# LincRNA sequences are biased to counteract their translation

Anneke Brümmer[1,3], Rene Dreos[2], Ana Claudia Marques[1,*], Sven Bergmann[1,3,4,*]

[1]: Department of Computational Biology (DBC), University of Lausanne, Lausanne, Switzerland

[2]: Center for Integrative Genomics (CIG), University of Lausanne, Lausanne, Switzerland

[3]: Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

[4]: Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa

*: equal last authors

**Abstract**

Long intergenic non-coding RNAs (lincRNAs) represent a large fraction of transcribed loci in the human genome. Although classified as non-coding, the vast majority of lincRNAs contain putative open reading frames (pORFs), and the molecular mechanisms hindering their translation are currently unclear.

Here, we analysed the sequences of lincRNAs in six species, ranging from human to fission yeast. We found that the trinucleotide ("codon") composition of their pORFs is distinct from that of mRNA. Strikingly, relative to mRNAs and nuclear lincRNAs, pORFs in highly expressed cytosolic lincRNAs tend to use codons corresponding to less frequent tRNAs and are less frequently bound by ribosomes than coding RNA types. For the majority of cytosolic lincRNAs, we found that their variation in ribosome binding across five human cell lines can partially be explained by cell line-specific tRNA abundances.

Our observations are consistent with lincRNA pORFs having adapted their sequences to less abundant tRNAs. We propose that this serves to counteract ribosome binding and translation of cytoplasmic lincRNAs.

**Keywords:** lincRNA, tRNA, codon usage bias, computational sequence analysis, ribosome binding, subcellular localization, small protein-coding ORF, tRNA adaptation, translation regulation, cell type-specificity

## Introduction

Long intergenic noncoding RNAs (lincRNAs) form a functionally heterogeneous class of RNAs longer than 200 nucleotides and lacking protein coding potential (Frankish et al., 2019; Ulitsky and Bartel, 2013). Despite being classified as non-coding, some lincRNAs contain start and stop codons flanking potential open reading frames (pORFs). Some pORFs have been found to associate with ribosomes (Guttman et al., 2013; Ingolia et al., 2014; Ji et al., 2015), however detailed analysis of lincRNA and ribosome interactions, using sequencing of ribosome protected fragments (Ribo-Seq) data, revealed differences between the association of the translation machinery with protein-coding mRNAs and lincRNAs. This included differences in the tri-nucleotide periodicity of binding (Calviello et al., 2016; Ji et al., 2015) or in ribosome release (Guttman et al., 2013). Furthermore, fewer than 10% of pORFs in human lincRNAs produce experimentally detectable peptides (Bánfai et al., 2012). With more and better suited experimental data being generated, the number of functional small protein-encoding ORFs continues to increase (Chen et al. 2020; Ouspenskaia et al. ; Martinez et al. 2020), but the majority of lincRNA pORFs remains untranslated suggesting yet unknown mechanisms to counteract their translation.

The genetic code is degenerate and most amino acids, except methionine and tryptophan, are encoded by multiple synonymous codons. Eukaryotic cells express between 40 and 60 tRNA anticodon types, each of which can recognize one to three different codons (due to wobble-base pairing and tRNA editing). The abundances of complementary tRNAs that decode each codon vary between codons due to differences in tRNA gene copy numbers, expression levels, or tRNA modifications. The rate at which a codon is translated correlates with the abundance of the decoding tRNAs (Dana and Tuller, 2014), and mRNA codon usage is a strong regulator of translation efficiency and speed (Tuller et al., 2010a). For example, the adaptation of codon usage to tRNA abundance allows fine tuning of translation efficiencies of genes required for different cellular processes. Such specialized translation programs have been observed in several contexts, e.g. under cellular stress conditions, during proliferation, or in cancer (Goodarzi et al., 2016; Guimaraes et al., 2020; Torrent et al., 2018). Thus, mRNA codon usage is biased and adapted to cellular tRNA abundances to ensure optimal protein output. The strength of the mRNA codon usage bias varies between species, and it is usually more pronounced in species evolving under stronger selection, due to a larger population size or shorter generation time (Subramanian, 2008). Among eukaryotes, the mRNA codon usage bias is strongest in yeast and lowest in complex species such as human and mouse.

Given the established role of mRNA codon usage in regulating translation, we hypothesize that the trinucleotide preferences within lincRNA pORFs could potentially be a way to attenuate their chances of being processed by the translation machinery. Although global and cell type-specific properties of lincRNAs with and without ribosome binding have been described previously (Wang et al., 2017), and many functional small ORFs were recently identified in long non-coding RNAs (Chen et al., 2020; Martinez et al., 2020; Ouspenskaia et al.), these studies did not investigate the reasons for these ORFs being translated. In particular the codon composition and its relationship with cellular tRNA abundances has not been analysed before, and we propose that this could be one of the mechanistic bases for these observations.

Here, we analysed lincRNA sequences of six species and detected a bias in the trinucleotide ("codon") composition of pORFs in lincRNAs. pORFs in lincRNAs tend to have codons corresponding to less abundant tRNAs, in particular in highly expressed cytosolic lincRNAs and for species under stronger evolutionary selection. We further related this bias to ribosome binding (measured by Ribo-Seq) for different coding and non-coding RNA types and across human cell lines. Our analysis supports that the usage of codons corresponding to less abundant tRNAs contributes to attenuating translation of cytoplasmic lincRNAs.

## Results

*Putative open reading frames in lincRNAs are frequent, but shorter than mRNA coding regions*

To investigate properties of lincRNA putative ORFs (pORFs), we focussed on six species (human, mouse, fish, fly, worm and fission yeast). These species were selected because they cover a range of mRNA codon bias strengths (Subramanian, 2008) and contain a sizable number (>1000) of annotated lincRNA genes in their genomes.

For each lincRNA, we predicted all pORFs longer than 30 nucleotides (=10 codons), starting with a canonical start codon (AUG) and ending at the first in-frame stop codon (UAG, UAA, UGA). As a control, we also predicted ORFs in randomly selected intronic and intergenic control regions of the same length (see Methods). Most lincRNAs (>94% for all species) had at least one pORF (Figure S1A). With the exception of fission yeast, the fraction of lincRNAs with pORFs is significantly larger than what we found for control regions (Figures 1A and S1A). The median number of pORFs per lincRNA (6-10 pORFs in different species) was also significantly larger than in control regions, in most species (Figure S1B), with the exception of worm and yeast. In the following, we considered the longest pORF for each lincRNA (median length 45-71 codons in different species; Figure S1C). The longest pORFs in lincRNAs were significantly longer than the longest pORFs in control regions in most species, except for worm and yeast (Figures 1B and S1C). In all species, mRNA coding regions were on average 5-8 times longer than longest lincRNA pORFs.

The analysis of pORFs in lincRNAs and control regions suggests that there is no selection to reduce their number or length in lincRNAs in most analysed species, potentially with the exception of worm and fission yeast.

*Sequence and structure context around lincRNA start codons differs from those of mRNAs and control pORFs*

In the following, for each coding and non-coding gene we consider only the isoform harboring the longest coding region for mRNAs or longest pORF for lincRNAs (Table S1). As a control, we considered, for each lincRNA, the longest predicted ORF (>10 codons) in each of 10 randomly selected intronic and intergenic regions of the same length as the lincRNA isoform harboring the longest pORF.

Since the RNA sequence and structure context around start codons was identified as a regulator of translation in mRNA before (Eraslan et al., 2019), we examined if the start codon context in lincRNA pORFs could be responsible for them not being translated. We found that in all species the lincRNA start codon sequence context was different from the consensus mRNA sequence motif (Figure 1C and S2; see Methods), which showed similarity with the Kozak sequence motif. However, for most species, with the exception of worm and yeast, the lincRNA start codon sequence context was significantly more similar to the consensus mRNA sequence motif than the region around start codons of control pORFs (Figure S2).

Comparing the RNA structure around mRNA and lincRNA start codons across species (Figures 1D and S3), we found that in human and mouse lincRNA sequences appear on average more accessible than sequences around mRNA start codons. This might be a consequence of a lower GC content in lincRNAs in these species (Haerty and Ponting, 2015). In contrast, in fly and fission yeast the structural accessibility upstream of lincRNA start codons was reduced in comparison with mRNAs, which may hinder an interaction between the ribosome and lincRNA start codons in these species.

Together, this analysis indicates that in fission yeast, and to some extent in worm and fly, the RNA sequence and structure context around lincRNA start codons may be unfavourable for translation initiation. Yet, in the remaining species, in particular human and mouse, we found less evidence for this, suggesting that alternative mechanisms may be in place in these species to hinder lincRNA translation.

*Codon composition of lincRNA pORFs is distinct from mRNA coding regions and control pORFs*

Codon usage of mRNA is biased and a way to regulate their translation (Hanson and Coller, 2018; Tuller et al., 2010a). We hypothesized that biased trinucleotides, which for simplicity we refer to as codons, in lincRNA pORFs can contribute to counteracting their translation. To gain initial insight into the characteristics of codons in lincRNA pORFs we compared their frequencies with those found in mRNA coding regions and control pORFs (Table S2). In all species studied, with the exception of fish, codon frequencies of lincRNA pORFs were more similar to those of control pORFs than to mRNA codons (Figure 2A-C). We then performed a principal component analysis of the corresponding codon frequency matrix (see Methods, Figure 2D). For vertebrates and flies we found a clear separation between RNA types in the first two principal components, with lincRNA codon composition being in-between, but distinct from mRNA codon usage and control codon frequencies. For worm, and more clearly for fission yeast, lincRNA and mRNA codon usages were located on opposite sides from control codon frequencies. These results confirm that codon preferences are different between lincRNA and mRNA. In particular, in species with stronger mRNA codon bias (worm and yeast), lincRNA and mRNA codon usages appear to have diverged in opposite directions relative to control codon frequencies, suggesting that lincRNA codon usage may have a functional role.

In agreement with the overall GC content difference between lincRNAs and mRNAs (Haerty and Ponting, 2015), the most enriched codons in mRNAs were often GC-rich, while the most enriched codons in lincRNA ORFs were often AU-rich (Figure 2A). To test if differences in GC content between lincRNAs and mRNAs impact codon usage, we compared the frequencies of codons with the same GC content (Figure S4A). We found that the correlation strengths between the different types of sequences were similarly distributed, and a separation between mRNA, lincRNA and control codon usages was detectable in the first two principal components for most GC content codon groups. We concluded that the difference in codon frequencies between mRNA and lincRNA is not driven by GC content differences, and that lincRNA codon usage is distinct by itself.

The relative frequencies of amino acids encoded by mRNA coding regions and lincRNA pORFs are more similar to each other than the relative codon frequencies of mRNAs and lincRNAs (Figure S4B and C). To account for the different numbers of amino acids and codons (19 and 60, respectively, excluding start and stop codons), we subsampled 100000 sets of 19 randomly selected codons and calculated correlations between frequencies of selected codons. The correlation between amino acids was also significantly stronger than the correlations between subsampled codons (Figure S4C). This indicates an increased divergence in codon usage compared to amino acid usage between RNA types, further demonstrating the distinct codon usage in lincRNA pORFs. Interestingly, for the majority of the 18 amino acids encoded by multiple codons (for 10-18 in different species), the codon that is preferred by mRNAs is less used in lincRNA pORFs (Figure S5). Moreover, the rarest mRNA codon for an amino acid is used more often by lincRNA pORFs (for 10-18 amino acids in different species).

*Relative to mRNA coding regions, codons in lincRNA pORFs are less adapted to tRNA abundances*

Next, we investigated how differences in codon usage between mRNA and lincRNA sequences relate to tRNA abundances. As a first estimate of the tRNA abundances in different species we used the relative number of tRNA genes coding for the same tRNA anticodon (see Methods) that has been previously shown to correlate well with tRNA abundances (Tuller et al., 2010b). We used previously determined wobble-base pairing and tRNA editing efficiencies (dos Reis et al., 2004) to estimate effective tRNA anticodon abundances for all codons including those that lack a complementary tRNA encoded in the genome (see Methods). We found that mRNA codon frequencies correlated better and more significantly with relative tRNA abundances than codon frequencies of lincRNAs, in all species (Figure 3A) including human (Figure 3B) and fission yeast (Figure 3C). Strikingly, in the two species with the strongest mRNA codon usage bias (worm and fission yeast), relative tRNA frequencies correlated less with the codon usage in lincRNA pORFs than with control codon frequencies, suggesting adaptation of lincRNA pORFs towards using codons corresponding to less abundant tRNAs in these species (Figure 3A).

The tRNA adaptation index (tAI) was previously defined as a measure for the adaptation of codon usage to tRNA abundance (dos Reis et al., 2003). tAI ranges from 0, no codon adaptation, to 1, for perfectly adapted codon usage, when the codons used are decoded by the most abundant tRNA anticodon type. We will use the word adaptation in this sense throughout this paper. We computed tAIs for each mRNA coding region and lincRNA and control pORFs (see Methods). We found that, in all species studied, mRNAs were significantly ($p<1e-300$, Wilcoxon rank-sum test) better adapted to relative tRNA abundances than lincRNAs (Figure 3D, E and S6). Compared to control sequences, lincRNAs were better adapted to relative tRNA abundances for vertebrates and fly (Figure 3D left panel for human; summary statistics for all species in Figure 3E). In contrast, in worm and fission yeast, tAIs were lower for lincRNA pORFs than for control pORFs (Figure 3D right panel) indicating that lincRNA pORFs adapted to avoid codons recognized by abundant tRNAs in these species.

To further quantify the extent and direction of sequence adaptation, we related tAIs to those of three types of randomized sequences (see Methods), each preserving different properties of the original ORFs, such as nucleotide content, nucleotide sequence or amino acid sequence (see Supplementary Text; Figure S7). Overall, this analysis supports lincRNAs exhibiting a distinct and non-random tRNA adaptation, with stronger signatures in worm and yeast, i.e. species with generally stronger selection.

*Codons in cytoplasmic lincRNAs are less adapted to tRNA abundances than those in nuclear or non-expressed lincRNAs*

In the previous section, we found that lincRNA codon frequencies are less adapted to relative tRNA abundances than mRNA codon usage, and less than control codon frequencies for worm and yeast. We hypothesized that, if lincRNA sequences evolved to reduce the likelihood of being translated, this would be more pronounced for highly expressed and cytoplasmic lincRNAs, since these lincRNAs are more frequently exposed to the translation machinery and might therefore experience a stronger evolutionary pressure for lowering their tAIs. To test this hypothesis, we examined codon frequencies in expressed and cytoplasmic RNAs focussing on species and cell types, for which we found suitable experimental data publicly available (Table S3). In particular, we downloaded data for five human cell lines (GM12878, HeLa-S3, HEK293, HepG2 and K562), mouse embryonic stem cells (mESCs), two fly cell types (S2 and early embryos) and fission yeast grown on rich medium (YES).

Thus far, we used tRNA gene counts to estimate relative tRNA abundances. This metric is not cell type-specific and therefore not suitable to analyse tRNA adaptation in different cell types. Quantification of tRNA abundances using dedicated experimental approaches (DM-tRNA-Seq (Zheng et al., 2015) and hydro-tRNA-Seq (Gogakos et al., 2017)) is available only for a limited number of cell

types, including HEK293 cells. Previously tRNA abundances have also been quantified using H3K27ac-ChIP-seq (Gingold et al., 2014) and smallRNA-seq data (Hernandez-Alias et al., 2020; Shi et al., 2018), which are available for a large number of cell types to be used in this study. We evaluated the performance of these metrics in HEK293 cells and found that the smallRNA-seq-based tRNA quantification agreed better with those from dedicated approaches than the H3K27ac-ChIP-seq-based one (Figure S8; see Methods). Therefore, we used smallRNA-seq data to estimate cell type-specific tAIs in different cell types and species.

Overall, cell type-specific tAIs confirmed that tRNA adaptation of mRNA codons was better than that of lincRNAs, for all nine cell types (Figure 4A, S9 and S10). We also found that tAIs of expressed mRNAs were significantly higher than those of non-expressed mRNAs (Figure S9A, C and S10A), as observed previously (Waldman et al., 2010). In contrast, tAIs of expressed lincRNAs were significantly lower than tAIs of non-expressed lincRNAs for most cell types (Figure S9A, C and S10A). Furthermore, whereas the tAIs of mRNAs increased with increased cytoplasmic abundance, tAIs decreased for lincRNAs that were more abundant in the cytoplasm (Figure 4A, S9 and S10). For HEK293, using cell type-specific tAIs calculated based on data from dedicated experimental approaches, we found similar trends, but the significance values (for mRNA) were strongest when using smallRNA-Seq to estimate tRNA abundances (Figure S10A).

