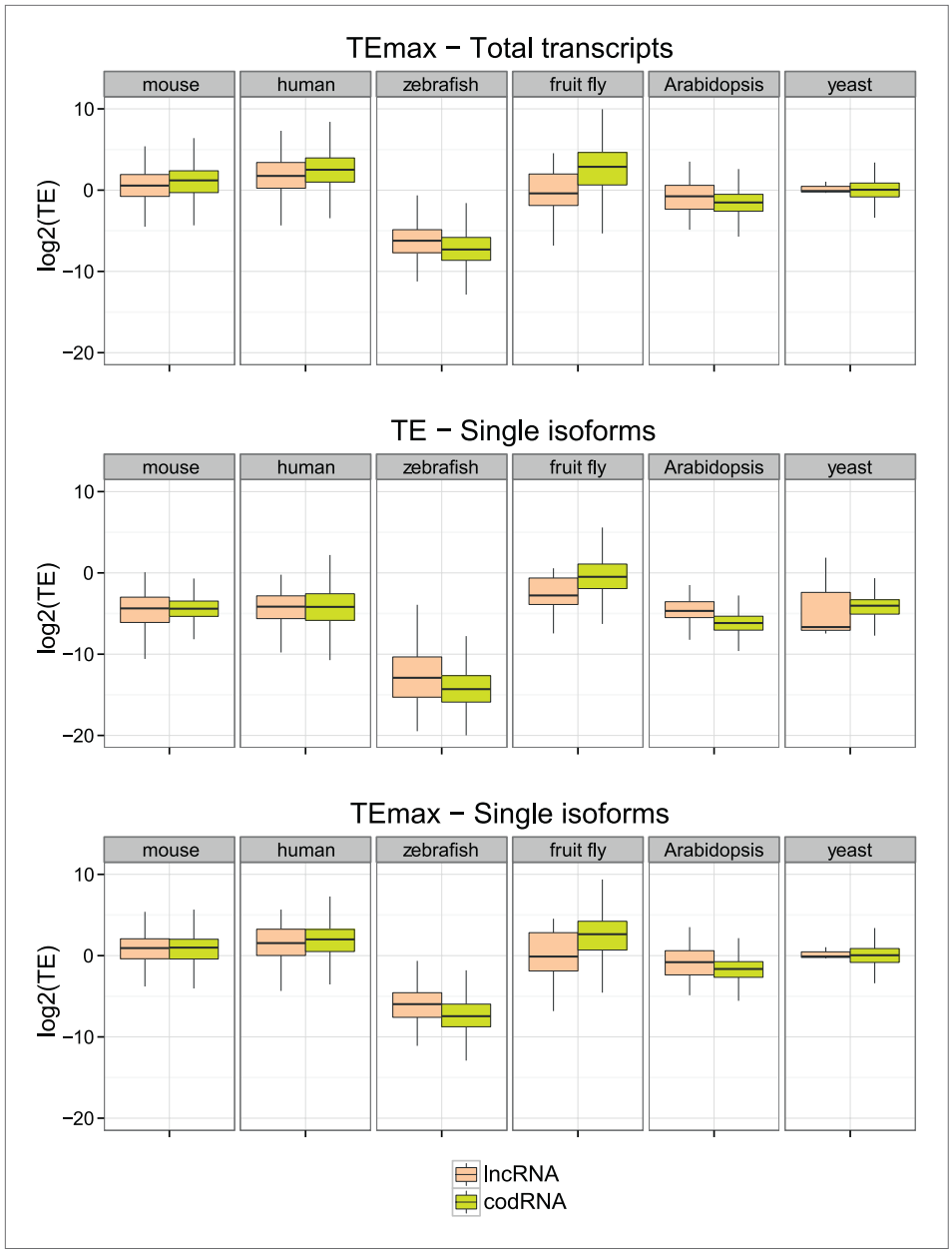

Figure 4—figure supplement 1. Additional translational efficiency (TE) measures. Single isoforms correspond to data for genes with a single transcript. The number of such genes was 2961 codRNA and 246 lncRNA_ribo for mouse, 2853 codRNA and 150 lncRNA_ribo for human, 9352 codRNA and 412 lncRNA_ribo for zebrafish, 836 codRNA and 18 lncRNA_ribo for fruit fly, and 3024 codRNA and 92 lncRNA_ribo for Arabidopsis. In the case of yeast, all genes were taken. TE max is the maximum TE value taking 90 nucleotide windows.

