SUPPLEMENTARY FILE 1. Ruiz-Orera et al. Long non-coding RNAs as a source of new peptides. May 16th 2014.

Table 1. Properties of transcripts and open reading frames (ORFs). a) Length of transcripts in different species. b) Length of transcripts associated with ribosomes. c) Density and length of non-overlapping ORFs in different species and transcript classes (in nucleotides). d) Density and length of non-overlapping ORFs associated with ribosomes (in nucleotides). ORFs were defined in all sequences as starting with ATG, finishing with a STOP codon, and being at least 24 amino acids long. In the cases that two ORFs overlapped, the longest one was considered. Annotated coding transcripts (codRNA) correspond to mRNAs encoding experimentally validated proteins except for zebrafish, for which all transcripts annotated as coding were considered. Only transcripts with at least one ORF were considered. In c and d, yeast was not considered here because most codRNAs are covered by a single ORF and UTR annotations are incomplete. In all cases, only transcripts expressed at > 0.5 FPKM were considered. See Table 1 in main manuscript file for details on the experimental data. Av: average; Med: median; SD: standard deviation

a)

Transcript length									
All transcripts	**Novel lncRNA**			**Annotated lncRNA**			**codRNA**		
	Av	**Med**	**SD**	**Av**	**Med**	**SD**	**Av**	**Med**	**SD**
Mouse	1645	672	2425	1322	832	1239	3010	2406	2232
Human	1131	636	1301	1167	666	1221	3162	2772	1990
Zebrafish	1364	670	850	961	714	1732	2415	1978	1674
Fruit fly	371	252	320	685	517	369	2686	2240	1950
Arabidopsis	319	286	131	749	706	353	1825	1605	1019
Yeast	361	303	166	382	369	134	1566	1302	1156

b)

Transcript length									
Transcripts bound to ribosomes	**Novel lncRNA**			**Annotated lncRNA**			**codRNA**		
	Av	**Med**	**SD**	**Av**	**Med**	**SD**	**Av**	**Med**	**SD**
Mouse	1996	867	2660	1399	911	1281	3162	2772	1990
Human	629	629	298	1269	803	1156	3020	2419	2235
Zebrafish	2151	1340	2167	1280	872	1019	2480	2031	1698
Fruit fly	553	361	451	708	527	370	2686	2240	1950
Arabidopsis	368	330	157	781	720	344	1828	1607	1019
Yeast	399	359	184	446	446	108	1579	1306	1164

c)

Open Reading Frame length												
All ORFs	Novel lncRNA				Annotated lncRNA				codRNA			
	ORFs/ kb	ORF length			ORFs/ kb	ORF length			ORFs/ kb	ORF length		
		Av	Med	SD		Av	Med	SD		Av	Med	SD
Mouse	2.66	173	130	133	2.65	158	132	173	1.37	516	148	949
Human	2.65	136	118	59	2.69	166	127	187	1.36	541	148	1059
Zebrafish	2.64	194	124	346	2.59	204	124	299	1.33	598	151	992
Fruit fly	3.94	141	115	94	2.87	137	110	56	3.11	992	240	1467
Arabidopsis	3.80	130	114	46	2.64	136	114	63	0.95	1060	870	1027

d)

Open Reading Frame length												
ORFs bound to ribosomes	Novel lncRNA				Annotated lncRNA				codRNA			
	ORFs/ kb	ORF length			ORFs/ kb	ORF length			ORFs/ kb	ORF length		
		Av	Med	SD		Av	Med	SD		Av	Med	SD
Mouse	1.49	177	142	245	1.81	213	142	167	0.91	805	225	1180
Human	0.14	159	137	69	0.98	216	151	319	0.71	1251	807	1522
Zebrafish	0.53	460	216	687	1.09	302	154	453	0.66	1347	1065	1288
Fruit fly	2.01	154	118	114	2.58	139	110	57	0.92	1192	507	1558
Arabidopsis	2.50	136	114	50	2.02	138	114	66	0.90	1120	951	1030

Table 2. Details on the number of coding transcripts associated with ribosomes. codRNAe: Transcripts encoding experimentally validated proteins. codRNAne: Transcripts encoding non experimentally validated proteins. In the case of zebrafish codRNAe and codRNAne were grouped together due to the low number of experimentally validated proteins. In all cases, only transcripts expressed at > 0.5 FPKM were considered.

	codRNAe	**codRNAne**	**Pseudogene**
Mouse	5,462/5,465 (99.95%)	8,734/8,780 (99.48%)	1,429/1,587 (90.04%)
Human	10,830/10,902 (99.34%)	5,800/6,109 (94.94%)	2,657/3,424 (77.60%)
Zebrafish	11,643/12,595 (92.4%)		11/23 (47.82%)
Fruit fly	1,450/1,450 (100.00%)	6,581/6,595 (99.79%)	0/1 (0.00%)
Arabidopsis	3,874/3,888 (99.64%)	15,005/15,274 (98.24%)	70/94 (74.47%)
Yeast	4,530/4,649 (97.44%)	579/824 (70.27%)	5/7 (71.42%)

Table 3. Details on the number of non-coding transcripts associated with ribosomes. The main lncRNA classes annotated in Ensembl annotations (v.70) are displayed in this table for mouse, human and zebrafish. Annotated and novel lncRNA classes are shown separately. In all cases only transcripts expressed at > 0.5 FPKM were considered.

	Annotated in Ensembl					Novel
Species dataset	**lincRNA**	**Processed transcript**	**Non-coding retained intron**	**Antisense**	**Sense-intronic / overlapping**	
Mouse	130 / 152	60 / 67	9 / 9	19 / 25	-	172 / 223
Human	152 / 399	164 / 238	22 / 30	65 / 162	0 / 21	2 / 86
Zebrafish	17 / 51	108 / 256	4 / 8	1 / 4	-	596 / 2070

Table 4. Number of "single isoform" genes. Number of genes with a unique transcript isoform and associated with ribosomes. For these genes the full detected ribosome profiling signal belongs to a single transcript. In yeast, all genes have a single isoform and therefore this class was not defined.

Species	codRNA	lncRNA_ribo
Mouse	2961	246
Human	2853	150
Zebrafish	9352	412
Fruit fly	836	53
Arabidopsis	3024	92

Table 5. Translational properties of transcript 3' UTRs. Number of transcripts (total, annotated and single-isoform) with UTR and ORF regions expressed at > 0.2 FPKM and with 5'UTR and 3'UTR longer than 30 nucleotides. TEorf/TEutr: ratio that quantifies the differences in median translational efficiency in primary ORF and 3' UTR region. In parentheses, percent of transcripts in which at least one mapped RPF was found in the 3' UTR region and therefore the ratio was computed. In fruit fly, ratio was not computed in lncRNAs due to the low number of considered transcripts.