Compared to tAIs of cytoplasmic mRNAs, median tAIs among the 50% most highly expressed lincRNAs in the cytoplasm were more than one standard deviation lower in all nine cell types investigated (Figure 4B, left panel). Median tAIs of these lincRNAs were also lower than tAIs of control sequences for most data sets (Figure 4B, right panel), except for mESC and HepG2 (see Discussion). Interestingly, ranking the top 5% mRNAs most abundant in the cytoplasm by their median tAI (z-scored relative to control sequences) reflected the expected increasing strength of mRNA codon bias in different species (Figure 4B, right panel). Human and mouse cell lines ranked among those with smallest deviation from control sequences (around 1 standard deviation), then fly cell lines (~1-2 standard deviations) and fission yeast with the largest deviation from control sequences (~3 standard deviations). This observation suggests evolutionary forces acting on mRNAs and lincRNAs codon usage. Whereas mRNAs favor codons recognized by more abundant tRNAs, as seen previously (Frumkin et al., 2018; Hanson and Coller, 2018), lincRNAs expressed in the cytoplasm prefer codons corresponding to less abundant tRNAs.

To investigate the relationship between tAI and RNA expression further, we ranked RNAs by their expression level and then calculated the median tAI for an increasing, cumulative fraction of RNAs starting with the most highly expressed ones. We observed that for mRNAs the median tAI decayed steadily as a function of this fraction (Figure S9B, D and S10B, blue line). The same observation was made when ranking mRNAs by their cytoplasmic expression level. In contrast, median tAIs for lincRNAs ranked by cytoplasmic expression were lower for the most highly expressed lincRNAs in the cytoplasm and increased for lincRNAs with lower cytoplasmic abundances, for most cell types studied (Figure S9B, D and S10B, orange lines). This relationship was also indicated by negative correlations between tAIs and cytoplasmic expression levels for lincRNAs (Figure S9B, D and S10B). Yet, this trend was less pronounced for lincRNAs ranked by total expression levels, and the correlation was even weaker when ranking lincRNAs by the ratio of cytoplasmic to nuclear expression levels, in most human cell lines (not shown).

Together, these results support the hypothesis that codons of abundant cytosolic lincRNAs tend to correspond to less frequent tRNAs, likely to reduce translation, in contrast to codons in lincRNAs that are not or less present in the cytoplasm.

*tRNA adaptation differences between lincRNA pORFs and small protein-encoding ORFs agree with differences in ribosome binding*

To investigate the impact of tAI on ribosome binding, we focussed on three types of cytoplasmic RNAs, mRNAs, lincRNAs, and RNAs with small protein-coding ORFs (smORFs, see Methods). We compared the tAIs and ribosome binding within human cell lines for these three RNA classes. We observed that tAIs of lincRNAs and smORFs are significantly different from those of mRNAs in HEK293 (Figure 5A) and other human cell lines (Figure S11A). This difference in tAIs is concordant with differences in ribosome binding, estimated from Ribo-seq data, in two ways. First, the fraction of transcripts with Ribo-seq reads was significantly larger for mRNAs than for smORFs, and both of these were larger than the fraction of lincRNA pORFs with Ribo-seq reads (Figure 5B and S11B). Second, relative ribosome-binding, calculated as the ratio of normalized Ribo-seq reads to cytoplasmic expression level of the ORF harboring transcript (see Methods), was significantly lower for lincRNAs than for other cytoplasmic RNA types, in HEK293 (Figure 5C) and most human cell lines (Figure S11B).

The difference in relative ribosome-binding between lincRNAs and smORFs in most cells appears to be more pronounced than the difference in tAIs. Previously, the codon usage immediately downstream of mRNA start codons was shown to differ from the rest of the mRNA coding region and was suggested to facilitate efficient translation initiation (Bentele et al., 2013; Tuller et al., 2010b). Thus, we investigated tRNA adaptation of the first codons downstream of start codons for different RNA types. Indeed, tAIs calculated for the first 10 codons (first-10-codon tAIs) were significantly higher for mRNAs and smORFs compared to lincRNAs (Figure 5D), and first-10-codon tAIs of smORFs were not different from first-10-codon tAIs of mRNAs (Figure 5D and S11D). We also investigated if the start codon RNA sequence and structure context could explain ribosome-binding differences between these cytoplasmic RNA types, but found no evidence for that (Figure S11E,F).

These results suggest that lower tRNA adaptation of codons in cytoplasmic lincRNA pORFs is associated with reduced ribosome-binding, and our results are in line with codons at the beginning of ORFs being more important for facilitating (or hindering) translation (Bentele et al., 2013; Tuller et al., 2010b).

*Cell line-specific tAIs are concordant with ribosome-binding differences across human cell lines for most cytoplasmic lincRNAs*

The comparison between RNA types, suggested that reduced codon adaptation can attenuate ribosome binding to cytoplasmic lincRNAs. To quantify this effect, we focussed on 45 lincRNAs that are cytoplasmic in all five human cell lines and analysed to what extent changes in tRNA abundances between cell lines affect ribosome binding to these lincRNA pORFs. We first confirmed differences in relative tRNA abundances between the human cell lines studied. We found that relative tRNA abundances in HepG2 and HEK293 cells differed from those in GM12878, Hela-S3 and K562 cells, which are very similar to each other (Figure 6A and S12A). As expected, different relative tRNA abundances between cell lines resulted in different cell line-specific tAIs (z-scored relative to tAIs of control pORFs in each cell line) for pORFs in 45 cytoplasmic lincRNAs (Figure 6B). Notably, in HepG2 and HEK293 cells, cytoplasmic lincRNAs appeared to be significantly better adapted to relative tRNA abundances than in other cell lines. Analysing ribosome-binding differences for the same cytoplasmic lincRNAs across cell lines, we found that differences in ribosome binding mostly agreed with differences in cell line-specific tAIs (Figure 6C). In particular in HepG2 cells, relative ribosome-binding to cytoplasmic lincRNAs was significantly stronger than in other cell lines.

For the majority of cytoplasmic lincRNAs (~60%) we found a positive correlation between their cell line-specific tAIs and relative ribosome-binding across cell lines (Figure 6D). The correlation coefficients were similarly distributed as for cytoplasmic mRNAs (Figure 6D and S12B), with p values largely more significant than expected (Figure 6E), similar to mRNAs (Figure S12C). The relationship between cell line-specific tAIs and relative ribosome-binding across human cell lines is shown for four cytoplasmic lincRNA pORFs in Figure 6F.

Together, this good agreement between cell line-specific tAIs and ribosome binding across human cell lines provides strong evidence for the codon bias in cytoplasmic lincRNAs functioning as a mechanism to attenuate ribosome binding.

## Discussion

In contrast to mRNAs, lincRNAs are classified as not coding for proteins, and many lincRNAs have been shown to have regulatory functions (Ulitsky and Bartel, 2013). However, the mechanisms counteracting translation of putative ORFs in cytoplasmic lincRNAs are unclear. Here, we carried out a comprehensive analysis of lincRNA sequences in six different species, covering a wide range of mRNA codon usage bias strengths. We provide evidence that many pORFs within lincRNAs, in particular those found in the cytoplasm, are biased towards codons recognized by lower abundance tRNAs. We propose that this codon bias attenuates translation of cytosolic lincRNA pORFs.

In the absence of any functional role of peptides resulting from lincRNA translation, reducing stable association of lincRNAs with ribosomes is likely to be advantageous for the cell for several reasons. First, unwanted lincRNA translation would waste energy, reduce the pool of ribosomes available for mRNA translation and lead to the synthesis of possibly toxic peptides. Given the association between translation and transcript stability for mRNAs (Presnyak et al. 2015; Tuck et al. 2020), unwanted lincRNA translation may also interfere with their stability and function.

By definition lincRNAs lack coding potential, i.e. they have no ORFs longer than 300 nucleotides and no ORFs with any homology to known protein domains (Harrow et al., 2012). Thus, it may not be surprising that pORFs within lincRNAs present codon compositions different from mRNA coding regions. However, our analysis also revealed that the codon composition of lincRNA pORFs is distinct from that of intronic and intergenic control pORFs. Strikingly, in species with stronger mRNA codon bias, i.e. worm and fission yeast, lincRNA codon frequencies were less correlated to the matching tRNA abundances than control codons, while for mRNA codons this correlation was stronger than for controls (Figure 3D,E). This suggests that, like codons in mRNA, lincRNA codons likely underwent some adaptation, but, in contrast to mRNA, to reduce matching with abundant tRNAs.

Also in species with weaker mRNA codon usage bias, such as human and mouse, lincRNA codon frequencies correlate less with tRNA abundances than mRNA codons (i.e. lincRNAs tend to have smaller tAIs than mRNAs), yet on average lincRNA tAIs are larger than that of control sequences. This observation may reflect an overall weaker evolutionary pressure on codons within the majority of lincRNAs in these species, such as those that are very lowly expressed or retained in the nucleus. It may also be an indication for lincRNA sequences originating recently from coding regions, e.g. pseudogenes (Hezroni et al., 2017) or becoming coding sequences (Knowles and McLysaght, 2009). Indeed, focussing on highly expressed cytoplasmic lincRNAs in nine different cell types and species, we found that their tAIs were significantly lower than that of cytoplasmic mRNAs (>1 standard deviation for all nine cell types) and of control tri-nucleotides (except for mESCs and HepG2, see below; Figure 4B), suggesting a stronger evolutionary pressure for adapting these lincRNA sequences.

Finally, detailed analysis of 45 lincRNAs that are found in the cytoplasms of five human cell lines confirmed that differences in tRNA abundances between cell lines result in tAI variation that correlates with the variation in ribosome-binding. This strongly suggests that codon adaptation can provide a mechanism for counteracting ribosome binding in lincRNAs. Among the five human cell lines we analysed, tAIs and ribosome binding of cytoplasmic lincRNAs were particularly high for the liver cancer cell line HepG2 and increased for the kidney cell line HEK293 (Figure 6). This suggests enhanced translation of cytoplasmic lincRNA pORFs in these cell lines, which was indeed reported before for liver and kidney cells (van Heesch et al., 2019). Higher tAIs were also observable for mESCs (Figure 4B), which might be related to the observation that rare tRNA anticodons are more frequent during cellular proliferation (Guimaraes et al., 2020), and mESCs having a very short doubling time (Wang et al., 2011).

Our analysis of a codon bias in lincRNAs does not exclude that other mechanisms may be in place to counteract ribosome binding and translation of cytoplasmic lincRNAs. Indeed, consistent with fission yeast being the species with the strongest evolutionary selection, we found that lincRNAs in yeast have fewer and shorter than expected pORFs and we predicted the structural accessibility in the vicinity of lincRNA pORF start codons to be reduced in this species. Thus the reduction of pORF number and length and a stronger folding around pORFs start codons might be additional ways to reduce the chance of cytoplasmic lincRNAs to be translated.

In conclusion, in this study we provided a comprehensive analysis of putative codons in lincRNA pORFs. Our results suggest that these are biased towards those recognised by less abundant tRNAs. We propose that this is a mechanism to counteract unwanted translation and ribosome binding. The resulting sequence signatures may help in distinguishing bona-fide lincRNAs with regulatory roles in the cytoplasm from those coding for peptides, yet more work will be needed to distill such signals in the context of other potential constraints on lincRNA sequences related to their regulatory function, such as structure or binding motifs. Another interesting aspect is that tRNA concentrations can vary across cell types and conditions, imposing potentially different constraints on the evolution of codon frequencies to either curb ribosome binding of true lincRNAs, or promote it for peptide-coding transcripts. Analysis of how natural genetic variations or targeted mutations in lincRNA sequences impact on ribosome binding and translation might shed light onto these questions.

## Methods

*Gene annotations*

Gene annotations and genomic fasta sequences were downloaded from GENCODE (www.gencodegenes.org; v19 (corresponding to hg19) for human and vM16 (corresponding to mm10) for mouse) (Frankish et al., 2019), from Ensembl (www.ensembl.org; dr11 for zebrafish, dm6 for fly, and ce11 for worm), and from EnsemblFungi (http://fungi.ensembl.org; ASM294v2 for fission yeast). In the case of worm (C. elegans), we additionally included recently identified lincRNA genes by Akay et al. (Akay et al., 2019), which were not contained the Ensembl gene annotation. We chose fission yeast as opposed to the more commonly studied budding yeast, because the number of annotated lincRNA genes was much larger in fission yeast (>1000) than in budding yeast (<100). For all species, we excluded genes on mitochondrial chromosomes from our analysis, as these are translated by mitochondrial tRNAs.

*mRNA coding regions*

For each mRNA gene the transcript isoform with the longest coding region starting with a canonical start codon (ATG) was considered.

*Identification of putative ORFs (pORFs) in lincRNAs*

LincRNA genes that did not overlap with any other gene annotation on either strand were considered. All pORFs (longer than 30 nucleotides) were identified that start with a canonical start codon (AUG) and end at the first in-frame stop codon (UAG, UAA, UGA) using a custom python script. For each lincRNA gene, the transcript isoform with the longest pORF was considered for further analysis.

*Intronic and intergenic control sequences*

For each considered lincRNA transcript, a region of the same length was randomly selected in mRNA introns (that did not overlap any other genes on the opposite strand and excluding 10 nucleotides flanking exons) and in intergenic regions (that did not overlap any gene annotation on both strands) using bedtools shuffle. The random selection was repeated 10 times for each considered lincRNA transcript and longest pORFs (>10 codons) were identified as for lincRNA sequences. For the statistical comparison of pORF identification in lincRNA sequences with that in random intronic and intergenic sequences (Figure S1), the random selection of size-matched control regions in introns and intergenic regions was repeated 100 times.

*Sequence and structure context analysis around start codons*

To analyse the sequence context around start codons the fraction of nucleotides (A,C,G,T) at each position in a region of +/- 12 nucleotides around start codons was counted. The information content (*I*) was calculated as: $I = 2 + \sum_{n=A,C,G,T} f_n \, log2(f_n)$, where $f_n$ is the fraction of occurrences of nucleotide n. The probability (*P*) for (or similarity with) the mRNA start codon region motif was calculated as: $P = exp((\sum_{l=-12}^{12} log(f_{n(l)})) / L)$, where *L* is the total length of the start codon region and $f_{n(l)}$ is the fraction of nucleotide n (based on the mRNA motif) at position l around the start codon of an ORF.

The structure context in Figure S3 was calculated as the probability for a nucleotide being accessible (not paired) in a sliding window of 70 nucleotides (sliding by 1 nucleotide) using RNAplfold (Bernhart et al., 2011) with the parameters -u 1 -W 70.

In Figure 1D and S11E, the minimum free energy of sequence regions +/- 25 nucleotides around start codons was calculated using RNAfold (Hofacker, 2004).

*Principal component analysis*

The principal component analysis in Figures 2D and S4A was performed on z-scored codon frequencies for different species and RNA types in Python 2.7 using matplotlib.mlab.PCA.

*Estimation of relative tRNA abundances based on tRNA gene counts*

tRNA gene predictions were downloaded from GtRNAdb (http://gtrnadb.ucsc.edu/GtRNAdb2/) (Chan and Lowe 2016) for each species studied. The numbers of annotated tRNA genes coding for the same tRNA anticodon type were counted considering the high confidence set of predicted tRNA genes. Effective tRNA abundances were calculated using previously determined weights for tRNA editing and wobble-base pairing efficiencies (dos Reis et al., 2004). In particular, the weights *w* are *w*(G:U)=0.41, *w*(I:C)=0.28, *w*(I:A)=0.999, and *w*(U:G)=0.68, where the first letter denotes the first nucleotide of a tRNA anticodon nucleotide triplet and the second letter the third nucleotide of a codon. Effective tRNA abundances were then calculated as:

tRNA-NNT = tRNA-NNT + (1 - w(G:U)) * tRNA-NNC

tRNA-NNC = tRNA-NNC + (1 - w(I:C)) * tRNA-NNT

tRNA-NNA = tRNA-NNA + (1 - w(I:A)) * tRNA-NNT

tRNA-NNG = tRNA-NNG + (1 - w(U:G)) * tRNA-NNA

The above indicated nucleotide triplet is the reverse complement of the tRNA anticodon sequence (i.e. the codon it recognizes). N stands for any nucleotide.

*Estimation of cell type-specific relative tRNA abundances*

Due to the repetitive nature of tRNAs, their strong secondary structure, and the high frequency of post-transcriptional tRNA modifications, high-throughput experimental quantification of tRNA expression levels is still challenging. Two dedicated experimental high-throughput approaches for the quantification of tRNA expression, hydro-tRNA-seq (Gogakos et al., 2017) and DM-tRNA-seq (Zheng et al., 2015), have been proposed, but have only been applied to a limited number of cell types including HEK293 cells. On the other hand, smallRNA-seq and H3K27ac-ChIP-seq data were used previously to quantify tRNA expression (Gingold et al., 2014; Hernandez-Alias et al., 2020; Ji et al., 2015), and these data are more widely available for different species and cell types. We established which of the latter two experimental approaches is better suited to quantify relative tRNA expression levels by comparing relative tRNA expression levels based on smallRNA-seq and H3K27ac-ChIP-seq in HEK293 with those from hydro-tRNA-seq and DM-tRNA-seq. Since smallRNA-seq-based relative tRNA abundances correlate better with those from hydro-tRNA-seq and DM-tRNA-seq data (Figure S8), we used smallRNA-Seq data to calculate cell type-specific relative tRNA abundances in our analysis (for sources of smallRNA-Seq data see Table S3). Effective tRNA abundances for tRNA anticodon types not encoded in a genome were calculated as described above.

