LincRNA sequences are biased to counteract 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 eukaryotic genomes. Although classified as non-coding, most lincRNAs contain putative open reading frames (pORFs), and it is unclear what hinders translation of cytoplasmic lincRNAs.

Here, we analysed lincRNA sequences from six eukaryotes, including human and fission yeast, and identified signatures of translational repression. In species under strong selective pressure, i.e. yeast and mostly worm, we detected fewer and shorter pORFs in lincRNAs than in intronic and non-transcribed control regions, and the sequence context around their putative start codons appeared suboptimal for translation initiation. Moreover, trinucleotides ("codons") in lincRNA pORFs tended to be decoded by tRNAs with fewer gene copies than codons in mRNAs or control regions, likely to impede translation elongation.

In order to better understand how in multicellular species varying tRNA and cytoplasmic lincRNA expression levels between cell types affect lincRNA translation regulation, we analysed data from human cell lines. In three out of five cases, we detected signatures of translational repression of cytoplasmic lincRNAs, i.e. a codon usage that corresponded to lower expressed tRNAs than control trinucleotides. Importantly, the degree of codon-tRNA correspondence was in agreement with the ribosome-association strengths of cytoplasmic lincRNAs across cell lines.

Together, the codon usage of lincRNA pORFs is biased, and it appears to function to repress ribosome binding. This mechanism is capable of counteracting lincRNA translation in a cell type-specific manner.

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

Significance statement

Although long intergenic non-coding RNAs (lincRNAs) are classified as non-coding, mechanistic insight into what hinders translation of cytoplasmic lincRNAs is still lacking. We investigated signatures in lincRNA sequences counteracting translation in six eukaryotes and five human cell lines. The detected signatures were more pronounced in species evolving under stronger selective pressures, i.e. worm and yeast, and for lincRNAs with higher cytoplasmic expression. In human, signatures were cell type-specific. Our observations suggest that lincRNA sequences have been shaped to counteract translation. Mechanistically, we provide evidence that a poor correspondence between lincRNA tri-nucleotide and tRNA frequencies, which is less than for coding RNAs and non-coding controls, reduces ribosome-binding to lincRNAs. The identified signatures may assist distinguishing peptide- from real non-coding lincRNAs.

Introduction

Long intergenic noncoding RNAs (lincRNAs) form a functionally heterogeneous class of RNAs longer than 300 nucleotides and lacking protein coding potential (Frankish et al. 2019; Ulitsky and Bartel 2013). Despite being classified as non-coding, most lincRNAs contain potential open reading frames (pORFs) flanked by start and stop codons. Although, small ORFs encoding functional peptides within annotated human lincRNAs have been identified recently (Chen et al. 2020; Ouspenskaia et al., n.d.; Martinez et al. 2020), the majority of lincRNA pORFs remains untranslated and the mechanisms counteracting translation are unclear.

Differences between coding and non-coding RNAs have been investigated recently. By analysing sequencing of ribosome protected fragments (Ribo-Seq), differences in ribosome interaction patterns have been reported, including in the tri-nucleotide periodicity of binding (Calviello et al. 2016; Ji et al. 2015a) or in ribosome release (Guttman et al. 2013). Furthermore, discriminating sequence features between human mRNAs and lincRNAs (Niazi and Valadkhan 2012) and between lincRNAs with and without ribosome-association have also been analysed, in human and mouse (Zeng and Hamada 2018; Wang et al. 2017). These studies identified a poor start codon context of lincRNA pORFs for translation initiation and reported cell type-specific associations between human lincRNAs and ribosomes. Yet, these studies considered only a relatively small set of lincRNAs in only one or two species, and provided little insight into how translation of lincRNAs is prevented.

mRNA translation to proteins is known to be regulated at initiation and during elongation (Riba et al. 2019; Eraslan et al. 2019; Tuller, Carmi, et al. 2010; Nieuwkoop et al. 2020). While the RNA sequence and secondary structure around start codons are important for translation initiation, it has been demonstrated that the codon usage is a regulator of translation elongation. Specifically, the rate at which a codon is translated correlates with the abundance of its decoding tRNA (Dana and Tuller 2014). Consequently, mRNAs using codons corresponding to more abundant cellular tRNAs tend to be translated more efficiently. Evidence for this mechanism to increase translation efficiency has been observed in several contexts, e.g. under cellular stress conditions, during proliferation or meiosis, or in cancer (Goodarzi et al. 2016; Torrent et al. 2018; Guimaraes et al. 2020; Sabi and Tuller 2019). The strength of the mRNA codon usage bias varies between species, and it is usually more pronounced in species 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 lower in species such as human and mouse. In most species, the codon usage bias is stronger for highly expressed mRNAs, likely because there is more selection pressure on these mRNAs.

Given these well-studied connections between mRNA sequence features and translation efficiency, here we investigate signatures in lincRNA sequences counteracting translation, in six species and five human cell lines. We provide evidence for their existence and functionality.

Results

Putative open reading frames in lincRNAs are frequent

To investigate signatures in lincRNAs sequences counteracting translation, 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 putative ORFs (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 predicted pORFs in randomly selected intronic and non-transcribed intergenic regions of the same length as lincRNAs (see Methods). These nuclear (intronic) and non-transcribed (intergenic) sequences are never in contact with the translation machinery and therefore provide a reference, against which we evaluated the sequence bias of lincRNAs in each species. We found that most lincRNAs (>94% for all species) had at least one pORF (Figure 1A). While the fraction of lincRNAs with pORFs in fission yeast was significantly smaller than in control regions, pORFs were more often found in lincRNAs than in control regions of other species. The median number of pORFs per lincRNA was between 6 (human, fly, worm) and 10 (fish) for different species. In the following, we considered only the longest pORF for each lincRNA (median length between 45 to 71 codons in different species; Figure 1B). The longest pORFs in lincRNAs were significantly shorter than the longest pORFs in control regions for yeast, and shorter than pORFs in intronic regions for worm, but longer than control pORFs in other species (Figure 1B). In all species, mRNA coding regions were on average 5-8 times longer than longest lincRNA pORFs.

Thus, lincRNA pORFs are reduced and shortened compared to those in control regions, in yeast and to some extent in worm. In other species, we found no signs for that.

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.

RNA sequence and structure around lincRNA start codons differs from those of mRNAs

Since the RNA sequence and structure around start codons was identified as a regulator of translation in mRNA (Eraslan et al. 2019), we examined if the start codon context of lincRNA pORFs could be less optimal for initiating translation. We found that in all species the sequence context around lincRNA start codons was less specific than the consensus mRNA sequence motif, which showed similarity with the Kozak sequence motif (Kozak 1989) (Figure S1). For worm and yeast, the sequence around lincRNA start codons was less similar to this consensus mRNA sequence motif than those of control pORFs, and this trend was more pronounced among highly abundant cytoplasmic lincRNAs in yeast (Figure 1C; see Methods). For vertebrates and fly, we found the opposite, namely a higher similarity to the consensus mRNA sequence motif for lincRNAs than in control sequences. However, among highly abundant cytoplasmic lincRNAs in fly the similarity was not different between lincRNAs and controls.

We next compared the RNA structure around mRNA and lincRNA start codons across species (Figure S2A) and found that in human, mouse, fish and worm this region had a higher minimum free energy, calculated using RNAfold (Hofacker 2004) (see Methods), in lincRNAs than in mRNAs and thus was

more accessible in lincRNAs. This region also was more accessible in control pORFs than in mRNAs, which could be a consequence of a higher GC content in mRNAs in these species (Haerty and Ponting 2015; Niazi and Valadkhan 2012). In contrast, in fly and fission yeast the structural accessibility upstream of lincRNA start codons was reduced in comparison with mRNAs (Figure S2B), which may hinder an interaction between the ribosome and lincRNA start codons in these species. Overall, it appears that the regulation by the RNA secondary structure around translation start sites is complex and might differ between species.

Thus, in some species the RNA sequence and structure around lincRNA start codons appears to be less favourable for translation initiation. Yet, in particular for 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 contributes to translation regulation (Hanson and Coller 2018; Tuller, Waldman, et al. 2010). We investigated if biased frequencies of 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 such codon usage in lincRNA pORFs in comparison with those in mRNA coding regions and control pORFs, we performed a multiple correspondence analysis of the corresponding codon frequencies in different RNA types and species (Table S2). We found a clear separation between RNA types in the first two components (Figure 2A). For worm and fission yeast, lincRNA and mRNA codon usages were located on opposite sides relative to controls, while for vertebrates and fly, lincRNA codon composition was positioned between those of mRNAs and control pORFs in the first two components space. The same observation was obtained using a principal component analysis (Figure S3A), and it was reflected in the pattern of correlation strengths between codon frequencies in different RNA types across species (Figure 2B).

These results show that codon preferences differ between lincRNA and mRNA. In particular, in species known to have a strong mRNA codon bias (worm and yeast), lincRNA and mRNA codon usages appear to have diverged in opposite directions relative to control codon frequencies.

To rule out that differences in GC content between lincRNAs and mRNAs (Haerty and Ponting 2015) explain the observed differences in codon usage, we compared frequencies of codons with the same GC content (Figure S3B). We found that the correlation strengths between the different types of RNAs 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 conclude that the difference in codon composition between mRNA and lincRNA is not driven by GC content differences.

Codons in lincRNA pORFs correspond to less abundant tRNAs than codons in mRNAs

Next, we investigated how codon usages of mRNA and lincRNA relate to tRNA abundances. As a first estimate of the tRNA abundances in different species we used the number of annotated tRNA genes for each tRNA anticodon type (see Methods), which has been shown to correlate well with tRNA abundances (Tuller, Carmi, et al. 2010). We used previously determined wobble-base pairing and tRNA editing efficiencies (dos Reis, Savva, and Wernisch 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 with relative tRNA

abundances than codon frequencies of lincRNAs, in all species (Figure 3A). In yeast and fly, this discrepancy was even stronger for codon frequencies in cytoplasmic mRNAs and lincRNAs. For the two species with the strongest mRNA codon usage bias, worm and fission yeast, lincRNA codon frequencies correlated even less with tRNA abundances than control codon frequencies, indicating a lincRNA codon bias towards codons corresponding to lower abundance tRNAs in these species.

Previously, the tRNA adaptation index (tAI) was defined as a measure for the correspondence between codon usage frequencies in an ORF and relative tRNA abundances (dos Reis, Wernisch, and Savva 2003). The tAI ranges from 0 to 1, where higher values indicate a preferential use of codons decoded by more abundant tRNAs. Since amino acid identities are not expected to constrain lincRNA pORFs, we computed tAIs relative to the overall maximum tRNA abundance (as opposed to the maxima among synonymous tRNAs; see Methods). We found that, in all species studied, mRNAs had significantly ($p<1e-300$, Wilcoxon rank-sum test) higher tAI values than lincRNAs (Figure 3B and S4). This was unrelated to the difference in number and length of mRNA coding regions and lincRNA pORFs, since for subsampled mRNAs, accounting for number and lengths of lincRNA pORFs, we also found significantly larger tAIs than for lincRNAs (data not shown). Compared to control pORFs, lincRNA pORFs had significantly lower tAIs for worm and fission yeast, while in vertebrates and fly, tAIs were higher for lincRNA pORFs than for control pORFs (Figure 3B, p values on the right, and S4). These tAI differences between lincRNAs and control pORFs reflect the results from the correlation analysis (Figure 3A), potentially indicating that lincRNA pORFs adapted towards using codons corresponding to less abundant tRNAs in these species.

To further quantify the extent and direction of codon bias in lincRNAs, we related tAIs to those of randomized sequences. In particular, we compared with tAIs for tri-nucleotides in shuffled ORFs, preserving the nucleotide content, and with tAIs for tri-nucleotides in frame-shifted ORFs, preserving the nucleotide sequence, including potential RNA sequence or structure motifs, of the original ORFs (see Methods; Figure S5). We performed the same comparisons with tAIs for randomized sequences for control pORFs and used them as a reference to account for the overall nucleotide and di-nucleotide biases in each species. In worm and yeast, the difference between tAIs from the original ORFs and those from shuffled or frame-shifted ORFs was significantly lower for lincRNA pORFs than for control pORFs. In fly, the trend was similar for lincRNAs in comparison with frame-shifted pORFs. In human and mouse, lincRNA pORFs had tAIs that were overall higher than those of shuffled pORFs, relative to control pORFs, and there was no difference in the comparison with frame-shifted pORFs in these species.

Overall, the comparison with tAIs of randomized sequences supports that lincRNAs exhibit a distinct codon composition which corresponds to less abundant tRNAs, in worm and yeast. In particular, the observation that lincRNA tAIs are lower than the tAIs of their corresponding frame-shifted sequence significantly more often than for control pORFs, strengthens the hypothesis that lincRNA pORF sequences are biased to preferential use of codons corresponding to tRNAs with lower abundances as opposed to maintaining specific RNA sequence or structure motifs, in these species.

A test for translational selection in a species, was proposed by dos Reis et al. (dos Reis, Savva, and Wernisch 2004). In particular, the correlation between mRNA tAI and a measure for the deviation of the mRNA codon usage from an equal usage of synonymous codons, accounting for the GC-bias at the third codon position, was proposed. Here, we adapted the correlation test from dos Reis et al. to test for a correlation between overall tAIs and the overall codon usage bias, among all codons not just among synonymous ones, as amino acid identities are not expected to constrain lincRNA pORFs (same assumption as before for calculating overall tAIs; see Methods). We then calculated Spearman correlation coefficients between tAIs and this adapted codon usage bias measure for different RNA types and species (Figure 3C). Indeed, we observed that, correlation coefficients were larger for mRNAs than for lincRNAs ($rho>=0.39$ and $rho<=0.11$, respectively), and while the correlation

increased for cytoplasmic mRNAs (rho>=0.53), it decreased for cytoplasmic lincRNAs (rho<=0.06), for all analysed species. Strikingly, for 4 out of 6 species (mouse, fly, worm and yeast) the correlation coefficients for lincRNAs (cytoplasmic, where available) were smaller than those for control pORFs. Thus, these observations are supportive of a selection of codons counteracting translation in eukaryotic lincRNA sequences.

Cytoplasmic lincRNA codons correspond to less abundant tRNAs than control tri-nucleotides in 3 out of 5 human cell lines

Since in multicellular eukaryotes tRNA expression and cytoplasmic lincRNA expression are often tissue- and cell type-specific (Pinkard et al. 2020; Dittmar, Goodenbour, and Pan 2006; Ransohoff, Wei, and Khavari 2018), we next investigated cell type-specific repression of lincRNA translation focussing on five human cell lines (GM12878, HEK293, HeLa-S3, HepG2, and K562), for which extensive experimental data are available ((Consortium and The ENCODE Project Consortium 2004; Cenik et al. 2015; Martinez et al. 2020; Solomon et al. 2017; Huang et al. 2019; Kishore et al. 2013; Aktaş et al. 2017; Subtelny et al. 2014; Calviello et al. 2016); see Methods and Table S3).