DOI: 10.7554/eLife.03523.010

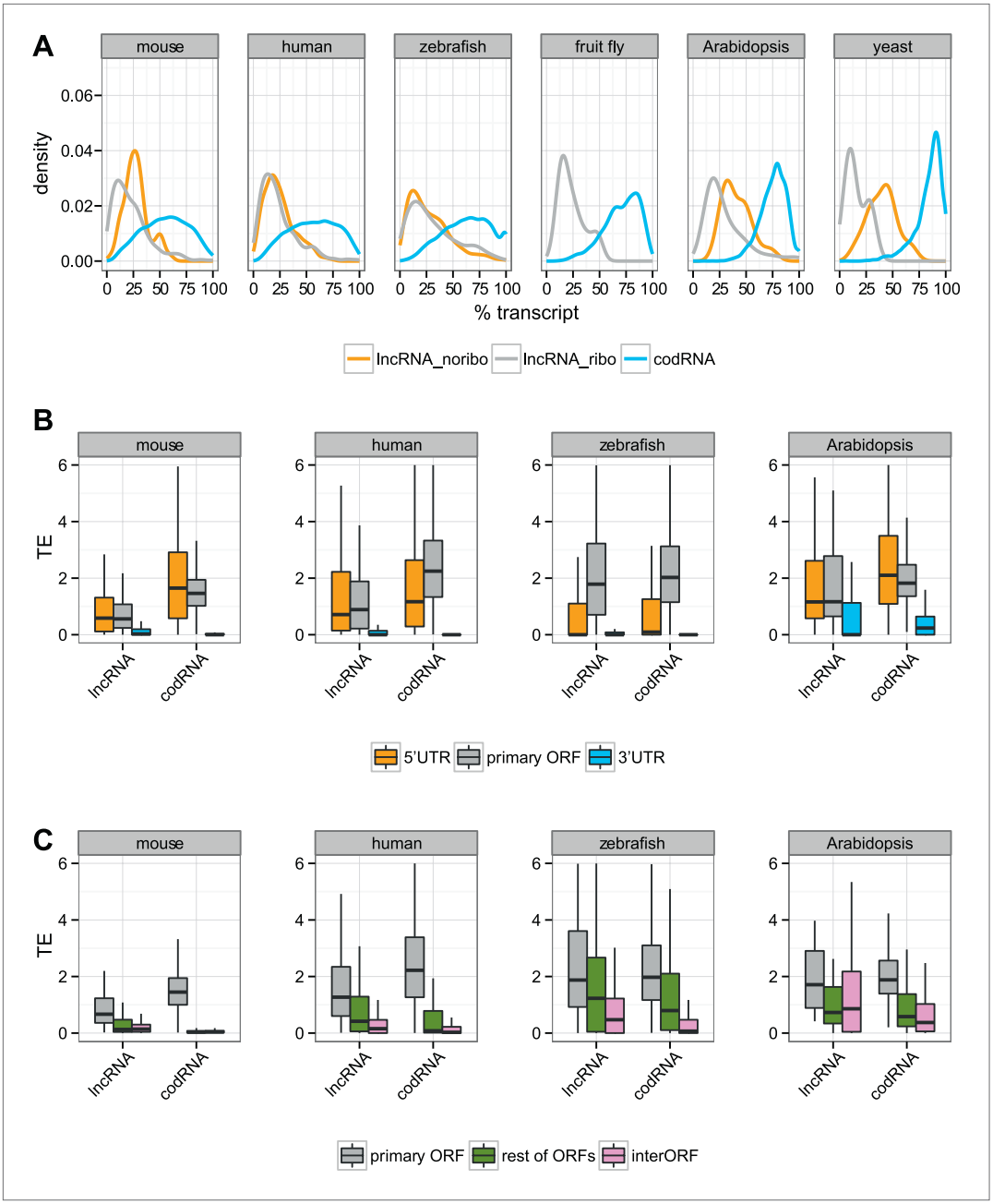

Figure 5. Ribosome association in different transcript regions. (**A**) Density plot of the relative length of the primary ORF in lncRNA_ribo and codRNA with respect to transcript length. For comparison data for the longest ORF in lncRNA_noribo is also shown (except for fruit fly due to insufficient data). (**B**) Box-plots of TE distribution in primary ORF, 5′UTR, and 3′UTR regions. The area within the box-plot comprises 50% of the data, and the line represents the median value. The analysis considered all transcripts with 5′UTR and 3′UTR longer than 30 nucleotides and >0.2 FPKM in all three regions. The number of transcripts was 1956 codRNA and 159 lncRNA_ribo in mouse, 3558 codRNA and 139 lncRNA_ribo in human, 5216 codRNA and 252 lncRNA_ribo in zebrafish, and 2019 codRNA and 33 lncRNA_ribo in Arabidopsis. (**C**) Box-plots of TE distribution in primary ORFs, rest of ORFs with ribosome profiling reads and non-ORF regions (interORF). The analysis considered all transcripts with at least two ORFs and more than 30 nucleotides interORF. The number of transcripts was 3264 codRNA and 204 lncRNA_ribo in mouse, 3104 codRNA and 168 lncRNA_ribo in human, 1646 codRNA and 212 lncRNA_ribo in zebrafish, and 1098 codRNA and 25 lncRNA_ribo in Arabidopsis. Fruit fly and yeast were not included in the last two analyses due to insufficient data (<8 lncRNA_ribo meeting the conditions).

DOI: 10.7554/eLife.03523.011

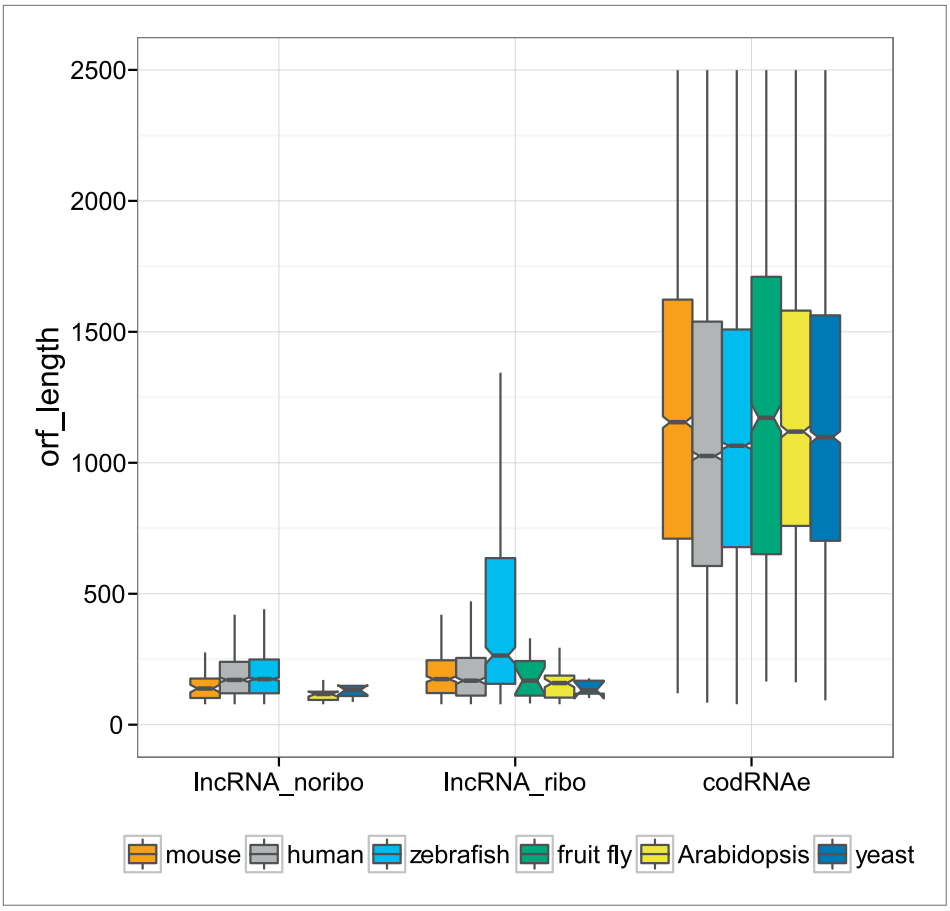

Figure 5—figure supplement 1. Absolute nucleotide length of ORFs in different kinds of transcripts. In codRNAs and lncRNA_ribo, we selected the primary ORF (the ORF with the largest number of ribosome profiling reads), whereas in lncRNA_noribo we selected the longest ORF.