In the following, the tRNA quantification based on different experimental data is explained in detail:

(a) relative tRNA quantification based on smallRNA-seq data:

SmallRNA-seq reads (fastq files) for different species and cell types were downloaded from various sources (see Table S3). Reads were pre-processed using the fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and then mapped to native and mature tRNA sequences using segemehl v0.2 (Hoffmann et al., 2009). Of the mapped reads, only those with a minimum length of 15 nucleotides were retained. To account for the high frequency of tRNA modifications, which may result in mapping mismatches, the allowed mismatch ratio (mismatched nucleotides / read length) was set to <=10%. Other mismatch ratio cutoffs (<7% and <15%) were tested, but did not improve the correlation with tRNA sequencing approaches (hydro-tRNA-seq (Gogakos et al., 2017) and DM-tRNA-seq (Zheng et al., 2015)), or resulted in a smaller fraction of reads mapping to tRNA sequences in sense direction. The number of smallRNA-seq reads mapping to each tRNA anticodon type divided by the total number of mapped reads was taken as an estimate for the relative tRNA abundance.

(b) relative tRNA quantification based on H3K27ac ChIP-seq data for HEK293:

Bedfiles of identified H3K27ac-ChIP-seq peaks in HEK293 were downloaded from ENCODE (www.encodeproject.org) (Consortium and The ENCODE Project Consortium, 2004) (see Table S3). H3K27ac ChIP-seq peaks that overlapped tRNA gene annotations extended by 500 nucleotides up- and downstream were determined using bedtools (Quinlan, 2014). Relative tRNA abundances were estimated by the ratio of the sum of peak enrichment values (log2 fold enrichment H3K27ac-ChIP-seq over background, in column 7 of the bedfiles) for each tRNA anticodon type to the peak enrichment values of all peaks overlapping extended tRNA genes.

(c) relative tRNA quantification from experimental tRNA-seq methods applied in HEK293:

Hydro-tRNA-seq-based tRNA quantifications were taken from Table S5 of Gogakos et al. (Gogakos et al., 2017). DM-tRNA-seq (Zheng et al., 2015) read counts of two replicates were downloaded from GEO (www.ncbi.nlm.nih.gov/geo, accession numbers GSM1624820 and GSM1624821) and tRNA abundances were calculated as the average of the two replicates.

*tRNA adaptation index (tAI)*

As proposed by dos Reis et al. (dos Reis et al., 2004), the tAI of an ORF was calculated as the geometric mean of normalized, effective tRNA abundances complementary to codons in the ORF:

$$tAI = \sqrt[n]{\prod_{i}^{n} w_i},$$

where $n$ is the total number of codons in an ORF, and $w_i$ is the normalized effective tRNA abundance of the tRNA anticodon complementary to the codon at position $i$.

Normalized tRNA abundances were obtained by dividing each effective tRNA abundance by the maximum effective tRNA abundance:

$$w_i = f_{tRNA_i} / max(f_{tRNA}),$$

where $f_{tRNA_i}$ is the frequency of the tRNA complementary to the codon at position $i$.

First-10-codon (or first-20-codon or first-30-codon) tAIs were calculated by considering only the first 10 (or 20 or 30, respectively) codons downstream of AUG start codons.

*tAIs of random control sequences for each mRNA coding region and lincRNA longest putative ORF*

(a) shuffled sequences:

Shuffled sequences were generated by random permutations of the nucleotides in the ORF. The shuffled tAI was calculated as the average over 100 random permutations of the original ORF sequence.

(b) frame-shifted sequences:

tAIs of frame-shifted sequences were calculated by counting codons starting one or two nucleotides downstream of start codons and ending two or one nucleotides, respectively, upstream of stop codons of the original ORF.

(c) random codons coding for the same amino acid sequence:

For random codon tAIs, each codon in an ORF was replaced by a random codon coding for the same amino acid. This was repeated 100 times and the average tAIs was returned.

*Quantification of RNA expression based on polyA-selected RNA-Seq data*

To quantify transcript expression levels in cytosol and total cells polyA-selected RNA-seq data were used from several sources. The sources and accession codes are listed in Table S3. In case processed transcript quantifications were publicly available, these were downloaded and used for further analysis. Otherwise, polyA-selected RNA-seq reads were preprocessed using the fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), and transcript expression levels were quantified with RSEM (Li and Dewey, 2011) using the STAR aligner (Dobin et al., 2013). The threshold to distinguish between expressed and unexpressed RNAs is given in Table S3 for each cell type. The cutoff for distinguishing nuclear and cytoplasmic lincRNAs was set to the first-quartile (25%) of the cytosolic expression values for expressed mRNAs, for each species and cell type. We excluded histone mRNAs, as these are not usually polyadenylated, and result in incorrect expression and cytosolic expression values in polyA-selected RNA-Seq.

*Quantification and analysis of Ribo-Seq data*

Ribo-Seq data for five human cell lines were downloaded from several sources (see Table S3). Adapter sequences were trimmed from read ends using cutadapt v1.8 (Martin, 2011), and reads were retained with a certain length (16 to 35 nt) and minimum quality score of 30 in at least 90% of read bases. Reads were discarded that mapped to human rRNAs or tRNAs (ENSEMBL database v91, (Zerbino et al., 2018)) using bowtie2 v2.3.0 (-L 15 -k 20) (Langmead and Salzberg, 2012). Reads were further removed if mapping to mRNA coding regions or longest lincRNA pORFs of two or more gene loci annotated in the human transcript database (ENSEMBL v91).

The position of the ribosome P site relative to the Ribo-Seq read start (read offset) was determined by analysing Ribo-Seq reads overlapping mRNA start codons. For each Ribo-Seq read length, the three

most frequent read offsets were considered, if these were found in more than 500 reads and in more than 1% of all reads overlapping mRNA start codons.

Ribo-Seq reads were counted for each longest mRNA coding region and lincRNA pORF, if the corresponding ribosome P site was in-frame to the longest coding region for mRNAs and longest pORF for lincRNAs.

Relative ribosome-binding was calculated for each longest coding region/pORF as the log2 ratio of the in-frame Ribo-Seq read count (normalized by coding region/pORF length and the total number Ribo-Seq reads mapping in-frame to mRNA coding regions and multiplied by 1e+9) to the cytosolic expression level (FPKM) of transcripts harboring the longest coding region/pORF.

We excluded snoRNA genes from this analysis, as these may associate with ribosomes during translations, making it difficult to distinguish if these are bound by ribosomes themselfs or associate with ribosomes that are translating other RNAs.

*Analysis of small protein-encoding ORFs (smORFs)*

We excluded in our analysis of longest pORFs those lincRNA transcripts that overlapped with likely small protein-coding regions. Likely small protein-coding regions were defined as regions with high PhyloCSF score (Lin et al., 2011), downloaded from https://data.broadinstitute.org/compbio1/PhyloCSFtracks/ for human (hg19), mouse (mm10), and fly (dm6), and regions encoding small proteins that were identified experimentally before in three human cell lines based on Ribo-Seq data. Regions were taken from Supplementary Table 1 of Martinez et al. (Martinez et al. 2020).

In Figure 5, smORFs were defined as annotated lincRNAs, whose longest pORF overlapped with genomic regions of increased PhyloCSF score (Lin et al., 2011) or with previously identified small protein-encoding ORFs (based on Ribo-seq data by (Martinez et al. 2020); see above).

**Acknowledgements**

We would like to thank the ENCODE consortium and other experimental labs for generating the publicly available data used in this study. We also thank two anonymous reviewers for their comments on a previous version of the manuscript. This work was supported by the Swiss National Science Foundation (grant FN 310030_152724/1).

**Author Contributions**

AB, ACM and SB designed the project. AB carried out the computational analysis and prepared the results. RD processed and mapped Ribo-Seq data. AB, ACM, and SB discussed results and wrote the paper.

**Declaration of Interests**

The authors declare no competing interests.

**Figure 1**

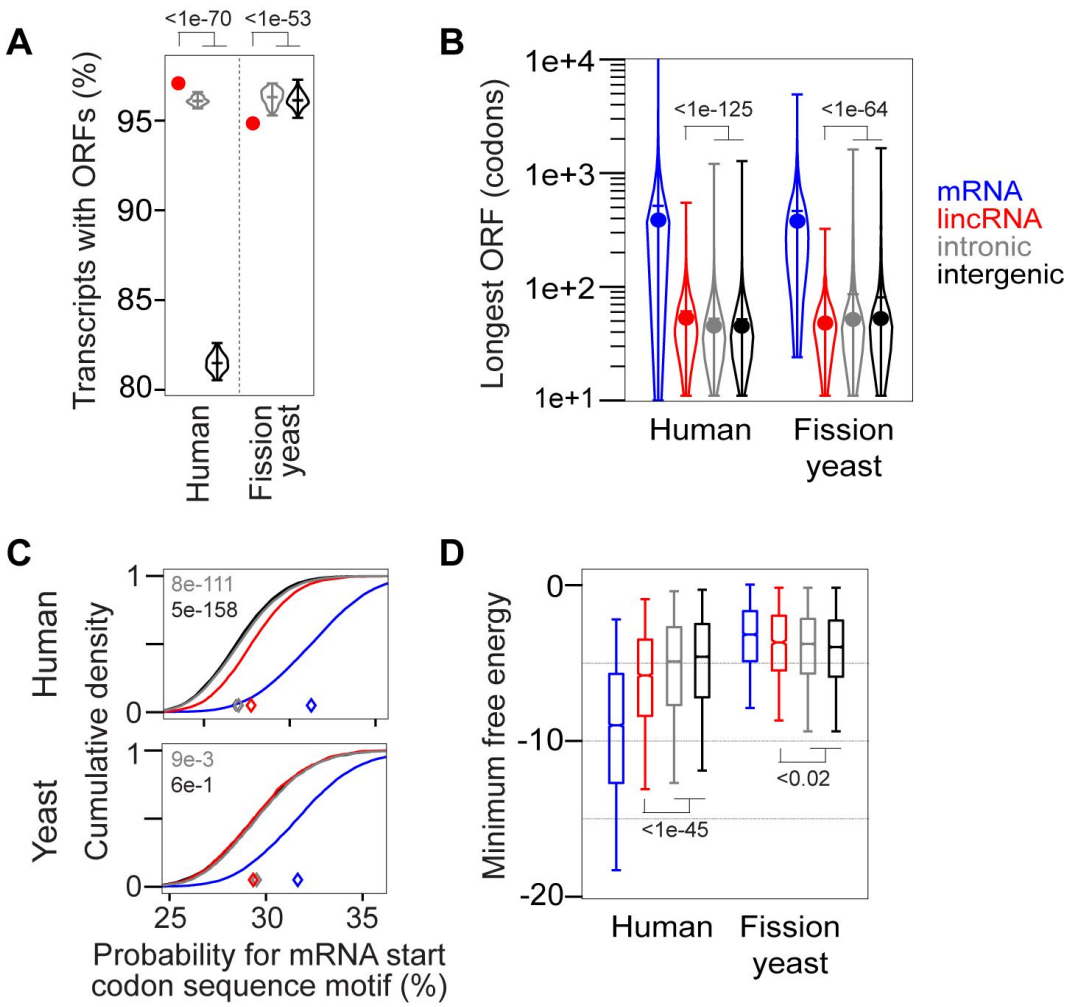
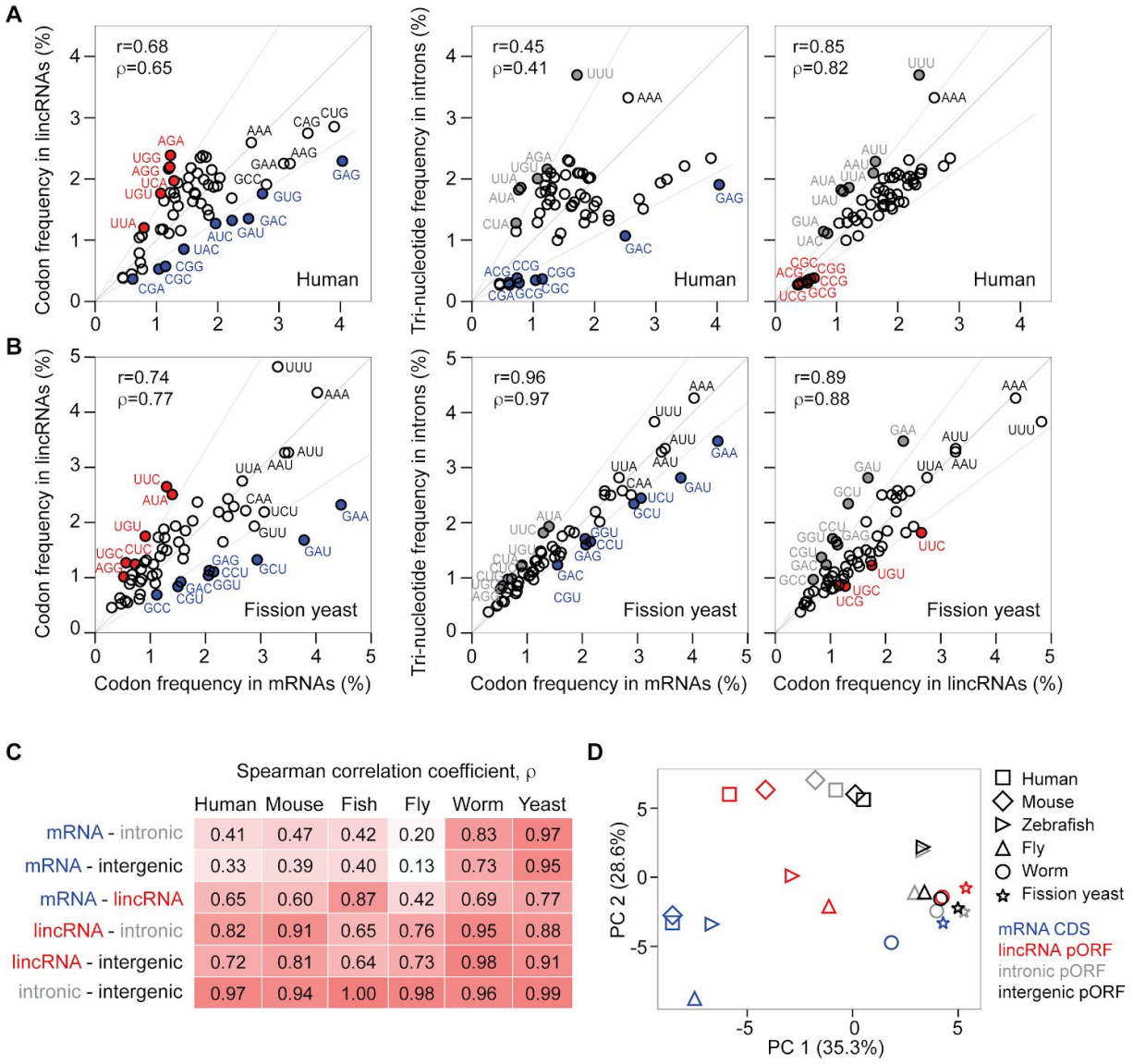



**Figure 1: Prevalence and lengths of lincRNA pORFs and their RNA sequence and structure context around start codons.**

(A) Median percentage of lincRNA transcripts with pORF (red dot) and violin plots of medians of percentages of size-matched, randomly selected intronic (gray) and intergenic (black) control regions with pORF for human and fission yeast. Random selection of control regions was repeated 100 times. P values are indicated from a one-sample t test to compare the median for lincRNAs with those from 100 randomly selected control regions. (B) Violin plots of length distributions of longest mRNA coding regions (blue), longest lincRNA pORFs (red), and longest pORFs in intronic (gray) and intergenic (black) control regions. P values are indicated from a one-sample t test to compare the median length of longest lincRNA pORFs with those of longest pORFs in 100 randomly-selected control regions. (C) Similarity (see Methods) of the start codon sequence context of longest lincRNA pORFs (red), intronic (gray) and intergenic (black) longest control pORFs with the consensus mRNA start codon sequence motif. P values are indicated from Wilcoxon's rank-sum test to compare the similarities for lincRNA pORFs with those for intronic and intergenic control pORFs. (D) Box plots of minimum free energies for regions of +/- 20 nucleotides around start codons of longest coding region in mRNA (blue), pORFs in lincRNAs (red), and control pORFs in intronic (gray) and intergenic (black) regions. P values are indicated from Wilcoxon's rank-sum test to compare minimum free energies for longest lincRNA pORFs with those for longest control pORFs.