Since tRNA gene counts are not cell type-specific, previously various methods were used to estimate cell type-specific tRNA abundances, such as H3K27ac-ChIP-seq (Gingold et al. 2014) and smallRNA-seq (Shi et al. 2018; Hernandez-Alias et al. 2020). For human HEK293 cells there are also data from two dedicated experimental approaches to quantify tRNA abundances, DM-tRNA-Seq (Zheng et al. 2015) and hydro-tRNA-Seq (Gogakos et al. 2017). We evaluated the concordance of these methods in HEK293 and found that the smallRNA-seq-based tRNA quantification agreed better with those from dedicated experimental approaches than the H3K27ac-ChIP-seq-based one (Figure S6A; see Methods). Therefore, we used smallRNA-seq data to estimate cell type-specific tRNA abundances in human cell lines.

We found considerable differences in tRNA abundances between cell lines (Figure 4A), with GM12878, HeLa-S3 and K562 having relatively similar tRNA abundances (Spearman correlation >0.87), while those in HEK293 and HepG2 were more different (Spearman correlation <0.75). We then used these estimated tRNA abundances to calculate cell line-specific tAIs (Figure 4B). We confirmed that cytoplasmic lincRNAs had significantly smaller tAIs than cytoplasmic mRNAs and also than mRNAs with cytoplasmic expression levels matching those of lincRNAs (see Methods). For 3 out of 5 cell lines (GM12878, HeLa-S3 and K562), tAIs of lincRNAs were significantly lower than tAIs of control sequences. The deviation of lincRNA tAIs from control tAIs was in the range (but opposite) of the deviation between tAIs of cytoplasmic expression-matched mRNAs and control tAIs, in these cell lines. In HEK293 and HepG2 tAIs of cytoplasmic lincRNAs were not different from those of control pORFs.

We next focussed on 45 lincRNAs that we classified as cytoplasmic in all five cell lines (see Methods). We confirmed that tAIs for these lincRNAs vary between cell lines, with significantly higher values in HepG2 and HEK293 ($p<0.02$, Wilcoxon's rank sum test) compared to tAIs in GM12878, HeLa and K562. These differences in cell line-specific tAIs were largely reflected in ribosome-binding differences (estimated from Ribo-seq data; see Methods) for the same lincRNAs across cell lines (Figure 4C). In particular, ribosome-binding in HepG2 was significantly higher than in other cell lines ($p<0.01$, Wilcoxon's rank sum test). Interestingly, an increased translation of lincRNAs was observed before in liver and kidney (van Heesch et al. 2019).

Together, the analysis of human cell lines revealed tAI differences for lincRNAs between cell lines, and confirmed tAIs of lincRNAs to be smaller than those of control pORFs in 3 out of 5 cell lines, in agreement with reduced ribosome-binding in these cell lines.

Ribosome-binding reflects tAI differences between lincRNAs and small protein-coding ORFs, in particular for codons at ORF starts

To better understand the relationship between tAI and ribosome binding, we focussed on three types of cytoplasmic RNAs: mRNAs, lincRNAs, and transcripts with small protein-encoding ORFs (smORFs, see Methods), and compared tAIs and ribosome binding between these RNA types. We observed that tAIs of lincRNAs and smORFs are significantly different from those of mRNAs in K562 (Figure 5A) and other human cell lines (Figure S7A). 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 S7B). 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 K562 (Figure 5C) and most human cell lines (Figure S7C).

Despite this concordance, the differences in relative ribosome-binding between lincRNAs and smORFs were more pronounced than their differences in tAIs, for most cell lines (Figures S7A,C). Previously, the codon usage immediately downstream of mRNA start codons was found to be distinct and was proposed to facilitate efficient translation initiation and elongation (Bentele et al. 2013; Tuller, Carmi, et al. 2010). Thus, we investigated tAIs calculated for the first codons downstream of start codons in different RNA types. Indeed, tAIs calculated for the first 10 or 20 codons (first-10-codon or first-20-codon tAIs) were higher for smORFs compared to lincRNAs, while first-10-codon or first-20-codon tAIs of smORFs were more similar to those of mRNAs (Figure 5D and S7D). Notably, we found no indication for the RNA sequence or structure context around start codons of smORFs and lincRNA pORFs to explain the observed ribosome-binding differences between these cytoplasmic RNA types (Figure S7E,F).

These observations support that lower tAIs of cytoplasmic lincRNA pORFs are associated with reduced ribosome-binding.

Discussion

In the absence of any functional role of peptides resulting from lincRNA translation, it likely would be advantageous to reduce stable associations between lincRNAs and ribosomes for several reasons: unwanted lincRNA translation wastes energy (Wagner 2005), reduces the pool of ribosomes available for mRNA translation (Raveh et al. 2016), and may lead to the synthesis of peptides with possibly harmful interference. Furthermore, it may hinder a potential regulatory function of lincRNAs, and, given the association between translation and transcript stability for mRNAs (Presnyak et al. 2015; Tuck et al. 2020), unwanted lincRNA translation may interfere with their stability.

Here, we provide multiple lines of evidence for lincRNA sequences being biased to counteract translation. We detected a reduction in pORF occurrence and length, a start codon context that appears suboptimal for initiating translation and a preferred usage of codons corresponding to lower abundant tRNAs likely impeding translation elongation. All corresponding measures were significantly reduced for lincRNAs compared to non-coding control sequences in fission yeast, the unicellular eukaryote with a sufficient number of lincRNA genes in its genome to be included in our analysis. Worm lincRNAs showed similar trends, but the comparison with control sequences was only significant for the case of lincRNAs using codons corresponding to tRNAs with fewer gene copies compared to control codons. In the other four species, lincRNA measures were not lower than those for control sequences, but in all six species, lincRNA measures were significantly reduced relative to those for mRNAs.

There could be several explanations for these different observations across species. First, the selection pressure is known to be higher in species with a large population size and a small generation time, such as yeast and worm, compared to higher eukaryotes, such as human, where translational selection on mRNA is much weaker (dos Reis and Wernisch 2009). Thus, it could be that translational counter-selection on lincRNAs is also stronger (and therefore better detectable) in yeast and worm than in other eukaryotes. Second, it is known that some lincRNA genes resulted from duplication of mRNA genes that eventually lost their coding capacity (Hezroni et al. 2017). Thus, it could be that for some of those lincRNAs the selection pressure may have been too weak or the time to evolve too short to readapt their codon usage. Finally, in multicellular organisms tRNA expression levels vary between cell types, and the comparison of lincRNA codon usage with tRNA abundances quantified from gene copy numbers may not be appropriate there. Indeed, cell type-specific tRNA quantification showed in three out of five human cell lines that lincRNA codons corresponded to less abundant tRNAs than tri-nucleotides in control regions (while for the remaining two cell lines it was similar). A conclusion that could be drawn from this is that in unicellular organisms lincRNA translation is often repressed at both, the level of initiation and elongation, while in multicellular organisms translation of some lincRNAs is only suppressed in a cell-type specific manner, either because peptides cause no or little harm in these cell types, or, in some cases, because they actually serve a function, consistent with previous observations of cell type- or condition-specific translation of human lincRNAs (Wang et al. 2017; Martinez et al. 2020; Chen et al. 2020; Ouspenskaia et al., n.d.; van Heesch et al. 2019).

Mechanistically, we linked the correspondence between tRNA abundances and codon composition of lincRNA pORFs to their ribosome-association. To the best of our knowledge this has not yet been reported and it would enable a cell type-specific regulation of lincRNA translation. Interestingly, tAIs for the first 10-20 codons of lincRNA pORFs and coding ORFs appeared to better match differences in ribosome-binding between these RNA types, in most cell lines. This result suggests an increased importance of the first codons in an ORF for ribosome engagement and translation (or its avoidance in case translation would be detrimental).

In conclusion, in this study we provided a comprehensive analysis of signatures of translational repression in lincRNA sequences of six species and in five human cell lines. The identified sequence

signatures may help in distinguishing bona-fide lincRNAs with regulatory roles in the cytoplasm from those coding for peptides. An interesting aspect is how cell type- and condition-specific expression of tRNAs and cytoplasmic lincRNAs imposes different constraints on the evolution of lincRNA sequences to either curb ribosome-binding in certain cell types, or promote it to enable peptide translation in others. Analysis of the impact of natural genetic variation or targeted mutations in lincRNA sequences on ribosome-binding and translation might further shed light onto these questions.

Material and Methods

Gene annotations

Gene annotations and genomic fasta sequences were downloaded from GENCODE (Frankish et al. 2019) (www.gencodegenes.org) for human (v19 corresponding to hg19) and mouse (vM16 corresponding to mm10), from Ensembl (www.ensembl.org) for zebrafish (dr11), fruit fly (dm6) and worm (ce11), and from EnsemblFungi (http://fungi.ensembl.org) for fission yeast (ASM294v2). For worm (*C. elegans*), we additionally considered recently identified lincRNA genes by Akay et al. (Akay et al. 2019), which are not included 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 only one coding region was analysed, namely the longest coding region starting with a canonical start codon (AUG) among all coding regions of different transcript isoforms.

Identification of putative ORFs (pORFs) in lincRNAs

LincRNA genes were required to not overlap with any other gene annotation on either strand. We also excluded lincRNAs that overlapped with regions of high PhyloCSF score (Lin, Jungreis, and Kellis 2011) (downloaded from https://data.broadinstitute.org/compbio1/PhyloCSFtracks/ for human, mouse, fly, and worm), and with regions that were identified experimentally in human cells to likely encode small proteins based on Ribo-Seq data (downloaded from Supplementary Table 1 of Martinez et al. (Martinez et al. 2020)). 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 longest pORF was considered for further analysis among all pORFs of different transcript isoforms.

Intronic and intergenic control sequences

For each lincRNA transcript harboring the longest pORF, 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 (Quinlan and Hall 2010). In each region, the longest ORFs (>10 codons) was identified. To avoid a bias from a particular random selection, 10 random selections were considered for each lincRNA transcript. 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.

Small protein-encoding ORFs (smORFs)

In Figure 5, human smORFs were defined as annotated lincRNAs, whose longest pORF overlapped with small protein-encoding regions, which were defined as regions with high PhyloCSF score (Lin, Jungreis, and Kellis 2011) (downloaded from https://data.broadinstitute.org/compbio1/PhyloCSFtracks/ for human) and regions that were identified experimentally before in three human cell lines based on Ribo-Seq data (downloaded from Supplementary Table 1 of Martinez et al. (Martinez et al. 2020)).

Sequence and structure context around start codons

To analyse the sequence context around start codons the fraction of nucleotides (A,C,G,T) at each position in the region +/- 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 frequency of nucleotide *n*. The probability (*P*) for the mRNA start codon region motif (also referred to as similarity with it) 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 frequency of nucleotide *n* in mRNAs at position *l* around the start codon of an ORF.

For the structure context analysis in Figures S2A and S8E, the minimum free energy of the sequence region +/- 25 nucleotides around start codons was calculated using RNAfold (Hofacker 2004). In Figure S2B the probability for a nucleotide being accessible (i.e. not paired) was calculated using RNAplfold (Bernhart, Mückstein, and Hofacker 2011), in a sliding window of 70 nucleotides (sliding by 1 nucleotide) with the parameters -u 1 -W 70.

Multiple correspondence and principal component analysis

The multiple correspondence analysis in Figure 2A was implemented in python and carried out on codon counts, and the principal component analysis in S3 was performed on codon frequencies for each species and RNA type using matplotlib.mlab.PCA in Python 2.7.

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 all species studied. The number of annotated tRNA genes coding for the same tRNA anticodon type were counted considering high confidence tRNA gene predictions. Effective tRNA abundances were calculated using previously determined weights accounting for tRNA editing and wobble-base pairing efficiencies (dos Reis, Savva, and Wernisch 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:

$(\text{tRNA-NNU})_{eff}$ = tRNA-NNU + (1 - w(G:U)) * tRNA-NNC

$(\text{tRNA-NNC})_{eff}$ = tRNA-NNC + (1 - w(I:C)) * tRNA-NNU

$(\text{tRNA-NNA})_{eff}$ = tRNA-NNA + (1 - w(I:A)) * tRNA-NNU

$(\text{tRNA-NNG})_{eff}$ = tRNA-NNG + (1 - w(U:G)) * tRNA-NNA

The above nucleotide triplets are the reverse complements of the tRNA anticodon sequences (i.e. the recognized codon sequence). N stands for any nucleotide.

Estimation of cell type-specific 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. smallRNA-seq and H3K27ac-ChIP-seq data were used previously to quantify tRNA expression (Ji et al. 2015b; Hernandez-Alias et al. 2020; Gingold et al. 2014), and these data are more widely available for different species and cell types. We established which of the latter two experimental approaches performs better by comparing relative tRNA expression levels based on smallRNA-seq and H3K27ac-ChIP-seq with those from hydro-tRNA-seq and DM-tRNA-seq in HEK293 (Figure S7A). Since smallRNA-seq-based relative tRNA abundances correlate better with those from hydro-tRNA-seq and DM-tRNA-seq data, we used smallRNA-Seq data to calculate cell type-specific relative tRNA abundances in our analysis. As a confirmation of this choice, HEK-specific tAIs calculated using smallRNA-seq-based tRNA abundances showed the most significant difference between expressed mRNAs and mRNAs with highest cytoplasmic expression, while tAIs calculated using H3K27ac-ChIP-seq-based tRNA abundances showed no difference between both (Figure 7B). Increased tAIs for mRNAs with highest cytoplasmic expression is expected, based on previous analyses of mRNA codon biases.

In the following, relative tRNA quantification based on different experimental data is explained in detail. Effective tRNA abundances were calculated from relative tRNA abundances as described above.

(a) relative tRNA quantification based on smallRNA-seq data: Fastq files with smallRNA-seq reads were downloaded for different species and cell types 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 also tested, but did not improve the correlation with 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; not shown.) 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 as 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 dedicated 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, Savva, and Wernisch 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 of all effective tRNA abundances:

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

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

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

Randomized control sequences

Shuffled sequences were generated by randomly permuting nucleotides in an ORF. Shuffled tAIs were calculated for codon frequencies from 100 random permutations.

Frame-shifted tAIs were calculated for codon frequencies in sequences starting one and two nucleotides downstream of start codons and ending two and one nucleotide upstream of stop codons of ORFs.