DOI: 10.7554/eLife.03523.012

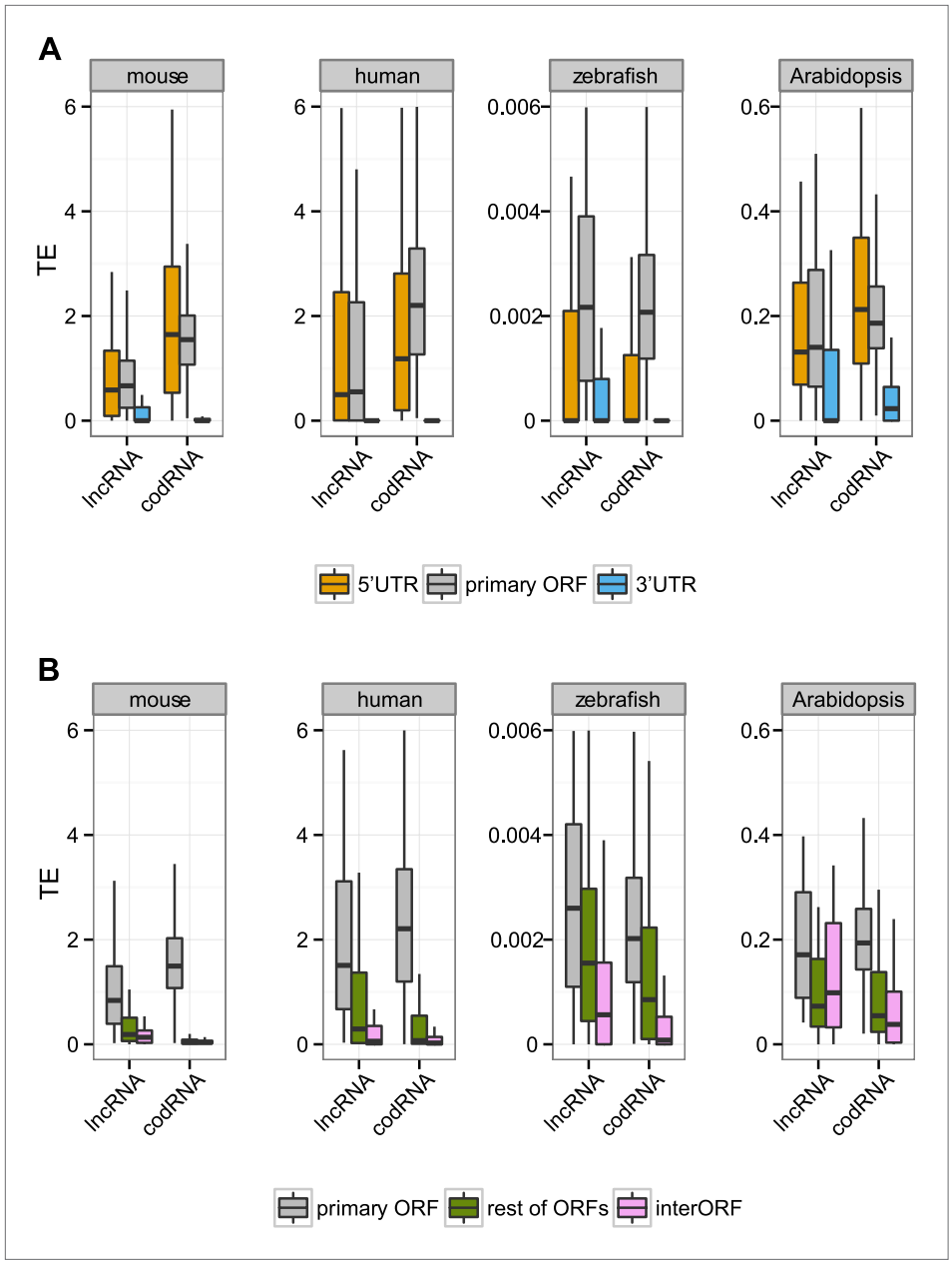

Figure 5—figure supplement 2. Translational efficiency in single-isoform genes. (**A**) Box-plots of TE distribution in primary ORF, 5'UTR, and 3'UTR regions. The analysis considered only genes with one isoform, with UTR and ORF regions expressed at >0.2 FPKM and with 5'UTR and 3'UTR longer than 30 nucleotides. The number of transcripts was 980 codRNA and 97 lncRNA_ribo in mouse, 758 codRNA and 36 lncRNA_ribo in human, 3763 codRNA and 117 lncRNA_ribo in zebrafish, and 1495 codRNA and 32 lncRNA_ribo in Arabidopsis. (**B**) Box-plots of TE distribution in primary ORFs, other ORFs with ribosome profiling reads and non-ORF regions (interORFs). The analysis only considered genes with one isoform in which these regions were longer than 30 nucleotides and with expression >0.2 FPKM. The number of transcripts was 1691 codRNA and 113 lncRNA_ribo in mouse, 763 codRNA and 54 lncRNA_ribo in human, 1170 codRNA and 108 lncRNA_ribo in zebrafish, and 817 codRNA and 25 lncRNA_ribo in Arabidopsis.