Number of transcripts

	Total transcripts		Annotated transcripts		Single-isoform transcripts	
Species	**codRNA**	**lncRNA_ribo**	**codRNA**	**lncRNA_ribo**	**codRNA**	**lncRNA_ribo**
Mouse	1956	159	1956	92	980	97
Human	3558	139	3558	138	758	36
Zebrafish	5216	252	5216	54	3763	117
Fruit fly	875	5	875	5	458	2
Arabidopsis	2019	33	2019	22	1495	32

Ratio TEorf / TEutr

	Total transcripts		Annotated transcripts		Single-isoform transcripts	
Species	**codRNA**	**lncRNA_ribo**	**codRNA**	**lncRNA_ribo**	**codRNA**	**lncRNA_ribo**
Mouse	41.94 (48.67%)	3.50 (54.08%)	41.94 (48.67%)	4.32 (59.78%)	40.77 (44.90%)	2.72 (47.42%)
Human	44.92 (15.68%)	3.09 (36.69%)	44.92 (15.68%)	3.28 (36.23%)	39.77 (11.35%)	0.45 (16.67%)
Zebrafish	5.77 (8.64%)	1.14 (32.54%)	5.77 (8.64%)	0.96 (45.59%)	4.79 (7.54%)	1.30 (38.46%)
Fruit fly	149.94 (70.51%)	- (80.00%)	149.94 (70.51%)	- (80.00%)	112.94 (76.86%)	- (100.00%)
Arabidopsis	3.98 (70.63%)	0.82 (54.55%)	3.98 (70.63%)	0.82 (40.91%)	4.02 (68.22%)	0.83 (56.25%)

Table 6. Number of transcripts with multiple ORFs associated with ribosomes. Number of transcripts (total, annotated and single-isoform) with at least two predicted ORFs associated with ribosomes and an interORF region longer than 30 nucleotides. interORF is the sum of all regions between non-overlapping ORFs except for transcript UTR regions. In yeast, ORFs from codRNAs spanned the whole transcript in annotations isoform and therefore this class was not defined since no additional ORFs may be detected in this class of transcripts.

	Total transcripts		Annotated transcripts		Single-isoform transcripts	
Species	**codRNA**	**lncRNA_ribo**	**codRNA**	**lncRNA_ribo**	**codRNA**	**lncRNA_ribo**
Mouse	3264	204	3264	128	1691	113
Human	3104	168	3104	167	763	54
Zebrafish	1646	212	1646	58	1170	108
Fruit fly	618	9	618	6	276	7
Arabidopsis	1098	25	1098	18	817	25

Table 7. NMD candidates. Percent of primary ORFs that are candidates to be regulated via nonsense mediated decay surveillance mechanisms in lncRNAs associated with ribosomes and in codRNAs, including cases annotated as NMD in Ensembl that were not considered in this study. We defined a primary ORF as NMD candidate if it spanned a non-terminal exon and its stop codon was located ≥ 55 nucleotides upstream of a splice junction site.

	lncRNA_ribo	codRNA
Mouse	9.25	5.59
Human	14.11	13.33
Zebrafish	6.86	0.77
Fruit fly	13.33	0.34
Arabidopsis	7.14	0.45

Table 8. Homology hits for lncRNAs. Primary ORFs from lncRNAs with homologues in each of the species studied. Homology was inferred using BlastP and an E-value < 10^{-4}. All cases matched to primary ORFs in codRNAs from the same or other species.

	Primary ORFs					
	Mouse	**Human**	**Zebrafish**	**Fruit Fly**	**Arabidopsis**	**Yeast**
Mouse	17	14	10	7	8	10
Human	16	22	19	9	7	10
Zebrafish	115	102	161	62	52	56
Fruit fly	0	0	0	0	0	1
Arabidopsis	0	0	0	0	0	0
Yeast	0	0	0	0	0	0

Table 9. GC content (%) in ORFs and complete sequences. In transcripts associated with ribosomes the ORF corresponds to the primary ORF (see main manuscript file). We also defined different gene deserts located in two different chromosomes (chr7 and chr14) and with different evolutionary properties (stable and flexible).

ORFs	Mouse	Human	Zebrafish	Fruit fly	Arabidopsis	Yeast
Flexible gene desert – chr4	-	37.52	-	-	-	-
Stable gene desert – chr4	-	40.31	-	-	-	-
Flexible gene desert – chr17	-	37.64	-	-	-	-
Stable gene desert – chr17	-	40.01	-	-	-	-
introns	45.38	45.63	38.88	42.37	33.12	34.37
lncRNA_noribo	45.89	47.21	43.26	39.94	34.17	36.02
lncRNA_ribo	47.62	48.52	49.55	43.49	38.94	35.62
pseudogene	49.5	48.04	-	-	-	-
codRNAne	51.57	50.82	-	53.91	44.44	39.58
codRNAe	52.01	50.61	50.34	53.85	45.09	39.56

Sequences	Mouse	Human	Zebrafish	Fruit fly	Arabidopsis	Yeast
Flexible gene desert – chr4	-	34.54	-	-	-	-
Stable gene desert – chr4	-	36	-	-	-	-
Flexible gene desert – chr17	-	35.58	-	-	-	-
Stable gene desert – chr17	-	38.49	-	-	-	-
introns	42.91	41.66	35.24	39.17	32.41	32.68
lncRNA_noribo	44.22	44.69	38.82	39.48	35.22	32.52
lncRNA_ribo	46.14	48.42	42.74	40.91	37.71	35.68
pseudogene	48.78	47.53	-	-	-	-
codRNAne	49.24	48.94	-	49.7	42.43	39.54
codRNAe	49.57	48.02	45.67	49.61	42.95	39.52

Table 10. Number of codRNAs in different gene age datasets that were found associated with ribosomes in our study.

Mouse (Neme & Tautz, 2013)	
Rodent-specific	68
Mammalian-specific	127
Metazoan-specific	11203
Human (Neme & Tautz, 2013)	
Primate-specific	72
Mammalian-specific	123
Metazoan-specific	13423
Zebrafish (Neme & Tautz, 2013)	
Fish-specific	162
Metazoan-specific	9812
Arabidopsis (Donoghue et al., 2011)	
Crucifera-specific	208
Yeast (Ekman, Björklund, & Elofsson, 2007)	
S.Cerevisiae-specific	28
Saccharomyces-specific	84

Table 11. PN and PS values for different sequence subsets. PN/PS: PN/PS ratios and 95% confidence intervals. Fields marked with a * remark cases in which PN/PS is not a reliable metric due to low SNP data. Youngest codRNA correspond to rodent-specific genes for mouse, primate-specific genes for human, *Actinopterygii*-specific genes for zebrafish, See main manuscript file for more details.

PN	**Mouse**	**Human**	**Zebrafish**
lncRNA_noribo	362	812	94
lncRNA_ribo	1017	811	170
codRNAne	43064	59515	9131
codRNAe	23270	193293	
youngest codRNA	254	159	709

PS	**Mouse**	**Human**	**Zebrafish**
lncRNA_noribo	142	271	57
lncRNA_ribo	511	305	209
codRNAne	87910	30059	18027
codRNAe	68969	107672	
youngest codRNA	113	73	449

PN/PS	**Mouse**	**Human**	**Zebrafish**
lncRNA_noribo	2.55	2.66	1.65
lncRNA_ribo	1.99	3.00	0.81
codRNAne	0.49	1.98	0.51
codRNAe	0.34	1.80	
youngest codRNA	2.25	2.18	1.58

95% conf intervals	**Mouse**	**Human**	**Zebrafish**
lncRNA_noribo	2.40-2.68	2.56-2.75	1.44-1.84
lncRNA_ribo	1.91-2.06	2.89-3.10	0.72-0.91
codRNAne	0.49-0.50	1.97-1.99	0.50-0.52
codRNAe	0.33-0.34	1.80-1.81	
youngest codRNA	2.08-2.40	1.98-2.36	1.51-1.65

Figure 1. Comparison between the length of coding and annotated lncRNA transcripts. The difference with Figure 1 in main manuscript file is that here novel lncRNAs were not considered. Density plots of transcript length. lncRNA_ribo: lncRNAs associated with ribosomes; lncRNA_noribo: lncRNAs for which association with ribosomes was not detected. codRNA: coding RNAs transcripts encoding experimentally validated proteins except for zebrafish in which all transcripts annotated as coding were considered. Only transcripts annotated in Ensembl and with expression level > 0.5 FPKM were considered. Fruit fly and yeast had few annotated lncRNAs and were not analysed.

