LincRNA sequences are biased to counteract their translation

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

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

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

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

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

*: equal last authors

Abstract (183 words)

Long intergenic non-coding RNAs (lincRNAs) account for a large fraction of transcribed loci in the human genome. While many lincRNAs are retained in the cell nucleus, preventing their association with ribosomes, binding of cytosolic lincRNAs to ribosomes has been observed, but rarely results in translation. This raises the question of how translation of short open reading frames (ORFs) within cytosolic lincRNAs is hindered. Here, we investigate the content of nucleotide triplets in lincRNA putative ORFs (i.e. “codons”) and its potential impact on ribosome binding and translation.

We find that lincRNA and mRNA ORFs have distinct codon frequencies, that are well conserved between human and mouse. In lincRNAs, codon frequencies are less correlated with the corresponding tRNA abundance measures than in mRNAs. This correlation is weaker for cytoplasmic lincRNAs and lowest for those without experimental evidence for ribosome binding.

Our results suggest that putative lincRNA codons are a substrate of evolutionary forces modulating them to counteract unwanted ribosomal binding and translation. The resulting sequence signatures may help in distinguishing bona-fide lincRNAs with regulatory roles in the cytoplasm from transcripts coding for peptides.

Keywords: lincRNA, tRNA, codon bias, computational sequence analysis, ribosome, subcellular localization, peptide

Introduction

Long intergenic noncoding RNAs (lincRNAs) form a functionally heterogeneous class of RNA, that is defined based on their transcript length being longer than 200 nucleotides and their lack of protein coding potential (Frankish et al., 2019; Ulitsky and Bartel, 2013). Despite being classified as non-coding, many lincRNAs contain short open reading frames (ORFs) and some have been shown to associate with ribosomes (Guttman et al., 2013; Ingolia et al., 2014; Ji et al., 2015). While a small fraction of short ORFs in human lincRNAs (fewer than 10%) does translate into experimentally detectable peptides (Bánfai et al., 2012), the general consensus is that most ribosome-lincRNA associations are non-productive. In support of this, detailed analysis of lincRNA and ribosome interactions, using sequencing of ribosome protected fragments (Ribo-seq) revealed marked differences between the association of the translation machinery with protein-coding mRNAs and lincRNAs, including differences in the tri-nucleotide periodicity of binding (Calviello et al., 2016; Ji et al., 2015) or in ribosome release (Guttman et al., 2013). However, the mechanism(s) preventing translation of putative ORFs within lincRNAs, particularly of those located in the cytosol, are unclear.

The genetic code is degenerate and multiple synonymous codons can code for the same amino acid. Each of these codons is decoded by different tRNAs, whose abundances vary. The rate at which a codon is translated correlates with the abundance of the decoding tRNA (Dana and Tuller, 2014). In mRNA, codon usage is a strong regulator of translation efficiency and speed (Tuller et al., 2010a). For example, codon usage has been shown to differ between mRNAs functioning during proliferation or differentiation, in agreement with subsets of tRNAs that are induced during each of these processes (Gingold et al., 2014). This co-adaptation of codon usage to tRNA abundance has the consequence that protein translation efficiency is enhanced for genes required for a certain cellular process, while it is reduced for genes with opposing function. Such specialized translation programs have been observed in several contexts, e.g. under cellular stress conditions, in different tissues or in cancer (Goodarzi et al., 2016; Plotkin et al., 2004; Torrent et al., 2018). On the other hand, codons that are decoded by less abundant tRNAs are required at certain inter-protein domains to slow down amino acid synthesis. This allows a protein domain that is already synthesized to fold before the next protein domain is being synthesized, which is important for ensuring correct co-translational folding of functional protein domains (Buhr et al., 2016; Komar et al., 1999; Walsh et al.; Yu et al., 2015). Thus, mRNA codon usage is fine-tuned and adapted to tRNA abundances to ensure optimal protein output.

Given the established role of codon usage in modulating translation rate and efficiency in mRNA, we hypothesised that the codon preferences within lincRNA putative ORFs would be a potential mechanism to counteract their translation. The prevention or early abortion of unwanted lincRNA translation is important for reducing energy waste and synthesis of possibly harmful peptides. Furthermore, ribosomes would not be blocked by lincRNA transcripts, but, instead, be available for protein synthesis.

Here, we analysed the tri-nucleotide (i.e. codon) composition of putative ORFs in lincRNA transcripts and detected a bias in the frequencies of putative codons in many lincRNA putative ORFs. We further related the bias in putative lincRNA codon frequencies to ribosome binding measured by Ribo-seq and propose that the usage of putative codons that are decoded by rare tRNAs is a mechanism to prevent cytosolic lincRNA translation.

Results

Codon usage is distinct between mRNA coding-regions and lincRNA putative open reading frames

We considered all long intergenic noncoding RNAs (lincRNAs) annotated by GENCODE (Frankish et al., 2019) in human (v19) and mouse (vM16). We predicted, for each lincRNA transcript, all possible open reading frames (ORFs) longer than 30 nucleotides, starting with a canonical start codon (AUG) and ending at the first in-frame stop codon (UAG, UAA, UGA). Most lincRNAs (97.5%) had at least one such predicted ORF. On average lincRNA transcripts had 10.1 predicted ORFs with an average length of 32.4 codons (8.5 times shorter than the average mRNA coding region of 476.2 codons). In the following, we considered for each lincRNA transcript its longest putative ORF (average length of 61.1 codons; Figure S1A).

To gain initial insight into the characteristics of codons in putative lincRNA ORFs we compared their relative frequencies with mRNA codon frequencies. Interestingly, the correlation coefficients for human (r^2=0.51; Figure 1A) and mouse (r^2=0.46; Figure S1B) were much lower than the correlation of lincRNA codon frequencies between species (r^2=0.94), which was similar to what we found for mRNA codon frequencies (r^2=0.99; Figure 1B). These results indicate that codon preferences are different between lincRNA and mRNA, and that the codon usage in lincRNA putative ORFs is also conserved, and may thus have a functional role.

In agreement with the overall GC content difference between lincRNAs and mRNAs (Haerty and Ponting, 2015), the most enriched codons in mRNAs were often GC-rich, while the most enriched codons in lincRNA ORFs were often AU-rich (Figure 1A). To test how different GC contents between lincRNA and mRNA impact the correlation in codon usage, we compared frequencies of codons with the same CG content. Regardless of the GC content codon class, we found that the correlations between lincRNA and mRNA codon frequencies were much lower in mouse and human (Figure S1C) than the correlations of lincRNA codon frequencies between human and mouse (r^2>0.96 for all four lincRNA GC content codon classes; Figure S1C), which was comparable to the correlations between mRNA codon frequencies in mouse and human (r^2>0.98 for all four GC content codon classes; Figure S1C). Thus, we concluded that the difference in codon frequencies between mRNA and lincRNA is not driven by GC content differences, and that lincRNA codon usage is distinct by itself.

The relative frequencies of amino acids encoded by mRNA coding regions and lincRNA putative ORFs are more similar to each other than the relative codon frequencies between mRNA and lincRNA (r^2=0.78 for amino acid usage versus r^2=0.51 for codon usage; Figure S1D), further suggesting that the different codon frequencies between the two types of transcripts have a functional role. Interestingly when we compare codon preferences for lincRNAs and mRNAs, we found that for most amino acids encoded by multiple codons the codon that is preferred by mRNAs is less used by lincRNA putative ORFs (94%, 17 out of 18 amino acids), and for 39% the most common codon in lincRNAs is different from the one in mRNA (Figure S1E). Moreover, the rarest mRNA codon for an amino acid is used more often by lincRNA putative ORFs (72%, 13 out of 18), but lincRNAs use a different rarest codon for only 17% of amino acids.

LincRNA putative codons are less adapted to tRNA abundances than mRNA coding regions

Next, we investigated how the codon bias between mRNA and lincRNA sequences relates to tRNA abundances. We first used the relative number of tRNA genes coding for the same tRNA anticodon as an estimate for the relative tRNA abundances in human and mouse (see Methods). This measure was previously shown to correlate well with tRNA abundances (Tuller et al., 2010b). We used previously

determined wobble-base pairing and tRNA editing efficiencies (dos Reis et al., 2004) to calculate effective tRNA anticodon abundances for codons without genetically encoded, complementary tRNA anticodon genes (Figure S2; see Methods). We compared codon frequencies with relative tRNA abundances and found that their correlation is stronger and more significant for mRNA as compared to lincRNA (Spearman correlation $\rho=0.57$ with $p<2e-6$ for mRNA versus $\rho=0.32$ with $p<2e-2$ for lincRNA; Figure 2A).