Adaptation of the correlation test for translational selection from dos Reis et al. (2004)

To test for translational selection, dos Reis et al. (dos Reis, Savva, and Wernisch 2004) proposed a correlation test, in particular the correlation between tAI and the effective number of codons, adjusted for the GC bias at the third codon position in mRNAs of a species. Here, we hypothesized that amino acid identities should not constrain lincRNA codon usage. Thus, we adapted the calculation of the effective number of codons to be independent of amino acid identities, treating all codons as synonymous. Also, we adjusted for the GC bias at all codon positions. All calculations were implemented in python.

Quantification of cytoplasmic gene expression levels per species

In Figures 1C, 3, S2A, S5 and S6, we sought to estimate cytoplasmic RNA expression levels for three species (human, fly and fission yeast) by combining cytosolic gene expression quantifications from different studies (see Table S3 for sources and accession codes). For human, we used data for five cell lines, GM12878, HeLa-S3, HEK293, HepG2 and K562 (Subtelny et al. 2014; Consortium and The ENCODE Project Consortium 2004), for fly we used data for two cell lines, S2 and ML-DmD17-c3, and early embryos (Aspden et al. 2014; Bouvrette et al. 2018; H. Li et al. 2016), and for yeast we used data for two different growth conditions (Subtelny et al. 2014; Herzel, Straube, and Neugebauer 2018). For four human cell lines, except HEK293, we downloaded gene quantifications from ENCODE (www.encodeproject.org). For HEK293 and other species, we downloaded fastq files with polyA-selected RNA-seq reads obtained from cytosolic RNA fractions (see Table S3), preprocessed

them using the fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), and quantified gene expression levels using RSEM (B. Li and Dewey 2011) and STAR (Dobin et al. 2013). Gene expression levels from replicates were averaged. To combine expression data from different experiments for each species, we first ranked genes by their cytoplasmic expression level in each experiment. We then combined ranks for each gene by determining its median and maximum rank across experiments. Combined ranks were normalized to the maximum combined rank, so that genes with highest cytosolic expression levels had ranks close to 1. In Figure 1C and S2A, top cytoplasmic lincRNAs, were those with a maximum rank > 0.85. In Figures 3, S5 and S6, top-cytoplasmic mRNAs were those with a maximum rank > 0.99 and cytoplasmic lincRNAs were those with a median rank > 0.5.

Quantification of total and cytosolic transcript expression levels in human cell lines

In Figures 4, 5, S7 and S8, we used total and cytoplasmic transcript expression levels for five human cell lines (GM12878, HeLa-S3, HEK293, HepG2 and K562; see Table S3 for sources and accession codes). Transcript quantifications for four cell lines, except HEK293, were downloaded from ENCODE (Consortium and The ENCODE Project Consortium 2004). For HEK293, polyA-selected RNA-seq reads generated by (Aktaş et al. 2017; Subtelny et al. 2014) preprocessed using the fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), and transcript expression levels were quantified using RSEM (B. Li and Dewey 2011) and STAR (Dobin et al. 2013). Expression levels were averaged over replicates. RNAs with TPM > 0.1 were considered expressed. The threshold for defining cytoplasmic RNAs was set to the first-quartile (25%) of the mRNA cytosolic expression values, for each cell type. We excluded histone mRNAs, as these are not usually polyadenylated and result in incorrect expression values in polyA-selected RNA-Seq.

In Figure 4B, mRNAs with cytoplasmic expression levels matching those of lincRNAs were selected as follows. First, for each cytoplasmic lincRNA all mRNAs were identified that had similar cytoplasmic expression level (identical values after log10-transforming and rounding to 2-digits after comma). Then 50 mRNAs were randomly sampled with replacement for each cytoplasmic lincRNA.

Quantification and analysis of Ribo-Seq data in human cell lines

Ribo-Seq data for five human cell lines were downloaded from several studies ((Cenik et al. 2015; Martinez et al. 2020; Subtelny et al. 2014; Solomon et al. 2017; Huang et al. 2019); 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 the reads' 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 for Ribo-Seq reads of different length (read offset) was determined from 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 estimated ribosome P site (calculated using the determined read offsets) was in-frame with 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 length of coding region/pORF, total number Ribo-Seq reads mapping in-frame to mRNA coding regions in a sample and multiplied by 1e+9) to the cytosolic expression level (FPKM) of transcripts harboring the longest coding region/pORF. In Figure 5C, to estimate relative ribosome-binding for cytoplasmic lincRNAs without Ribo-Seq reads mapping in-frame to their ORFs, a pseudo count of 0.1 reads was used.

We excluded from this analysis histone mRNAs, as these are not usually polyadenylated and result in incorrect expression values in polyA-selected RNA-Seq, and snoRNAs, as these may associate with ribosomes that translate other RNAs.

Acknowledgements

We would like to thank the ENCODE consortium and other experimental labs for generating the publicly available data used in this study. 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 analyses and prepared the results. RD mapped Ribo-Seq data. AB, ACM, and SB discussed the results and wrote the paper.

Declaration of Interests

The authors declare no competing interests.

Figure 1

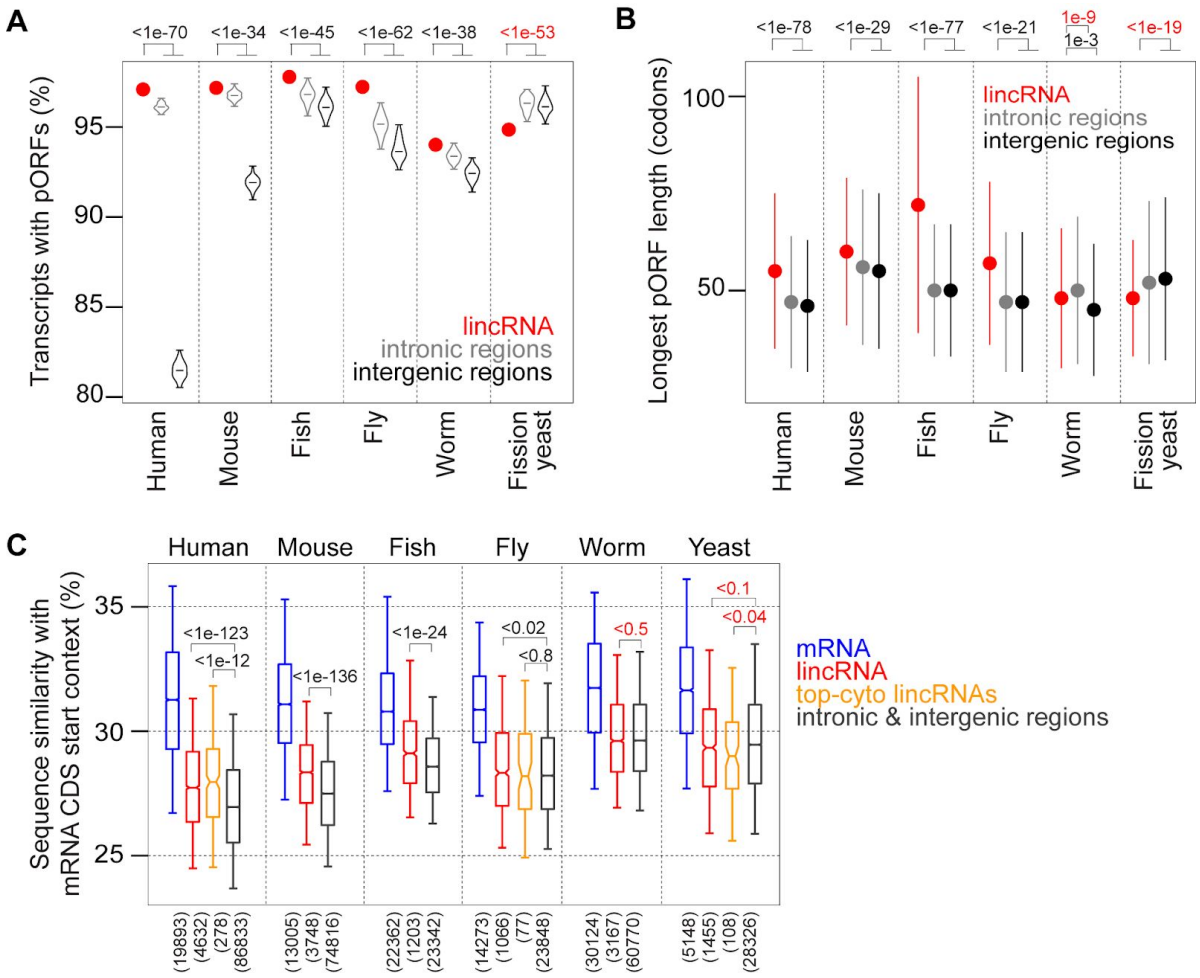
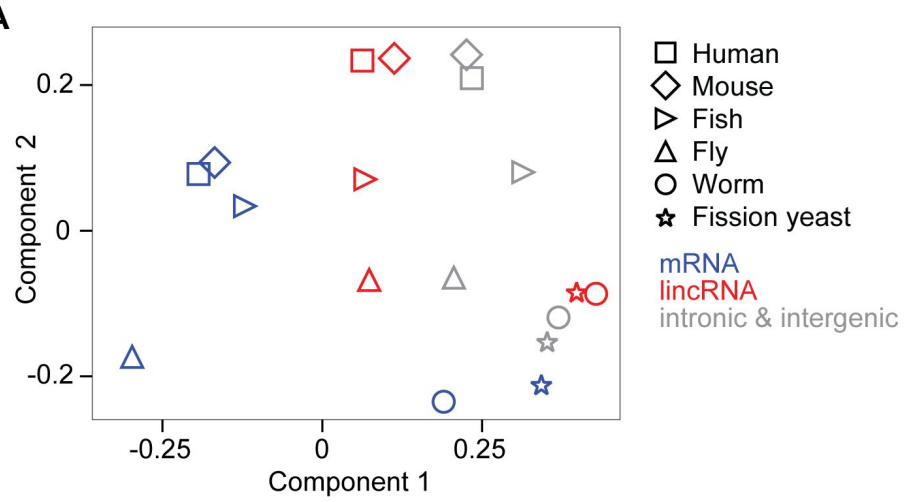


Figure 1: Prevalence, length, and start codon context of lincRNA pORFs.

(A) Median percentage of lincRNA transcripts with pORF (red dot) and violin plots of median percentages from 100 sets of size-matched, randomly selected intronic (gray) and intergenic (black) control regions with ORFs for six species. P values are indicated from a one-sample t test to compare the median for lincRNAs with those from the control sets. (B) Median lengths of longest lincRNA (red), and intronic (gray) and intergenic (black) control ORFs. Error bars represent median absolute deviations. P values are indicated from Wilcoxon's rank-sum test to compare the lengths of longest lincRNA pORFs with those of the longest control ORFs. (C) Sequence similarity (see Methods) of the region +/-12 nucleotides around start codons with the consensus mRNA sequence motif for that region in different species, for mRNAs (blue), lincRNA pORFs (red), top cytoplasmic-expressed lincRNA pORFs (median rank >85%, see Methods; orange), and intronic and intergenic control pORFs (black). For better visualization, intronic and intergenic control pORFs were combined. P values are indicated from Wilcoxon's rank-sum test to compare the similarities for (top-cytoplasmic) lincRNA pORFs with those of control ORFs. P values are marked red, if median lincRNA values were below the ones for control regions.

Figure 2

A



B

	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: Comparison of tri-nucleotide ("codon") frequencies in mRNAs, lincRNAs and control regions.

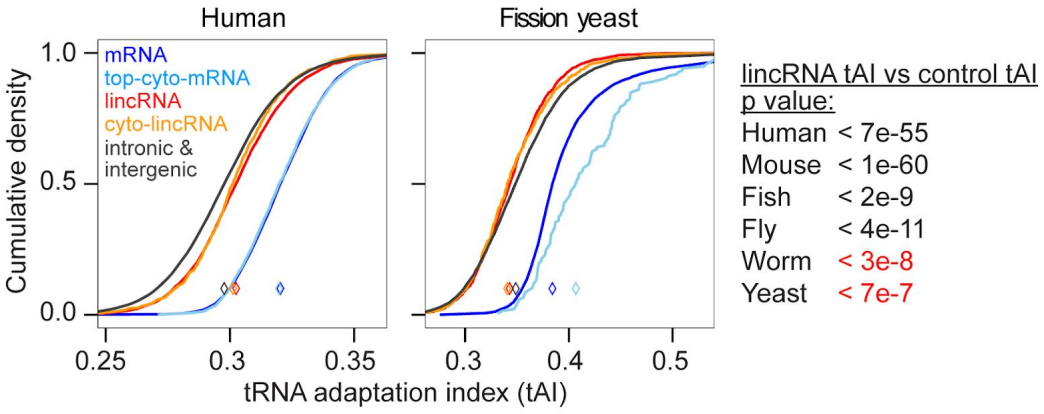
(A) First two components from a multiple correspondence analysis performed on tri-nucleotide ("codon") counts (excluding start and stop codons) in mRNA, lincRNA, and intronic and intergenic control regions for six species. For better visualization, intronic and intergenic control regions were combined. Similar results from a principal component analysis are shown in Figure S3A. (B) Spearman correlation coefficients between tri-nucleotide frequencies in mRNA, lincRNA, intronic and intergenic control regions for six species.

Figure 3

A

	Human	Mouse	Fish	Fly	Worm	Yeast
mRNA	0.58	0.53	0.46	0.58	0.58	0.46
top-cyto-mRNA	0.58			0.59		0.59
lincRNA	0.28	0.24	0.38	0.15	0.12	0.10
cyto-lincRNA	0.32			0.08		0.10
intronic & intergenic	0.17	0.10	0.21	0.04	0.21	0.34

B



C

	Human	Mouse	Fish	Fly	Worm	Yeast
mRNA	0.51	0.50	0.39	0.67	0.59	0.49
top-cyto-mRNA	0.53			0.75		0.61
lincRNA	0.05	-0.05	0.09	0.05	0.01	0.11
cyto-lincRNA	0.03			-0.06		0.06
intronic & intergenic	0.00	-0.02	-0.02	-0.03	0.05	0.11

Figure 3: Correspondence between tRNA abundances and codon frequencies in different RNA types.