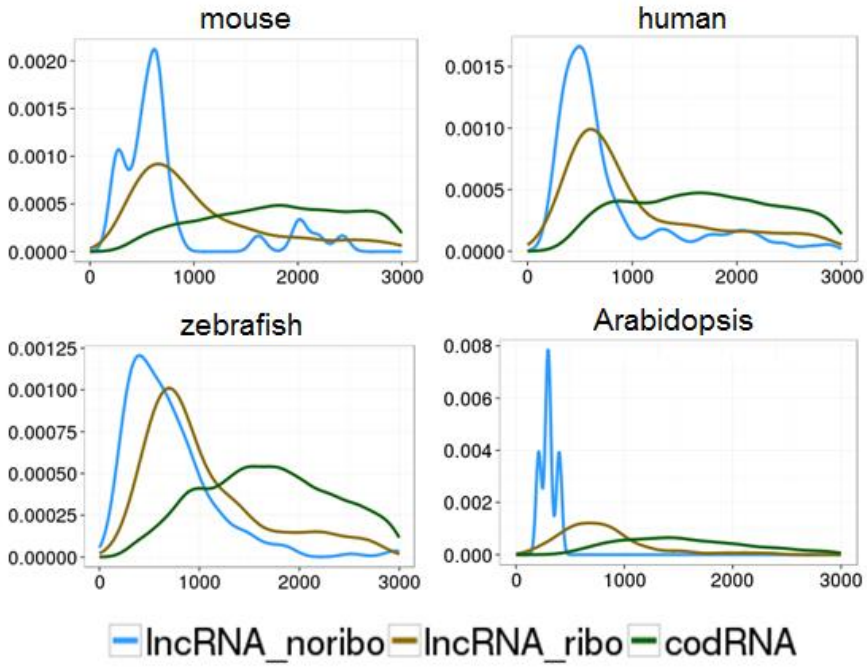
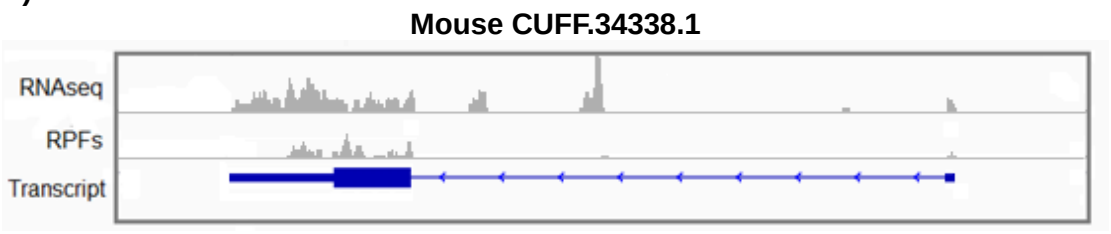
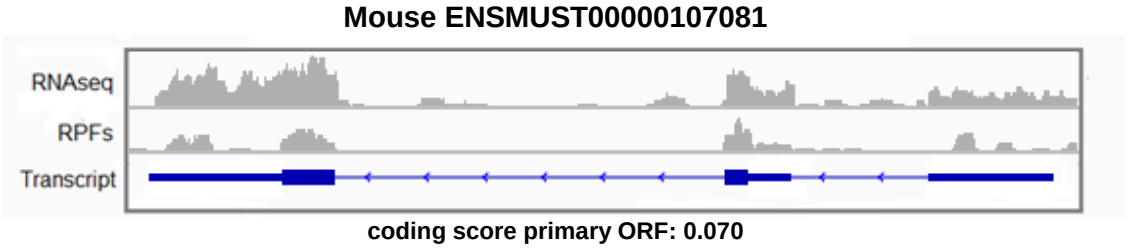


Figure 2. Examples of ribosome association in assembled transcripts. A) mouse CUFF.34338.1 (chr5:113183493-113188347) is a novel lncRNA containing a 169 amino acid primary ORF associated with ribosomes and with protein-coding homologues in human, zebrafish and yeast. **B)** ENSMUST00000107081 is an annotated codRNA in mouse which evolved recently since no homologs 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, that may result in a polycistronic translation since two redundant ORFs (in red) coding the sequence: MGLGFVN(V/F)LLGM are also associated with ribosomes. RNAseq: coverage profile of RNAseq reads. RPFs: coverage profile of ribosome profiling reads. Exon-intron transcript structures are also represented; the thickest boxes on them are the primary ORFs.

A)



B)



C)

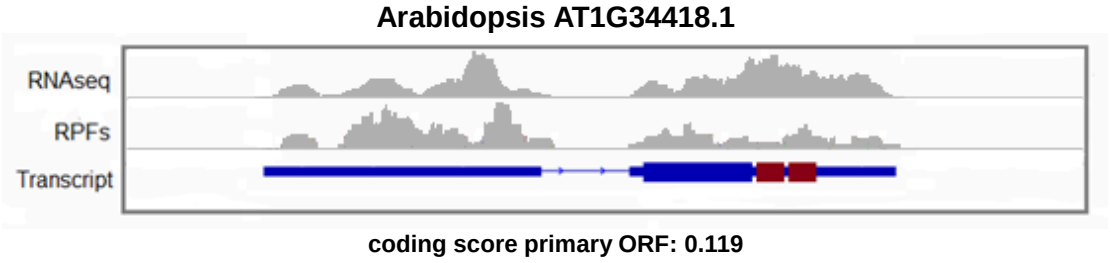


Figure 3. Fraction of the transcript covered by ribosome profiling reads. This fraction tended to be larger in codRNAs than in lncRNAs in the six species studied. For example in mouse this percentage was 67% for codRNAs and 37% for lncRNAs, and in zebrafish 29% and 18%, respectively. When we controlled for transcript expression level the differences between codRNAs and lncRNAs were still significant in all the species (Wilcoxon rank-sum test, p-value $<10^{-14}$ for transcripts expressed at >2 FPKM).

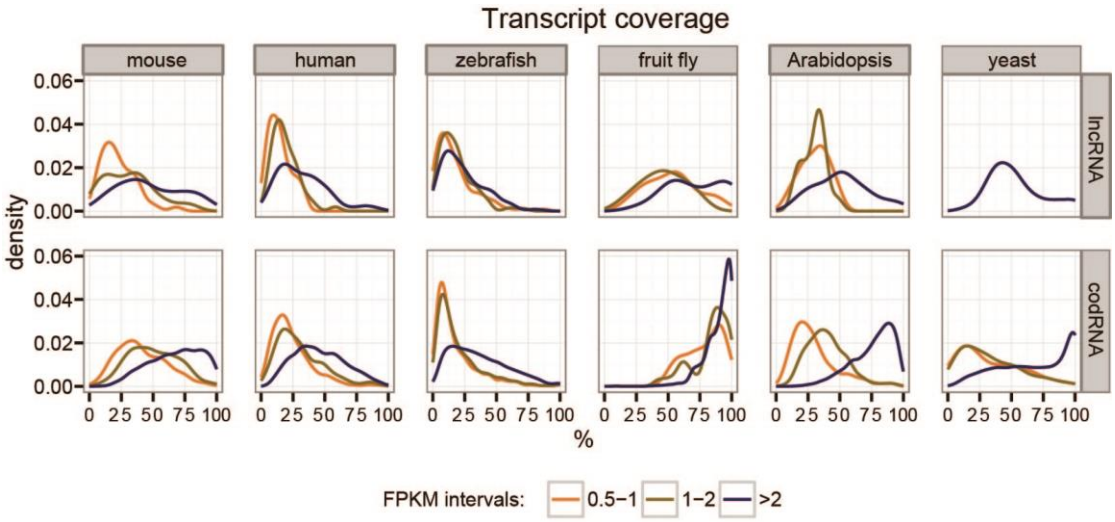


Figure 4. Mean and max translational efficiency (TE) for coding and lncRNA transcripts. Single isoforms corresponds to data for genes with a single transcript. Mean TE for all the transcripts is shown in the main manuscript file.