**Figure 2**




| | Human | Mouse | Fish | Fly | Worm | Yeast |
|---|---|---|---|---|---|---|
| mRNA - intronic | 0.41 | 0.47 | 0.42 | 0.20 | 0.83 | 0.97 |
| mRNA - intergenic | 0.33 | 0.39 | 0.40 | 0.13 | 0.73 | 0.95 |
| mRNA - lincRNA | 0.65 | 0.60 | 0.87 | 0.42 | 0.69 | 0.77 |
| lincRNA - intronic | 0.82 | 0.91 | 0.65 | 0.76 | 0.95 | 0.88 |
| lincRNA - intergenic | 0.72 | 0.81 | 0.64 | 0.73 | 0.98 | 0.91 |
| intronic - intergenic | 0.97 | 0.94 | 1.00 | 0.98 | 0.96 | 0.99 |

**Figure 2: Distinct tri-nucleotide ("codon") frequencies for lincRNAs.**

(A) Comparison of tri-nucleotide ("codon") frequencies (excluding start and stop codons) in human mRNA coding regions and lincRNA pORFs (first panel), mRNA coding regions and intronic control pORFs (second panel), and lincRNA pORFs and intronic control pORFs (last panel). Pearson (r) and Spearman ( $\rho$ ) correlation coefficients are indicated. Codons with absolute z-score >1.2 (gray dotted lines) of their log2 frequency ratio are labeled in red, blue or gray, if higher in lincRNA, mRNA or introns, respectively. Codons with high frequencies (>2%) in both RNA types are labeled in black. (B) Same as (A), but for tri-nucleotide ("codon") frequencies in fission yeast. (C) Spearman correlation coefficients ( $\rho$ ) between tri-nucleotide frequencies in mRNA, lincRNA, intronic and intergenic regions for six species. (D) Principal component analysis for the same tri-nucleotide frequencies as in (C).

**Figure 3**

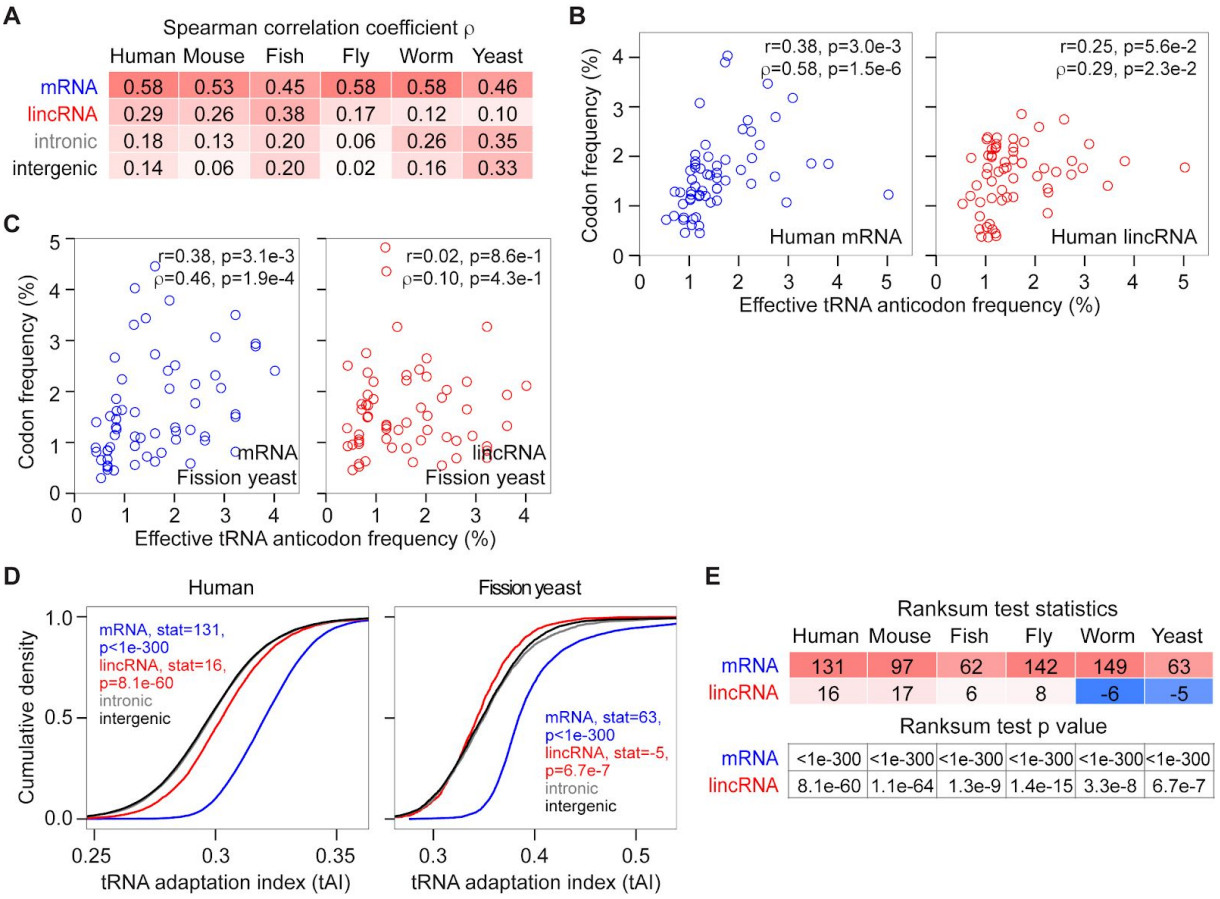



**Figure 3: LincRNA codons are less adapted to relative tRNA gene counts than mRNA codons for all species studied.**

(A) Spearman correlation coefficients ( $\rho$ ) between effective tRNA anticodon frequencies based on tRNA gene counts (see Methods) and codon frequencies in mRNA, lincRNA, intronic and intergenic control regions. (B) Scatter plots of effective tRNA anticodon frequencies based on tRNA gene counts and codon frequencies in mRNA coding regions (left panel) and lincRNA pORFs (right panel) in human. Pearson (r) and Spearman ( $\rho$ ) correlation coefficients with p values are indicated. (C) Same as (B) but for fission yeast. (D) Cumulative density of tRNA adaptation indices (tAIs) for mRNA (blue), lincRNA (red), intronic (gray) and intergenic (black) control sequences. Test statistics with p values are indicated from Wilcoxon's rank-sum test to compare tAIs of mRNAs and lincRNAs with those of control sequences. (E) Test statistics (top) and p values (bottom) from Wilcoxon's rank-sum test to compare tAIs of mRNAs and lincRNAs with those of control sequences for six species.

**Figure 4**

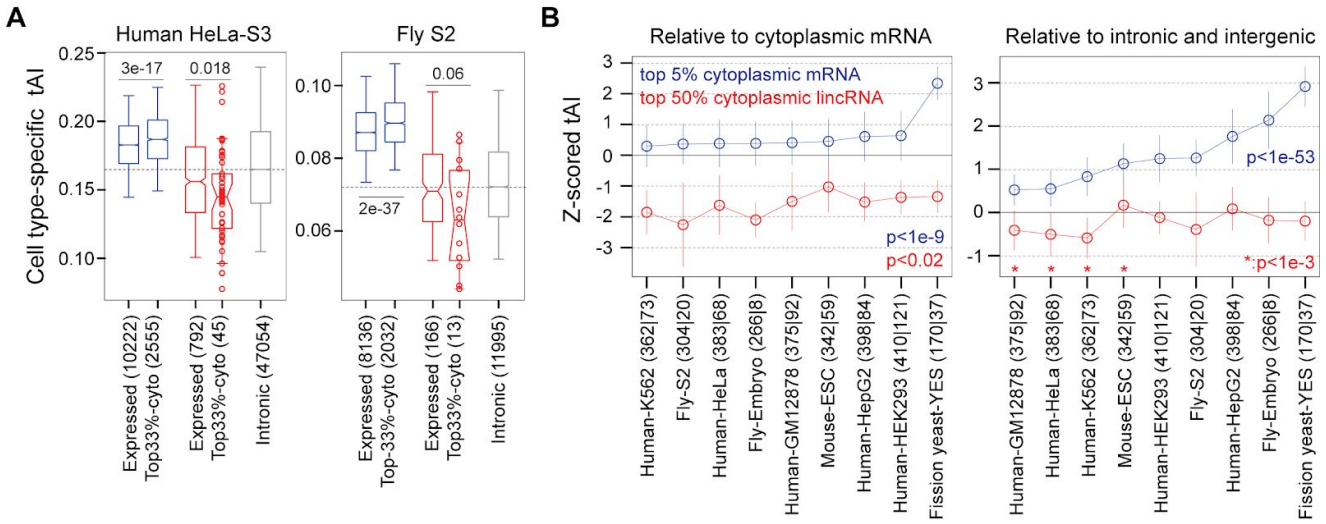



**Figure 4: Codon usage of highly expressed, cytoplasmic lincRNAs is less adapted to tRNA abundances.**

(A) Boxplots of cell type-specific tAIs for mRNA coding regions (blue) and lincRNA pORFs (red), for all expressed (first box) and top one-third RNAs with highest cytosolic expression (second box), for two cell types: human HeLa-S3 cells (first panel) and fly S2 cells (second panel). Red circles indicate tAIs of individual lincRNAs among the top one-third most highly expressed in the cytoplams. tAIs of intronic control pORFs are shown in gray. P values are calculated using Wilcoxon's rank-sum test. (B) Z-score of tAIs of the top 5% mRNAs (blue) and top 50% lincRNAs (red) with highest cytosolic expression for nine different cell types from four different species, sorted by median z-score for mRNAs (blue circles). Circles represent median values and error bars the median absolute deviation. Z-scores were calculated relative to tAIs of coding regions of all cytoplasmic mRNAs (left panel), and relative to tAIs of intronic and intergenic pORFs (right panel). P values are calculated using a one sample t test to compare the distribution of z-scored tAIs with 0.0.

**Figure 5**

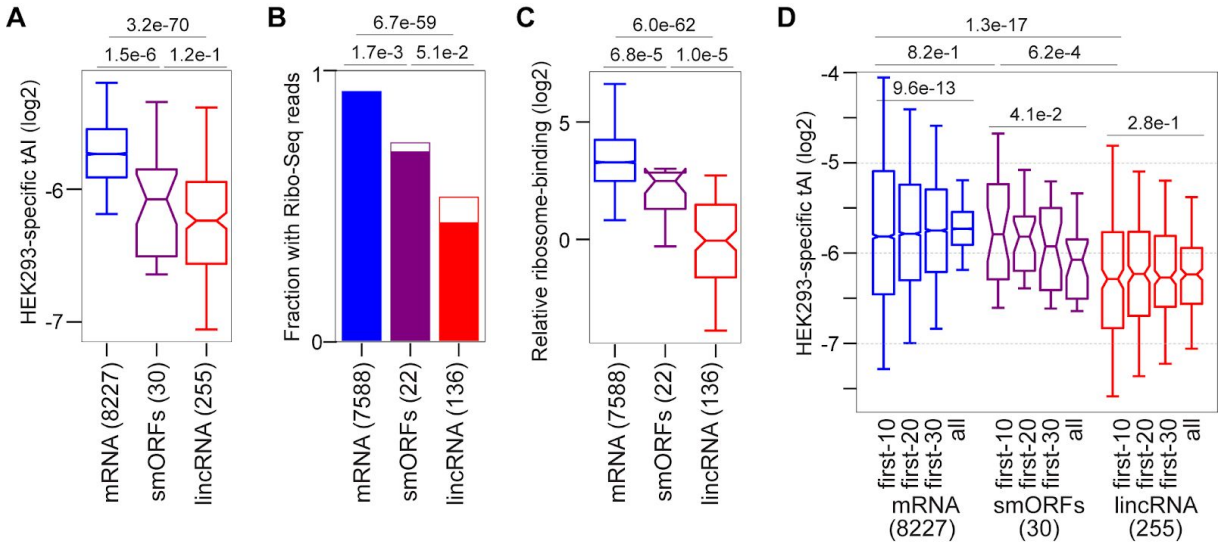



**Figure 5: Different tRNA adaptation between lincRNA pORFs and small protein-encoding ORFs, especially for codons close to ORF starts**

(A) Cell type-specific tAIs of three types of cytoplasmic RNAs in HEK293 cells: mRNA coding regions (blue), small protein-encoding ORFs (smORFs; blue) and lincRNA pORFs (red). P values are indicated from Wilcoxon's rank-sum test. (B) Fraction of ORFs from (A) with in-frame Ribo-Seq reads (filled bars: more than one read, empty bars: one read). P values are indicated from Fisher's exact test to compare fractions of RNAs with at least one in-frame Ribo-Seq read. (C) Relative ribosome-binding (calculated as the ratio of normed in-frame Ribo-Seq reads to the cytoplasmic expression level; see Methods) in HEK293 for ORFs with Ribo-Seq reads. P values are indicated from Wilcoxon's rank-sum test. (D) Cell type-specific tAIs for different ORF regions: the first 10, 20 or 30 codons downstream of start codons (first three boxes of each RNA type), and entire ORFs (last boxes of each RNA type). P values are indicated from Wilcoxon's rank-sum test.

**Figure 6**

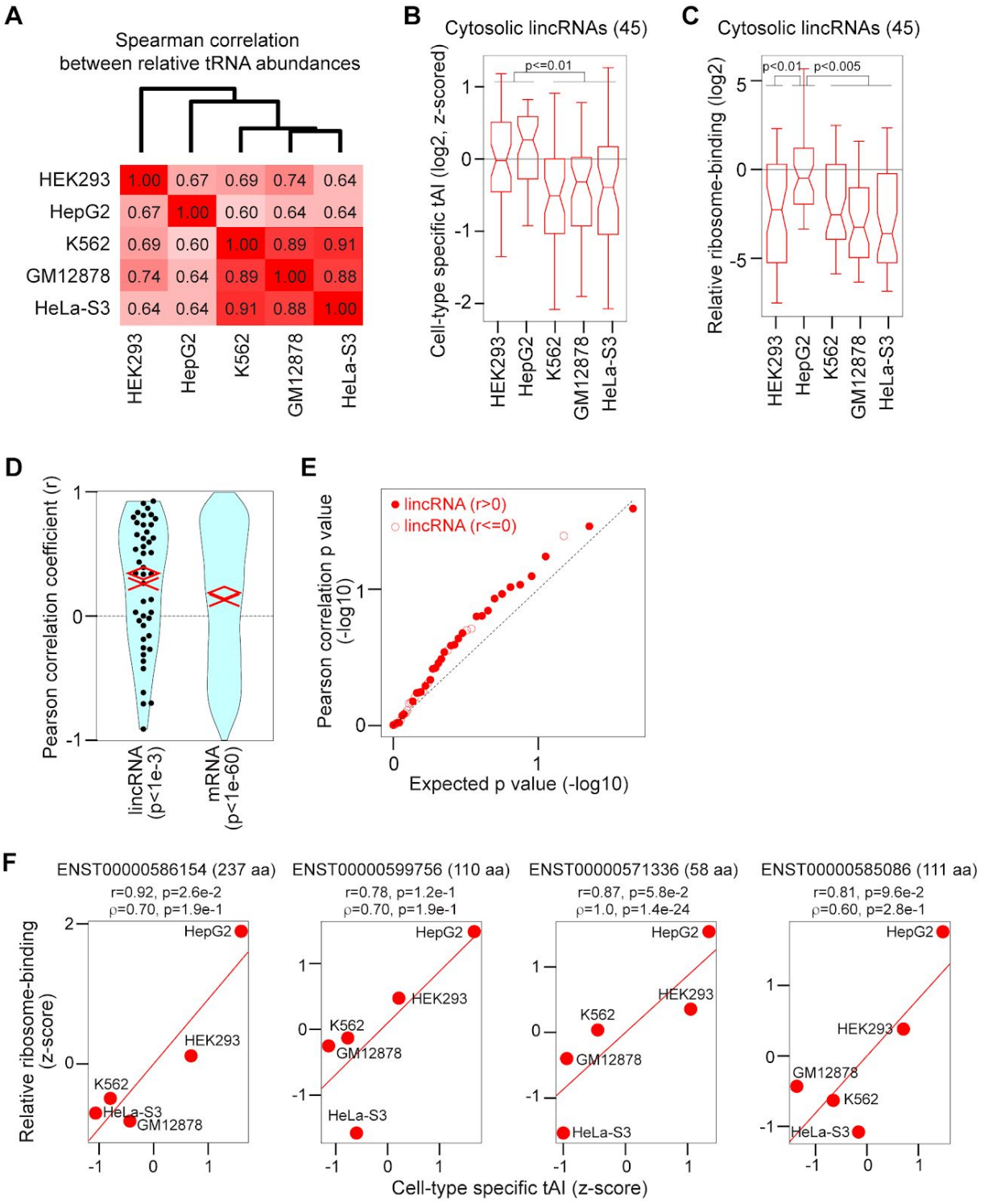
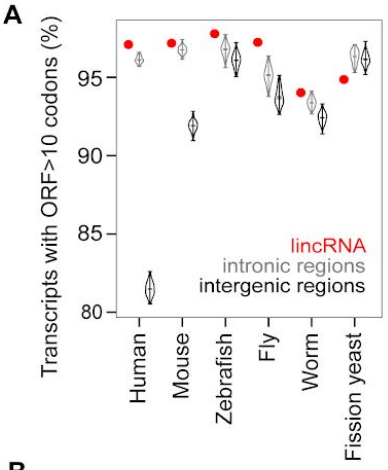



**Figure 6: Cell type-specific tRNA adaptation of cytoplasmic lincRNAs is reflected in ribosome-binding differences between five human cell lines.**

(A) Clustered pairwise Spearman correlation coefficients between cell type-specific tRNA abundances, estimated from smallRNA-Seq data, in five human cell lines, with linkage tree on top. (B) Cell type-specific tAIs of longest pORFs in 45 lincRNAs, which are cytoplasmic in all five cell lines. tAIs

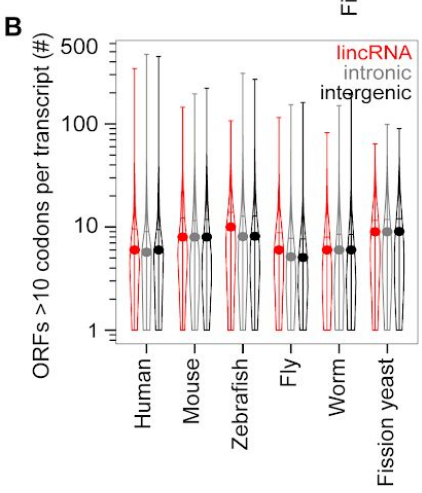
were log2-transformed and z-scored relative to log2-transformed tAIs of intronic and intergenic control pORFs for each cell line. (C) Relative ribosome-binding (calculated as ratio of normed in-frame Ribo-Seq reads (+ pseudo count of 0.1) to cytoplasmic expression level; see Methods) in each cell line. P values are indicated from Wilcoxon's rank-sum test. (D) Violin plots of Pearson correlation coefficients between relative ribosome-binding and cell type-specific tAI across five cell lines for cytoplasmic lincRNAs (45) and mRNAs (5250). Correlation coefficients for individual lincRNA pORFs are indicated by black dots, and median and mean values are marked by red diamonds and crosses, respectively. P values indicate a significant difference of correlations coefficients from 0.0 for lincRNAs and mRNAs, based on a one sample t test. (E) QQ-plot comparing p values of Pearson correlations between cell type-specific tAIs and relative ribosome-binding for cytoplasmic lincRNAs with a uniform distribution of p values. (F) Four examples for the correlation between z-scored cell type-specific tAIs and relative ribosome-binding across five cell lines. Pearson (r) and Spearman ( $\rho$ ) correlation coefficients with p values are indicated.