(A) Spearman correlation coefficient between effective tRNA anticodon frequencies based on tRNA gene counts (see Methods) and codon frequencies in mRNA, lincRNA, intronic and intergenic control pORFs for six species; and in cytoplasmic mRNAs and lincRNAs for human, fly and yeast (see Methods). (B) Cumulative density of tRNA adaptation indexes (tAIs) for mRNA (blue), lincRNA (red), intronic and intergenic control sequences (gray) and cytoplasmic mRNAs (cyan) and lincRNAs (orange) for human (left panel) and fission yeast (right panel). P values from Wilcoxon's rank-sum test to compare tAIs of lincRNAs with those of control sequences are indicated for all six species on the right. P values are marked red, if median tAIs of lincRNA pORFs are lower than those of control ORFs. (C) Spearman correlation coefficient between tAI and the adjusted effective number of codons (see Methods), for the same ORF types and species as in (A).

Figure 4

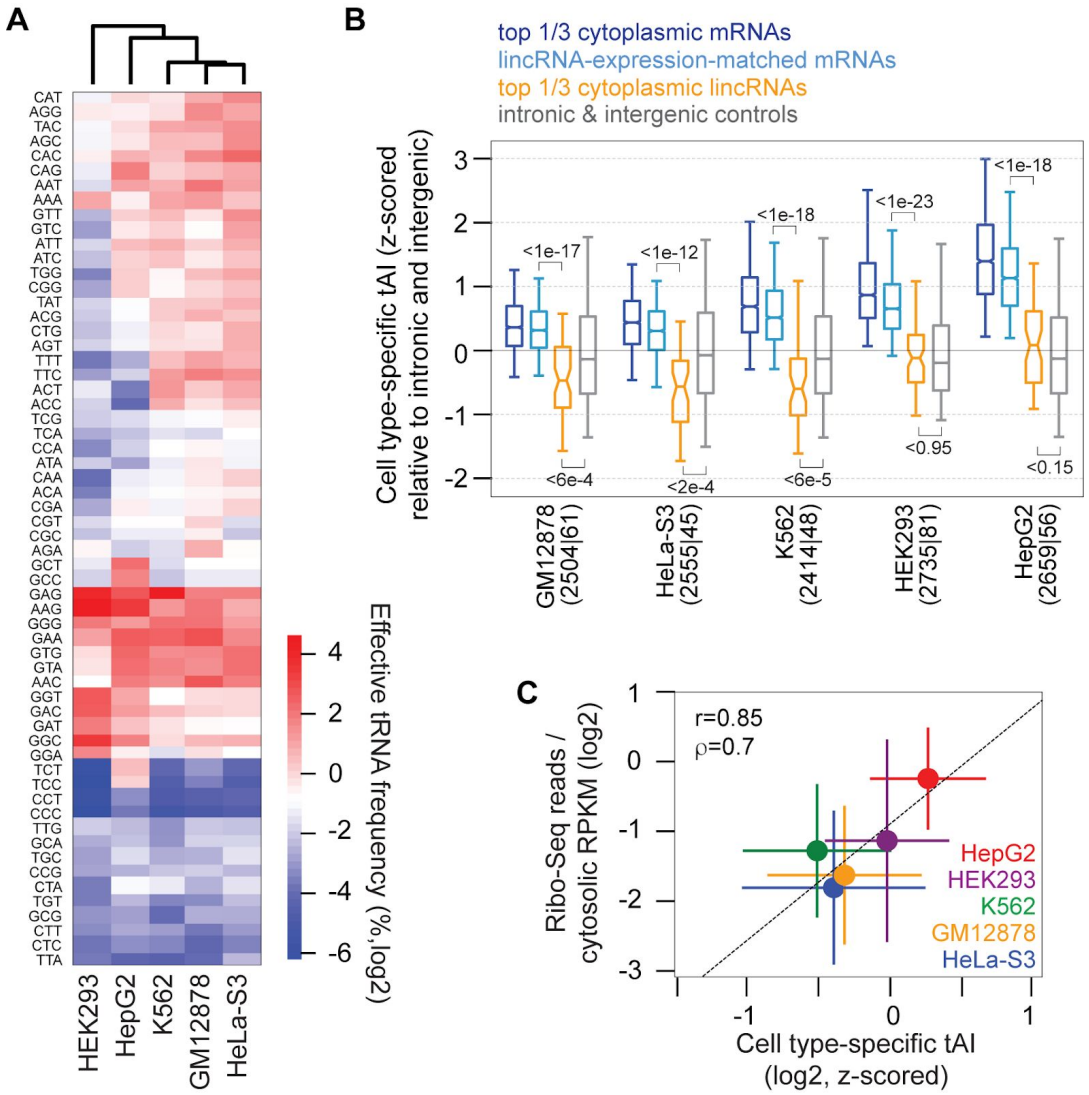


Figure 4: Cell type-specific tAIs and ribosome-binding in five human cell lines.

(A) Clustering (based on Euclidean distance) of effective tRNA anticodon frequencies, estimated from smallRNA-Seq data, in five human cell lines. (B) Boxplot of cell type-specific tAIs (z-scored relative to cell type-specific tAIs of intronic and intergenic control ORFs) of the top-1/3 cytoplasmic expressed mRNAs (blue) and lincRNAs (orange), and mRNAs that matched lincRNA cytoplasmic expression levels (light blue), for five human cell lines (label at bottom). The number of top-1/3 cytoplasmic mRNAs and lincRNAs are indicated in parentheses. Indicated p values were calculated using Wilcoxson's rank sum test. (C) Relationship between z-scored cell type-specific tAIs (relative to controls) and z-scored relative ribosome-binding (relative to mRNAs) across five human cell lines. Dots show median values and error bars median absolute deviations for 45 lincRNAs that are located in the cytoplasms of all five cell lines. Pearson (r) and Spearman (ρ) correlation coefficients between median values are indicated.

Figure 5

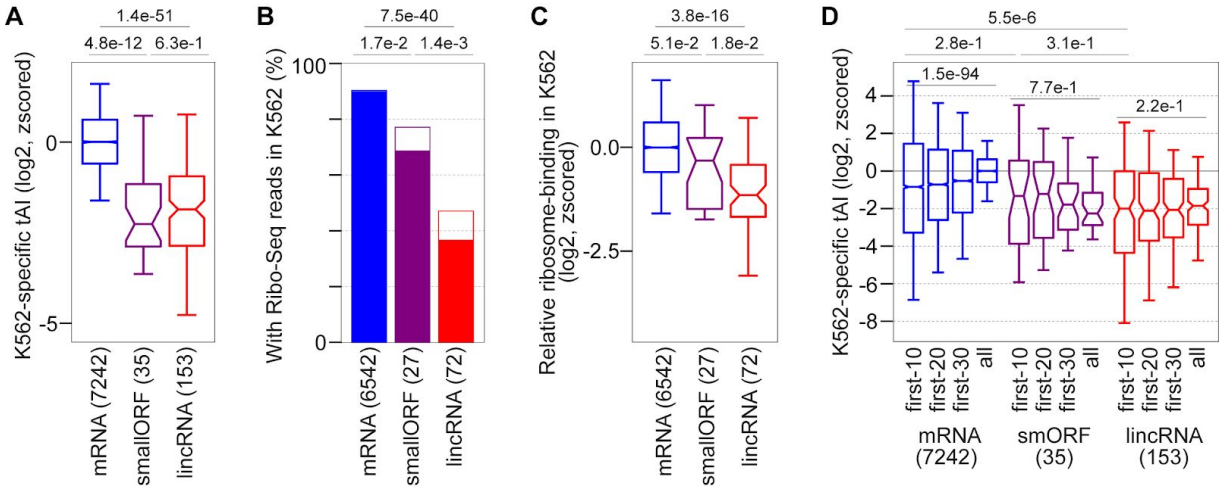


Figure 5: tAI, ribosome-binding and tAI of first codons in ORFs for lincRNAs, smORFs and mRNAs in K562 cells.

(A) Cell type-specific tAIs of three types of cytoplasmic RNA in K562 cells: mRNAs (blue), small protein-encoding ORFs (smORFs; purple) 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: with more than one read, empty bars: with 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 normalized in-frame Ribo-Seq reads to the cytoplasmic expression level; see Methods) for ORFs with Ribo-Seq reads. P values are indicated from Wilcoxon's rank-sum test. (D) Cell type-specific tAIs for different regions at ORF starts: the first 10, 20 or 30 codons downstream of start codons (first three boxes for each RNA type), and entire ORFs (last box for each RNA type). P values are indicated from Wilcoxon's rank-sum test.

Figure S1

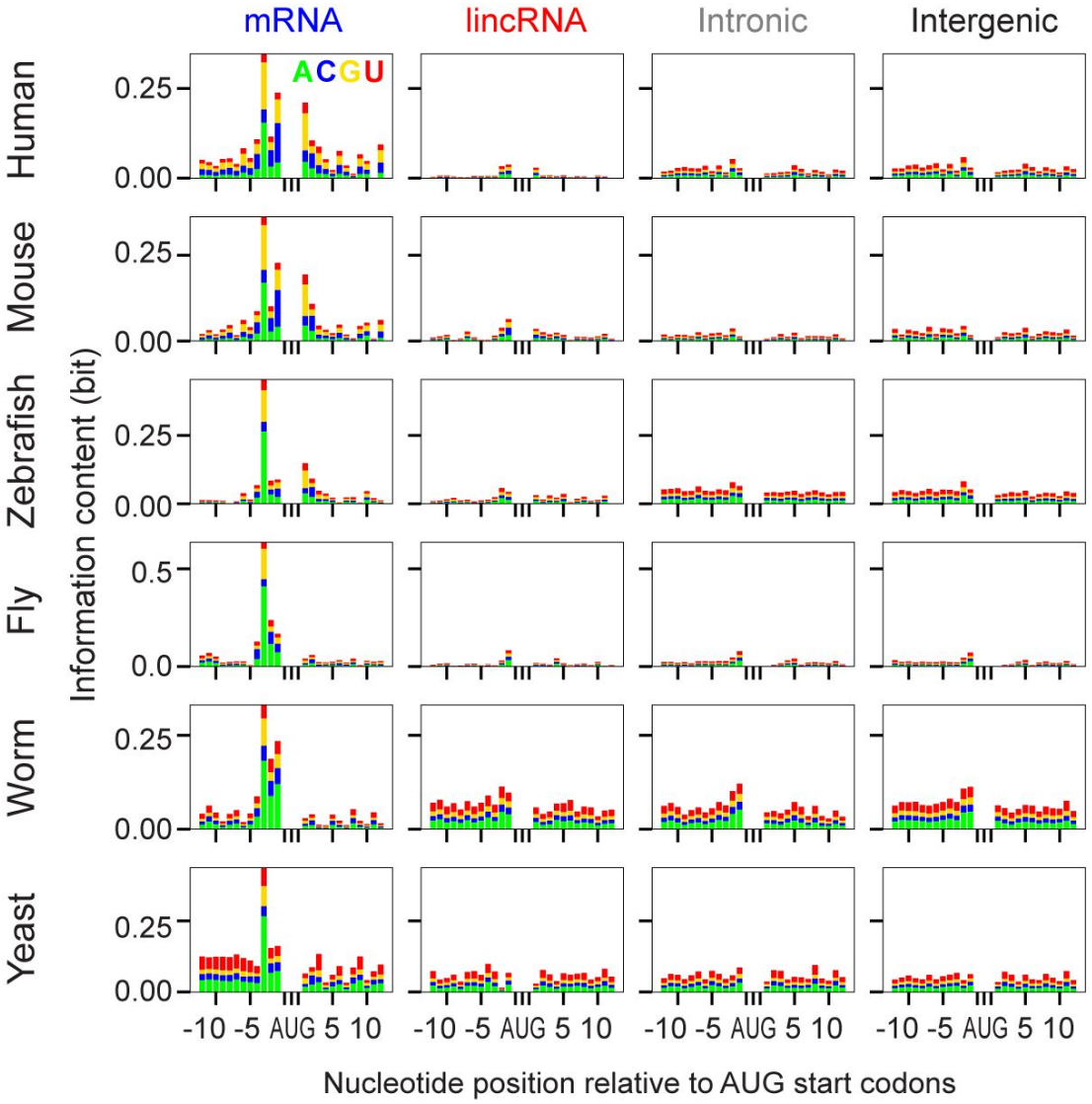


Figure S1: Consensus sequence motif around AUG start codons in different RNA types and species.

Information content (see Methods) for the region +/-12 nucleotides around AUG start codons in different RNA types (columns, indicated on top) and species (rows, indicated left). The sequence motif around mRNA start codons shows some similarity with the Kozak consensus sequence (gcc(A/G)ccAUGG).

Figure S2

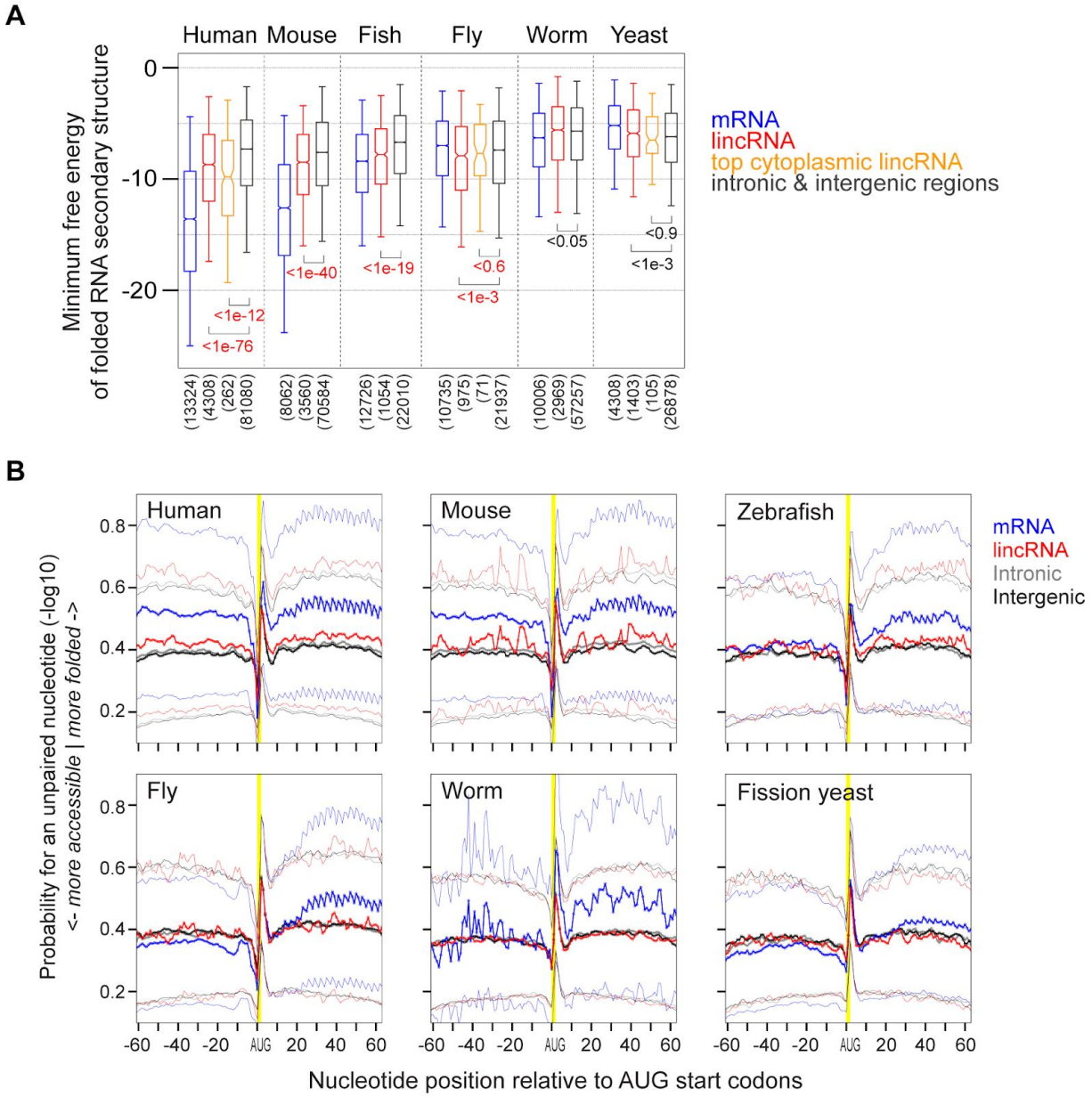


Figure S2: Structure context around AUG start codons in different RNA types and species.