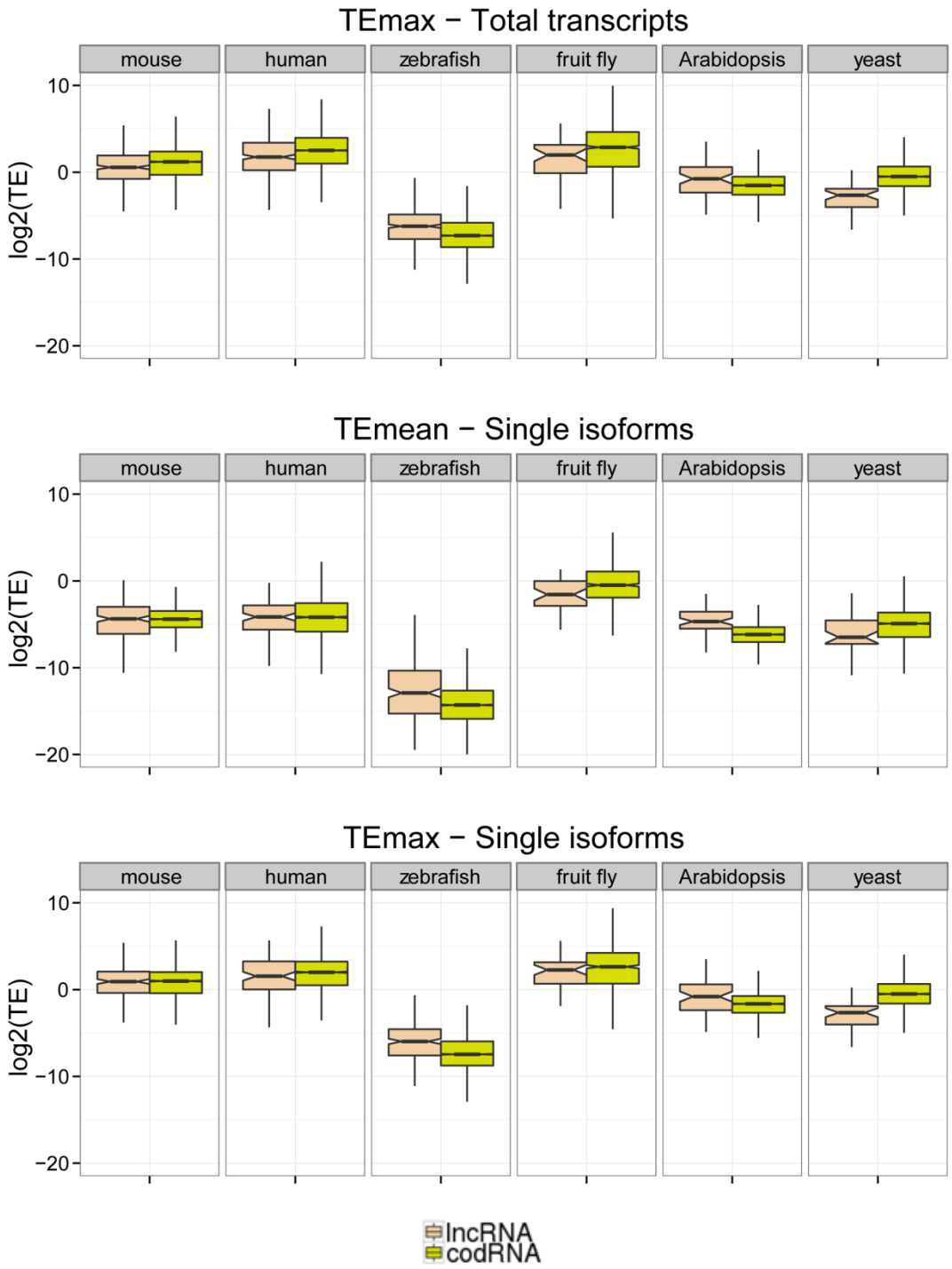


Figure 5. Definition of primary ORF in transcripts. Cases in which annotated CDS from protein-coding transcripts (codRNAe and codRNAne) were correctly (TP) or incorrectly (FP) identified using different ORF metrics. We also defined a third case in which the identified ORF was embedded into the CDS (Contained in TP). ORF maxRC, based on the ORF with the largest number of RPFs respect to the total number of RPFs in the transcript, was the metrics selected to define the primary ORF.

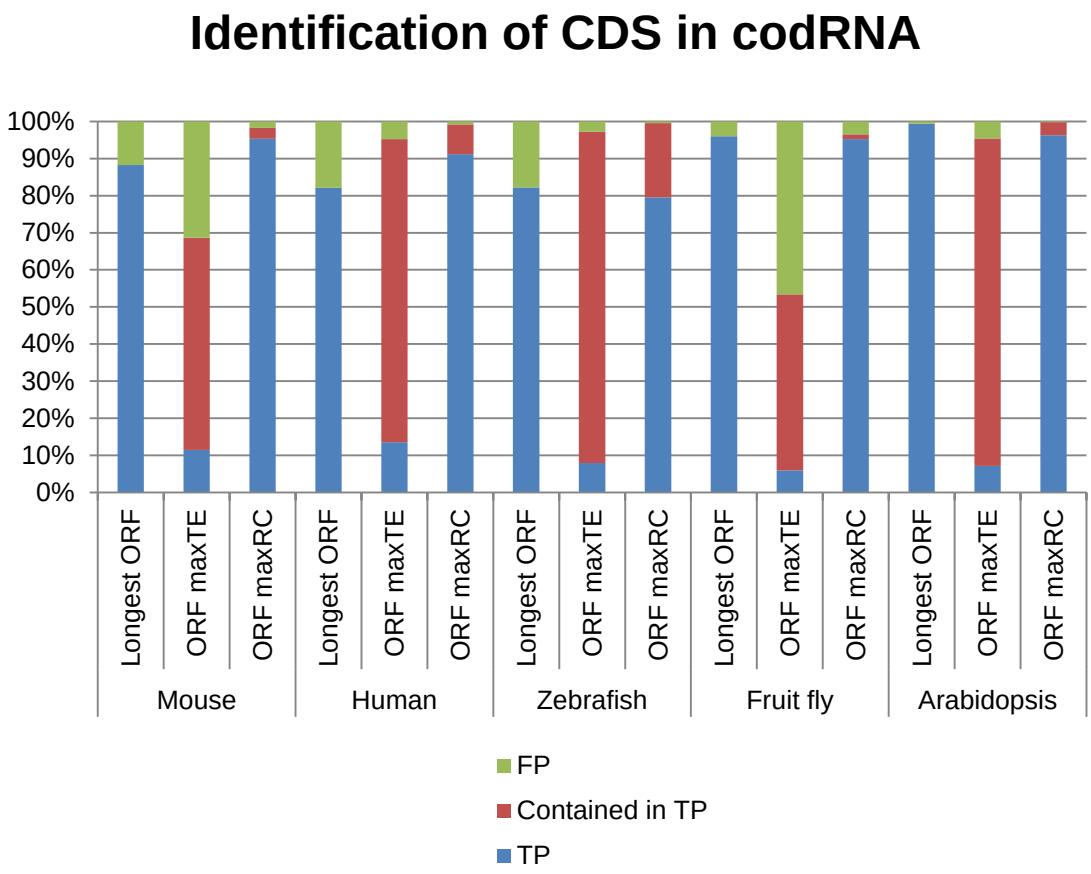


Figure 6. Length of ORFs in different kinds of transcripts (in nucleotides). In codRNAs and lncRNA_ribo, we selected the primary ORF (the ORF with the largest number of ribosome profiling reads), whereas in lncRNA_noribo we selected the longest ORF.