Previously, the tRNA adaptation index (tAI) was defined as a measure for the adaptation of codon usage to tRNA abundance (dos Reis et al., 2003). (We will use the word adaptation in this sense throughout this paper.) tAI ranges from 0 (no codon adaptation) to 1 (perfect codon adaptation: only codons decoded by the most abundant tRNA anticodon type are used). We computed tAIs for each mRNA coding sequence and lincRNA longest putative ORF (see Methods). We found that mRNA coding-regions were significantly ($p<1e-300$, Wilcoxon ranksum test) better adapted to tRNA abundances than lincRNA putative ORFs (median tAI 0.328 for mRNA versus 0.315 for lincRNA; Figure 2B).

To better understand the extent of adaptation to tRNA abundance, we related tAIs for mRNA and lincRNA sequences to tAIs of three different types of control sequences (see Methods; Figure 2C), each assuming different constraints on the nucleotide sequence: 1) shuffled control sequences, which preserve the nucleotide frequencies, account for constraints in nucleotide content; 2) frame-shifted control sequences, which preserve the nucleotide sequence but use ORFs that are shifted by one nucleotide upstream and downstream, account for sequence constraints to preserve functional sequence elements (e.g. DNA- or RNA-binding motifs or RNA secondary structure); and 3) “random codon” control sequences, in which each codon was replaced by a random codon coding for the same amino acid, account for constraints in amino acid identity.

To quantify the extent and direction of sequence adaptation, we calculated, for each transcript, the difference in tAIs (ΔtAI) between the original and each type of control sequence (Figure 2C). On average, ΔtAIs for both mRNAs and lincRNAs, were significantly greater than 0 (p values $<1e-300$, one-sample t-test) for all types of control sequences, and ΔtAIs for mRNAs were significantly more positive than those for lincRNAs (p values $< 1e-245$, Wilcoxon ranksum test), suggesting that, as expected, mRNA codons are better adapted to tRNA abundance than lincRNA putative codons. Interestingly, the distributions of ΔtAIs were broader for lincRNAs, in particular compared to shuffled (standard deviation 0.017 for mRNA and 0.027 for lincRNA) and frame-shifted (standard deviation 0.024 for mRNA and 0.042 for lincRNA) control sequences. Furthermore, a considerable number of lincRNA transcripts had negative ΔtAIs (284, 4076 and 2802 for shuffled, frame-shifted and random codon control sequences, respectively; Figure 2C), and the proportion of transcripts with negative ΔtAI was significantly greater for lincRNAs than for mRNAs (Fisher exact test p values $< 1e-72$ in comparison with all three types of control sequences; Figure 2C). In conclusion, lincRNAs exhibit a larger variability in their tRNA adaptation, which is likely related to a larger functional heterogeneity among lincRNAs than among mRNAs, with a significantly larger fraction of lincRNAs showing lower adaptation to tRNA abundances than expected. Similar observations were obtained for mouse (Figure S3).

Putative codons in cytoplasmic expressed lincRNAs result in lower tAIs than those in non-expressed lincRNAs

In the previous section, we found that lincRNA putative codon usage is less adapted to tRNA abundance than mRNA codon usage, and identified a subset of lincRNAs showing lower adaptation to tRNA abundance than expected. We hypothesized that, if lincRNA sequences had evolved to reduce the likelihood of being translated, this would be more pronounced for highly expressed, cytoplasmic

lincRNAs, since these lincRNAs are more frequently exposed to ribosomes and tRNAs and might therefore experience a stronger evolutionary pressure for lowering their tAIs. To test this hypothesis, we examined the codon usage of expressed transcripts in three ENCODE cell lines, GM12878, HeLa-S3, and K562, for which comprehensive experimental data are available to quantify mature mRNA and lincRNA expression levels in cytoplasm and whole cells, as well as relative tRNA expression levels.

Thus far, we have used tRNA gene counts to estimate relative tRNA abundances. However this metrics is not cell type-specific. Since here, we wanted to calculate tAIs for expressed lincRNAs and mRNAs in different cell lines, we evaluated different approaches for quantifying cell type-specific tRNA abundances (in particular based on H3K27ac ChIP-seq and smallRNA-seq; see Methods), and concluded that smallRNA-seq allows the best quantification of tRNA abundances to use for ENCODE cell lines.

Overall, tAIs calculated using smallRNA-seq-based tRNA quantification confirmed that mRNAs had on average higher tAIs than lincRNAs, for all three cell lines (Figure 3B, horizontal blue and red lines for mRNAs and lincRNAs, respectively). We also found that tAIs of expressed mRNAs were significantly higher than those of non-expressed mRNAs in a cell line (Figure 3A), as observed previously (Waldman et al., 2010). In contrast, the tAIs of expressed lincRNAs were significantly lower than tAIs of non-expressed lincRNAs for all cell lines (Figure 3A). This suggests different evolutionary forces acting on codon frequencies in expressed mRNAs and expressed lincRNAs, causing expressed mRNAs to favor codons recognized by more abundant tRNAs and expressed lincRNAs to prefer codons corresponding to lower abundance tRNAs.

Interestingly, the fraction of lincRNA transcripts with negative ΔtAI was larger, when calculating cell-type specific ΔtAIs, for shuffled (>4-fold increase) and random codon control sequences (>1.75-fold increase; Figure S5A). Furthermore, lincRNA transcripts with negative ΔtAI in comparison with shuffled control sequence were significantly enriched among expressed, as opposed to non-expressed, lincRNAs in a cell line (Figure S5B). In contrast, mRNA transcripts with negative ΔtAI in comparison with shuffled and frame-shifted control sequences were significantly enriched among non-expressed mRNAs in all cell lines (Figure S5B; shown for shuffled controls). This further strengthens the hypothesis that putative codons in expressed lincRNAs are biased towards codons decoded by rare tRNAs.

To investigate the relationship between tAI and RNA expression in more detail, we ranked the expressed transcripts by their expression level and then calculated the average tAI for an increasing fraction of the most highly expressed transcripts. We observed that for mRNAs the average tAI decayed steadily with the size of the fraction of transcripts (Figure 3B, blue circles). This was also indicated by a significant positive correlation between tAIs and mRNA expression levels (Spearman correlation coefficients 0.13, 0.08, 0.10 with p values < 7e-106, 3e-40, 2e-63 for GM12878, HeLa-S3 and K562, respectively). In contrast, the average tAIs for expressed lincRNAs were clearly below the average tAI for all lincRNAs in most cases, but there was no clear trend towards lower tAIs for more highly expressed lincRNAs (Figure 3B, red circles), and no significant correlation between tAIs and lincRNA expression levels (Spearman correlation coefficients -0.05, 0.01, -0.02 with p values > 1e-2, 5e-1, 2e-1 for GM12878, HeLa-S3 and K562, respectively). One reason for this could be that lincRNAs, although expressed in a cell, are more frequently located in the cell nucleus, where translation does not takes place, and thus, codon adaptation would not be required. To test this possibility, we restricted our analysis to lincRNAs that were expressed in the cytoplasm. Indeed, when we ranked lincRNAs by their cytoplasmic expression level, the decrease in average tAIs with the cytoplasmic expression became more apparent (Figure 3B, magenta circles). This was also indicated by a significant negative correlation between tAIs and cytoplasmic expression levels of lincRNAs for all

cell lines (Spearman correlation coefficients -0.12, -0.06, -0.11 with p values < 4e-9, 2e-2, 2e-6 for GM12878, HeLa-S3 and K562, respectively).

Together, these results support our hypothesis that putative codons in abundant cytosolic lincRNAs evolved to curb translation, in contrast to lincRNAs that are not expressed in the cytoplasm.

tRNA adaptation of putative codons in lincRNAs is able to counteract ribosome binding, with codons close to start codons being likely more important

To verify how tRNA adaptation relates to translation, and if low tAIs of putative ORFs in lincRNAs result in less ribosome binding, we analysed translation efficiencies (TEs) and ribosome binding using ribosome protected RNA fragment sequencing (Ribo-seq) data. We chose to focus on GM12878 because Ribo-seq data were available in replicates for this cell line (Cenik et al., 2015). TEs were computed for each gene as the ratio of Ribo-seq reads to RNA-seq reads covering annotated coding regions in mRNAs and identified longest putative ORFs in lincRNAs (see Methods). Similar to previous reports (Dana and Tuller, 2014), we found a significant correlation between TE and tAI for mRNAs (r=0.11, p<1e-30; Figure 4A). In the case of lincRNAs, we expected the influence of tAI on TE to be detectable only for those lincRNAs that were predominantly found in the cytoplasmic. We considered lincRNAs to be cytoplasmic if their relative cytoplasmic abundance (ratio of cytoplasmic to total expression) was higher than the median relative cytoplasmic abundance for mRNAs. Indeed, we found a significant correlation between TE and tAI for cytosolic lincRNAs (r=0.16, p=3.7e-2; Figure 4B), but not for the remaining lincRNAs (r=0.01, p=0.91), supporting our hypothesis that codon optimization can regulate ribosome association for lincRNAs and that this mechanism would predominantly affect cytosolic lincRNAs.