(A) Minimum free energy (MFE) of RNA structures folded from the region +/- 25 nucleotides around AUG start codons in different species (indicated on top) and RNA types (color coded). P values are indicated from Wilcoxon's rank-sum test to compare the MFEs for (cytoplasmic) lincRNA pORFs with those of intronic and intergenic control ORFs. P values are marked red, if lincRNA median values are below the ones for intronic and intergenic controls. (B) Probability that a nucleotide is unpaired in a region +/- 60 nucleotides around AUG start codons for mRNA coding regions (blue), lincRNA pORFs (red), and ORFs in intronic (gray) and intergenic (black) control regions, for different species. The AUG start codon region is highlighted in yellow. Thick lines represent mean values and thin lines +/- 0.1 standard deviations.

Figure S3

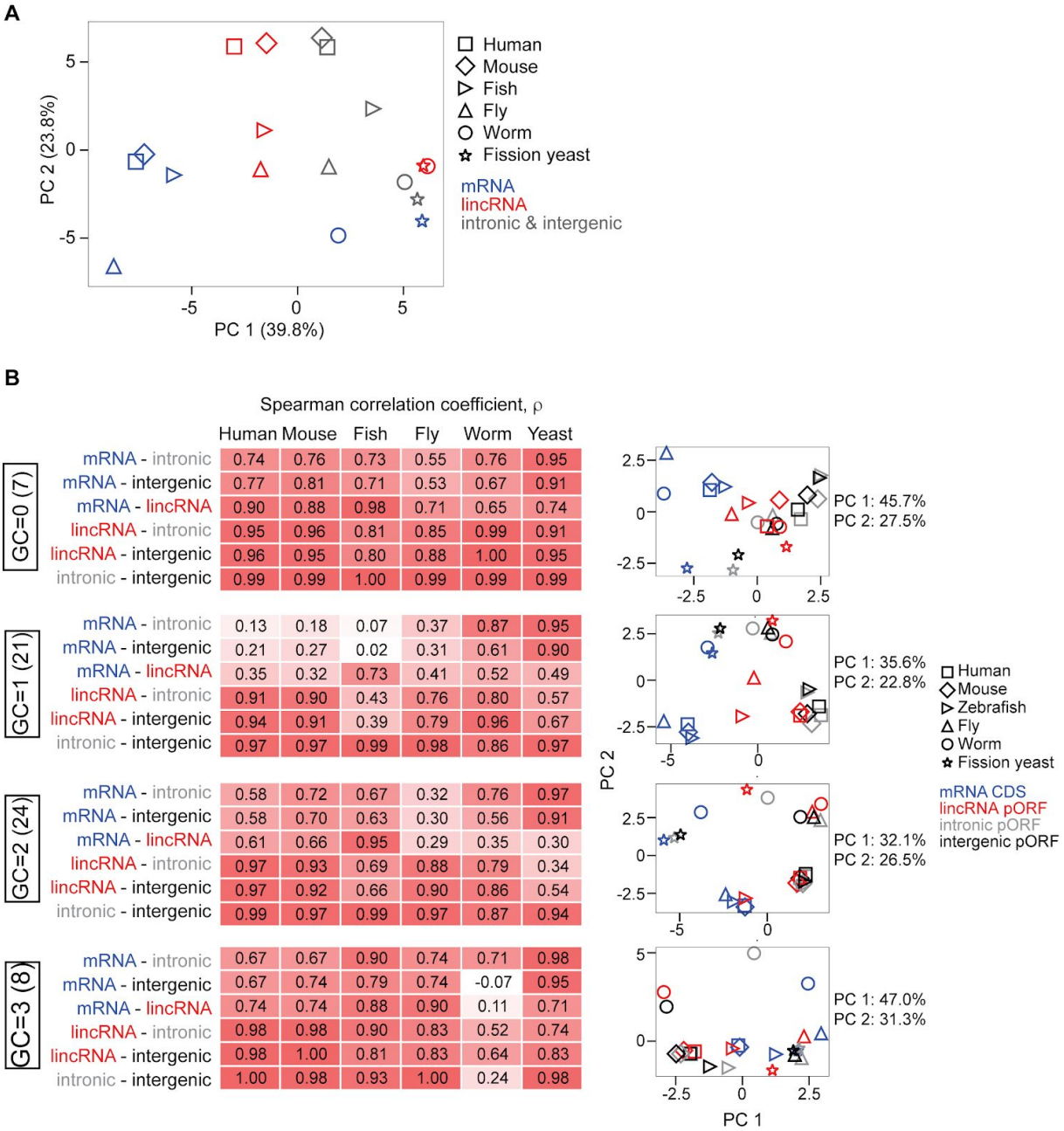


Figure S3: Influence of GC content on codon frequencies correlations.

(A) First two components from a principal component analysis performed on codon frequencies (excluding start and stop codons) in mRNA, lincRNA, intronic and intergenic regions for six species. (B) Left panels: Spearman correlation coefficients between codon frequencies in different RNA types for six species. Codons were separated into four groups according to the number of G and C nucleotides. First row shows correlations for 7 codons without G and C nucleotides, second row for 21 codons with one G or C nucleotide, thirds row for 24 codons with two G or 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 codon frequencies in different GC content groups for six species.

Figure S4

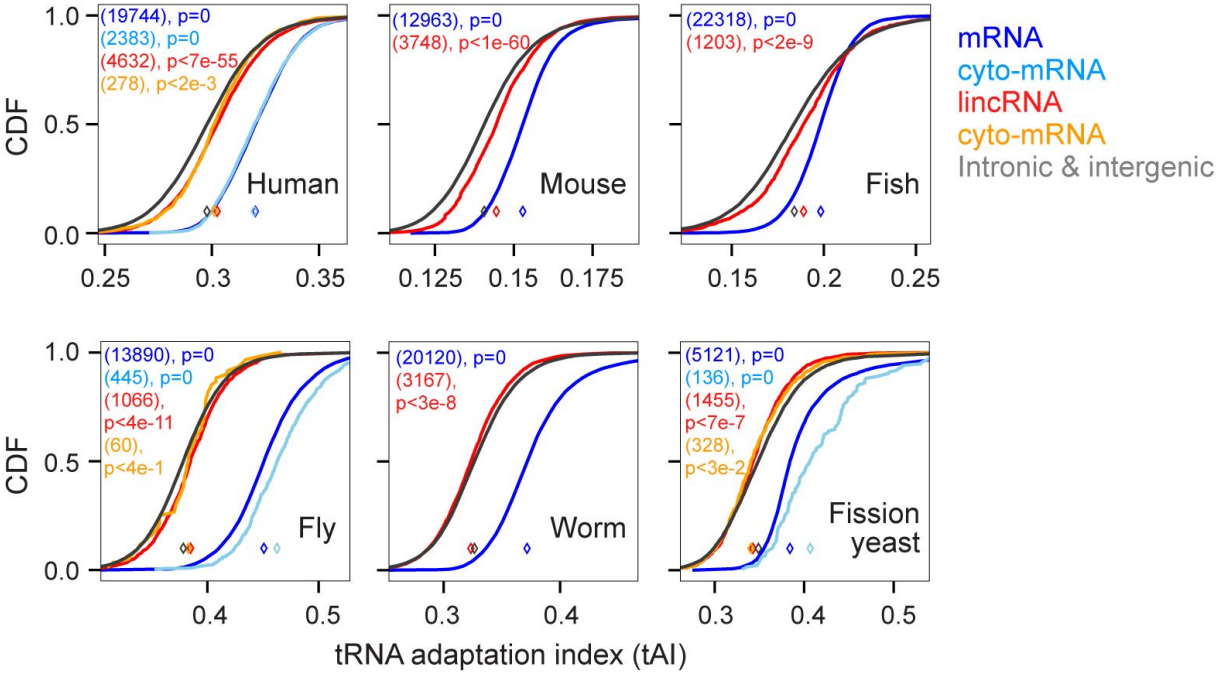


Figure S4: Cumulative distributions of tRNA adaptation indexes (tAIs) for codons in different RNA types and species.

tRNA adaptation index (tAI, calculated using tRNA gene counts; see Methods) for codons in mRNA coding regions (blue), lincRNA pORFs (red) and intronic and intergenic control ORFs (gray) in different species (see label inside each panel). In the case of human, fly, and yeast, tAIs of cytoplasmic mRNAs and lincRNAs are schown in addition. P values from Wilcoxon's rank-sum test are indicated to compare tAIs of control ORFs with those of mRNA and lincRNA (color-coded).

Figure S5

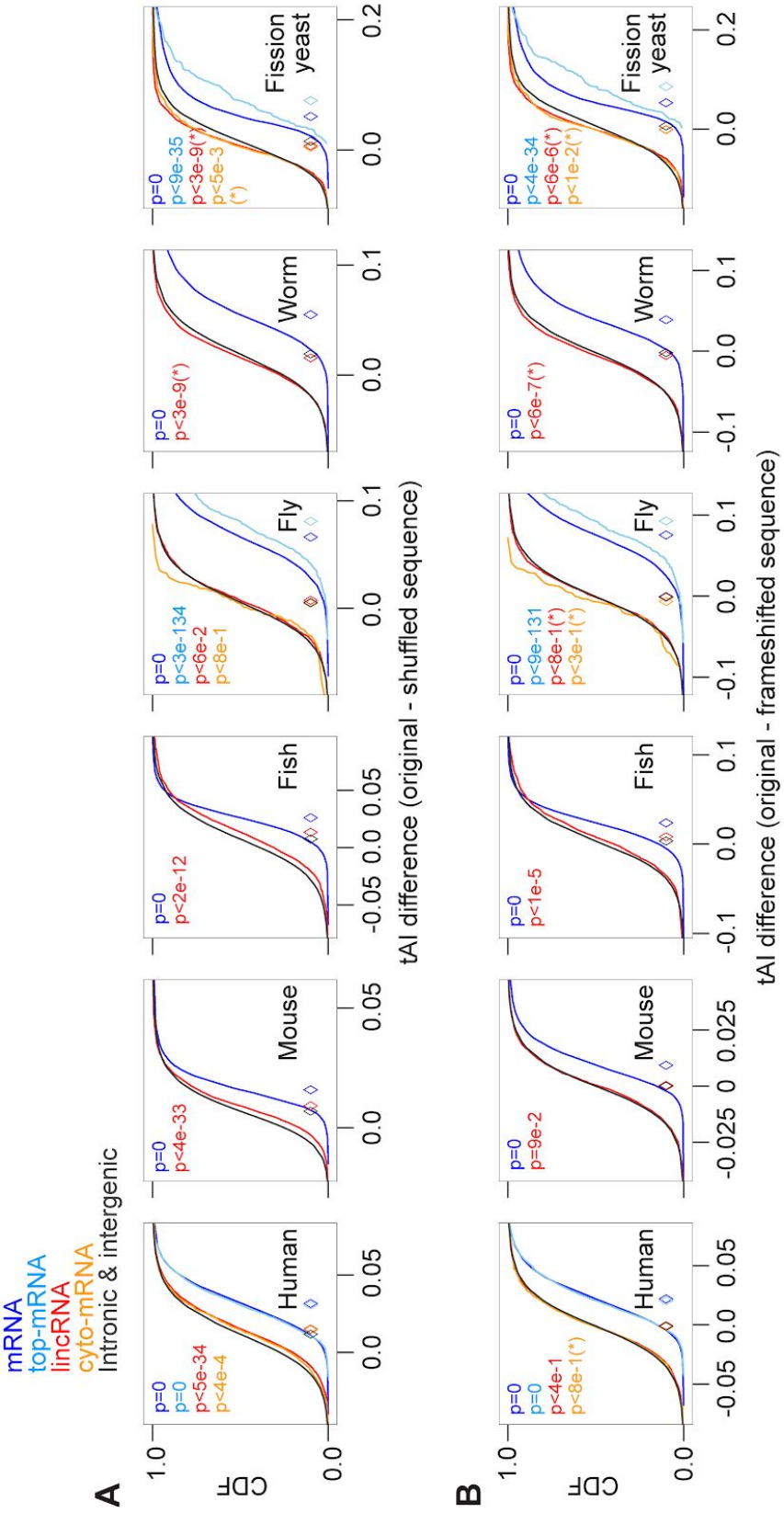


Figure S5: Comparison of tAIs of original sequences with those of shuffled and frame-shifted sequences in six species.

Cumulative distributions of tAI differences between original ORF and shuffled ORF sequences (A), and for original ORF - frame-shifted ORF sequences (B) for six species (indicated inside each panel). tAI differences were calculated for mRNA coding regions (blue), lincRNA pORF (red), and intronic and intergenic control ORFs (gray). For human, fly, and yeast, tAI differences of cytoplasmic mRNAs (light blue) and lincRNAs (orange) are shown in addition. P values were calculated using Wilcoxon's rank sum test to compare the distributions of tAI differences of intronic and intergenic controls with other RNA types (color-coded).