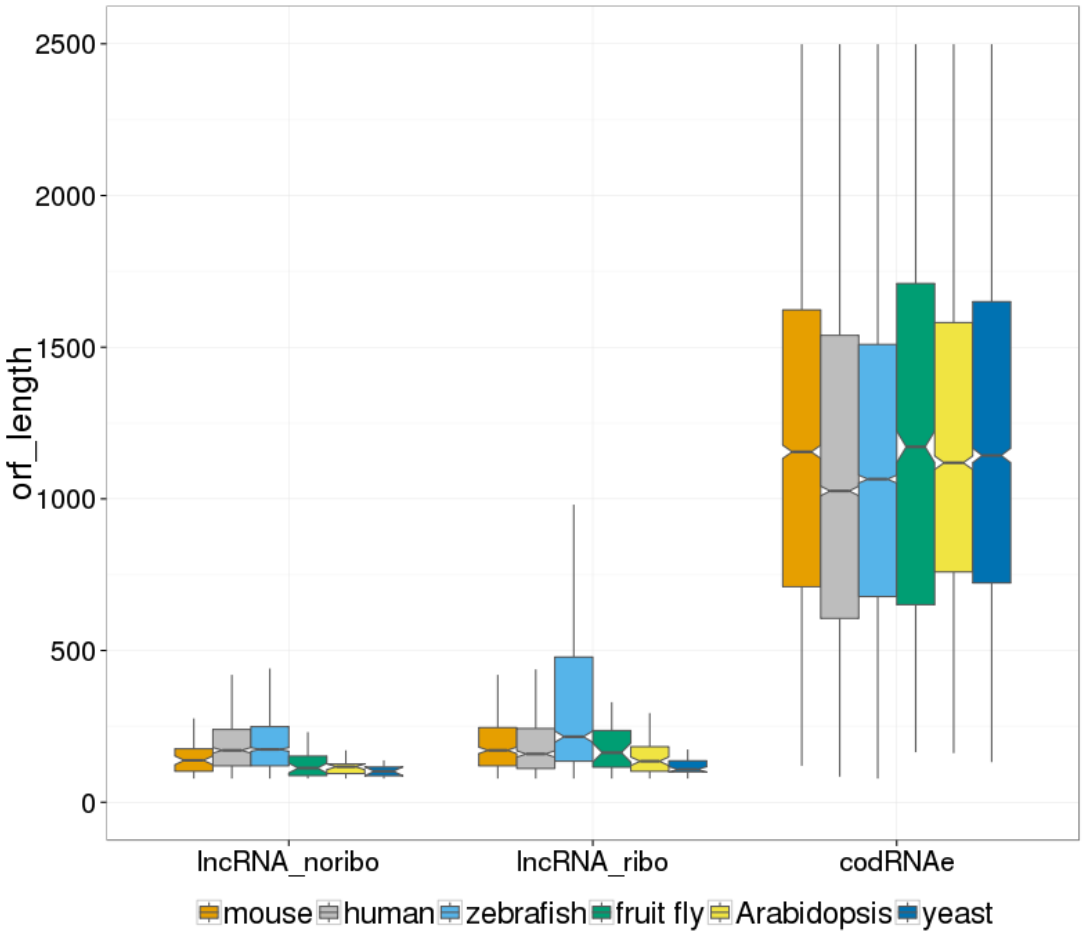
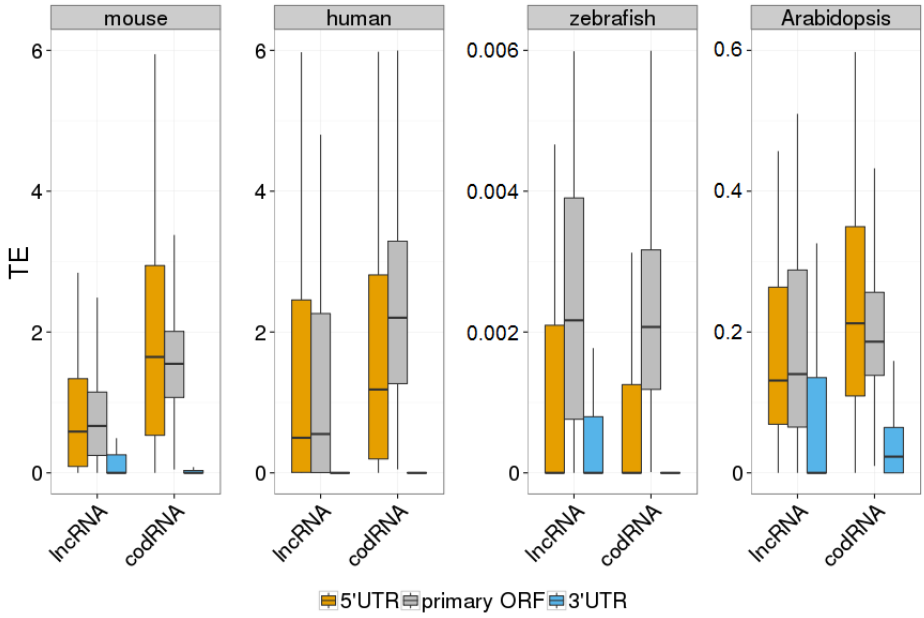


Figure 7. Translational efficiency in single-isoform transcripts. A) Box-plots of TE distribution in primary ORF, 5'UTR and 3'UTR regions. The analysis considered all unique transcript isoforms with UTR and ORF regions expressed at > 0.2 FPKM and with 5'UTR and 3'UTR longer than 30 nucleotides. B) Box-plots of TE distribution in primary ORFs, other ORFs with ribosome profiling reads and non-ORF regions (interORFs). The analysis considered all unique transcript isoforms with at least two ORFs and summed interORFs longer than 30 nucleotides. Both ORFs and interORFs regions had > 0.2 FPKM. Fruit fly and yeast were not represented here because number of transcripts meeting the previous requirements was low and the local TE could not be properly estimated.

a)



b)