DOI: 10.7554/eLife.03523.013

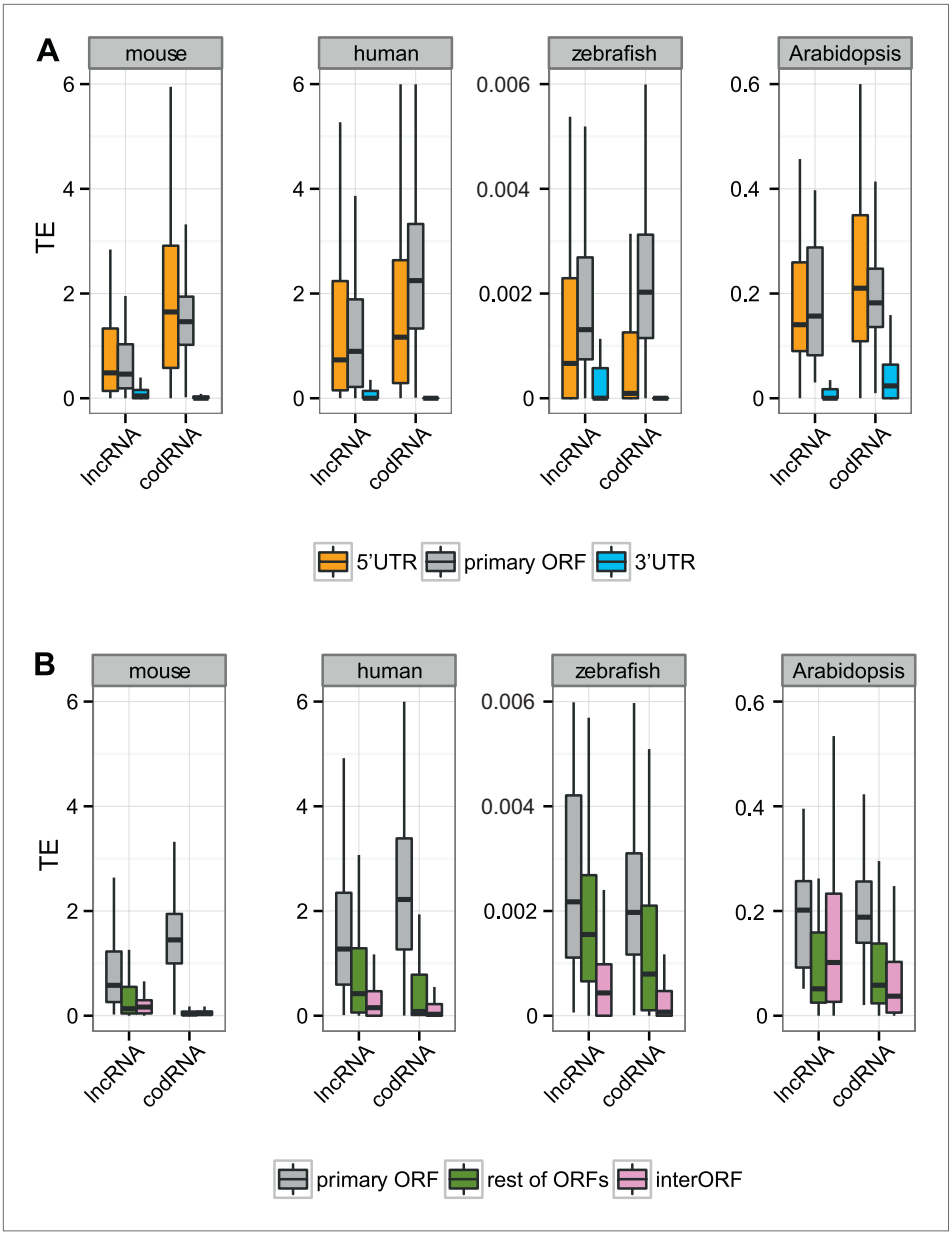

Figure 5—figure supplement 3. Translational efficiency in annotated transcripts. (**A**) Box-plots of TE distribution in primary ORF, 5'UTR, and 3'UTR regions. The analysis considered only annotated transcripts, with UTR and ORF regions expressed at >0.2 FPKM and with 5'UTR and 3'UTR longer than 30 nucleotides. The number of transcripts was 1956 codRNA and 92 lncRNA_ribo in mouse, 3558 codRNA and 138 lncRNA_ribo in human, 5216 codRNA and 54 lncRNA_ribo in zebrafish, and 2019 codRNA and 22 lncRNA_ribo in Arabidopsis. (**B**) Box-plots of TE distribution in primary ORFs, other ORFs with ribosome profiling reads (rest ORFs) and non-ORF regions (interORF). The analysis only considered annotated transcripts in which these regions were longer than 30 nucleotides and with expression >0.2 FPKM. The number of transcripts was 3264 codRNA and 128 lncRNA_ribo in mouse, 3104 codRNA and 167 lncRNA_ribo in human, 1646 codRNA and 58 lncRNA_ribo in zebrafish, and 1098 codRNA and 18 lncRNA_ribo in Arabidopsis.

DOI: 10.7554/eLife.03523.014

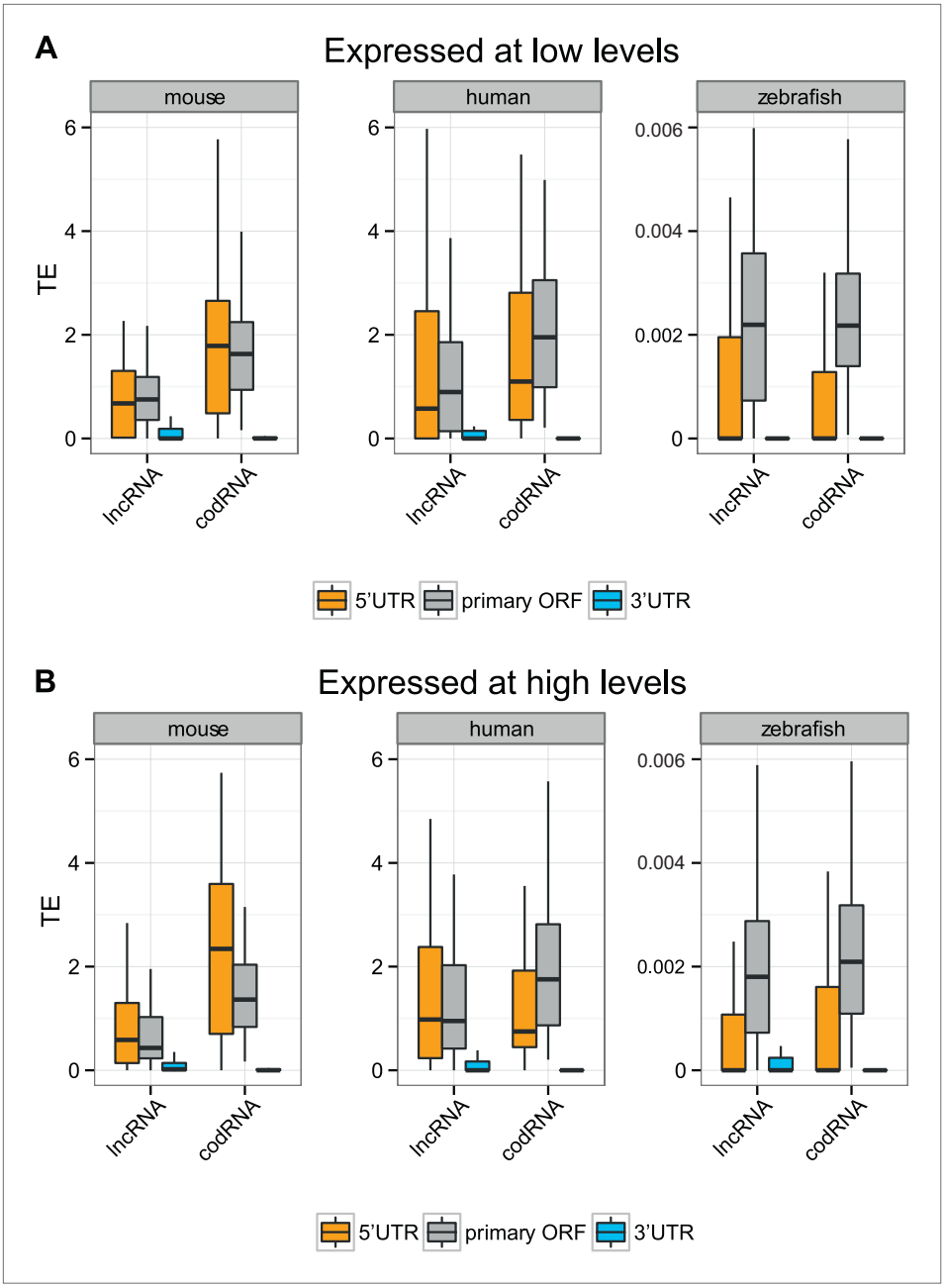

Figure 5—figure supplement 4. Translational efficiency in transcripts expressed at different levels. We restricted this analysis to transcripts with ORF and UTR regions expressed at >0.2 FPKM and with 5′UTR and 3′UTR longer than 30 nucleotides. (**A**) Expressed at low levels: transcripts expressed at 0.5–2 FPKM, (**B**) expressed at high levels: transcripts expressed at 2–10 FPKM. codRNAs were sampled in such a way as to have the same gene expression distribution as the corresponding lncRNA set. Results for species in which all sets contained at least 20 transcripts are shown.