Figure S6

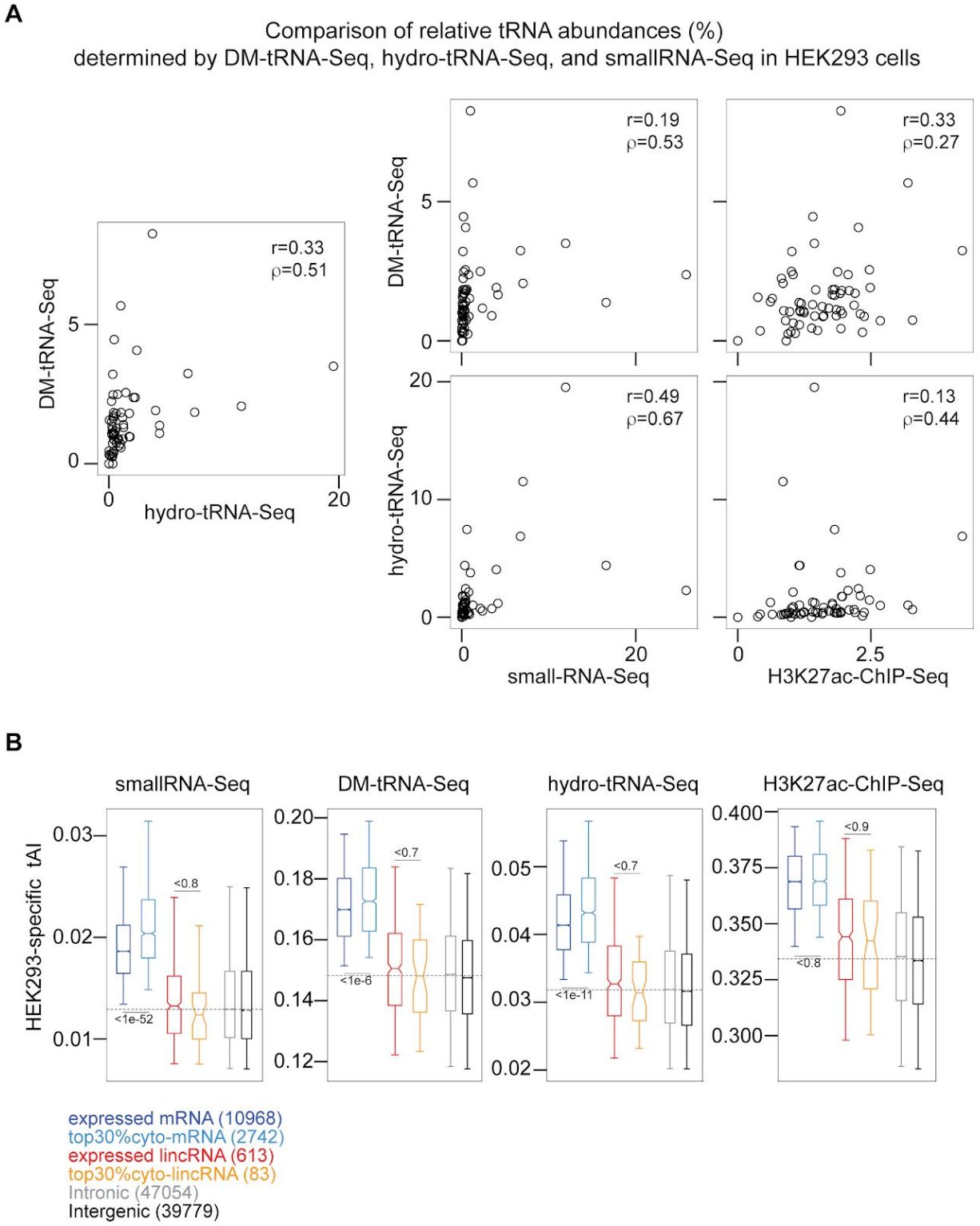


Figure S6: Comparison of tRNA frequencies and cell type-specific tAIs derived from different experimental approaches in HEK293.

(A) 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 (ρ) 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 those from smallRNA-Seq or H3K27ac-ChIP-Seq (see x-axis label; see Methods). Pearson (r) and Spearman (ρ) correlation coefficients are indicated. (B) Box plots of cell type-specific tAIs for HEK293 using different experimental approaches to quantify effective tRNA abundances: smallRNA-Seq (first panel), DM-tRNA-Seq (second panel), hydro-tRNA-Seq (third panel), and H3K27ac-ChIP-Seq (last panel). In each panel, cell type-specific tAIs are shown for expressed mRNAs (blue), top 30% cytoplasmic mRNAs (light blue), expressed lincRNAs (red), top 30% cytoplasmic lincRNAs (orange), and intronic (gray) and intergenic (black) control ORFs. The numbers of RNAs in each group are indicated at bottom. Indicated p values are calculated using Wilcoxon's rank-sum test. Cell type-specific tAIs in HEK293 show the most significant difference between expressed mRNAs and top cytoplasmic mRNAs with smallRNA-Seq-derived tRNA abundances, affirming the suitability of this measure.

Figure S7

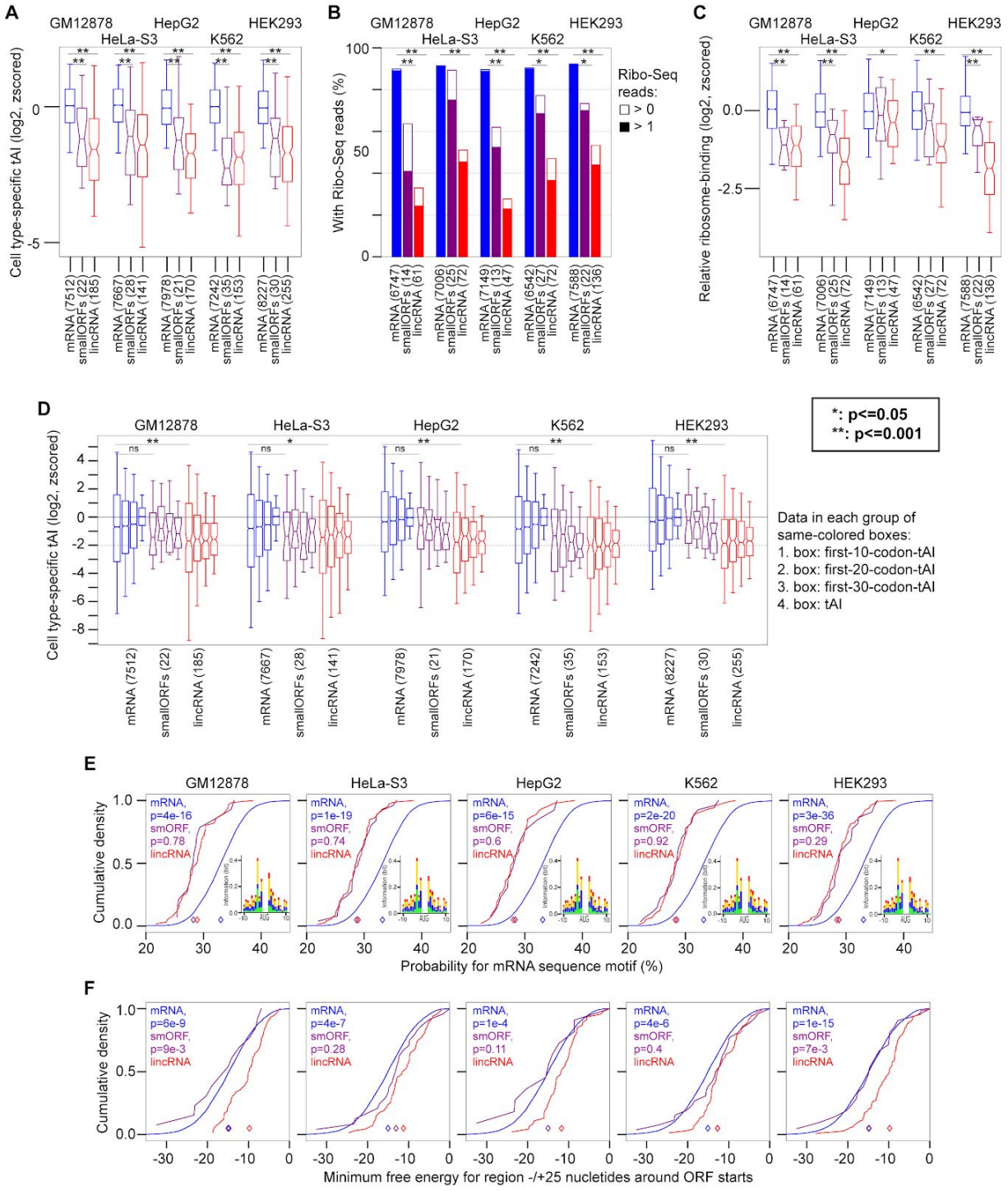


Figure S7: Comparisons of tAIs and ribosome-binding for mRNAs, small protein-coding RNAs and lincRNAs in five human cell lines.

(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 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 within ORFs in cytoplasmic RNAs: tAI for the first 10, 20, and 30 codons downstream of start codons shown in first, second, and third box, respectively; tAI of entire ORFs (fourth box) for each RNA type and in each cell line. 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 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 human cell line. P values are calculated using Wilcoxon's rank-sum test to compare lincRNAs with the two other types of RNA.

Table S1: Number of analyzed mRNA coding regions and lincRNA pORFs in different species

	Human	**Mouse**	**Fish**	**Fly**	**Worm**	**Fission yeast**
Number of mRNAs (isoform with longest CDS)	19744	12963	22318	13890	20120	5121
Number of lincRNAs (isoform with longest pORF)	4632	3748	1203	1066	3167	1455