According to previous studies (Ji et al., 2015), a fraction of lincRNAs may be missannotated and actually coding for small proteins or peptides. For these lincRNAs, we expected their codon frequencies to be more adapted to the tRNA abundance than those of bona-fide lincRNAs. To examine this, we specifically compared tAIs between cytoplasmic lincRNAs with no ribosomes bound (Ribo-seq reads=0) and cytoplasmic lincRNAs, whose longest putative ORFs overlapped with experimentally supported peptide coding sequences (see Methods) and, additionally, were bound by ribosomes in GM12878 (Ribo-seq reads>0). Indeed, we observed a significant difference in tAIs between peptide-encoding lincRNAs and cytoplasmic lincRNAs with no experimental evidence for ribosome binding (p=2.2e-5; Figure 4C).

Previously, an unusual codon usage immediately downstream of mRNA start codons was observed and connected with an efficient initiation of translation (Bentele et al., 2013; Tuller et al., 2010b). To investigate codon position dependent effects in tRNA adaptation, we calculated local-tAIs for each codon position downstream from start codons, as done previously (Tuller et al., 2010b). Overall, there was a strong difference in local-tAIs between likely misannotated and bona-fide lincRNAs at almost all codon positions within a window of 40 codons from start codons (Figure S6). Furthermore, local-tAIs of likely peptide encoding transcripts were more similar to those of mRNAs. In particular, the first 10 codons after the start codon of peptide encoding transcripts were well adapted to tRNA abundances, and tAIs were very similar to those of mRNAs, for local-tAIs (Figure S6) and for tAIs calculated considering the first 10 codons of each ORF only (Figure 4D).

These results suggest that tRNA adaptation of putative codons in lincRNAs is able to counteract ribosome binding, and that it is likely more important for codons at the beginning of putative ORFs, in agreement with previous studies highlighting the importance of translation initiation for mRNA translation efficiency (Nakahigashi et al., 2014).

Discussion

LincRNAs are similar to mRNAs with regard to their transcript length and biogenesis, but, in contrast to mRNAs, lincRNAs do not code for proteins and many have been shown to have regulatory functions (Ulitsky and Bartel, 2013). Whereas association between cytoplasmic lincRNAs and ribosomes has been reported, such interactions rarely give rise to detectable peptides, suggesting the presence of a mechanism counteracting the translation of lincRNA putative ORFs. Here, we carried out a comprehensive analysis of lincRNAs ORF sequences providing evidence that many of them are biased towards codons recognized by less abundant tRNAs. We propose that this codon bias contributes to preventing unwanted translation of putative ORFs in cytosolic lincRNAs.

By definition lincRNAs lack an apparent open reading frame and coding potential. Thus, it may not be surprising that putative ORFs within lncRNAs present codon compositions different from mRNA coding regions. However, the observed conservation of codon frequencies in lincRNAs between human and mouse to an extent comparable with the conservation observed for codon frequencies within mRNA (Figure 1B) indicates that lincRNA codon composition, while distinct from that of mRNAs, is not random.

Indeed, the comparison with randomized control sequences revealed a subset of lincRNAs composed of codons that were less adapted to tRNA abundances than expected, and the proportion of such lincRNAs was substantially larger than that in mRNAs (Figure 2). Importantly, these lincRNAs were enriched among cytosolic lincRNAs (Figure S5), and their codon composition was less adapted to tRNA abundances than those of non-expressed lincRNAs (Figure 3). It is important to note that the majority of lincRNAs are still better adapted to tRNA abundances than the control sequences we examined. This might be due to a coding capacity of some lincRNAs earlier during evolution (Hezroni et al., 2017). In general, lincRNAs are thought to be evolutionary younger than mRNAs and thus may have had less time to optimize their putative codons. Furthermore, some lincRNAs may have been mis-classified and actually code for peptides (Ma et al., 2014; Ruiz-Orera et al., 2014; Yeasmin et al., 2018).

Finally, relating the codon bias to translation efficiencies (TEs) we observed a significant correlation between tAI and TE for mRNA genes as well as for cytosolic lincRNA genes (Figures 4A and B). Closer examination of the set of cytosolic lincRNAs revealed a subset that is likely being misannotated as non-coding and may actually code for peptides. This subset showed markedly different tAIs compared to those found for bona fide lincRNAs that are non-coding and do not bind to ribosomes (Figures 4C, D and S6), providing further evidence for codon modulation to counteract translation of true lincRNAs. This suggests the tAI as a potential criterion for predicting likely misannotated lincRNAs that actually encode peptides, in addition to previous Ribo-seq-based methods (Calviello et al., 2016; Guttman et al., 2013).

The fact that the correlation between TE and tAI, while significant, was not very strong, might have several explanations. One reason could be that the evolutionary impact on codon bias in human is expected to be weaker than in species with larger population sizes or shorter generation times (Subramanian, 2008). Another explanation could be that there are additional factors influencing ribosome binding and translation efficiency, such as the RNA secondary structure or codon order, both of which have been reported previously for mRNA (Tuller et al., 2010a, 2010b). Finally, it could also be a consequence of the current inability of Ribo-seq methods to distinguish between different transcript isoforms expressed from the same gene locus (Figures 4A and B). This averaging over transcript isoforms likely causes a dilution of the actual correlation strength between TE and tAI. Indeed, the difference in tAIs was much more significant, when comparing between lincRNAs with experimental

evidence for being translated into peptides and lincRNAs without ribosome-binding on the transcript level (Figure 4C).

In conclusion, in this study we provided a comprehensive analysis of putative codons in lincRNA ORFs. Our results suggest that these codons are a substrate of evolutionary forces counteracting unwanted ribosomal binding and translation. The resulting sequence signatures may help in distinguishing bona-fide lincRNAs with regulatory roles in the cytoplasm from those transcripts coding for peptides, yet more work will be needed to distill such signals in the context of other potential constraints on lincRNA sequences related to their regulatory function, such as structure or binding motifs. Another interesting aspect is that tRNA concentrations can vary across cell types, imposing potentially different constraints on the evolution of codon frequencies to either curb ribosome binding of true lincRNAs, or promote it for peptide-coding transcripts. A further future direction could be to study natural genetic variations or targeted mutations in lincRNA sequences to establish an impact on ribosome binding and translation. Finally, it has not escaped our notice that similar signatures in putative codons may exist in ORFs of other classes of cytosolic non-coding RNA.

Methods

Identification of putative ORFs in lincRNAs

All lincRNA transcripts annotated in GENCODE v19 (for human) and vM16 (for mouse) (Frankish et al., 2019) that did not overlap any protein coding genes on the same strand were considered. Putative ORFs in lincRNAs starting with a canonical start codon (AUG) and ending at the first in-frame stop codon (UAG, UAA, UGA) were identified using a custom python script. The longest putative ORF in a lincRNA transcript was considered for further analysis, if it was longer than 30 nucleotides.

We excluded mitochondrially encoded transcripts, as these are translated using mitochondrially encoded tRNAs.

Estimation of relative tRNA abundances based on tRNA gene counts

The numbers of tRNA genes coding for the same tRNA anticodon type were counted based on genomic annotations of tRNAs from GENCODE v19 (for human) and vM16 (for mouse) (www.gencodegenes.org) (Frankish et al., 2019). For tRNA anticodon types that are not encoded in the human or mouse genomes, effective tRNA abundances were estimated using previously determined weights for tRNA editing and wobble-base pairing efficiencies (dos Reis et al., 2004). In particular, the weights, w, are w(G:U)=0.41, w(I:C)=0.28, w(I:A)=0.999, and w(U:G)=0.68, where the first letter denotes the first nucleotide in a tRNA anticodon nucleotide triplet and the second letter the third nucleotide of a codon.