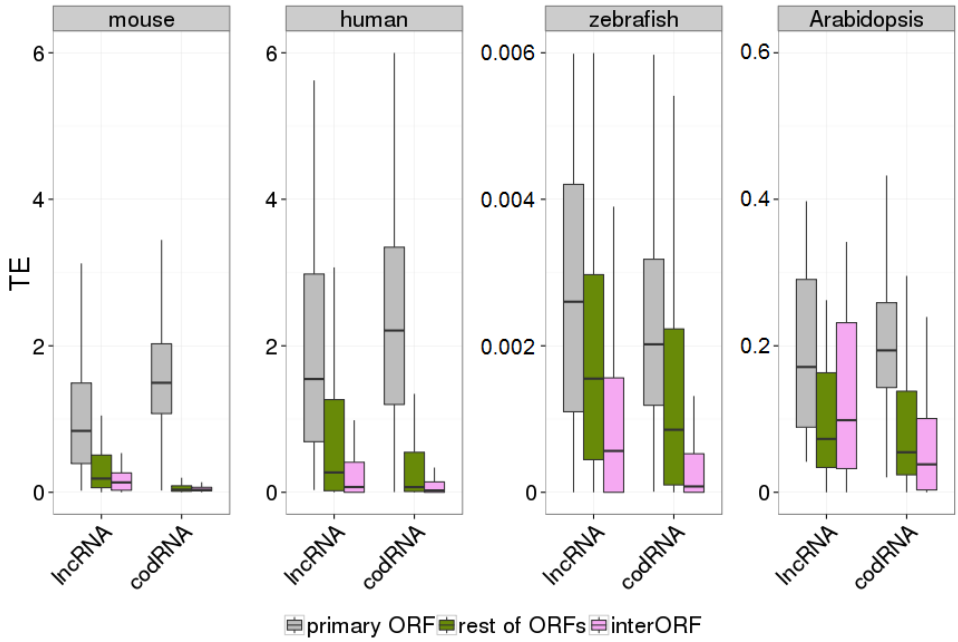
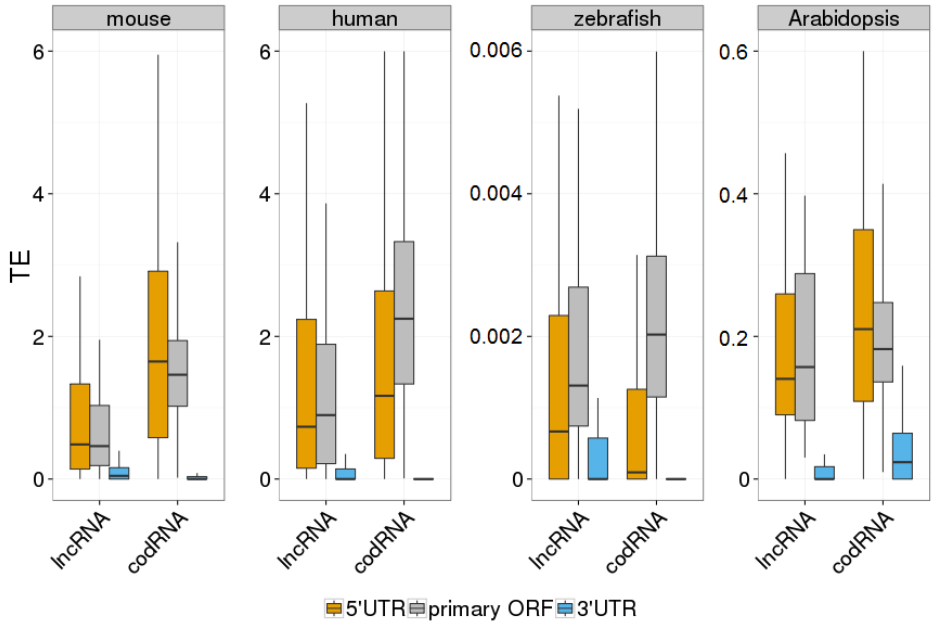


Figure 8. Translational efficiency in annotated transcripts. A) Box-plots of TE distribution in primary ORF, 5'UTR and 3'UTR regions. The analysis considered all annotated transcripts with UTR and ORF regions expressed at > 0.2 FPKM and with 5'UTR and 3'UTR longer than 30 nucleotides. B) Box-plots of TE distribution in primary ORFs, other ORFs with ribosome profiling reads and non-ORF regions (interORFs). The analysis considered all annotated transcripts with at least two ORFs and summed interORFs longer than 30 nucleotides. Both ORFs and interORFs regions had > 0.2 FPKM. Fruit fly and yeast were not represented here because number of transcripts meeting the previous requirements was low and the local TE could not be properly estimated.

a)



b)

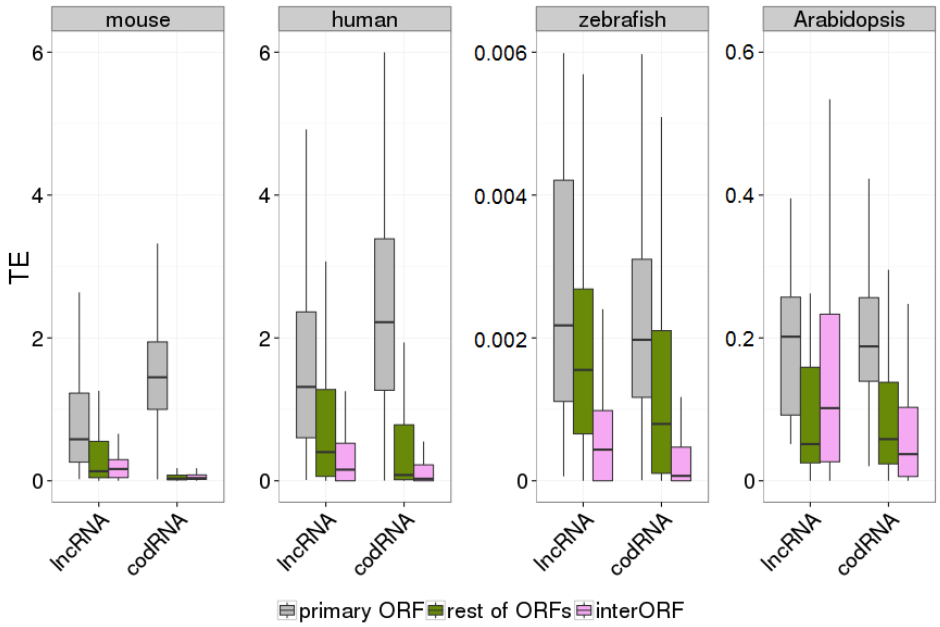
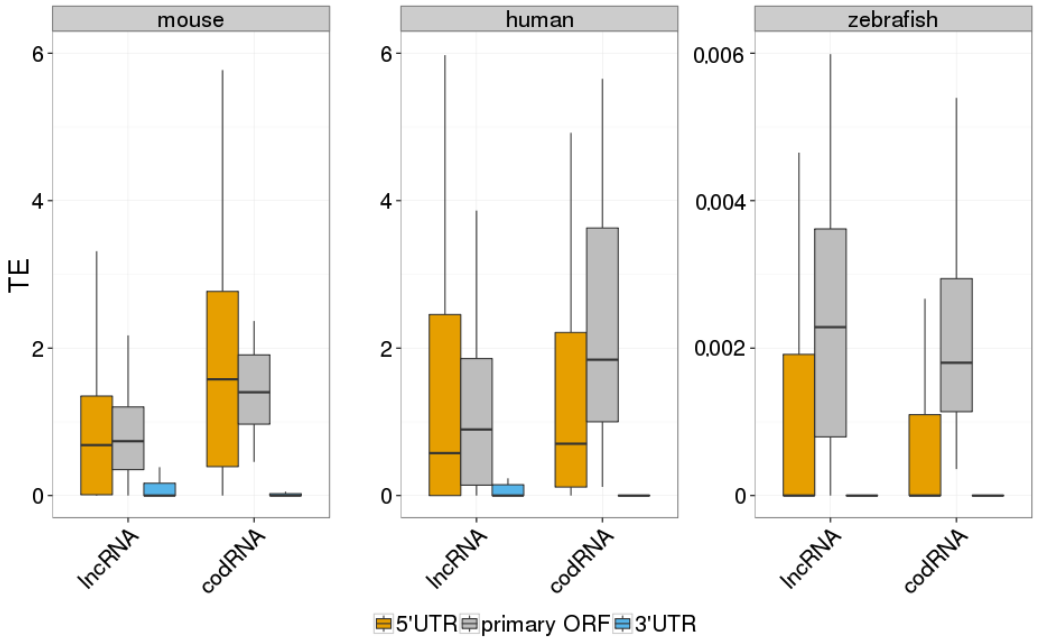