Table S2: Codon counts in different ORFs and species

Codon	hg19_mRNA	hg19_lincRNA	hg19_Intronic	hg19_Intergenic	mm10_mRNA	mm10_lincRNA	mm10_Intronic	mm10_Intergenic	dr11_mRNA	dr11_lincRNA	dr11_Intronic	dr11_Intergenic	dm6_mRNA	dm6_lincRNA	dm6_Intronic	dm6_Intergenic	ce11_mRNA	ce11_lincRNA	ce11_Intronic	ce11_Intergenic	sp_mRNA	sp_lincRNA	sp_Intronic	sp_Intergenic
AAA	285955	7566	84756	77179	152961	6117	70284	79684	364983	4816	26436	25834	132292	2452	25196	26373	312554	10187	101090	95355	95948	3289	52781	48012
AAC	208283	4054	36046	32086	131755	3989	35790	36419	282008	3033	12379	12006	186710	1414	11929	12366	150255	2944	30896	29806	42124	1533	22935	22446
AAG	356829	6494	50812	44236	216619	6262	51885	52927	359951	3677	11617	11243	286402	1235	12222	11970	210954	2843	34484	29483	57316	1593	31046	28694
AAT	193770	4674	53427	49215	107058	4364	42694	49736	202887	2996	17681	17706	162096	1622	17784	18282	252502	5792	63228	54627	81897	2466	40698	36799
ACA	171683	5918	52658	46757	108873	6080	61044	57347	229505	6323	18314	18227	87082	1525	12995	13099	168700	3105	31477	28969	34742	1460	19424	19016
ACC	205715	4332	32760	26195	121152	3808	31897	29025	191821	2669	7137	7217	156708	857	7259	7487	85405	1459	16656	16347	25137	937	12993	14033
ACG	66847	1240	7827	5534	36237	887	6829	5016	86376	1060	3839	4028	106640	750	6659	6563	73807	1281	16618	14380	15937	798	10053	10759
ACT	151817	4542	40233	33950	94448	4540	39712	38017	191367	4098	10843	11247	76194	917	9489	10053	159463	2389	27420	25396	55192	1244	24987	21897
AGA	137886	6930	55029	48567	88004	6830	58211	60817	198025	2829	12067	11924	42161	1014	9203	9583	126578	3102	30570	31299	27275	1308	16142	16964
AGC	223664	5363	39731	30570	130534	4396	37763	30288	231999	2806	9136	8991	150699	1198	10314	9813	68741	1541	18213	15466	22152	993	13493	14615
AGG	134436	6209	48332	37868	84945	5106	47257	43343	134776	1880	7398	7773	50601	876	7308	7597	31297	1576	14551	14066	12364	766	9858	10386
AGT	142852	4115	42208	34401	88916	3724	38628	37023	180332	2439	10612	10466	90308	856	9317	9382	100399	2374	24838	23335	36111	1242	18618	18356
ATA	85692	3153	46435	45122	53091	3136	43565	51366	97206	1994	18260	18149	75500	966	13452	13799	82009	3183	30683	31112	33302	1893	23923	23427
ATC	220493	3685	35036	31217	143740	3429	33827	37566	277121	2608	9959	10058	164948	926	9508	9460	156920	2274	29325	24793	29675	1418	18382	19176
ATT	178742	4787	58228	54109	104804	4059	45870	53758	195384	2790	18237	18474	127119	1424	18170	18562	273899	6137	65295	57457	83427	2468	41411	38752
CAA	144371	5173	46102	41185	84961	4898	41115	43544	165964	3046	13308	13102	120947	1669	15040	15072	222051	4008	44389	38659	65005	1672	31966	28276
CAC	169793	5382	44679	35376	102739	5173	47808	40213	192355	2354	12767	13058	116361	1245	10355	9596	73662	1914	20325	18463	14860	667	9084	9413
CAG	389288	7845	56347	43563	225811	7133	55372	46731	415749	4659	13010	12956	264079	1210	11121	10313	115671	1713	23181	18204	25968	677	13758	11669
CAT	125917	4816	45310	39689	76192	4967	44470	44115	144142	2318	13617	13470	80025	1233	11478	11251	114877	2799	28768	26044	39016	1397	21761	19455
CCA	196899	6852	53426	42656	118871	5660	47743	44545	205848	3143	9699	10529	105668	1342	11494	11713	213490	2013	27497	21877	30918	1127	15727	16441
CCC	228209	5652	41576	30689	120409	4178	39917	34236	148477	1914	5934	6445	132341	884	7915	7576	36852	1291	11946	14174	19011	461	8872	8305
CCG	82850	1669	9725	5979	41789	943	7167	4483	92747	1035	3332	3542	117006	827	6698	6409	80709	1338	16113	12070	11117	395	6307	6021
CCT	203120	6764	53178	41660	125279	6061	50303	45450	209425	2791	7753	8252	55146	696	7233	7891	73983	1667	16754	16174	51170	834	20527	17569
CGA	68906	1024	6899	4667	42754	906	5537	4412	81741	838	3394	3503	65422	1004	8021	7860	98685	1746	21987	16182	19426	699	10586	9969
CGC	116547	1351	8813	5191	60864	753	6406	3972	110704	944	4138	4273	127633	812	7606	6949	40462	1062	11519	9996	13959	414	7008	6511
CGG	128800	1526	9355	5603	65050	1015	7150	4516	78345	811	3297	3512	62325	793	6444	6213	38673	1265	14145	11143	7149	345	4728	4635
CGT	50494	1074	7435	5027	30568	934	6432	4940	80897	779	3597	3772	64131	578	6213	5963	89484	1410	16392	13455	35756	633	17004	13729
CTA	80788	3045	32570	28895	56093	2754	32412	33294	83169	1240	7504	7897	64870	530	6472	6944	67783	2023	17893	19202	21368	966	14934	12815
CTC	213355	6183	51239	40450	133038	5345	50411	46225	210670	2171	9181	9165	103003	853	8955	8532	120507	2190	25256	23133	17209	939	12185	12263
CTG	437608	8110	59604	45908	255773	7056	57713	50088	454577	4330	12987	13045	281280	1161	11724	10938	102001	1814	22309	18616	15581	722	12074	11941
CTT	151007	5625	51578	43644	94352	5557	51908	50416	162187	2296	11398	11618	70656	860	11842	11887	174138	3345	33992	31389	59867	1729	30934	28626
GAA	345234	6558	49349	44306	186463	6398	48628	52461	314918	4876	12084	12140	167391	1395	13300	13970	339413	4353	53625	43000	106185	1752	43116	38461
GAC	280707	3812	27281	22336	167761	3898	28366	26232	321689	3013	7497	7832	176516	967	7590	8189	140590	1501	19100	15777	36963	697	15319	14133
GAG	452574	6523	48573	38234	256565	5539	50235	46583	520610	5195	9420	9463	310826	971	9454	9054	200512	1985	28050	20123	49260	852	19829	19281
GAT	250587	3814	33734	28848	142226	3863	31340	33924	302165	3450	9595	9526	207751	944	9818	9895	298419	2265	35625	24226	90217	1269	34864	30150
GCA	181823	5065	39550	33221	107589	4050	38438	33705	200756	2885	10426	10769	96602	1497	12598	12854	167249	1948	24218	18888	38496	1138	18078	18316
GCC	313504	5241	38509	28166	168080	3630	31153	24622	215661	1929	6103	6118	239215	1158	10411	9597	102656	1290	15940	13047	26615	520	11986	11379
GCG	86287	1364	7695	4529	44104	734	6170	3941	92263	1037	3845	3995	102245	852	7740	7329	68895	961	12759	9926	12994	435	7087	6800
GCT	207315	5449	43224	33985	132855	4781	40301	35219	245920	2905	9279	9592	107733	960	10429	10264	185042	1459	22108	16527	69975	999	29052	23466
GGA	187314	5976	41317	33662	111979	5703	41350	39971	258981	2852	7616	7935	132308	1163	8959	8964	260091	1927	28535	20353	37979	1016	17347	16112
GGC	250177	4706	35733	25431	136344	3215	29448	23278	191053	1619	5581	5696	186020	1093	9995	9243	55599	1357	14120	11509	19449	526	9373	9298
GGG	183657	4875	40000	29619	98701	3716	36345	31642	120479	1239	5621	6243	35186	770	7733	7195	37166	1212	11503	11580	10713	475	6055	6404
GGT	119902	3341	32856	25592	76167	2960	30407	28826	157771	2018	6710	6927	96391	597	7371	7231	89654	1451	17560	16191	48929	785	21177	18546
GTA	80859	2290	29107	24664	51580	2631	28296	27989	83931	1674	8113	8253	49326	565	7176	7894	82520	1734	20868	19042	30042	1303	17981	18099
GTC	156559	3312	25410	20530	100514	2871	27463	23952	174945	2091	7489	7368	100303	710	7352	7571	110805	1482	18127	15979	24703	779	12758	12849
GTG	306244	5045	42620	33759	183092	4731	48747	41165	330948	3862	12570	12156	204523	964	10881	10135	121334	1964	23167	21386	20129	733	11464	12141
GTT	124287	3432	36732	31578	73648	3386	35263	35765	168283	2784	11821	11886	84916	834	11747	11798	200263	3087	34318	30771	68754	1459	31017	27229
TAC	162448	2485	28315	24098	101428	2628	28807	27657	188462	1692	8603	8539	136862	589	7400	7557	115766	1834	20194	19257	28042	1048	16374	14705
TAT	134748	3287	45811	42900	80736	3687	42872	46408	144041	2020	19157	18972	85769	820	13211	13109	149450	3300	33023	33134	53354	1654	27249	24404
TCA	144413	5739	53518	45525	84670	5449	46278	46430	192453	3631	13978	14200	61702	978	10602	10792	171309	3544	37559	34176	44161	1789	22551	21715
TCC	198991	5687	44562	36101	121063	4726	45867	44553	189155	2410	8114	8365	144881	1035	9284	9498	87938	2111	23327	20313	29057	1150	14937	15234
TCG	51197	1053	6842	4854	28632	805	5840	4611	65483	795	3323	3475	120771	873	9203	8727	100288	1758	22406	16844	19863	871	10642	11176
TCT	175462	6354	58821	50871	115476	6188	64736	66972	224067	4053	13363	13231	54768	697	9923	10669	139376	3628	32603	32862	73024	1654	30292	26623
TGC	137762	5101	40348	32594	84708	4430	37094	31689	140309	1827	10279	10239	99247	1231	12464	11621	76660	1997	20400	18383	13301	958	10534	10413
TGG	136672	6341	52070	42355	82028	5145	47096	44393	144066	1916	9351	9962	74486	1198	11575	10913	91548	1874	22021	20146	26602	808	13695	13573
TGT	120145	5122	51120	44181	81117	5408	61502	55607	155684	2525	17925	17414	43425	962	12475	12166	94840	3135	28207	28253	21647	1322	15226	15457
TTA	89406	3535	47362	42029	49461	3120	40838	44362	90616	1851	16568	16530	36710	912	12926	14298	82196	3737	32815	35071	63504	2078	34911	29321
TTC	216843	5418	51554	45729	141549	4784	51426	54229	242024	2599	12495	12553	160276	991	13506	13260	199108	4912	49611	44718	30831	2000	22539	23020
TTG	146279	4943	47130	41642	94215	4058	44285	45949	153146	2227	13274	13395	124921	1273	15969	15764	167086	4149	43726	40158	57341	1834	31889	30509
TTT	192449	6842	94213	82027	117473	6426	72810	76946	215475	3510	27585	27632	105458	1562	26114	25816	195380	11473	102740	101754	78844	3643	47474	45565

Table S3: Sources and accession codes of experimental data used in this study

Organism	Cell type	Experimental assay	Source	Accessions	Reference
Human	GM12878	smallRNA-Seq	www.encodeproject.org	ENCSR000CRE	The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004)
Human	GM12878	total polyA-RNA-Seq	www.encodeproject.org	ENCSR000COQ	The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004)
Human	GM12878	cytosolic polyA-RNA-Seq	www.encodeproject.org	ENCSR000COR	The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004)
Human	GM12878	Ribo-Seq	www.ncbi.nlm.nih.gov/geo/	GSM1609378, GSM1609379	C. Cenik, et al., Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. Genome Res. 25, 1610–1621 (2015)
Human	HeLa-S3	smallRNA-Seq	www.encodeproject.org	ENCSR000CRQ	The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004)
Human	HeLa-S3	total polyA-RNA-Seq	www.encodeproject.org	ENCSR000CPR	The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004)
Human	HeLa-S3	cytosolic polyA-RNA-Seq	www.encodeproject.org	ENCSR000CPP	The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004)
Human	HeLa-S3	Ribo-Seq	www.ncbi.nlm.nih.gov/geo/	GSM3566406, GSM3566407, GSM3566408, GSM3566409	T. F. Martinez, et al., Accurate annotation of human protein-coding small open reading frames. Nature Chemical Biology 16, 458–468 (2020)
Human	HepG2	smallRNA-Seq	www.encodeproject.org	ENCSR000CRX	The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004)
Human	HepG2	total polyA-RNA-Seq	www.encodeproject.org	ENCSR000CPE	The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004)
Human	HepG2	cytosolic polyA-RNA-Seq	www.encodeproject.org	ENCSR000CPF	The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004)
Human	HepG2	Ribo-Seq	www.ncbi.nlm.nih.gov/geo/	GSM2676361, GSM2676363	O. Solomon, et al., RNA editing by ADAR1 leads to context-dependent transcriptome-wide changes in RNA secondary structure. Nat. Commun. 8, 1440 (2017)
Human	HepG2	Ribo-Seq	www.ncbi.nlm.nih.gov/geo/	GSM3450419, GSM3450420	H. Huang, et al., Histone H3 trimethylation at lysine 36 guides m6A RNA modification co-transcriptionally. Nature 567, 414–419 (2019)
Human	K562	smallRNA-Seq	www.encodeproject.org	ENCSR000CRF	The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004)
Human	K562	total polyA-RNA-Seq	www.encodeproject.org	ENCSR000CPH	The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004)
Human	K562	cytosolic polyA-RNA-Seq	www.encodeproject.org	ENCSR000COK	The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004)
Human	K562	Ribo-Seq	www.ncbi.nlm.nih.gov/geo/	GSM3566410, GSM3566411, GSM3566412	T. F. Martinez, et al., Accurate annotation of human protein-coding small open reading frames. Nature Chemical Biology 16, 458–468 (2020)
Human	HEK293	smallRNA-Seq	www.ncbi.nlm.nih.gov/geo/	GSM1067868	S. Kishore, et al., Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. Genome Biol. 14, R45 (2013)
Human	HEK293	total polyA-RNA-Seq	www.ncbi.nlm.nih.gov/geo/	GSM2258993, GSM2258994, GSM2258996	T. Aktaş, et al., DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. Nature 544, 115–119 (2017)
Human	HEK293	cytosolic polyA-RNA-Seq	www.ncbi.nlm.nih.gov/geo/	GSM1276541	A. O. Subtelny, S. W. Eichhorn, G. R. Chen, H. Sive, D. P. Bartel, Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature 508, 66–71 (2014)
Human	HEK293	H3K27ac-ChIP-Seq	www.encodeproject.org	ENCSR000FCH	The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004)
Human	HEK293	Ribo-Seq	www.ncbi.nlm.nih.gov/geo/	GSM3566397, GSM3566398, GSM3566399	T. F. Martinez, et al., Accurate annotation of human protein-coding small open reading frames. Nature Chemical Biology 16, 458–468 (2020)
Human	HEK293	Ribo-Seq	www.ncbi.nlm.nih.gov/geo/	GSM1887643	A. O. Subtelny, S. W. Eichhorn, G. R. Chen, H. Sive, D. P. Bartel, Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature 508, 66–71 (2014)
Fly	S2	cytosolic polyA-RNA-Seq	www.ncbi.nlm.nih.gov/geo/	GSM1477474	J. L. Aspden, et al., Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. Elife 3, e03528 (2014)
Fly	ML-DmD17-c3	cytosolic polyA-RNA-Seq	www.encodeproject.org	ENCSR622ROA	L. P. B. Bouvrette, et al., CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in Drosophila and human cells. RNA 24, 98–113 (2018)
Fly	Early embryo	cytosolic polyA-RNA-Seq	www.ebi.ac.uk/arrayexpress/	E-MTAB-4571	H. Li, et al., Ultra-deep sequencing of ribosome-associated poly-adenylated RNA in earlyDrosophilaembryos reveals hundreds of conserved translated sORFs. DNA Research 23, 571–580 (2016)
Fission yeast	Rich growth medium (YES)	cytosolic polyA-RNA-Seq	www.ncbi.nlm.nih.gov/geo/	GSM2805397, GSM2805401, GSM2805405	L. Herzel, K. Straube, K. M. Neugebauer, Long-read sequencing of nascent RNA reveals coupling among RNA processing events. Genome Res. 28, 1008–1019 (2018)
Fission yeast		cytosolic rRNA-depleted RNA-Seq	www.ncbi.nlm.nih.gov/geo/	GSM1276537	A. O. Subtelny, S. W. Eichhorn, G. R. Chen, H. Sive, D. P. Bartel, Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature 508, 66–71 (2014)

References

Akay, Alper, David Jordan, Isabela Cunha Navarro, Tomasz Wrzesinski, Chris P. Ponting, Eric A. Miska, and Wilfried Haerty. 2019. “Identification of Functional Long Non-Coding RNAs in C. Elegans.” *BMC Biology* 17 (1): 14.

Aktaş, Tuğçe, İbrahim Avşar Ilık, Daniel Maticzka, Vivek Bhardwaj, Cecilia Pessoa Rodrigues, Gerhard Mittler, Thomas Manke, Rolf Backofen, and Asifa Akhtar. 2017. “DHX9 Suppresses RNA Processing Defects Originating from the Alu Invasion of the Human Genome.” *Nature* 544 (7648): 115–19.

Aspden, Julie L., Ying Chen Eyre-Walker, Rose J. Phillips, Unum Amin, Muhammad Ali S. Mumtaz, Michele Brocard, and Juan-Pablo Couso. 2014. “Extensive Translation of Small Open Reading Frames Revealed by Poly-Ribo-Seq.” *eLife* 3 (August): e03528.

Bentele, Kajetan, Paul Saffert, Robert Rauscher, Zoya Ignatova, and Nils Blüthgen. 2013. “Efficient Translation Initiation Dictates Codon Usage at Gene Start.” *Molecular Systems Biology* 9 (June): 675.

Bernhart, Stephan H., Ullrike Mückstein, and Ivo L. Hofacker. 2011. “RNA Accessibility in Cubic Time.” *Algorithms for Molecular Biology: AMB* 6 (1): 3.

Bouvrette, Louis Philip Benoit, Neal A. L. Cody, Julie Bergalet, Fabio Alexis Lefebvre, Cédric Diot, Xiaofeng Wang, Mathieu Blanchette, and Eric Lécuyer. 2018. “CeFra-Seq Reveals Broad Asymmetric mRNA and Noncoding RNA Distribution Profiles in Drosophila and Human Cells.” *RNA*. https://doi.org/10.1261/rna.063172.117.

Calviello, Lorenzo, Neelanjan Mukherjee, Emanuel Wyler, Henrik Zauber, Antje Hirsekorn, Matthias Selbach, Markus Landthaler, Benedikt Obermayer, and Uwe Ohler. 2016. “Detecting Actively Translated Open Reading Frames in Ribosome Profiling Data.” *Nature Methods* 13 (2): 165–70.