Estimation of cell-type specific relative tRNA abundances

Due to the repetitive nature of tRNAs, their strong secondary structure, and the high frequency of post-transcriptional tRNA modifications, high-throughput quantification of tRNA expression is still challenging. Recently, 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 they have only been applied to one cell line, HEK293. On the other hand, smallRNA-seq and H3K27ac-ChIP-seq data were previously used to quantify tRNA abundances (Ji et

al., 2015; Shi et al., 2018), and these data were generated for the selected ENCODE cell lines. To establish which of the latter two approaches performs better in estimating relative tRNA expression levels, we quantified relative tRNA expression levels based on smallRNA-seq and H3K27ac-ChIP-seq in HEK293 cells, and compared it with those from hydro-tRNA-seq and DM-tRNA-seq. Based on these comparisons in HEK293 cells (Figure S4), we chose the smallRNA-seq-based approach to estimate cell-type specific relative tRNA abundances in our analysis of ENCODE cell lines. Effective tRNA abundances for tRNA anticodons not encoded in the human genome were calculated as described above.

In the following, details of the different approaches for the quantification of tRNA abundances are given:

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

SmallRNA-seq reads (fastq files) for GM12878, HeLa-S3, and K562 cells were downloaded from the ENCODE data portal (www.encodeproject.org) (Consortium and The ENCODE Project Consortium, 2004). In the case of HEK293 cells, smallRNA-seq reads were downloaded from the GEO database (www.ncbi.nlm.nih.gov/geo, accession number GSM1067868) (Kishore et al., 2013). 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 ratios (7% and 15%) were tested, but these did not improve the correlation with tRNA sequencing approaches (hydro-tRNA-seq (Gogakos et al., 2017) and DM-tRNA-seq (Zheng et al., 2015)), or resulted in a smaller fraction of reads mapping to tRNA sequences in sense direction. The number of smallRNA-seq reads mapping to each tRNA anticodon type divided by the total number of mapped reads was taken as an estimate for the relative tRNA abundance.

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

Bedfiles of identified H3K27ac-ChIP-seq peaks in GM12878, HeLa-S3, HEK293, and K562 cells were downloaded from ENCODE (www.encodeproject.org) (Consortium and The ENCODE Project Consortium, 2004). H3K27ac ChIP-seq peaks that overlapped tRNA gene annotations extended by 500 nucleotides up- and downstream were determined using bedtools (Quinlan, 2014). Relative tRNA abundances were estimated by the ratio of the sum of peak enrichment values (log2 fold enrichment H3K27ac-ChIP-seq over background, in column 7 of the bedfiles) for each tRNA anticodon type to the peak enrichment values of all peaks overlapping extended tRNA genes.

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

Hydro-tRNA-seq-based tRNA quantifications were downloaded from the supplement (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 over the two replicates.

tRNA adaptation index (tAI)

As proposed by dos Reis et al. (dos Reis et al., 2004), the tAI of an ORF was calculated as the geometric mean over normalized abundances of tRNAs that are complementary to codons in an 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 abundance of the tRNA anticodon type that is complementary to the codon at position i.

Normalized tRNA abundances were obtained through division by the maximum of all tRNA abundances:

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

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

In Figures 4A and B, in order to compare translation efficiencies (TEs) with tRNA adaptation values on the gene-level, tAIs were calculated per gene by considering the union of codons in the ORFs of all annotated transcript isoforms encoded by a gene.

In Figure S6, local tAIs were calculated per codon position within a window of 40 nucleotides downstream of start codons. In this case, codons were considered at the same position of a set of ORFs. For better visualization, local-tAI values were smoothed by taking the geometric mean of local-tAI values over three consecutive codons (positions i to i+2; Figure S6B) or five consecutive codons (positions i to i+4; Figure S6C).

In Figure 4D, tAIs were calculated per transcript, but by considering only the first 10 codons after start codons.

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

(a) shuffled sequences:

Shuffled sequences were generated by random permutations of the nucleotides in the ORF. This was done 100 times. tAIs were then calculated for the union of codons resulting from all shufflings.

(b) frame-shifted sequences:

tAIs of frame-shifted sequences were calculated for codons in nucleotide sequences shifted by one and two nucleotides downstream and ending two and one nucleotides, respectively, upstream of the stop codon of the original ORF.

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

Each codon in an ORF was replaced by a random codon coding for the same amino acid. This was done 100 times. tAIs were calculated for the union of codons resulting from all randomizations.

Quantification of RNA expression based on ENCODE data

Transcript quantifications in cytosol and total cells based on polyA-selected RNA-seq were downloaded from ENCODE (www.encodeproject.org) (Consortium and The ENCODE Project Consortium, 2004) for GM12878, HeLa-S3, and K562 cells. TPM (transcripts per million) values were

used to quantify the relative expression of transcripts within a cell line. Transcripts with TPM>0.1 were considered as expressed.

Quantification and analysis of translational efficiencies in GM12878

Ribo-seq data (2 replicates) and polyA-selected RNA-seq data (3 replicates) for GM12878 cells were downloaded from GEO (www.ncbi.nlm.nih.gov/geo; accession number GSE65912) (Cenik et al., 2015). 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 for Ribo-seq and 35 to 60 for RNA-seq) and minimum quality score of 30 in at least 90% of read bases. Reads were further 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), or if mapping to coding regions or longest putative ORFs of two or more gene loci annotated in the human transcript database (ENSEMBL v91) using bowtie2 (v2.3.0, -L 15 -k 20). Remaining reads were summarized at gene level using an in-house script.

Translation efficiencies (TEs) were calculated for each gene as the log2 ratio of Ribo-seq to RNA-seq read counts, as proposed before (Ingolia et al., 2009), using DESeq2 (Love et al.).

Due to size selection of the ribosome-protected RNA fragments in the experimental Ribo-seq method, only RNA fragments that were covered by a single (isolated) ribosome will be sequenced, and longer fragments that were protected by multiple adjacent ribosomes will not be captured. This has the consequence that transcripts that are translated intensively (covered with many adjacent ribosomes) will end up with a low number of Ribo-seq reads (only those from single ribosomes), resulting in very low TE values. To exclude these cases, we restricted our analysis to genes with a log2 TE value of larger than -6.

We also excluded histone mRNA genes, which represented outliers with very high TEs. These high TE values are likely caused by the inability of the quantification of the total expression level of histone mRNAs based on polyA-selected RNA-seq. Since histone mRNAs are not usually polyadenylated, polyA-selected RNA-seq does not capture the true total expression of histone mRNAs.

Analysis of lincRNAs that likely code for peptides

We downloaded the genomic coordinates of experimentally supported peptides (<100 amino acids) in lincRNAs from the SmProt database (http://bioinfo.ibp.ac.cn/SmProt/) (Hao et al., 2018). We combined peptide-coding regions in annotated human lincRNAs that were supported by various experimental data, in particular from mass spectrometry, literature mining, ribosomal profiling, and known databases, as indicated in the SmProt database. We then overlapped these regions with longest putative lincRNA ORFs (requiring a minimum overlap of 10 codons) to obtain a set of experimentally supported peptide-coding lincRNA ORFs. In total, experimentally supported peptide-coding lincRNA ORFs were found in 222 lincRNA genes.

Acknowledgements

We would like to thank the ENCODE consortium for generating the data and making them publicly available.

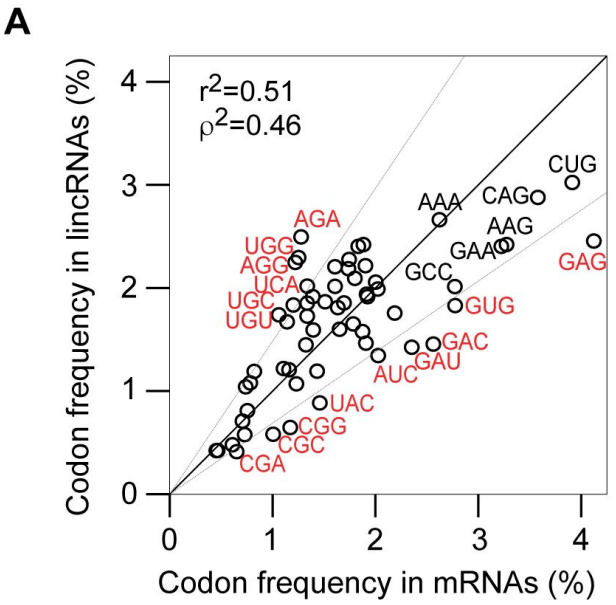
Author Contributions

AB, ACM and SB designed the project. AB carried out the computational analysis and prepared the results. RD quantified translation efficiencies based on Ribo-seq data. AB, ACM, and SB discussed results and wrote the paper.