DOI: 10.7554/eLife.03523.015

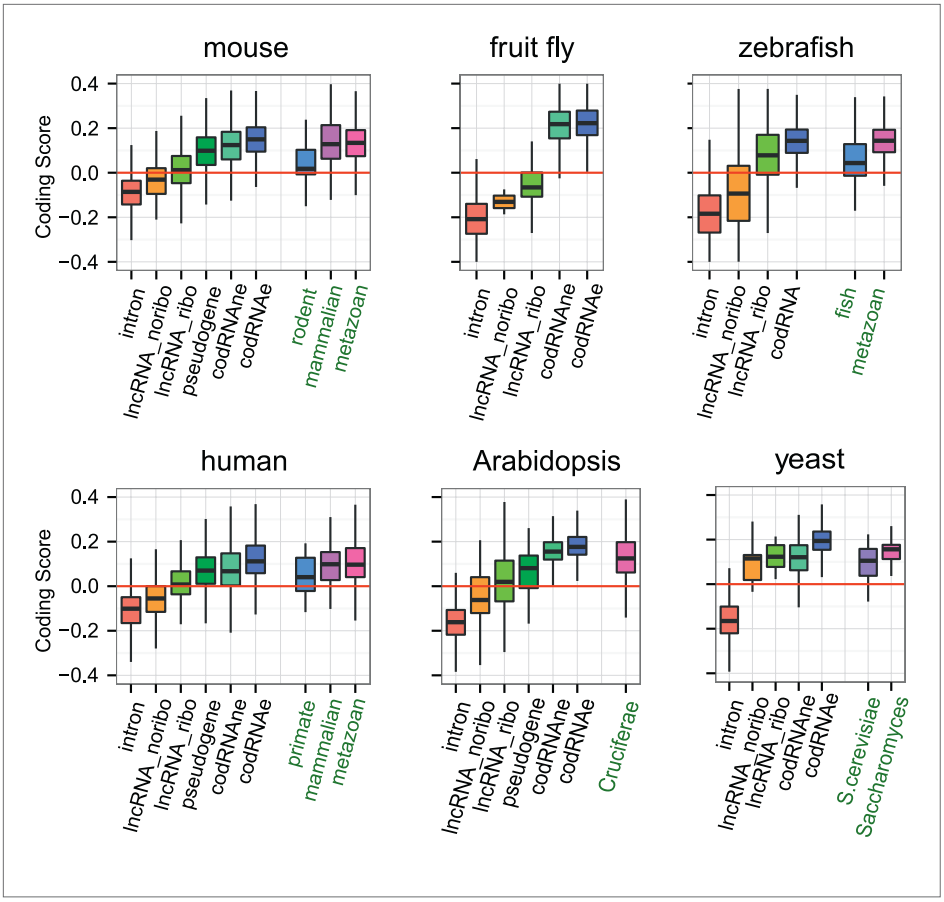

Figure 6. Coding scores in ORFs from different types of transcripts. Intron: randomly selected intronic regions; lncRNA_noribo: lncRNAs not associated with ribosomes; lncRNA_ribo: lncRNAs associated with ribosomes; pseudogene: pseudogenes associated with ribosomes; codRNAne: coding transcripts encoding non-validated proteins associated with ribosomes; codRNAe: coding transcripts encoding experimentally validated proteins. The coding score was calculated as the log ratio of hexamer frequencies in coding vs intronic sequences. In lncRNA_noribo and introns, we considered the longest ORF and in the rest of transcripts the primary ORF. The Class 'pseudogene' was only included in species with more than 20 expressed pseudogenes with mapped ribosome profiling reads. The coding score of the primary ORF in lncRNAs (lncRNA_ribo) was significantly higher than the coding score in ORFs defined in introns (Wilcoxon test, human, mouse, zebrafish, and Arabidopsis $p < 10^{-16}$; fruit fly and yeast $p < 10^{-4}$, Wilcoxon test) and in lncRNA_ribo it was significantly higher than in lncRNA_noribo in four species (Wilcoxon test, human, mouse and zebrafish $p < 10^{-5}$, and Arabidopsis $p < 0.05$). Transcripts from genes of different evolutionary age were taken from the literature (see manuscript text). The number of transcripts was 68 for rodent, 127/123 for mammalian (mouse/human as reference species), 11,203/13,423/9812 for metazoan (mouse/human/zebrafish), 162 for fish, 208 for Crucifera, 28 for *S. cerevisiae* and 84 for Saccharomyces. The youngest class of codRNAs displayed similar scores than lncRNA_ribo in mouse, zebrafish, and yeast (classes rodent, fish and *S. cerevisiae*, respectively), being only significantly higher in human and Arabidopsis (Wilcoxon test, $p < 0.005$; classes primate and Cruciferae). We did not analyze young genes in fruit fly due to lack of a suitable young set of codRNAs in this species.