Cenik, Can, Elif Sarinay Cenik, Gun W. Byeon, Fabian Grubert, Sophie I. Candille, Damek Spacek, Bilal Alsallakh, et al. 2015. “Integrative Analysis of RNA, Translation, and Protein Levels Reveals Distinct Regulatory Variation across Humans.” *Genome Research* 25 (11): 1610–21.

Chan, Patricia P., and Todd M. Lowe. 2016. “GtRNAdb 2.0: An Expanded Database of Transfer RNA Genes Identified in Complete and Draft Genomes.” *Nucleic Acids Research* 44 (D1): D184–89.

Chen, Jin, Andreas-David Brunner, J. Zachery Cogan, James K. Nuñez, Alexander P. Fields, Britt Adamson, Daniel N. Itzhak, et al. 2020. “Pervasive Functional Translation of Noncanonical Human Open Reading Frames.” *Science* 367 (6482): 1140–46.

Consortium, The Encode Project, and The ENCODE Project Consortium. 2004. “The ENCODE (ENCyclopedia Of DNA Elements) Project.” *Science*. https://doi.org/10.1126/science.1105136.

Dana, Alexandra, and Tamir Tuller. 2014. “The Effect of tRNA Levels on Decoding Times of mRNA Codons.” *Nucleic Acids Research* 42 (14): 9171–81.

Dittmar, Kimberly A., Jeffrey M. Goodenbour, and Tao Pan. 2006. “Tissue-Specific Differences in Human Transfer RNA Expression.” *PLoS Genetics* 2 (12): e221.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21.

Eraslan, Basak, Dongxue Wang, Mirjana Gusic, Holger Prokisch, Björn M. Hallström, Mathias Uhlén, Anna Asplund, et al. 2019. “Quantification and Discovery of Sequence Determinants of Protein-per-mRNA Amount in 29 Human Tissues.” *Molecular Systems Biology* 15 (2): e8513.

Frankish, Adam, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge, et al. 2019. “GENCODE Reference Annotation for the Human and Mouse Genomes.” *Nucleic Acids Research* 47 (D1): D766–73.

Gingold, Hila, Disa Tehler, Nanna R. Christoffersen, Morten M. Nielsen, Fazila Asmar, Susanne M. Kooistra, Nicolaj S. Christophersen, et al. 2014. “A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation.” *Cell* 158 (6): 1281–92.

Gogakos, Tasos, Miguel Brown, Aitor Garzia, Cindy Meyer, Markus Hafner, and Thomas Tuschl. 2017. “Characterizing Expression and Processing of Precursor and Mature Human tRNAs by Hydro-tRNAseq and PAR-CLIP.” *Cell Reports* 20 (6): 1463–75.

Goodarzi, Hani, Hoang C. B. Nguyen, Steven Zhang, Brian D. Dill, Henrik Molina, and Sohail F. Tavazoie. 2016. “Modulated Expression of Specific tRNAs Drives Gene Expression and Cancer Progression.” *Cell* 165 (6): 1416–27.

Guimaraes, Joao C., Nitish Mittal, Alexandra Gnann, Dominik Jedlinski, Andrea Riba, Katarzyna

Buczak, Alexander Schmidt, and Mihaela Zavolan. 2020. “A Rare Codon-Based Translational Program of Cell Proliferation.” *Genome Biology* 21 (1): 44.
Guttman, Mitchell, Pamela Russell, Nicholas T. Ingolia, Jonathan S. Weissman, and Eric S. Lander. 2013. “Ribosome Profiling Provides Evidence That Large Noncoding RNAs Do Not Encode Proteins.” *Cell* 154 (1): 240–51.
Haerty, Wilfried, and Chris P. Ponting. 2015. “Unexpected Selection to Retain High GC Content and Splicing Enhancers within Exons of Multiexonic lncRNA Loci.” *RNA* 21 (3): 333–46.
Hanson, Gavin, and Jeff Coller. 2018. “Codon Optimality, Bias and Usage in Translation and mRNA Decay.” *Nature Reviews. Molecular Cell Biology* 19 (1): 20–30.
Heesch, Sebastiaan van, Franziska Witte, Valentin Schneider-Lunitz, Jana F. Schulz, Eleonora Adami, Allison B. Faber, Marieluise Kirchner, et al. 2019. “The Translational Landscape of the Human Heart.” *Cell* 178 (1): 242–60.e29.
Hernandez-Alias, Xavier, Hannah Benisty, Martin H. Schaefer, and Luis Serrano. 2020. “Translational Efficiency across Healthy and Tumor Tissues Is Proliferation-Related.” *Molecular Systems Biology* 16 (3): e9275.
Herzel, Lydia, Korinna Straube, and Karla M. Neugebauer. 2018. “Long-Read Sequencing of Nascent RNA Reveals Coupling among RNA Processing Events.” *Genome Research* 28 (7): 1008–19.
Hezroni, Hadas, Rotem Ben-Tov Perry, Zohar Meir, Gali Housman, Yoav Lubelsky, and Igor Ulitsky. 2017. “A Subset of Conserved Mammalian Long Non-Coding RNAs Are Fossils of Ancestral Protein-Coding Genes.” *Genome Biology* 18 (1): 162.
Hofacker, Ivo L. 2004. “RNA Secondary Structure Analysis Using the Vienna RNA Package.” *Current Protocols in Bioinformatics*. https://doi.org/10.1002/0471250953.bi1202s04.
Hoffmann, Steve, Christian Otto, Stefan Kurtz, Cynthia M. Sharma, Philipp Khaitovich, Jörg Vogel, Peter F. Stadler, and Jörg Hackermüller. 2009. “Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures.” *PLoS Computational Biology* 5 (9): e1000502.
Huang, Huilin, Hengyou Weng, Keren Zhou, Tong Wu, Boxuan Simen Zhao, Mingli Sun, Zhenhua Chen, et al. 2019. “Histone H3 Trimethylation at Lysine 36 Guides m6A RNA Modification Co-Transcriptionally.” *Nature*. https://doi.org/10.1038/s41586-019-1016-7.
Ji, Zhe, Ruisheng Song, Aviv Regev, and Kevin Struhl. 2015a. “Many lncRNAs, 5’UTRs, and Pseudogenes Are Translated and Some Are Likely to Express Functional Proteins.” *eLife*. https://doi.org/10.7554/elife.08890.
Kishore, Shivendra, Andreas R. Gruber, Dominik J. Jedlinski, Afzal P. Syed, Hadi Jorjani, and Mihaela Zavolan. 2013. “Insights into snoRNA Biogenesis and Processing from PAR-CLIP of snoRNA Core Proteins and Small RNA Sequencing.” *Genome Biology* 14 (5): R45.
Kozak, M. 1989. “The Scanning Model for Translation: An Update.” *The Journal of Cell Biology* 108 (2): 229–41.
Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.
Li, Bo, and Colin N. Dewey. 2011. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-12-323.
Li, Hongmei, Chuansheng Hu, Ling Bai, Hua Li, Mingfa Li, Xiaodong Zhao, Daniel M. Czajkowsky, and Zhifeng Shao. 2016. “Ultra-Deep Sequencing of Ribosome-Associated Poly-Adenylated RNA in earlyDrosophilaembryos Reveals Hundreds of Conserved Translated sORFs.” *DNA Research*. https://doi.org/10.1093/dnares/dsw040.
Lin, Michael F., Irwin Jungreis, and Manolis Kellis. 2011. “PhyloCSF: A Comparative Genomics Method to Distinguish Protein Coding and Non-Coding Regions.” *Bioinformatics* 27 (13): i275–82.
Martinez, Thomas F., Qian Chu, Cynthia Donaldson, Dan Tan, Maxim N. Shokhirev, and Alan Saghatelian. 2020. “Accurate Annotation of Human Protein-Coding Small Open Reading Frames.” *Nature Chemical Biology*. https://doi.org/10.1038/s41589-019-0425-0.
Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.journal*. https://doi.org/10.14806/ej.17.1.200.
Niazi, Farshad, and Saba Valadkhan. 2012. “Computational Analysis of Functional Long Noncoding RNAs Reveals Lack of Peptide-Coding Capacity and Parallels with 3’ UTRs.” *RNA* 18 (4): 825–43.
Nieuwkoop, Thijs, Max Finger-Bou, John van der Oost, and Nico J. Claassens. 2020. “The Ongoing Quest to Crack the Genetic Code for Protein Production.” *Molecular Cell* 80 (2): 193–209.
Ouspenskaia, Tamara, Travis Law, Karl R. Clauser, Susan Klaeger, Siranush Sarkizova, François

Aguet, Bo Li, et al. n.d. “Thousands of Novel Unannotated Proteins Expand the MHC I Immunopeptidome in Cancer.” https://doi.org/10.1101/2020.02.12.945840.
Pinkard, Otis, Sean McFarland, Thomas Sweet, and Jeff Coller. 2020. “Quantitative tRNA-Sequencing Uncovers Metazoan Tissue-Specific tRNA Regulation.” *Nature Communications* 11 (1): 4104.
Presnyak, Vladimir, Najwa Alhusaini, Ying-Hsin Chen, Sophie Martin, Nathan Morris, Nicholas Kline, Sara Olson, et al. 2015. “Codon Optimality Is a Major Determinant of mRNA Stability.” *Cell* 160 (6): 1111–24.
Quinlan, Aaron R. 2014. “BEDTools: The Swiss-Army Tool for Genome Feature Analysis.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 47 (September): 11.12.1–34.
Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq033.
Ransohoff, Julia D., Yuning Wei, and Paul A. Khavari. 2018. “The Functions and Unique Features of Long Intergenic Non-Coding RNA.” *Nature Reviews. Molecular Cell Biology* 19 (3): 143–57.
Raveh, Alon, Michael Margaliot, Eduardo D. Sontag, and Tamir Tuller. 2016. “A Model for Competition for Ribosomes in the Cell.” *Journal of the Royal Society, Interface / the Royal Society* 13 (116). https://doi.org/10.1098/rsif.2015.1062.
Reis, Mario dos, Renos Savva, and Lorenz Wernisch. 2004. “Solving the Riddle of Codon Usage Preferences: A Test for Translational Selection.” *Nucleic Acids Research* 32 (17): 5036–44.
Reis, Mario dos, and Lorenz Wernisch. 2009. “Estimating Translational Selection in Eukaryotic Genomes.” *Molecular Biology and Evolution* 26 (2): 451–61.
Reis, Mario dos, Lorenz Wernisch, and Renos Savva. 2003. “Unexpected Correlations between Gene Expression and Codon Usage Bias from Microarray Data for the Whole Escherichia Coli K-12 Genome.” *Nucleic Acids Research* 31 (23): 6976–85.
Riba, Andrea, Noemi Di Nanni, Nitish Mittal, Erik Arhné, Alexander Schmidt, and Mihaela Zavolan. 2019. “Protein Synthesis Rates and Ribosome Occupancies Reveal Determinants of Translation Elongation Rates.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (30): 15023–32.
Sabi, Renana, and Tamir Tuller. 2019. “Novel Insights into Gene Expression Regulation during Meiosis Revealed by Translation Elongation Dynamics.” *NPJ Systems Biology and Applications* 5 (April): 12.
Shi, Junchao, Eun-A Ko, Kenton M. Sanders, Qi Chen, and Tong Zhou. 2018. “SPORTS1.0: A Tool for Annotating and Profiling Non-Coding RNAs Optimized for rRNA- and tRNA-Derived Small RNAs.” *Genomics, Proteomics & Bioinformatics* 16 (2): 144–51.
Solomon, Oz, Ayelet Di Segni, Karen Cesarkas, Hagit T. Porath, Victoria Marcu-Malina, Orel Mizrahi, Noam Stern-Ginossar, et al. 2017. “RNA Editing by ADAR1 Leads to Context-Dependent Transcriptome-Wide Changes in RNA Secondary Structure.” *Nature Communications* 8 (1): 1440.
Subramanian, Sankar. 2008. “Nearly Neutrality and the Evolution of Codon Usage Bias in Eukaryotic Genomes.” *Genetics* 178 (4): 2429–32.
Subtelny, Alexander O., Stephen W. Eichhorn, Grace R. Chen, Hazel Sive, and David P. Bartel. 2014. “Poly(A)-Tail Profiling Reveals an Embryonic Switch in Translational Control.” *Nature* 508 (7494): 66–71.
Torrent, Marc, Guilhem Chalancon, Natalia S. de Groot, Arthur Wuster, and M. Madan Babu. 2018. “Cells Alter Their tRNA Abundance to Selectively Regulate Protein Synthesis during Stress Conditions.” *Science Signaling* 11 (546). https://doi.org/10.1126/scisignal.aat6409.
Tuck, Alex Charles, Aneliya Rankova, Alaaddin Bulak Arpat, Luz Angelica Liechti, Daniel Hess, Vytautas Iesmantavicius, Violeta Castelo-Szekely, David Gatfield, and Marc Bühler. 2020. “Mammalian RNA Decay Pathways Are Highly Specialized and Widely Linked to Translation.” *Molecular Cell* 77 (6): 1222–36.e13.
Tuller, Tamir, Asaf Carmi, Kalin Vestsigian, Sivan Navon, Yuval Dorfan, John Zaborske, Tao Pan, Orna Dahan, Itay Furman, and Yitzhak Pilpel. 2010. “An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation.” *Cell* 141 (2): 344–54.
Tuller, Tamir, Yedael Y. Waldman, Martin Kupiec, and Eytan Ruppin. 2010. “Translation Efficiency Is Determined by Both Codon Bias and Folding Energy.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (8): 3645–50.
Ulitsky, Igor, and David P. Bartel. 2013. “lincRNAs: Genomics, Evolution, and Mechanisms.” *Cell* 154 (1): 26–46.
Wagner, Andreas. 2005. “Energy Constraints on the Evolution of Gene Expression.” *Molecular Biology*

and Evolution. https://doi.org/10.1093/molbev/msi126.
Wang, Hongwei, Yan Wang, Shangqian Xie, Yizhi Liu, and Zhi Xie. 2017. “Global and Cell-Type Specific Properties of lincRNAs with Ribosome Occupancy.” *Nucleic Acids Research* 45 (5): 2786–96.
Zeng, Chao, and Michiaki Hamada. 2018. “Identifying Sequence Features That Drive Ribosomal Association for lncRNA.” *BMC Genomics* 19 (Suppl 10): 906.
Zerbino, Daniel R., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, et al. 2018. “Ensembl 2018.” *Nucleic Acids Research* 46 (D1): D754–61.
Zheng, Guanqun, Yidan Qin, Wesley C. Clark, Qing Dai, Chengqi Yi, Chuan He, Alan M. Lambowitz, and Tao Pan. 2015. “Efficient and Quantitative High-Throughput tRNA Sequencing.” *Nature Methods* 12 (9): 835–37.