Declaration of Interests

The authors declare no competing interests.

Figure 1



B

r^2	human mRNA	mouse lincRNA
mouse mRNA	0.99	0.46
human lincRNA	0.51	0.94

ρ^2	human mRNA	mouse lincRNA
mouse mRNA	0.99	0.41
human lincRNA	0.46	0.92

Figure 1: Distinct codon usage in lincRNAs.

(A) Comparison of codon frequencies (excluding start and stop codons) in mRNAs and lincRNA longest putative ORFs. Squared Pearson (r^2) and Spearman (ρ^2) correlation coefficients are indicated. Codons with an absolute z-score >1.2 (gray lines) of their log2 frequency ratio are labeled in red, and codons with high frequencies (>2%) in both RNA classes are labeled in black. (B) Squared correlation coefficients (Pearson in top panel and Spearman in bottom panel) for comparing codon frequencies (excluding start and stop codons) between mouse and human within the same RNA class (diagonal elements in each panel), and between mRNA and lincRNA in either human or mouse (off-diagonal elements).

Figure 2

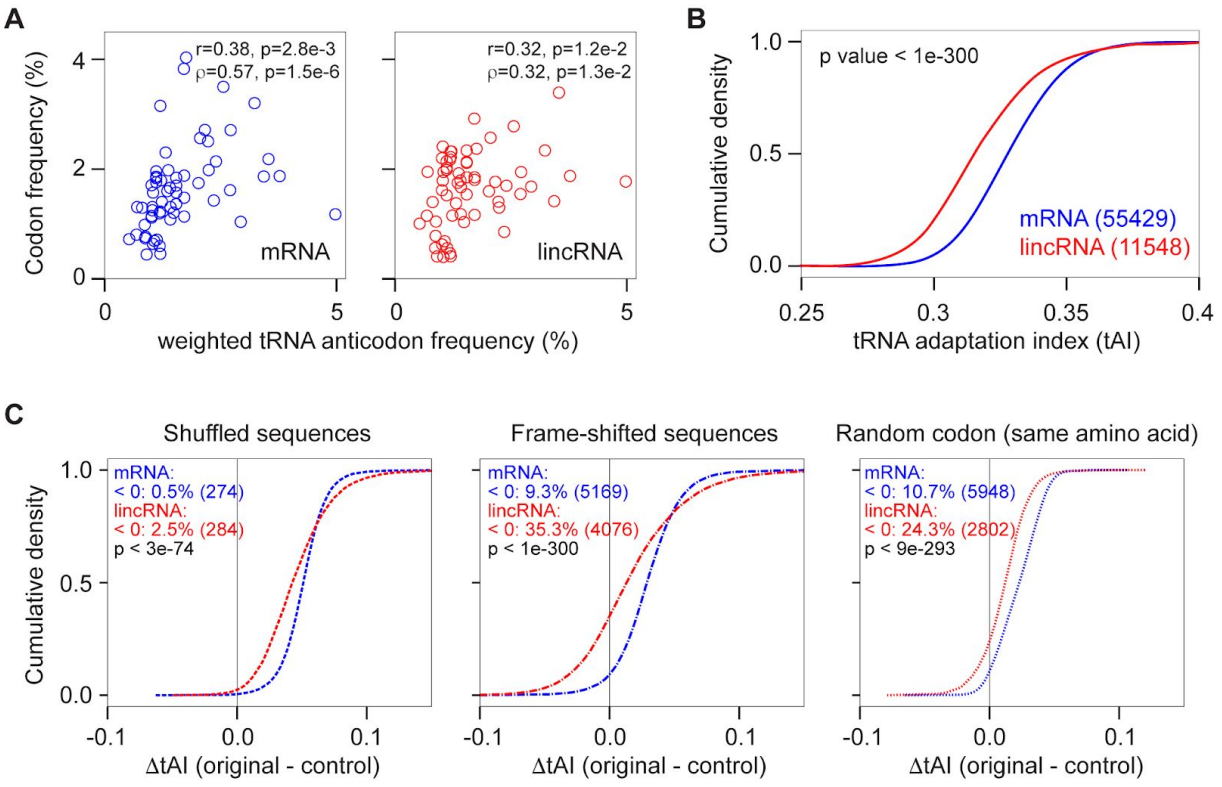


Figure 2: LincRNA codons are less adapted to relative tRNA abundances than mRNA codon usage.

(A) Scatter plots of relative tRNA anticodon frequencies and codon frequencies in mRNA coding regions (left panel) and lincRNA longest putative ORFs (right panel). Pearson (r) and Spearman (ρ) correlation coefficients with p values are indicated. (B) Cumulative density of tRNA adaptation indexes (tAIs) for mRNA (blue) and lincRNA sequences (red). The total number of transcripts for each class of RNA is indicated in parentheses. P value was calculated using a Wilcoxon ranksum test to compare tAIs of mRNAs with those of lincRNAs. (C) Cumulative densities of ΔtAIs calculated for each transcript (mRNA in blue, lincRNA in red) as the difference between tAIs of original and control sequence. Three types of control sequences are shown as indicated on the top of each panel. For each ΔtAI distribution the percentage and number of transcripts with negative ΔtAIs (ΔtAI < 0) is indicated. P values were calculated using Fisher's exact test to compare the proportions of transcripts with negative ΔtAI between mRNAs and lincRNAs.

Figure 3

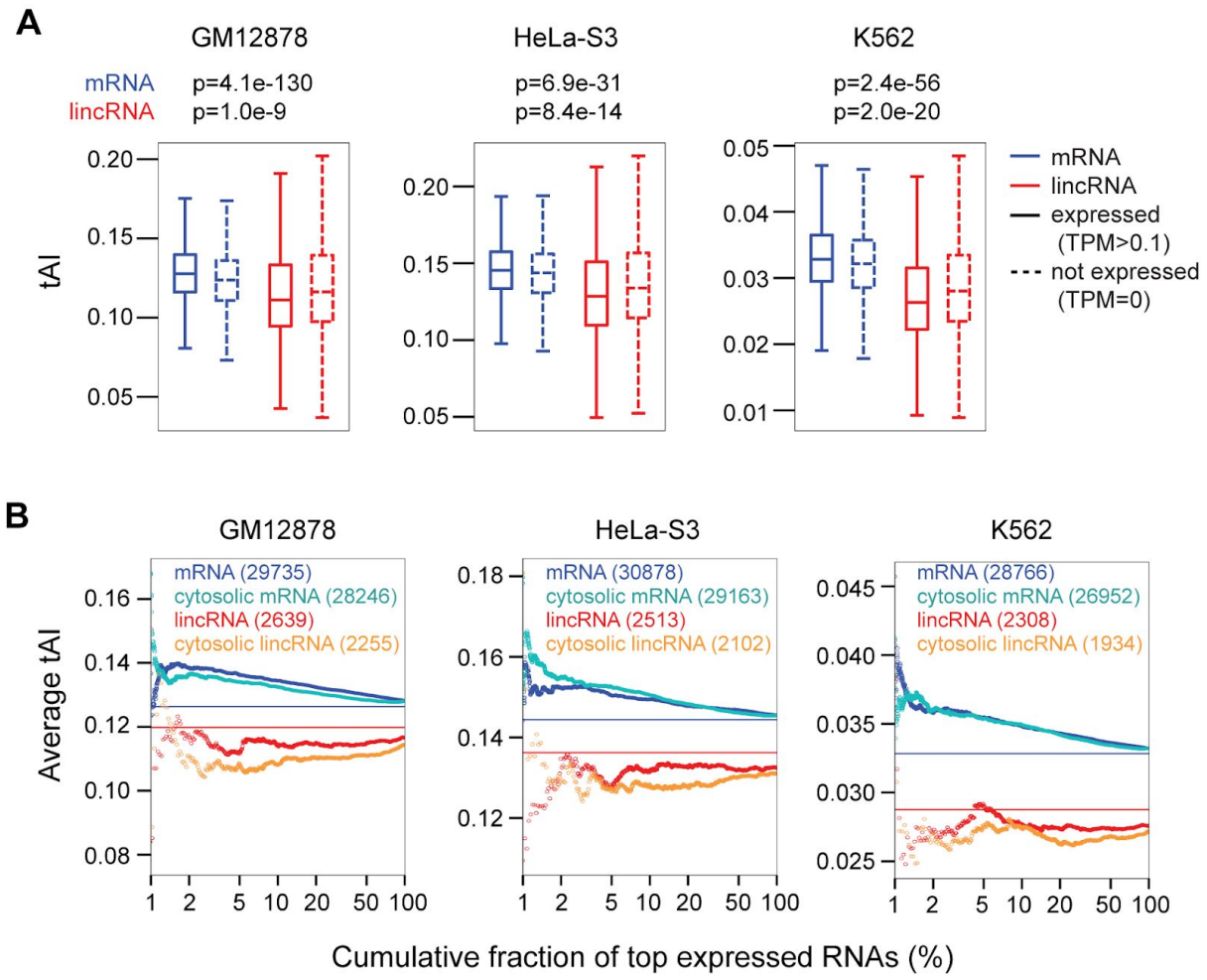


Figure 3: Codon usage of highly expressed, cytoplasmic lincRNAs is less adapted to tRNA abundances.