DOI: 10.7554/eLife.03523.016

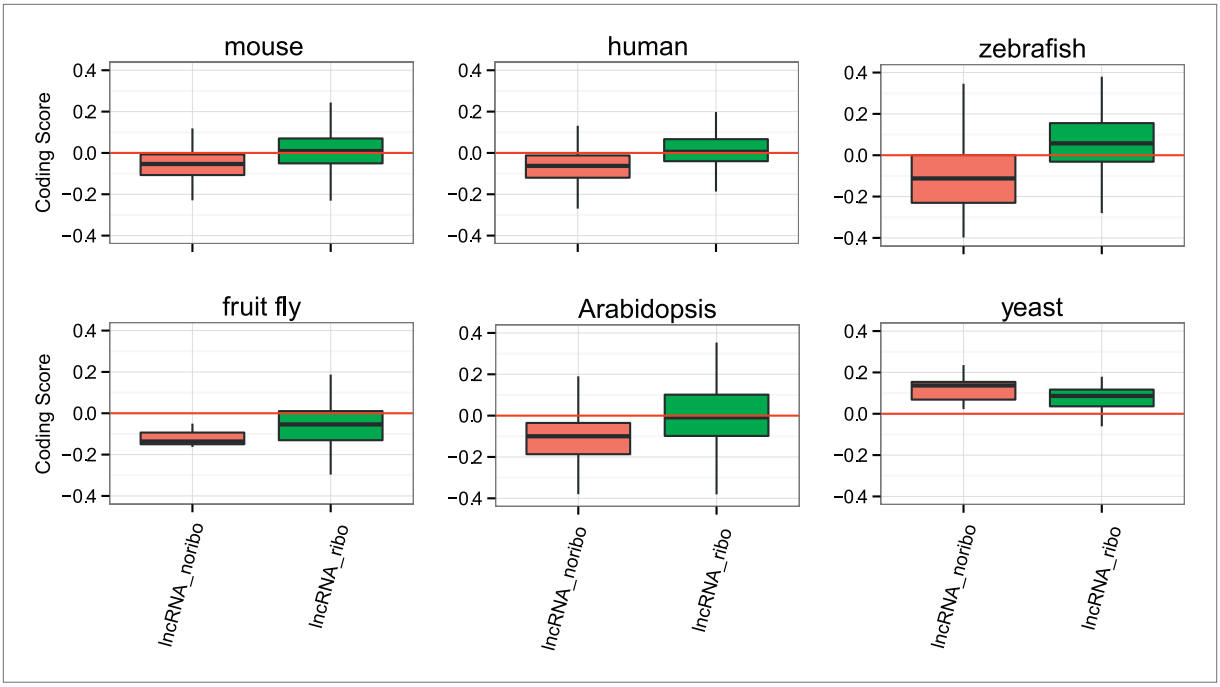

Figure 6—figure supplement 1. Coding scores for the longest ORF. Comparison between lncRNAs associated and not associated with ribosomes using the longest ORF in both cases (lncRNA_ribo and lncRNA_noribo, respectively). Differences between lncRNA_ribo and lncRNA_noribo are significant by a Wilcoxon test ($p < 10^{-10}$ in human, mouse, and zebrafish; $p < 0.005$ in Arabidopsis).

DOI: 10.7554/eLife.03523.017

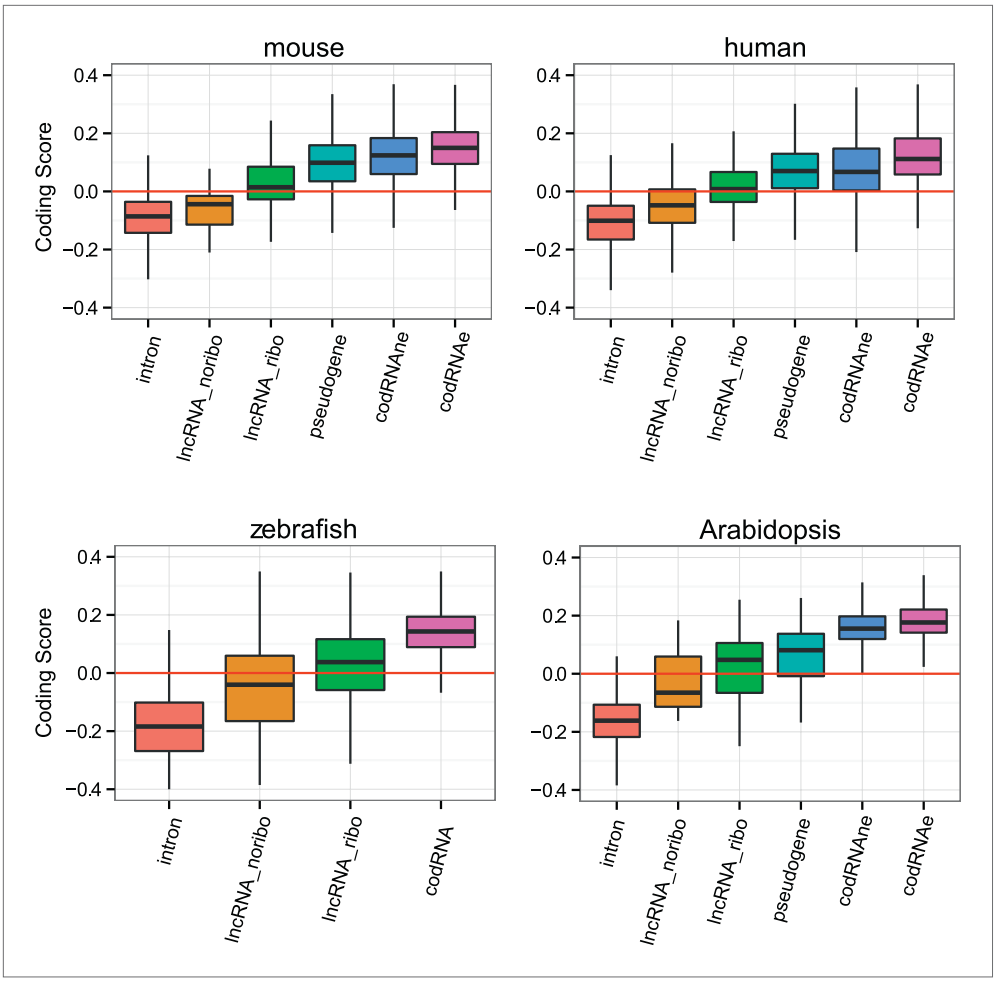

Figure 6—figure supplement 2. Coding scores in different classes of annotated sequences. Comparison between different transcript classes using only annotated lncRNAs. Yeast transcriptome is composed of very few annotated lncRNAs, and this analysis could not be performed.

DOI: 10.7554/eLife.03523.018

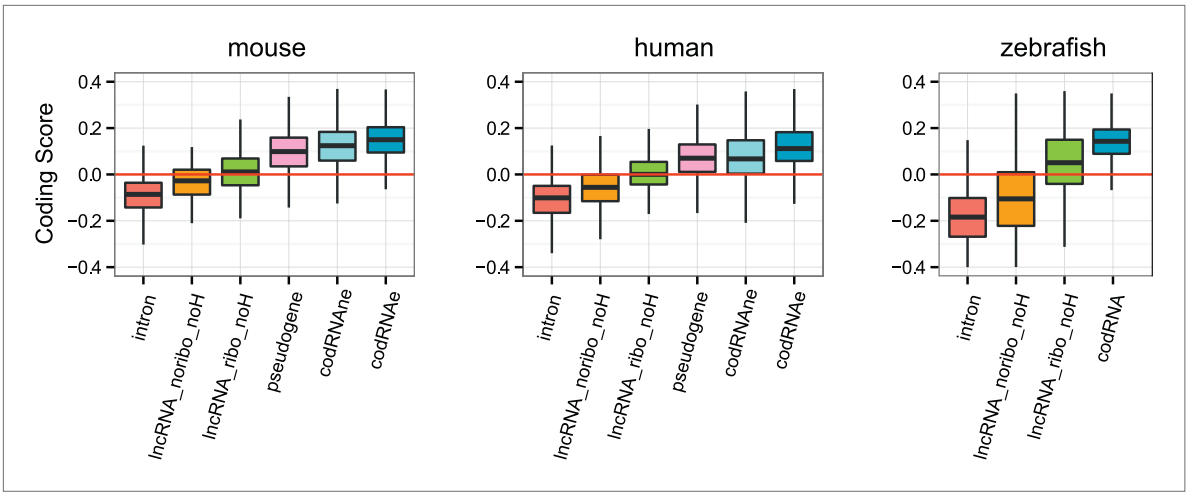

Figure 6—figure supplement 3. Coding scores in lncRNAs without homologues in other species. Comparison between different transcript classes using only lncRNA with no homologues (noH) in other species. Only species in which several lncRNA_ribo and lncRNA_noribo had homology matches were considered.