Figure 9. Translatonal efficiency in transcripts at different expression level. We restricted this analysis to transcripts with ORF and UTR regions expressed at > 0.2 FPKM and with 5'UTR and 3'UTR longer than 30 nucleotides for increased ribosome association detectability. a) expressed at low levels: Box-plots of transcripts expressed at 0.5-2 FPKM, b) expressed at high levels: Box-plots of transcripts expressed at 2-10 FPKM. codRNA represent a sampled subset a with similar gene expression distribution as the corresponding lncRNA set. Results for species in which all sets contained at least 20 transcripts are shown.

a) expressed at low levels



b) expressed at high levels

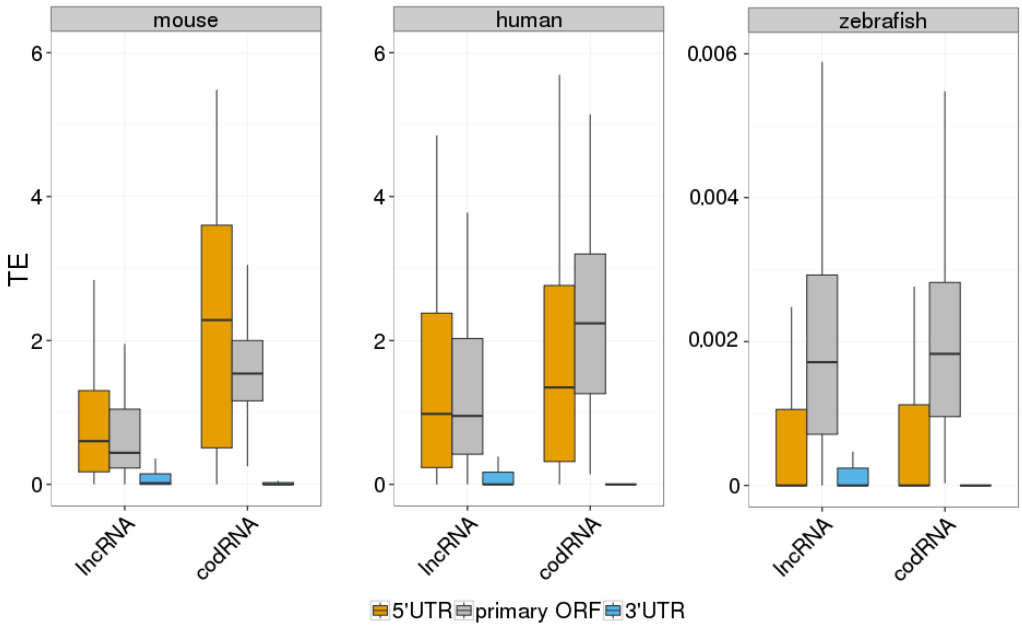


Figure 10. Coding scores in different classes of annotated sequences. The score was based on the differences in hexamer frequencies in coding and non-coding sequences. Here we only employed annotated lncRNAs. Fruit fly and yeast had very few annotated lncRNAs and this analysis could not be performed.

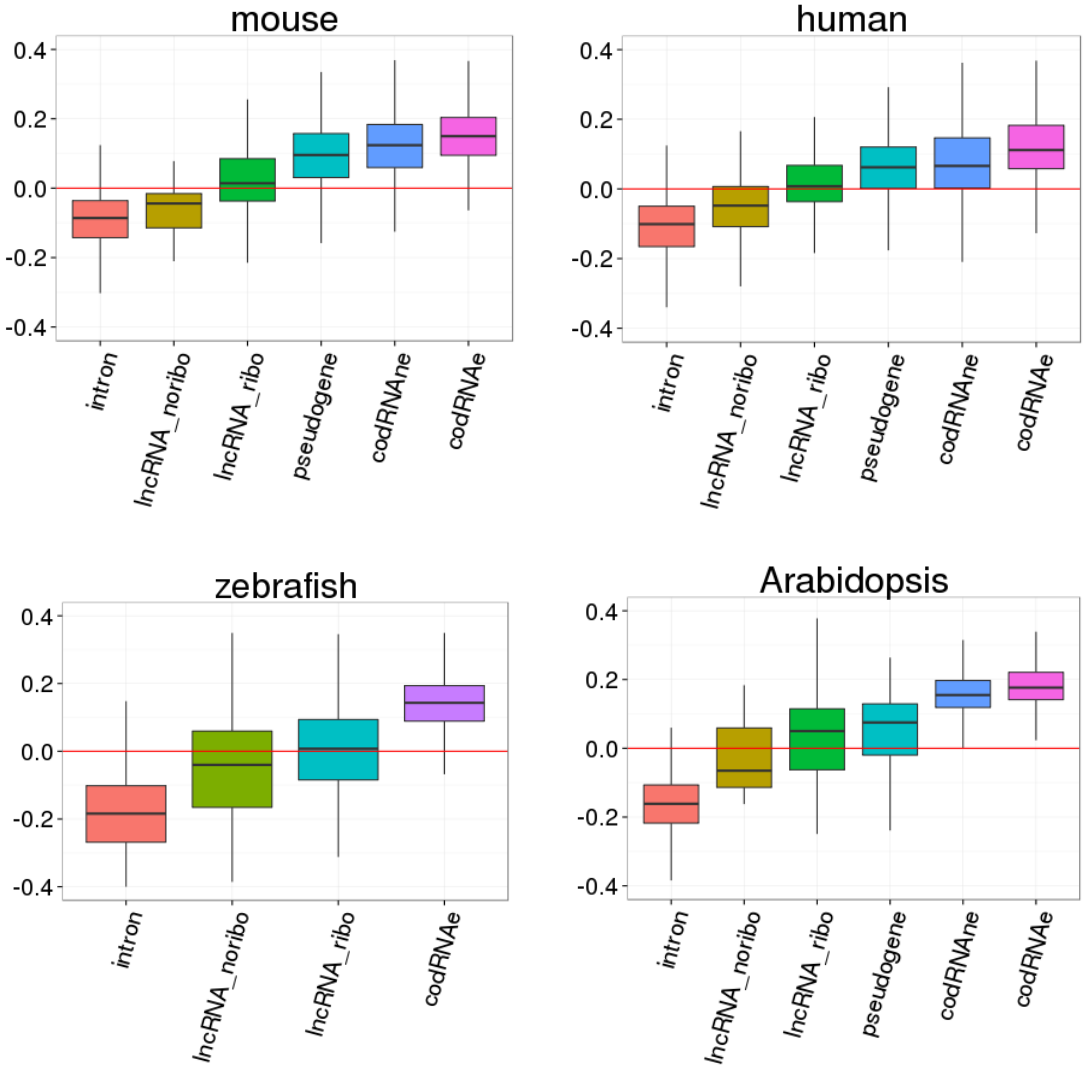


Figure 11. Coding scores in small ORFs from different types of transcripts. The score was based on the differences in hexamer frequencies in coding and non-coding sequences. Here we only employed lncRNAs in which the primary ORF was shorter than 100 amino acids. codRNA refers to joined codRNAe and codRNAne, since experimentally verified proteins are usually longer than 100 amino acid. Fruit fly and yeast had very few lncRNAs containing sORFs and this analysis could not be performed.