(A) Boxplots of tAIs for expressed (solid boxes) mRNAs (blue) and lincRNAs (red), and non-expressed sequences (dashed boxes) in three cell lines, GM12878, HeLa-S3, and K562, as indicated on top of each panel. P values to compare tAIs of expressed with those of non-expressed RNA sequences are indicated above each box and are calculated using a Wilcoxon ranksum test. Notably, the range of tAIs differs for each cell line, as it depends on the cell type-specific distribution of relative tRNA abundances. (B) Average tAIs (y-axis) for an increasing fraction of top expressed RNA sequences, as indicated on the x-axis. The RNA expression level was either measured in whole cells (mRNA blue and lincRNA red) or in the cytoplasmic compartment of cells (mRNA light blue and lincRNA orange). The total numbers of expressed transcripts are indicated in parentheses in the legend inside each panel. Horizontal lines (mRNA blue and lincRNA red) indicate the average tAIs for all transcripts (expressed and not expressed) calculated using cell type-specific tRNA abundance estimates.

Figure 4

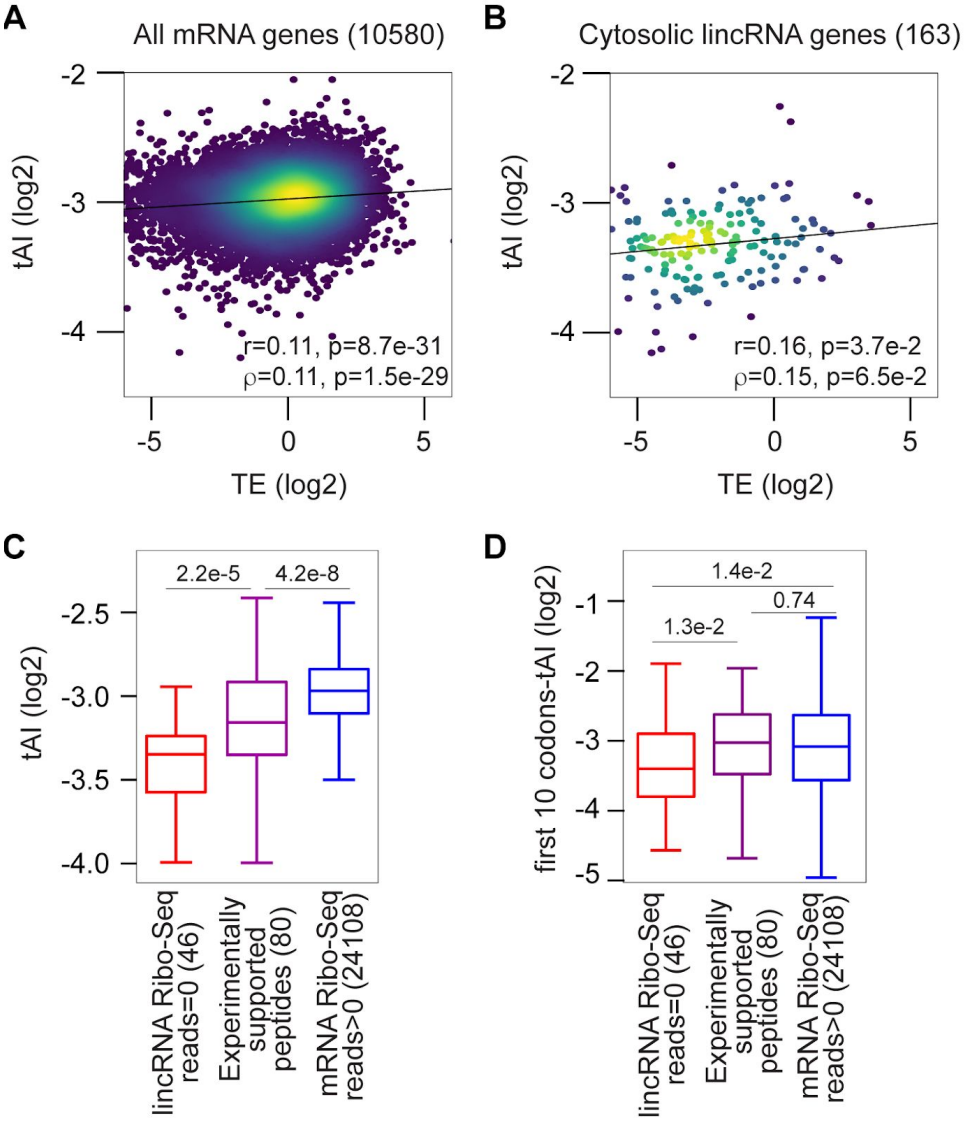
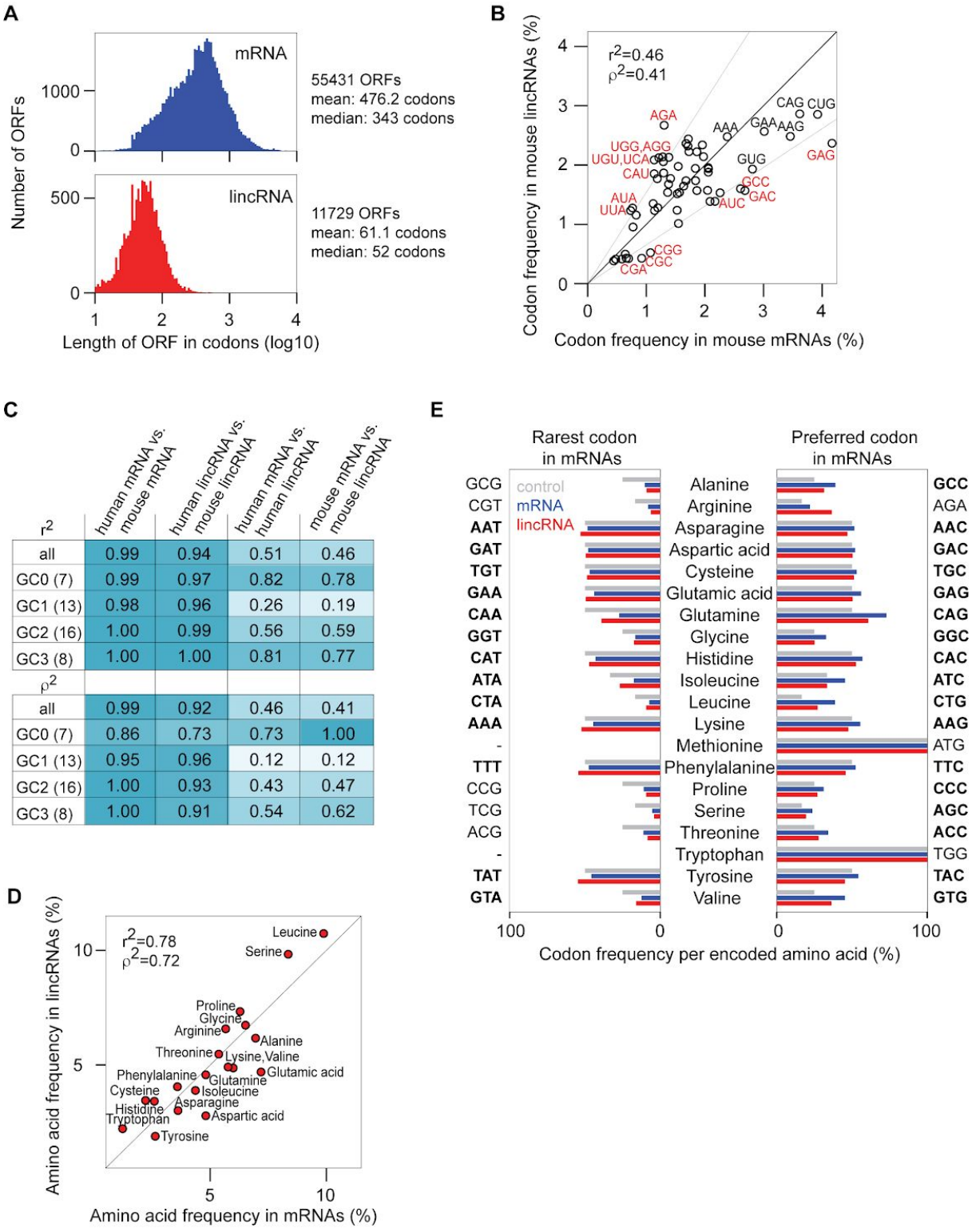


Figure 4: Lower tAIs correspond to lincRNAs with less or no ribosome-binding.