**Figure S1**



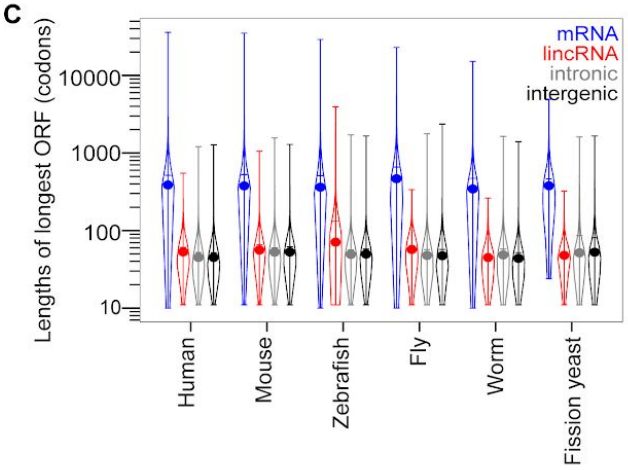

**Fraction of sequences with predicted ORF (>10 codons)**

| | lincRNAs | intronic (median) | intronic (p value) | intergenic (median) | intergenic (p value) |
|---|---|---|---|---|---|
| **Human** | 97.08 | 96.11 | 4.6e-71 | 81.48 | 2.5e-153 |
| **Mouse** | 97.19 | 96.74 | 1.1e-35 | 91.90 | 1.5e-120 |
| **Fish** | 97.83 | 96.75 | 3.0e-46 | 96.08 | 2.3e-62 |
| **Fly** | 97.29 | 95.11 | 9.5e-63 | 93.76 | 5.5e-81 |
| **Worm** | 94.03 | 93.38 | 7.9e-39 | 92.41 | 5.0e-67 |
| **Fission yeast** | 94.91 | **96.31** | 2.3e-56 | **96.15** | 2.4e-54 |




**Number of predicted ORFs (>10 codons) per sequence**

| | lincRNA (median) | lincRNA (mean) | intronic (mean median) | intronic (p value) | intergenic (mean median) | intergenic (p value) |
|---|---|---|---|---|---|---|
| **Human** | 6 | 9.58 | 5.70 | 1.8e-09 | 6.00 | 1.0 |
| **Mouse** | 8 | 12.26 | 7.99 | 7.3e-82 | 8.00 | 1.1e-89 |
| **Fish** | 10 | 13.87 | 8.09 | 8.5e-86 | 8.14 | 9.6e-75 |
| **Fly** | 6 | 8.79 | 5.14 | 4.6e-44 | 5.07 | 2.0e-61 |
| **Worm** | 6 | 7.98 | 6.00 | 1.0 | 6.00 | 1.0 |
| **Fission yeast** | 9 | 11.57 | 9.00 | 1.0 | 9.06 | 9.8e-03 |




**Length of longest predicted ORF (>10 codons)**

| | lincRNA (median) | lincRNA (mean) | intronic (mean median) | intronic (p value) | intergenic (mean median) | intergenic (p value) | mRNA (median) | mRNA (mean) |
|---|---|---|---|---|---|---|---|---|
| **Human** | 53 | 60.75 | 45.03 | 3.4e-127 | 45.15 | 6.2e-127 | 387 | 516.19 |
| **Mouse** | 56 | 65.44 | 52.88 | 8.1e-140 | 52.78 | 1.5e-119 | 377 | 528.51 |
| **Fish** | 71 | 131.70 | 49.31 | 3.5e-150 | 49.69 | 7.2e-144 | 363 | 511.49 |
| **Fly** | 57 | 63.09 | 47.82 | 2.5e-106 | 47.44 | 5.7e-104 | 466 | 657.63 |
| **Worm** | 45 | 49.71 | **48.58** | 1.7e-81 | 43.83 | 1.9e-46 | 344 | 472.01 |
| **Fission yeast** | 48 | 51.89 | **51.67** | 5.6e-66 | **52.58** | 2.1e-65 | 378 | 464.1 |

**Figure S1: Identification of pORFs in lincRNAs, and intronic and intergenic control regions**

(A) Left plot: Percentage of lincRNAs with pORF (>10 codons) in different species (red dots). As a control pORFs were predicted in randomly selected intronic and intergenic regions of the same length as lincRNAs (see Methods). Violin plots represent percentages of control sequences with pORF >10 codons obtained from 100 random selections of control regions. Right table: Percentages of lincRNA transcripts and randomly selected control regions with p values for comparing values for lincRNAs with randomly selected control regions calculated using a one-sample t test. Higher values for control regions than for lincRNAs are marked in bold. (B) Left plot: Violin plots of numbers of pORFs (<10 codons) per lincRNA transcript (red) and intronic (gray) and intergenic (black) control regions in different species. Dots represent median values. Right table: Number of pORFs (>10 codons) in lincRNA transcripts and in random control regions with p values comparing both calculated using Wilcoxon's rank-sum test. ("Mean median" refers to the average over median values for 100 random selections of control regions.) (C) Top plot: Violin plots of lengths of mRNA coding regions (blue), longest pORFs in lincRNAs (red) and in control regions (introns, gray, and intergenic, black). Dots represent median values. Bottom table: Lengths of longest pORFs (>10 codons) in lincRNA transcripts and in random control regions with p values comparing both calculated using Wilcoxon's rank-sum test, and for mRNA coding regions. ("Mean median" refers to the average over median values for 100 random selections of control regions.) Higher values for control regions than for lincRNAs are marked in bold.

**Figure S2**

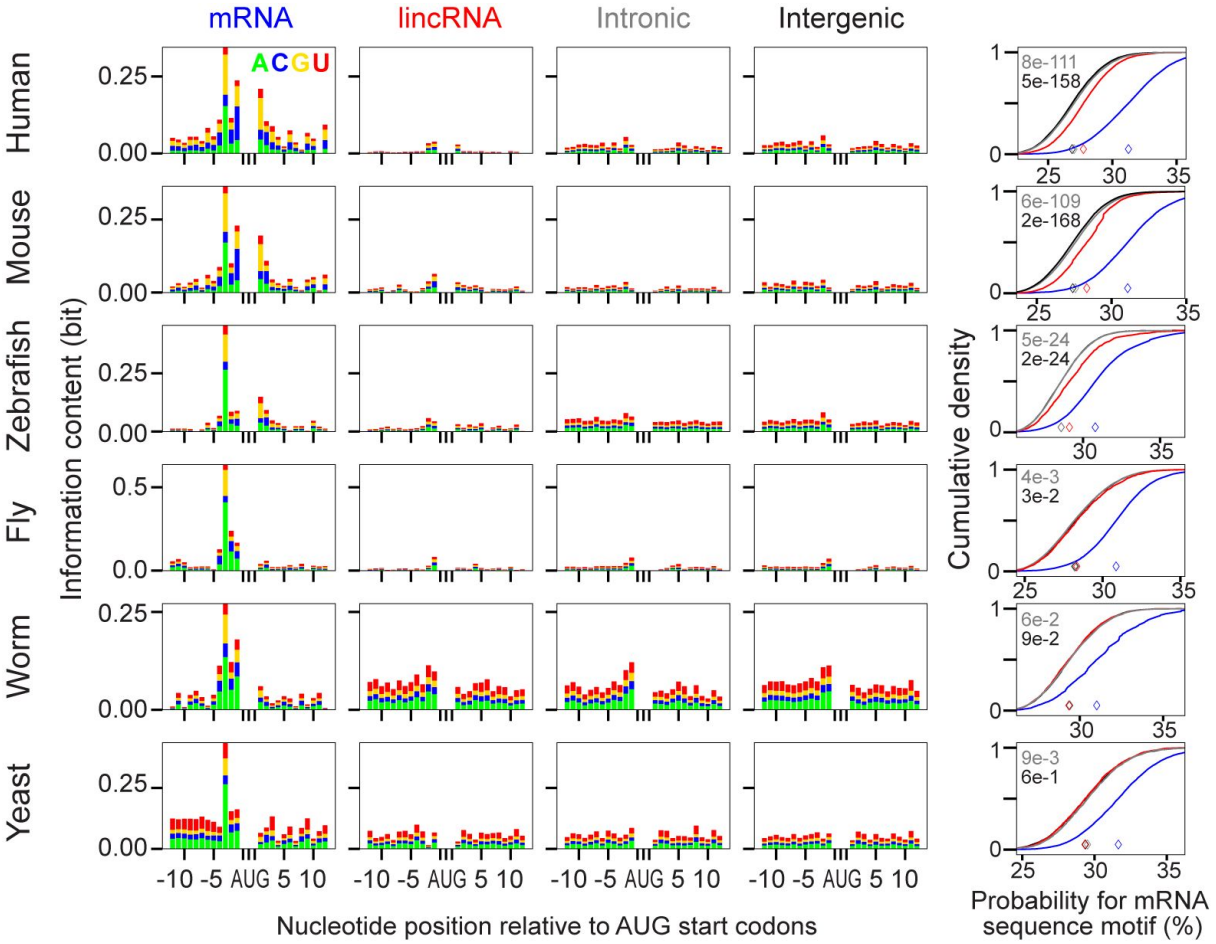



**Figure S2: Sequence context around AUG start codons in different RNA types and species**

Information content around AUG start codons in different RNA types (columns, indicated on top) and species (rows, indicated left). Right column of panels shows the cumulative density of probabilities for the mRNA sequence motif for individual transcripts. Numbers indicate p values, calculated using Wilcoxon's rank-sum test, to compare lincRNA probabilities with those for intronic (gray) and intergenic (black) control regions.

**Figure S3**

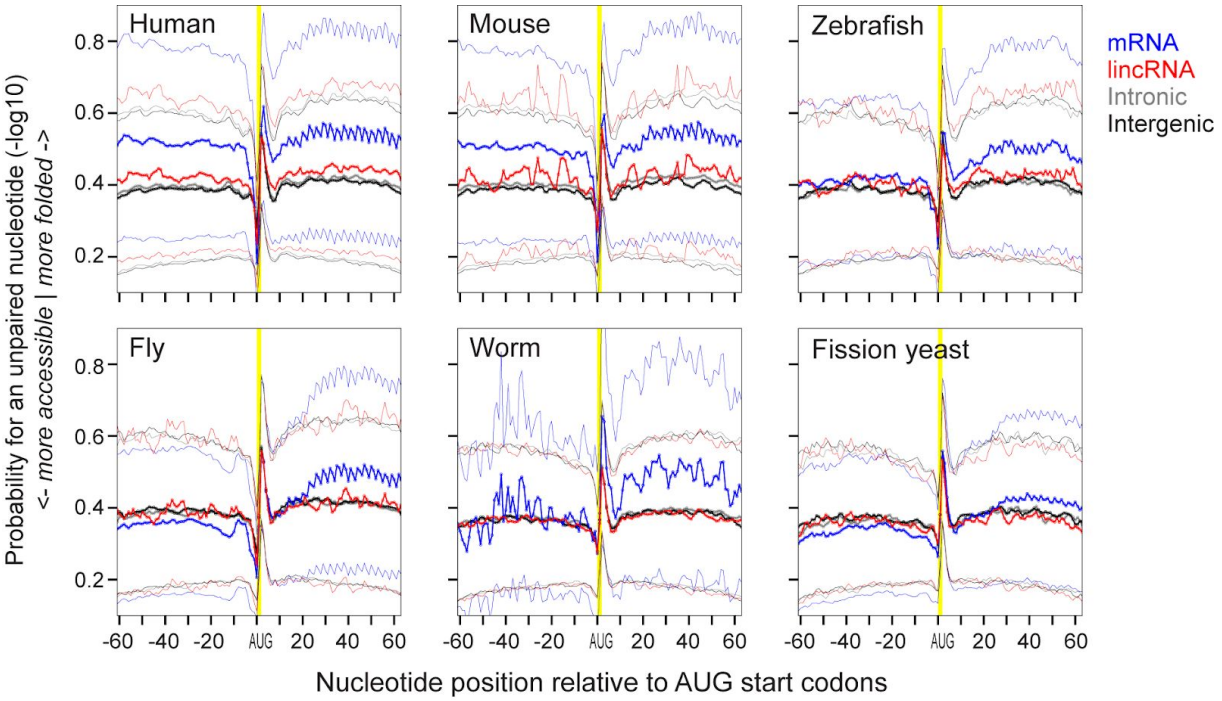



**Figure S3: Structure context around AUG start codons in different RNA types and species**

Probability that a nucleotide is unpaired for nucleotides within +/- 60 nucleotides around AUG start codons for mRNA coding regions (blue), lincRNA longest pORFs (red), and longest pORFs in intronic (gray) and intergenic (black) control regions, for different species. The AUG start codon is highlighted in yellow. Lines represent mean values +/- 0.1 standard deviations.