DOI: 10.7554/eLife.03523.019

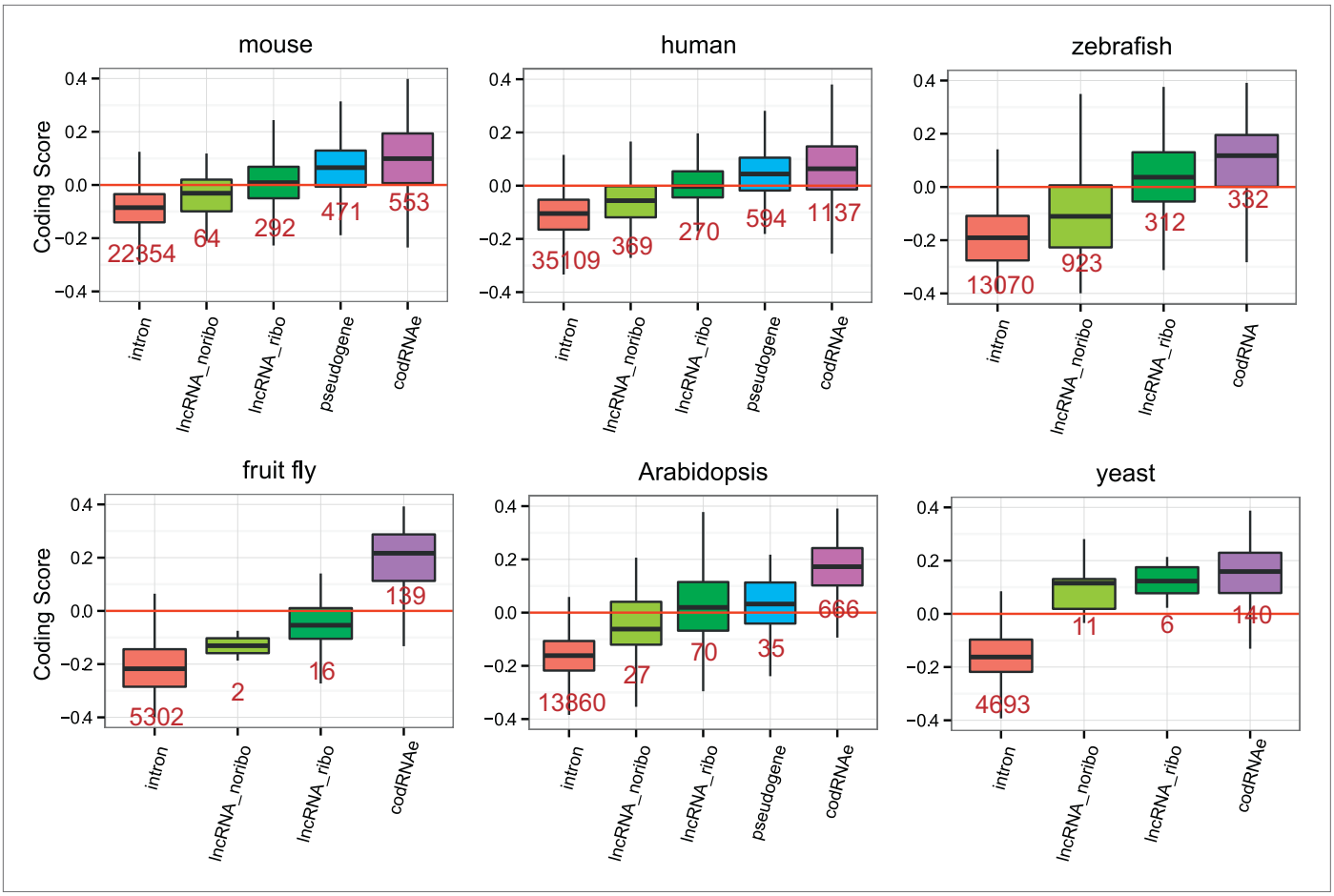

Figure 6—figure supplement 4. Coding scores in small ORFs from different types of transcripts. Here we only employed lncRNAs in which the primary ORF was shorter than 100 amino acids. codRNA refers to joined codRNAe and codRNAne sets, since experimentally verified proteins are usually longer than 100 amino acid. The number of transcripts is shown in red.

DOI: 10.7554/eLife.03523.020

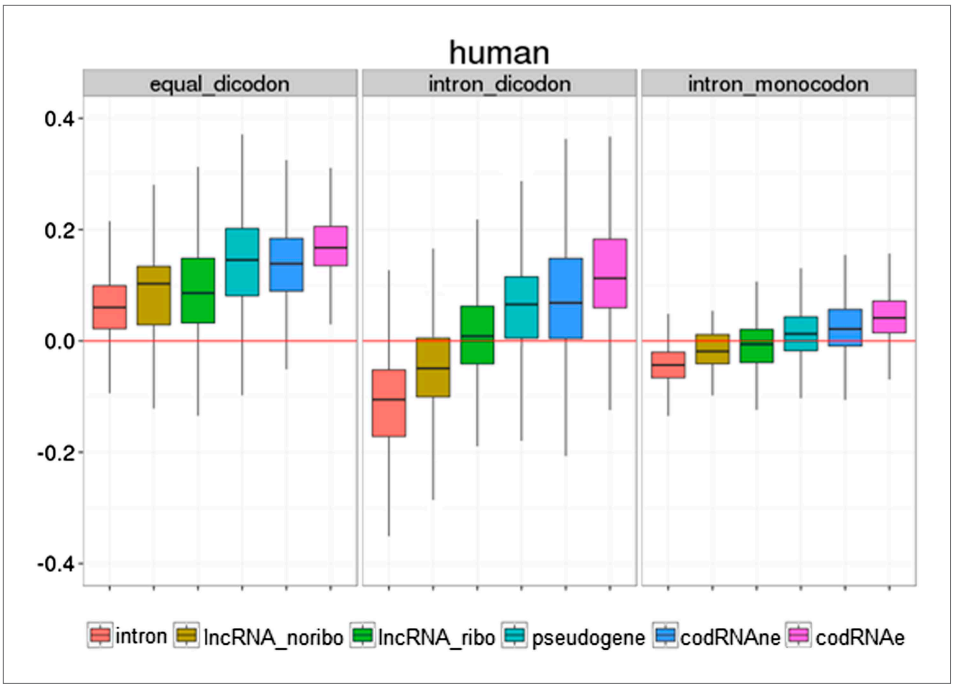

Figure 6—figure supplement 5. Use of different coding statistics in human transcripts. Equal dicodon was based on the observed hexamer frequencies in coding sequences vs hexamer equiprobability, intron dicodon was based on the differences between hexamer frequencies in coding vs non-coding sequences and intron_monocodon was based on the observed codon frequences in coding sequences vs codon equiprobability.