(A, B) Scatter plots of translation efficiencies (TEs) and tAI values for expressed mRNA genes (A), and for cytosolic lincRNA genes (B). The gene density is color-coded from low (dark blue) to high density (yellow). Pearson (r) and Spearman (ρ) correlation coefficients with p values are indicated. The total number of quantifiable genes is indicated in parenthesis above each panel. (C) Boxplots with tAIs for cytosolic lincRNA transcripts without ribosome-binding (Ribo-seq reads=0; red), for cytosolic lincRNA transcripts with bound ribosomes (Ribo-seq reads>0) that overlapped experimentally supported peptide-coding regions (purple), and for mRNA transcripts (blue). The total numbers of transcripts in each group are indicated in parentheses below each box. (D) Boxplots of tAIs calculated for the first 10 codons downstream of start codons for the same three groups of transcripts described in (C).

Figure S1



C

r^2	human mRNA vs. mouse mRNA	human lincRNA vs. mouse lincRNA	human mRNA vs. human lincRNA	mouse mRNA vs. mouse lincRNA
all	0.99	0.94	0.51	0.46
GC0 (7)	0.99	0.97	0.82	0.78
GC1 (13)	0.98	0.96	0.26	0.19
GC2 (16)	1.00	0.99	0.56	0.59
GC3 (8)	1.00	1.00	0.81	0.77
ρ^2				
all	0.99	0.92	0.46	0.41
GC0 (7)	0.86	0.73	0.73	1.00
GC1 (13)	0.95	0.96	0.12	0.12
GC2 (16)	1.00	0.93	0.43	0.47
GC3 (8)	1.00	0.91	0.54	0.62

Figure S1: Comparison of codon usages in mRNA and lincRNA.

(A) Histograms of the lengths of mRNA coding regions (top) and lincRNA longest putative ORFs (bottom). (B) Same as Figure 1A, but for mouse. Comparison of codon frequencies (excluding start and stop codons) in mRNAs and lincRNA longest putative ORFs. Squared Pearson (r^2) and Spearman (ρ^2) correlation coefficients are indicated. Codons with a z-score >1.2 of their log2 frequency ratio are labeled in red, and codons with high frequencies in both RNA classes are labeled in black. (C) Squared Pearson (r^2, top table) and Spearman (ρ^2, bottom table) correlation coefficients for comparing codon frequencies between mouse and human within the same RNA class (first two columns) and between mRNA and lincRNA in either human or mouse (last two columns). The first row indicates the correlation coefficients for all codons (excluding start and stop codons), and the following rows indicate correlation coefficients for codons stratified by their GC content: 7 codons with no G or C nucleotides (GC0), 13 codons with one G or C nucleotide (GC1), 16 codons with two G or C nucleotides (GC2), 8 codons with only G or C nucleotides (GC3). (D) Comparison of amino acid frequencies encoded by mRNA coding regions and lincRNA longest putative ORFs. (E) Frequencies of the rarest (left panel) and most preferred (right panel) codon coding for each amino acid in mRNA, for mRNA (blue) and lincRNA (red) and for a uniform codon usage as a control (gray). Amino acids are indicated in between panels and the respective codons are indicated at the side of each panel (rarest codons left and most preferred codons right). Codons are labeled bold when the rarest codon has a higher frequency in lincRNA (left), or when the preferred codon has a lower frequency in lincRNA (right).

Figure S2

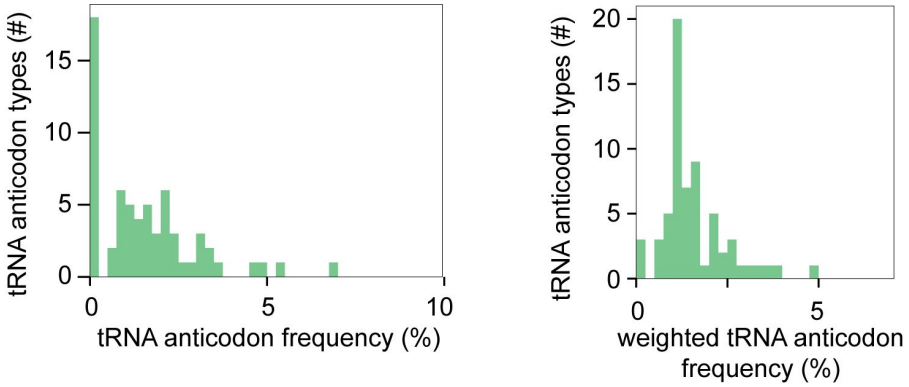


Figure S2: Distributions of tRNA abundances

Histogram of relative tRNA anticodon abundances based on tRNA gene counts (left panel), and for effective tRNA gene counts accounting for wobble base pairing and editing efficiencies (right panel; see Methods).

Figure S3

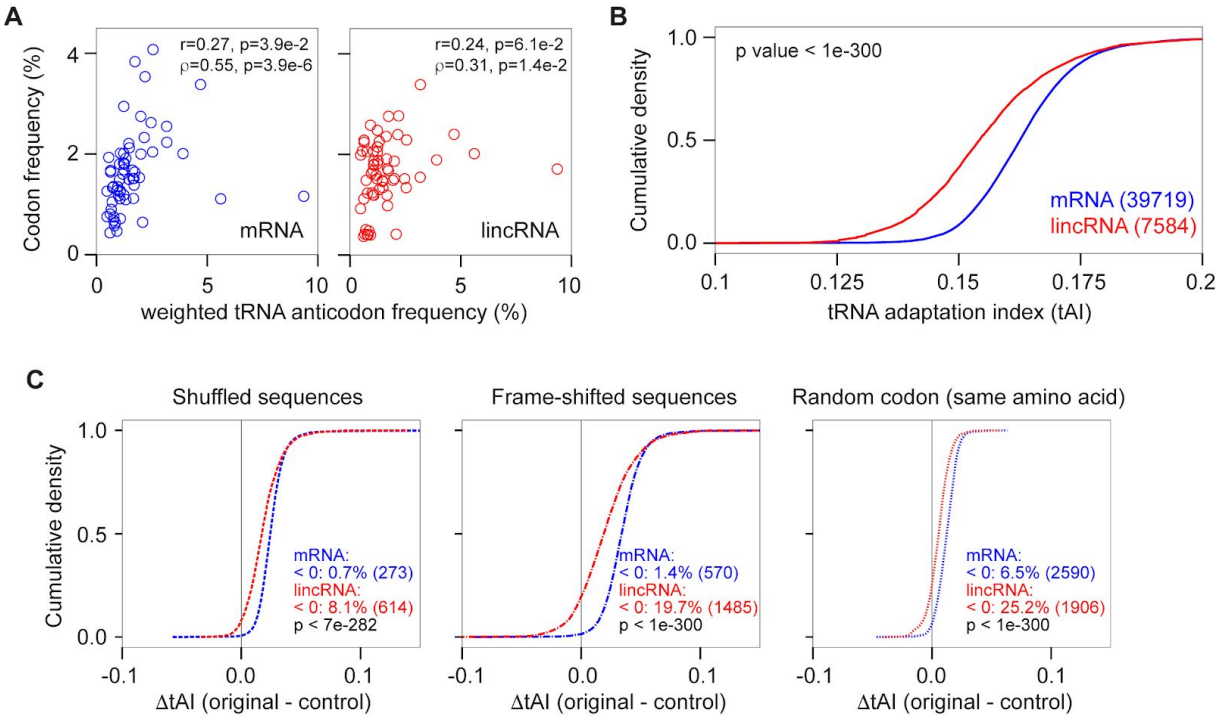


Figure S3: Same as Figure 2, but for mouse.