**Figure S4**

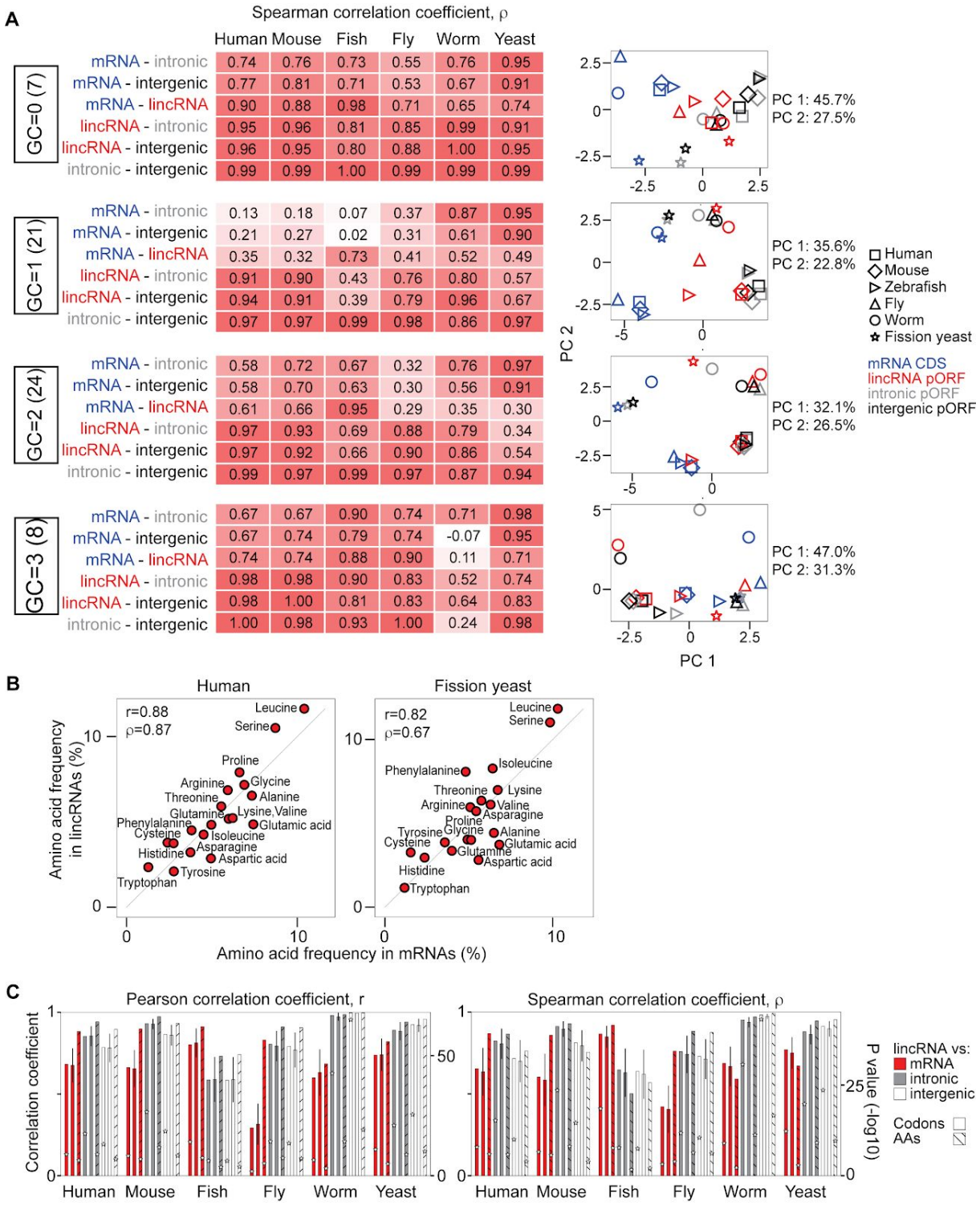



**Figure S4: Influence of GC content on codon frequencies correlations and variation of amino acid usage between RNA types**

(A) Left panels: Spearman correlation coefficients between codon frequencies in different RNA types for six species. Codons were separated into four groups according to their number of G and C nucleotides. First row shows correlations for 7 codons without a G and C nucleotide, second row for 21 codons with one G or C nucleotide, thirds row for 24 codons with two G and C nucleotides, and last

row for 8 codons consisting of only G or C nucleotides. Right panels: Principal component 1 versus 2 plot from a principal component analysis carried out with codons in different GC content groups. (B) Scatter plot of amino acid frequencies between mRNA and lincRNA in human (left panel) and fission yeast (right panel). Pearson (r) and Spearman ( $\rho$ ) correlation coefficients are indicated. (C) Comparison of Pearson (r; left panel) and Spearman ( $\rho$ ; right panel) correlation coefficients calculated for codon frequencies and amino acid frequencies between lincRNA pORFs and mRNA coding regions (red), lincRNA pORFs and intronic control pORFs (gray) and lincRNA pORFs and intergenic control pORFs (white), for six different species, indicated below. The first bar of a group of three bars with the same color shows the correlation coefficient between frequencies of 60 codons. The associated p value is indicated by a star with scale on the right y-axis. The second bar shows the average correlation coefficient between frequencies of 100000 random subsets of 19 codons. Error bars represent the standard deviation. The third bar shows the correlation coefficient between frequencies of 19 amino acids. The associated p value is indicated by a star with scale on the right y-axis. All differences between correlation coefficients for random subsets of codons and for amino acids were significant (p value from one sample t test <1e-300). Start and stop codons as well as methionine were omitted.

**Figure S5**

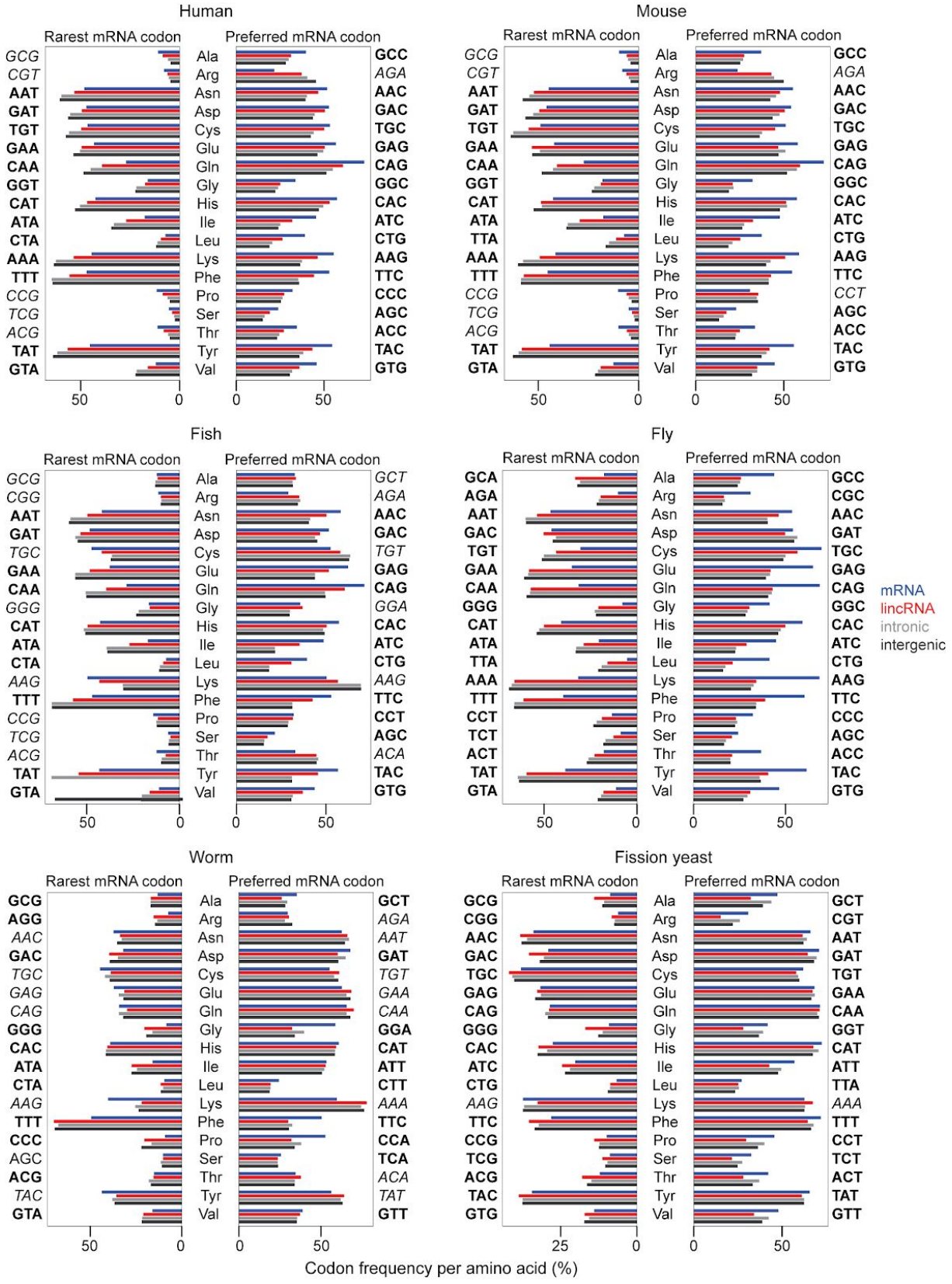



**Figure S5: Frequencies of rarest and most preferred codons in mRNA per amino acid.**

Frequencies of rarest and most preferred codons per amino acid in mRNAs for different species (indicated above each pair of panels). For each specie, left panels show the frequency of the codon, which is the rarest for each amino acid in mRNA coding regions, and right panels the frequency of the codon, which is most frequently used in mRNA coding regions for each amino acid. The frequencies per amino acid are shown for mRNA coding regions (blue), lincRNA pORFs (red) and intronic (gray) and intergenic (black) control pORFs. Amino acids are indicated in between both panels. Rarest codons are indicated left (bold, when the frequency for mRNAs was smaller than for lincRNAs) and most preferred codons are indicated right (bold, when the frequency for mRNAs was higher than for lincRNAs).

**Figure S6**

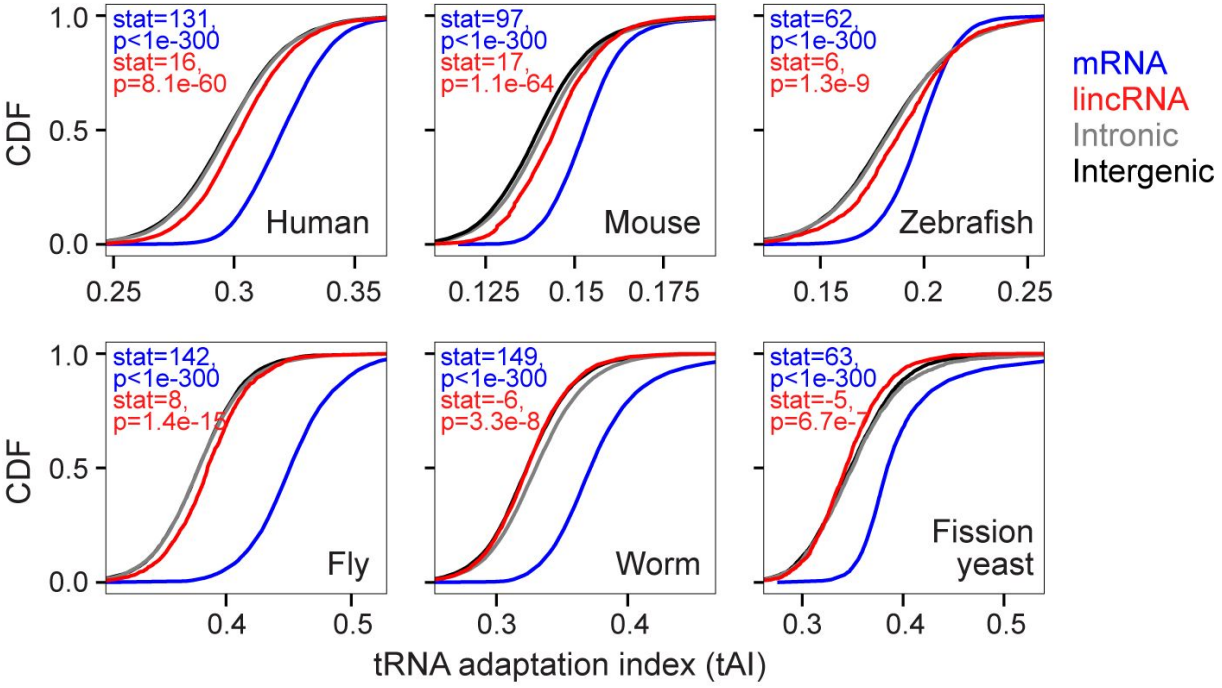



**Figure S6: Adaptation to tRNA abundances of tri-nucleotides in different RNA types and species**

Adaptation to effective tRNA abundances (tRNA adaptation index, tAI, calculated based on tRNA gene counts; see Methods) of codons in mRNA coding regions (blue), lincRNA pORFs (red) and intronic (gray) and intergenic (black) control pORFs in different specie (see label inside each panel). Wilcoxon's rank-sum test was used to compare tRNA adaptation indexes of mRNA and lincRNA with those of control sequences, and test statistics and p values indicated.

**Figure S7**

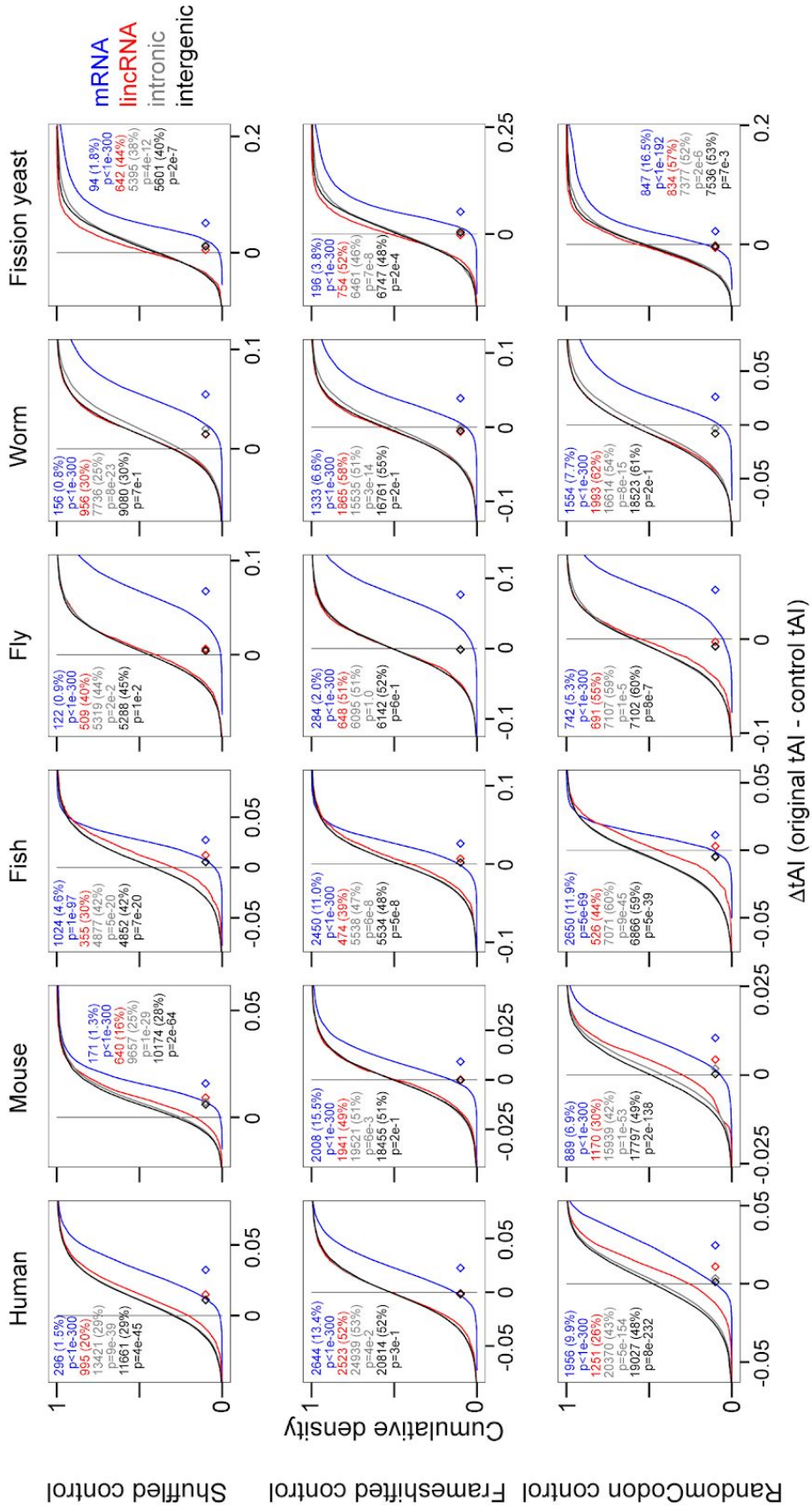

Human
Mouse
Fish
Fly
Worm
Fission yeast
mRNA
lincRNA
intronic
intergenic
Cumulative density
ΔtAI (original tAI - control tAI)
Shuffled control
Frameshifted control
RandomCodon control

**Figure S7: Comparison of tAIs between original and randomized sequences in different species**

Cumulative distributions of ΔtAIs (tAI of original ORF - tAI calculated for randomized ORFs) for different random controls: shuffled ORF sequences (top row), frame-shifted ORFs (middle row), and ORF sequences, in which each codon was replaced by a random codon coding for the same amino acid ("random codon" ORF; bottom row). ΔtAIs were calculated for mRNA coding regions (blue), lincRNA longest pORF (red), and longest pORFs in intronic (gray) and intergenic (black) regions. Numbers inside each panel represent the numbers and percentages of transcripts with negative ΔtAIs (ΔtAI<0) for each RNA type (color coded). P values were calculated using a Wilcoxon ranksum test to compare the ΔtAI distribution of lincRNA pORFs with those of other RNA types. Each column is a different specie indicated on top.