DOI: 10.7554/eLife.03523.021

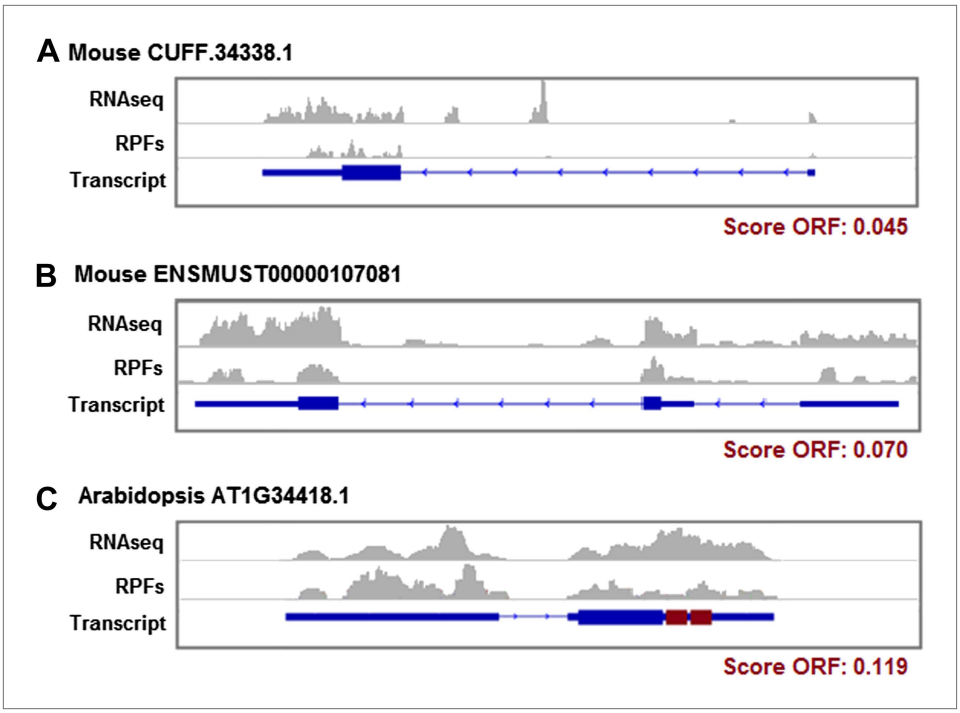

Figure 6—figure supplement 6. Ribosome protection patterns in transcripts containing short ORFs. (**A**) Mouse CUFF.34338.1 (chr5:113183493–113188347) is a novel lncRNA, it contains an ORF encoding a 169 amino acid protein associated with ribosomes and with protein-coding homologues in human, zebrafish, and yeast. (**B**) ENSMUST00000107081 is an annotated codRNA in mouse which evolved recently since no homologues were found in any other species. It has a small ORF that translates a 55 amino acid protein. (**C**) AT1G34418.1 is an annotated lncRNA in Arabidopsis showing abundant association with ribosomes in the 5′UTR region, the primary ORF (34 amino acid) and the final region of the transcript, which contains two redundant ORFs (in red) coding the sequence: MGLGFVN(V/F)LLGM. RNAseq: profile of RNAseq reads. RPFs: profile of ribosome profiling reads. Exon-intron transcript structures are represented; the thickest boxes on the exons are the primary ORFs.
DOI: 10.7554/eLife.03523.022

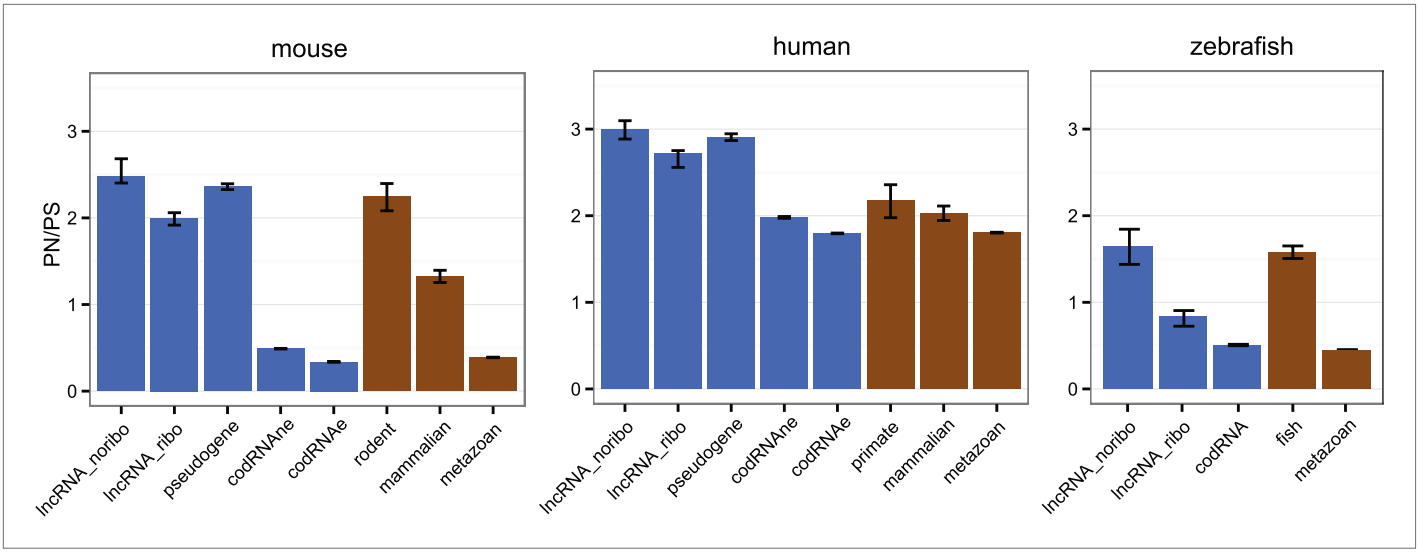

Figure 7. Selective pressure in ORFs from different types of transcripts. PN/PS: ratio between the number of non-synonymous and synonymous single nucleotide polymorphisms (SNPs) in the complete set of primary ORFs for a given class of transcripts (in lncRNA_noribo the longest ORF was considered). In blue, data for different coding and non-coding transcript classes. In brown, data for different age codRNA classes. The bars represent the 95% confidence interval for the PN/PS value. For the species not shown there was not sufficient data to perform this analysis.

DOI: 10.7554/eLife.03523.023