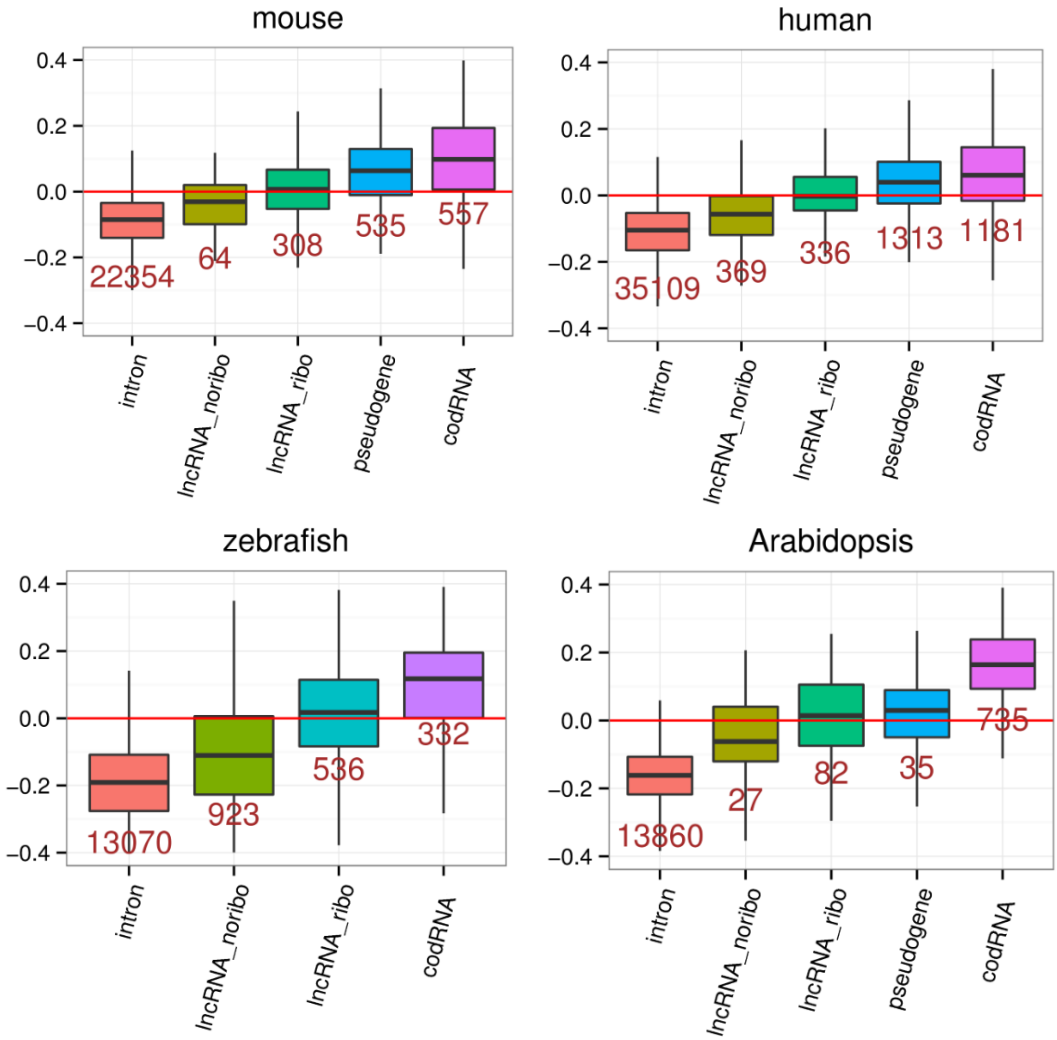


Figure 12. Coding scores for the longest ORF from lncRNAs associated and not associated with ribosomes. Whenever possible, values are shown separately for transcripts with and without homologous ORFs from codRNAs Differences between lncRNA_ribo_no_hom and lncRNA_noribo are significant by a Wilcoxon rank-sum test, p-value $<10^{-10}$ in human, mouse, and zebrafish; p-value <0.005 in *Arabidopsis.*

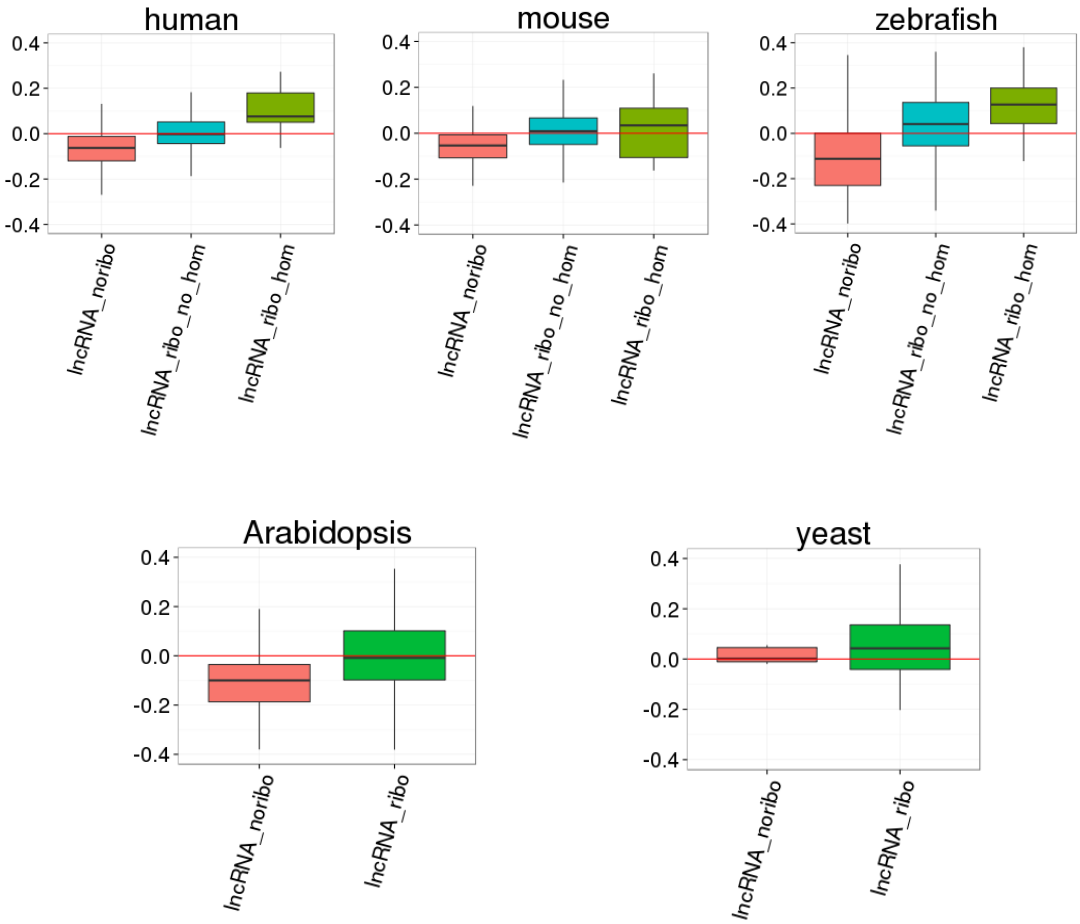


Figure 13. Coding statistics for different types of sequences in humans. Intron: randomly selected intronic regions; lncRNA_noribo: lncRNAs not associated with ribosomes; lncRNA_ribo: lncRNAs associated with ribosomes; pseudogene: pseudogenes associated with ribosomes; codRNAne: coding transcripts encoding non-validated proteins associated with ribosomes; codRNAe: coding transcripts encoding experimentally validated proteins. Equal dicodon was based on the observed hexamer frequencies in coding sequences versus hexamer equiprobability, intron dicodon was based on the differences between hexamer frequencies in coding versus non-coding sequences and intron_monocodon was based on the observed codon frequences in coding sequences versus codon equiprobability

.