**Figure S8**

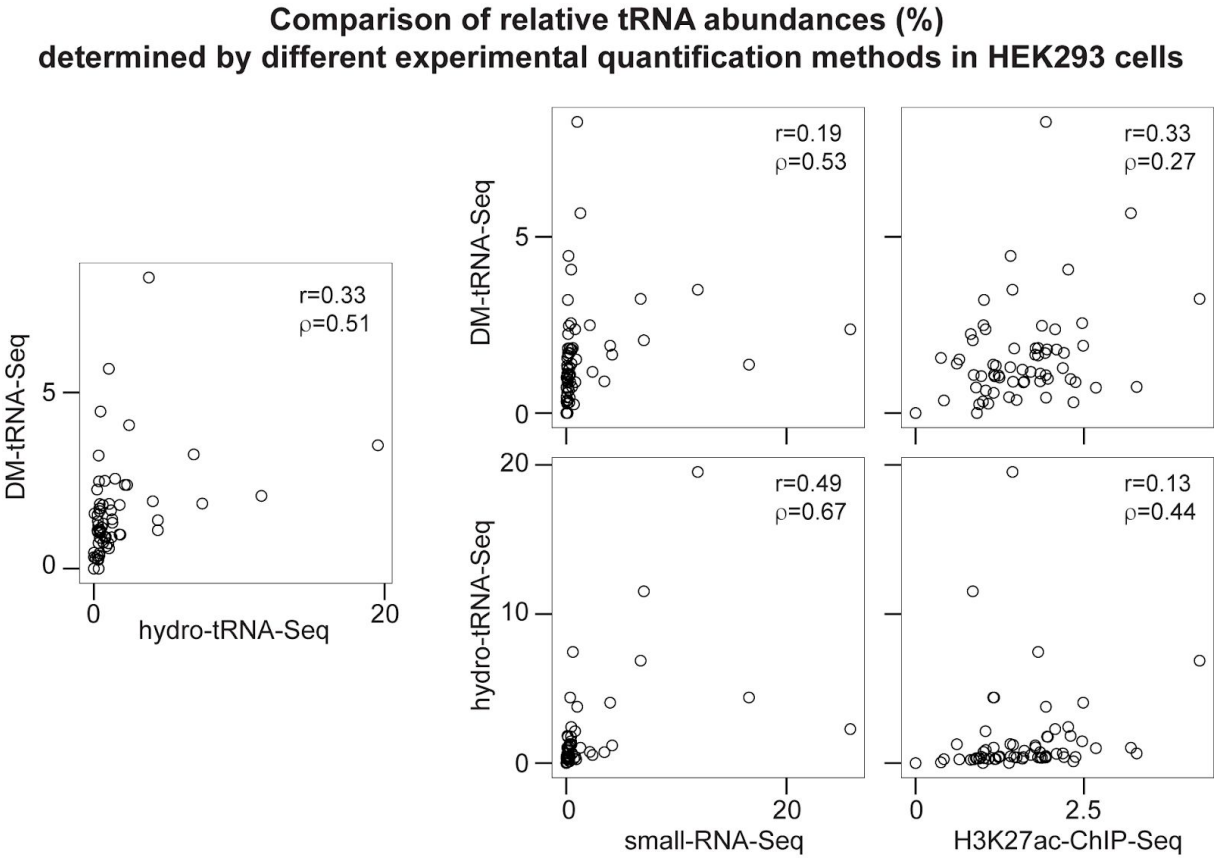



**Figure S8: Comparison of tRNA anticodon frequencies derived from different experimental approaches in HEK293**

Left panel: Comparison of tRNA anticodon frequencies derived from two dedicated experimental approaches, hydro-tRNA-Seq (Gogakos et al. 2017; x-axis) and DM-tRNA-Seq (Zheng et al. 2015; y-axis). Pearson (r) and Spearman ( $\rho$ ) correlation coefficients are indicated. Right four panels: Comparison between tRNA anticodon frequencies derived from one of the two dedicated experimental approaches (see y-axis labels) with tRNA anticodon frequencies derived from smallRNA-Seq or H3K27ac-ChIP-Seq (see x-axis label; see Methods). Pearson (r) and Spearman ( $\rho$ ) correlation coefficients are indicated.

**Figure S9**

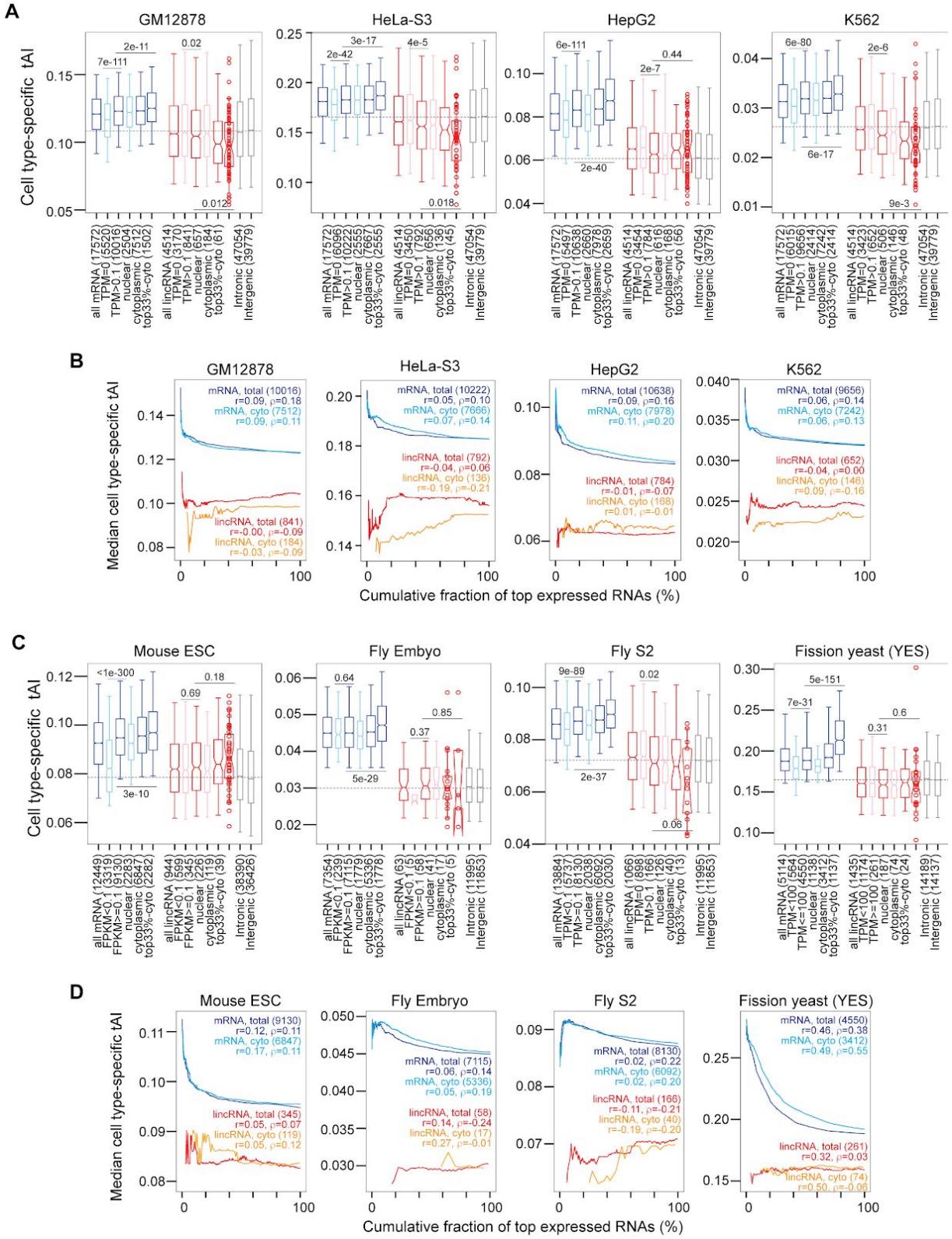



**Figure S9: Cell type-specific tAIs for various sets of RNAs in different species and cell types**

(A) and (C): Box plots of cell type-specific tAIs (using smallRNA-Seq data to quantify relative tRNA anticodon abundances) for mRNA coding regions (blue/light blue) and lincRNA longest pORFs

(red/pink). Boxes show cell type-specific tAIs for different sets of RNAs for each RNA type: first box, all RNAs; second box, non-expressed RNAs; third box, expressed RNAs; fourth box, expressed, non-cytoplasmic RNAs; fifth box, cytoplasmic RNAs; sixth box, top one-third cytoplasmic RNAs with highest cytoplasmic expression. Gray boxes indicate cell type-specific tAIs of intronic and intergenic control pORFs. The numbers of RNAs in each box are indicated below each panel. Indicated p values are calculated using Wilcoxon's rank-sum test. (B) and (D): Median cell-type specific tAIs for increasing sets of top ranked RNAs. RNAs are either ranked by total expression level (mRNA, blue; lincRNA red; including expressed RNAs) and by cytoplasmic expression level (mRNA, light blue; lincRNA, orange; including cytoplasmic RNAs). The numbers of RNAs in each set, at x-axis value of 100%, are indicated in parenthesis inside each panel. Pearson (r) and Spearman (ρ) correlation coefficients are given between tAIs and expression levels of RNAs in whole cells (total) or in the cytoplasmic fraction of cells (cyto). In (A) and (B) data for different human cell lines, and in (C) and (D) data for different species are shown, as indicated on top of each panel.

**Figure S10**

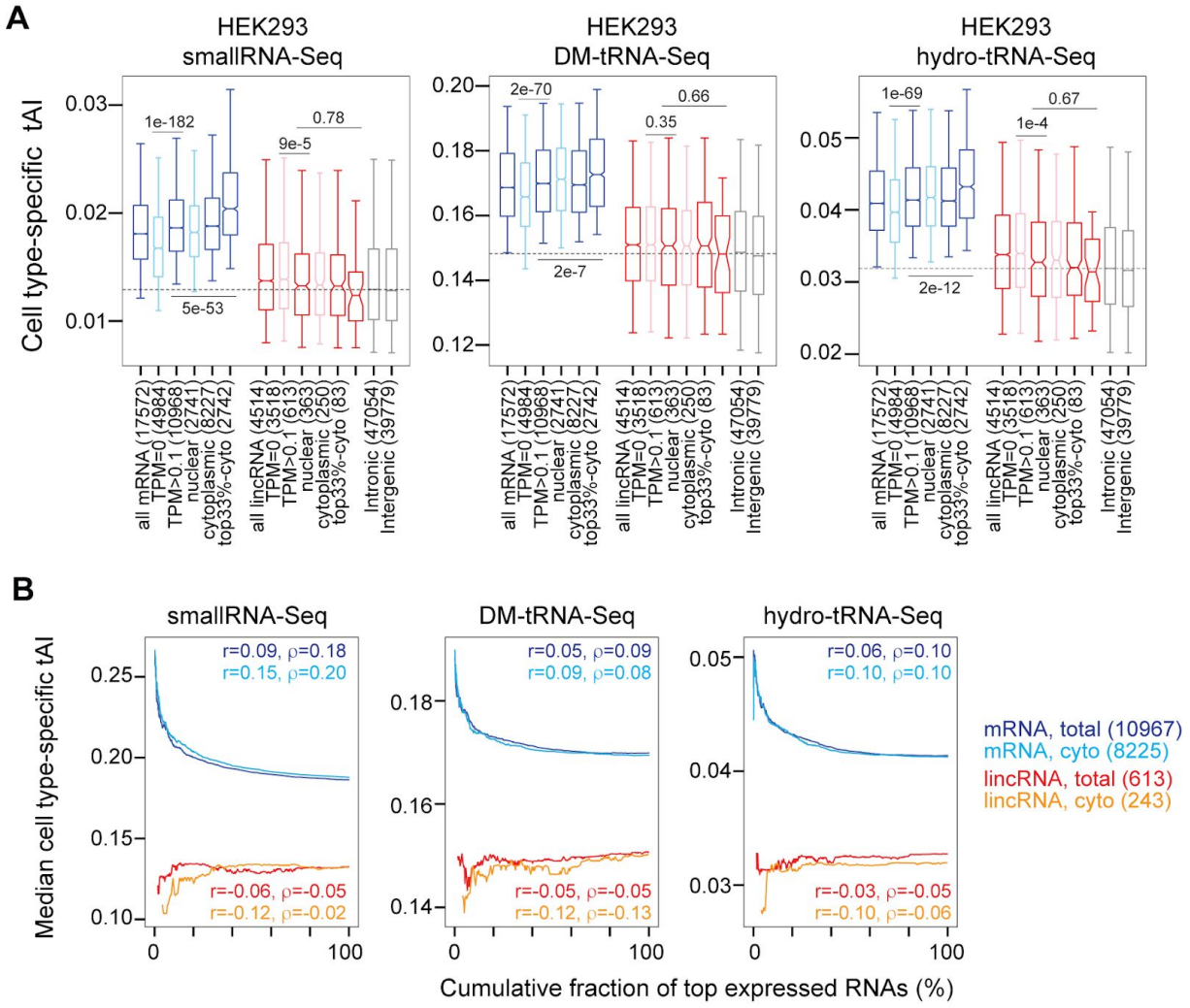



**Figure S10: Comparison of cell type-specific tAIs for HEK293, based on relative tRNA anticodon abundances derived from different experimental approaches**

(A) Box plots of cell type-specific tAIs for HEK293 cells using different experimental approaches to quantify relative tRNA anticodon abundances: first panel, smallRNA-Seq; second panel, DM-tRNA-Seq (Zheng et al. 2015); third panel, hydro-tRNA-Seq (Gogakos et al. 2017). In each panel, cell type-specific tAIs are shown for mRNA coding regions (blue/light blue) and lincRNA longest putative ORFs (red/pink). Boxes show cell type-specific tAIs for different sets of RNA for each RNA type: first box, all RNAs; second box, non-expressed RNAs; third box, expressed RNAs; fourth box, expressed, non-cytoplasmic RNAs; fifth box, cytoplasmic RNAs; sixth box, top one-third cytoplasmic RNAs with highest cytoplasmic expression. Gray boxes indicate cell type-specific tAIs of intronic and intergenic control pORFs. The numbers of RNAs in each box are indicated below each panel. Indicated p values are calculated using Wilcoxon's rank-sum test. (B) Median cell-type specific tAIs for increasing fractions of top ranked RNAs for different tRNA quantification methods in HEK293 cells (indicated above each panel). RNAs are either ranked by their total expression level (mRNA, blue; lincRNA red) or by their cytoplasmic expression level (mRNA, light blue; lincRNA, orange). The numbers of RNAs in the a set, at x-axis value of 100%, are indicated in parenthesis on the right. Pearson (r) and Spearman (ρ) correlation coefficients are given between cell type-specific tAIs and expression levels of RNAs in whole cells (total) or in the cytoplasmic fraction of cells (cyto).

**Figure S11**

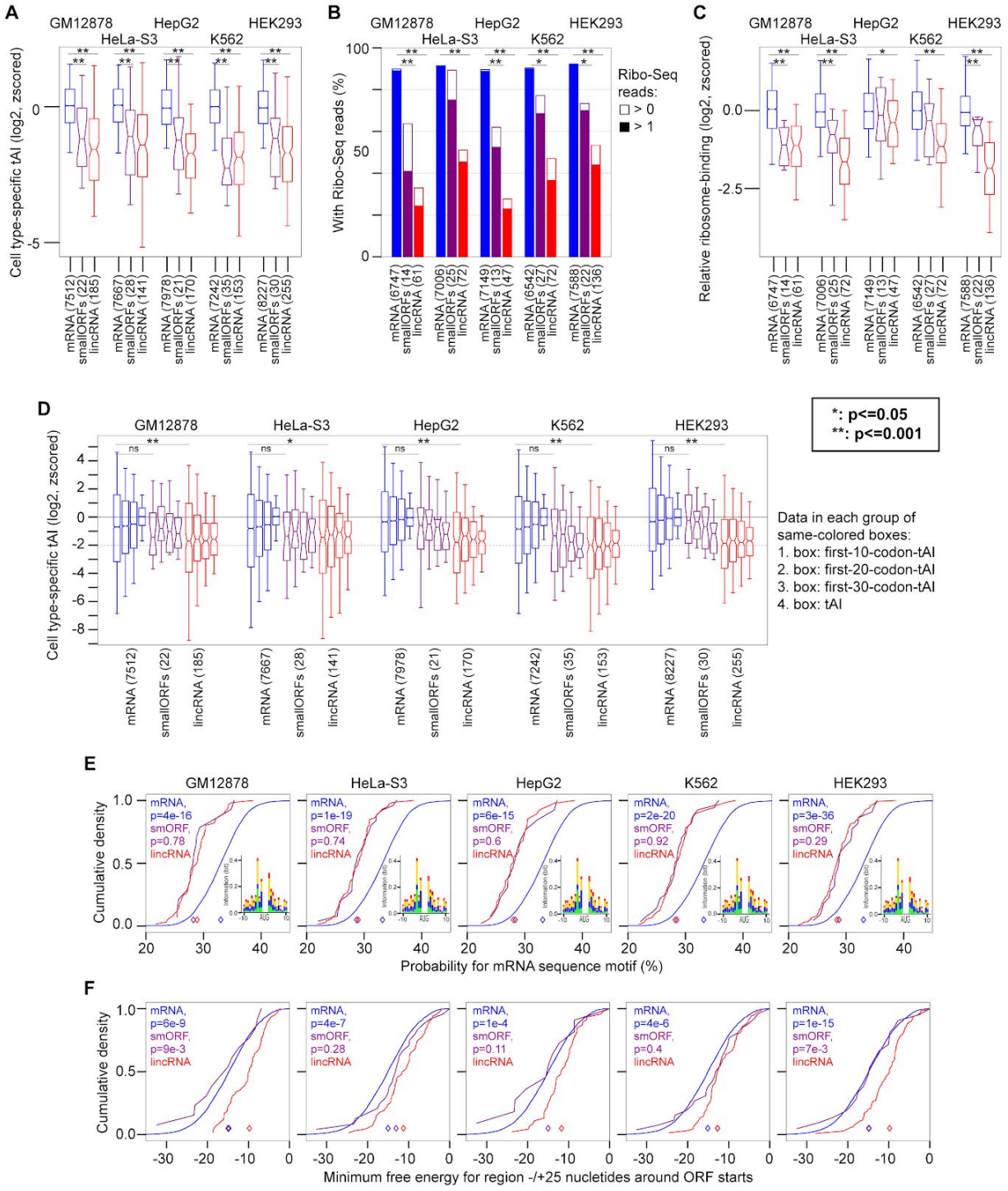



**Figure S11: Comparisons of tAIs and ribosome-binding between mRNAs, small protein-encoding RNAs and lincRNAs in five human cell lines**

(A) Boxplot of cell type-specific tAIs (log2-transformed and z-scored relative to tAIs of cytoplasmic mRNA coding regions) for mRNA coding regions (blue), small protein-encoding ORFs (smORFs; purple) and longest lincRNA pORFs (red) of cytosolic RNAs in five human cell lines, as indicated on top. Number of RNAs included in each box are indicated below each box. Asterisks represent p values from Wilcoxon's rank-sum test (*:<=0.05 and **:<=0.001). (B) Fraction of ORFs with in-frame Ribo-Seq

reads (filled bars with more than one Ribo-Seq reads and empty bars with one Ribo-Seq read). The number of ORFs with Ribo-Seq reads are indicated below each box. Asterisks represent p values from Fisher's exact test (*:<=0.05 and **:<=0.001). (C) Boxplot of relative ribosome-binding (calculated as the log2-ratio of normed Ribo-Seq reads to cytoplasmic expression, and z-scored relative to log2-ratios for mRNA coding regions) of the same types of cytoplasmic RNAs as in (A) with Ribo-Seq reads. Cell lines are indicated above. Number of RNAs included in each box are indicated below each box. Asterisks represent p values from Wilcoxon's rank-sum test (*:<=0.05 and **:<=0.001). (D) Cell type-specific tAIs (log2-transformed and z-scored as in (A)) for different regions of pORFs in cytoplasmic RNAs: first box of each RNA type and in each cell line, tAI of first 10 codons downstream of start codons; second box, tAI of first 20 codons; third box, tAI of first 30 codons downstream of start codons; and fourth box, tAI of entire ORF. Asterisks represent p values from Wilcoxon's rank-sum test (*:<=0.05 and **:<=0.001). (E) Sequence context around start codons of three types of cytoplasmic RNAs: mRNA (blue), smORFs (purple), and lincRNAs (red). Each panel shows the cumulative distribution of the probabilities for the mRNA sequence motif in a different human cell line. P values are calculated using Wilcoxon's rank-sum test to compare lincRNAs with the two other types of RNA. (F) RNA structure context around start codons of the three types of cytoplasmic RNAs. Each panel shows the cumulative distribution of the minimum free energy of the region -/+ 25 nucleotides around start codons for cytoplasmic RNAs in a different human cell line. P values are calculated using Wilcoxon's rank-sum test to compare lincRNAs with the two other types of RNA.

**Figure S12**

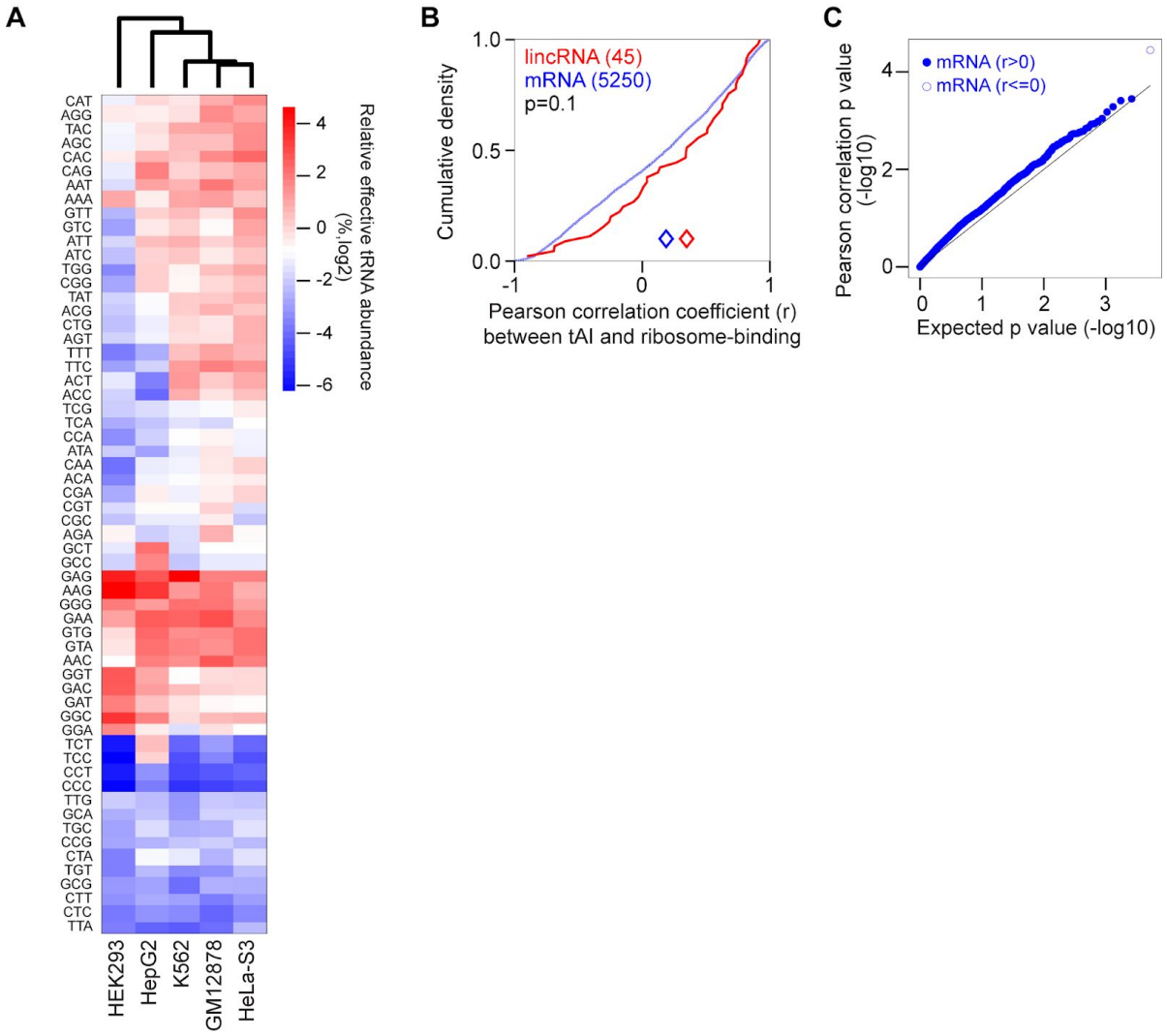



**Figure S12: Differences in tAIs and ribosome-binding between cytoplasmic lincRNAs in five human cell lines**

(A) Clustering of relative tRNA anticodon abundances, estimated from smallRNA-Seq data, in five human cell lines. (B) Cumulative distribution of Pearson correlation coefficients between cell line-specific tAIs and relative ribosome binding for 45 cytoplasmic lincRNAs (red) and 5250 mRNAs (blue). P value indicates no difference between both distributions, based on Wilcoxon's rank-sum test. (C) QQ-plot comparing p values of Pearson correlations between cell line-specific tAIs and relative ribosome binding for 5250 mRNAs with uniformly distributed p values.

**Supplementary Text**

To further quantify the extent and direction of sequence adaptation, we related tAIs to those of three types of randomized sequences (see Methods; Figure S7), each preserving different properties of the nucleotide sequence: 1) shuffled sequences preserved the nucleotide content; 2) frame-shifted sequences preserved the nucleotide content and sequence, accounting for functional sequence or structure elements (such as DNA- or RNA-binding motifs or RNA secondary structure); and 3) "random codon" sequences, in which each codon was replaced by a random codon coding for the same amino acid, preserved the amino acid sequence.

We then calculated, for each mRNA, lincRNA and control pORF, the difference in tAIs (ΔtAI) between the original and each type of randomized sequence (Figure S7). Overall, comparisons with randomized sequences reflected previous observations. ΔtAIs of lincRNA pORFs were in-between those of mRNA coding regions and control pORFs for vertebrates and fly, meaning that their putative codons deviate more from randomized sequences than control pORFs, but less than mRNA coding regions. The difference between ΔtAIs of lincRNAs and control sequences (and the significance thereof) decreases from human to fly. For worm and fission yeast, ΔtAIs of lincRNAs were significantly lower than control sequences (compared with intronic sequences for worm, and with intronic and intergenic sequences for fission yeast). The proportions of lincRNAs with negative ΔtAI were among the largests for worm and for fission yeast (>=30% and >=44%, respectively; Figure S7). This analysis provides a further indication for lincRNAs exhibiting a distinct and non-random tRNA adaptation, with stronger signatures in worm and yeast, which are those species with generally stronger selection.

## References


Akay, A., Jordan, D., Navarro, I.C., Wrzesinski, T., Ponting, C.P., Miska, E.A., and Haerty, W. (2019). Identification of functional long non-coding RNAs in C. elegans. BMC Biol. *17*, 14.

Bánfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W.E., Jr, Kundaje, A., Gunawardena, H.P., Yu, Y., Xie, L., et al. (2012). Long noncoding RNAs are rarely translated in two human cell lines. Genome Res. *22*, 1646–1657.

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., and Blüthgen, N. (2013). Efficient translation initiation dictates codon usage at gene start. Mol. Syst. Biol. *9*, 675.

Bernhart, S.H., Mückstein, U., and Hofacker, I.L. (2011). RNA Accessibility in cubic time. Algorithms Mol. Biol. *6*, 3.

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., and Ohler, U. (2016). Detecting actively translated open reading frames in ribosome profiling data. Nat. Methods *13*, 165–170.

Chen, J., Brunner, A.-D., Cogan, J.Z., Nuñez, J.K., Fields, A.P., Adamson, B., Itzhak, D.N., Li, J.Y., Mann, M., Leonetti, M.D., et al. (2020). Pervasive functional translation of noncanonical human open reading frames. Science *367*, 1140–1146.

Consortium, T.E.P., and The ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

Dana, A., and Tuller, T. (2014). The effect of tRNA levels on decoding times of mRNA codons. Nucleic Acids Res. *42*, 9171–9181.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Eraslan, B., Wang, D., Gusic, M., Prokisch, H., Hallström, B.M., Uhlén, M., Asplund, A., Pontén, F., Wieland, T., Hopf, T., et al. (2019). Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. Mol. Syst. Biol. *15*, e8513.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773.

Frumkin, I., Lajoie, M.J., Gregg, C.J., Hornung, G., Church, G.M., and Pilpel, Y. (2018). Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proc. Natl. Acad. Sci. U. S. A. *115*, E4940–E4949.

Gingold, H., Tehler, D., Christoffersen, N.R., Nielsen, M.M., Asmar, F., Kooistra, S.M., Christophersen, N.S., Christensen, L.L., Borre, M., Sørensen, K.D., et al. (2014). A dual program for translation regulation in cellular proliferation and differentiation. Cell *158*, 1281–1292.

Gogakos, T., Brown, M., Garzia, A., Meyer, C., Hafner, M., and Tuschl, T. (2017). Characterizing Expression and Processing of Precursor and Mature Human tRNAs by Hydro-tRNAseq and PAR-CLIP. Cell Rep. *20*, 1463–1475.

Goodarzi, H., Nguyen, H.C.B., Zhang, S., Dill, B.D., Molina, H., and Tavazoie, S.F. (2016). Modulated Expression of Specific tRNAs Drives Gene Expression and Cancer Progression. Cell *165*, 1416–1427.

Guimaraes, J.C., Mittal, N., Gnann, A., Jedlinski, D., Riba, A., Buczak, K., Schmidt, A., and Zavolan, M. (2020). A rare codon-based translational program of cell proliferation. Genome Biol. *21*, 44.

Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S., and Lander, E.S. (2013). Ribosome profiling

provides evidence that large noncoding RNAs do not encode proteins. Cell *154*, 240–251.

Haerty, W., and Ponting, C.P. (2015). Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. RNA *21*, 333–346.

Hanson, G., and Coller, J. (2018). Codon optimality, bias and usage in translation and mRNA decay. Nat. Rev. Mol. Cell Biol. *19*, 20–30.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

van Heesch, S., Witte, F., Schneider-Lunitz, V., Schulz, J.F., Adami, E., Faber, A.B., Kirchner, M., Maatz, H., Blachut, S., Sandmann, C.-L., et al. (2019). The Translational Landscape of the Human Heart. Cell *178*, 242–260.e29.

Hernandez-Alias, X., Benisty, H., Schaefer, M.H., and Serrano, L. (2020). Translational efficiency across healthy and tumor tissues is proliferation-related. Mol. Syst. Biol. *16*, e9275.

Hezroni, H., Ben-Tov Perry, R., Meir, Z., Housman, G., Lubelsky, Y., and Ulitsky, I. (2017). A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. Genome Biol. *18*, 162.

Hofacker, I.L. (2004). RNA Secondary Structure Analysis Using the Vienna RNA Package. Current Protocols in Bioinformatics.

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C.M., Khaitovich, P., Vogel, J., Stadler, P.F., and Hackermüller, J. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput. Biol. *5*, e1000502.

Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J.S., Jackson, S.E., Wills, M.R., and Weissman, J.S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep. *8*, 1365–1379.

Ji, Z., Song, R., Regev, A., and Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. eLife *4*.

Knowles, D.G., and McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. Genome Res. *19*, 1752–1759.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*.

Lin, M.F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics *27*, i275–i282.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10.

Martinez, T.F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M.N., and Saghatelian, A. (2020). Accurate annotation of human protein-coding small open reading frames. Nature Chemical Biology *16*, 458–468.

Martinez, T.F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M.N., and Saghatelian, A. An Improved Human smORF Annnotation Workflow Combining De Novo Transcriptome Assembly and Ribo-Seq.

Ouspenskaia, T., Law, T., Clauser, K.R., Klaeger, S., Sarkizova, S., Aguet, F., Li, B., Christian, E., Knisbacher, B.A., Le, P.M., et al. Thousands of novel unannotated proteins expand the MHC I

immunopeptidome in cancer.

Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R., et al. (2015). Codon optimality is a major determinant of mRNA stability. Cell *160*, 1111–1124.

Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr. Protoc. Bioinformatics *47*, 11.12.1–34.

dos Reis, M., Wernisch, L., and Savva, R. (2003). Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. Nucleic Acids Res. *31*, 6976–6985.

dos Reis, M., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. *32*, 5036–5044.

Shi, J., Ko, E.-A., Sanders, K.M., Chen, Q., and Zhou, T. (2018). SPORTS1.0: A Tool for Annotating and Profiling Non-coding RNAs Optimized for rRNA- and tRNA-derived Small RNAs. Genomics Proteomics Bioinformatics *16*, 144–151.

Subramanian, S. (2008). Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. Genetics *178*, 2429–2432.

Torrent, M., Chalancon, G., de Groot, N.S., Wuster, A., and Madan Babu, M. (2018). Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. Sci. Signal. *11*.

Tuck, A.C., Rankova, A., Arpat, A.B., Liechti, L.A., Hess, D., Iesmantavicius, V., Castelo-Szekely, V., Gatfield, D., and Bühler, M. (2020). Mammalian RNA Decay Pathways Are Highly Specialized and Widely Linked to Translation. Mol. Cell *77*, 1222–1236.e13.

Tuller, T., Waldman, Y.Y., Kupiec, M., and Ruppin, E. (2010a). Translation efficiency is determined by both codon bias and folding energy. Proc. Natl. Acad. Sci. U. S. A. *107*, 3645–3650.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010b). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell *141*, 344–354.

Ulitsky, I., and Bartel, D.P. (2013). lincRNAs: genomics, evolution, and mechanisms. Cell *154*, 26–46.

Waldman, Y.Y., Tuller, T., Shlomi, T., Sharan, R., and Ruppin, E. (2010). Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. Nucleic Acids Res. *38*, 2964–2974.

Wang, H., Wang, Y., Xie, S., Liu, Y., and Xie, Z. (2017). Global and cell-type specific properties of lincRNAs with ribosome occupancy. Nucleic Acids Res. *45*, 2786–2796.

Wang, J., Alexander, P., and McKnight, S.L. (2011). Metabolic specialization of mouse embryonic stem cells. Cold Spring Harb. Symp. Quant. Biol. *76*, 183–193.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Res. *46*, D754–D761.

Zheng, G., Qin, Y., Clark, W.C., Dai, Q., Yi, C., He, C., Lambowitz, A.M., and Pan, T. (2015). Efficient and quantitative high-throughput tRNA sequencing. Nat. Methods *12*, 835–837.