(A) Scatter plots of relative tRNA anticodon frequencies and codon frequencies in mouse mRNA coding regions (left panel) and in mouse lincRNA longest putative ORFs (right panel). Pearson (r) and Spearman (ρ) correlation coefficients with p values are indicated. (B) Cumulative density of tRNA adaptation indexes (tAIs) for mouse mRNA (blue) and lincRNA sequences (red). The total number of transcripts is indicated in parenthesis. P value was calculated using a Wilcoxon ranksum test to compare tAIs of mRNAs with those of lincRNAs. (C) Cumulative densities of ΔtAIs for each transcript calculated as the difference between tAIs of original and control sequence. Three types of control sequences are shown as indicated on the top of each panel. For each ΔtAI cumulative distribution the percentage and number of transcripts with negative ΔtAIs (< 0) is indicated. P values were calculated using Fisher's exact test to compare the proportions of transcripts with negative ΔtAI between mRNAs and lincRNAs.

Figure S4

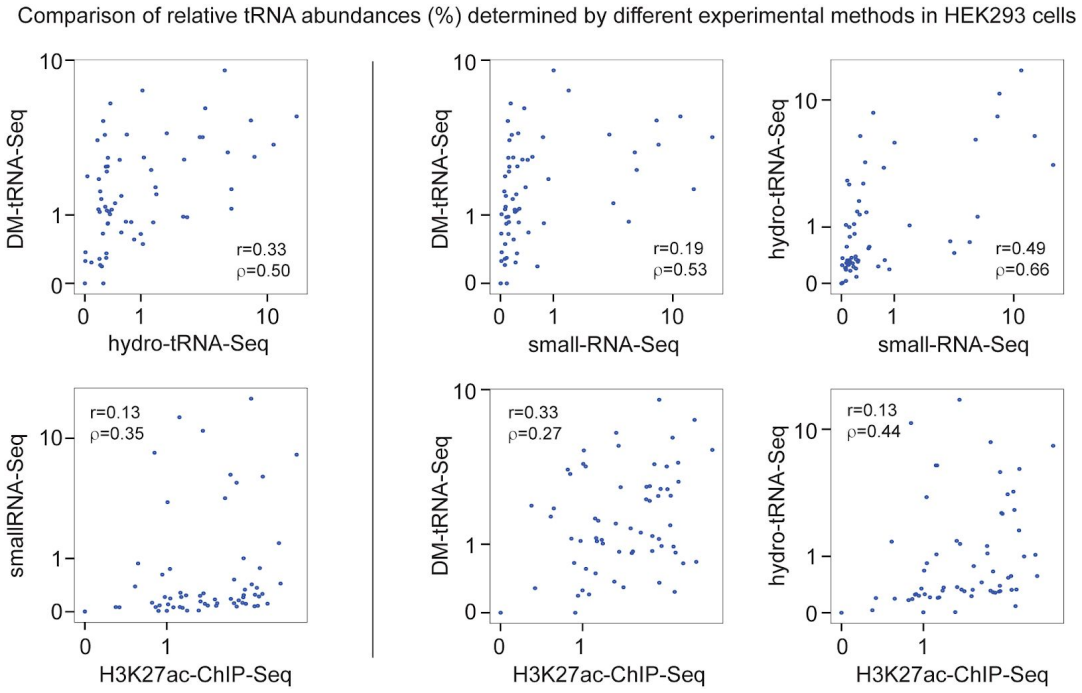


Figure S4: Comparison of relative tRNA expression quantification approaches in HEK293 cells (see Methods).

Scatter plots with Pearson (r) and Spearman (ρ) correlation coefficients for comparisons between different approaches for quantifying relative tRNA abundances (%) in HEK293 cells (see Methods), in particular based on H3K27ac-ChIP-seq data, smallRNA-seq data, hydro-tRNA-seq (Gogakos et al. 2017) and DM-tRNA-seq (Zheng et al. 2015), as indicated by the x- and y-axes labels.

Figure S5

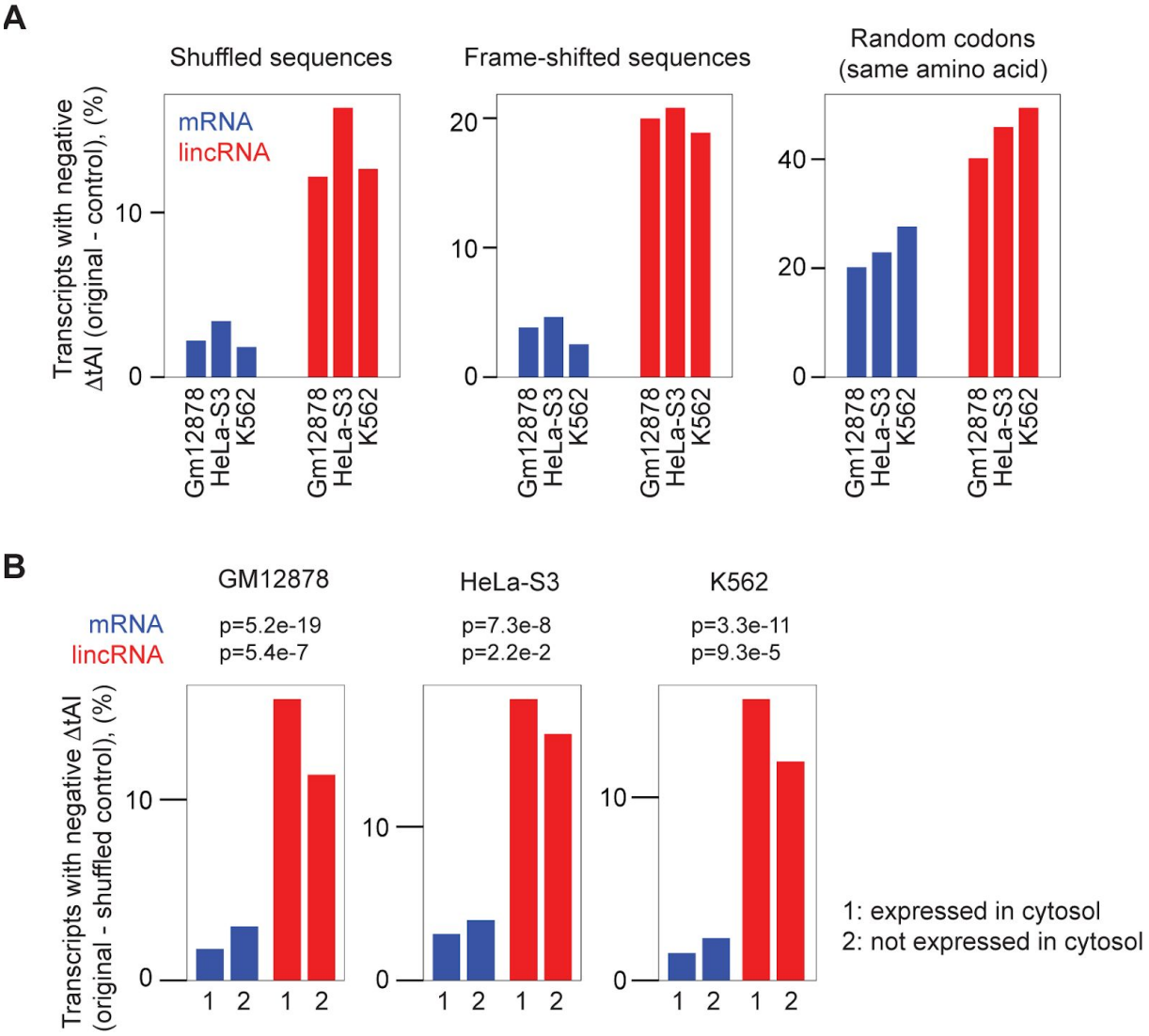


Figure S5: Analysis of cell-type specific ΔtAIs in GM12878, HeLa-S3 and K562 cells.

(A) Fraction of transcripts with negative ΔtAI, calculated using cell-type specific relative tRNA abundances, for the three types of control sequences as indicated on top of each panel. (B) Fraction of transcripts with negative ΔtAI in comparison with shuffled control sequences. Fractions are shown for mRNA (blue bars) and lincRNA (red bars) for three cell lines as indicated on top of each panel. The first bar of each type of RNA shows the percentage of transcripts with negative ΔtAI among expressed transcripts (TPM>0.1), and the second bar among non-expressed transcripts (TPM=0). P values comparing the fraction of transcripts with negative ΔtAI between expressed and non-expressed transcripts are calculated using Fisher's exact test and are indicated above each panel.

Figure S6

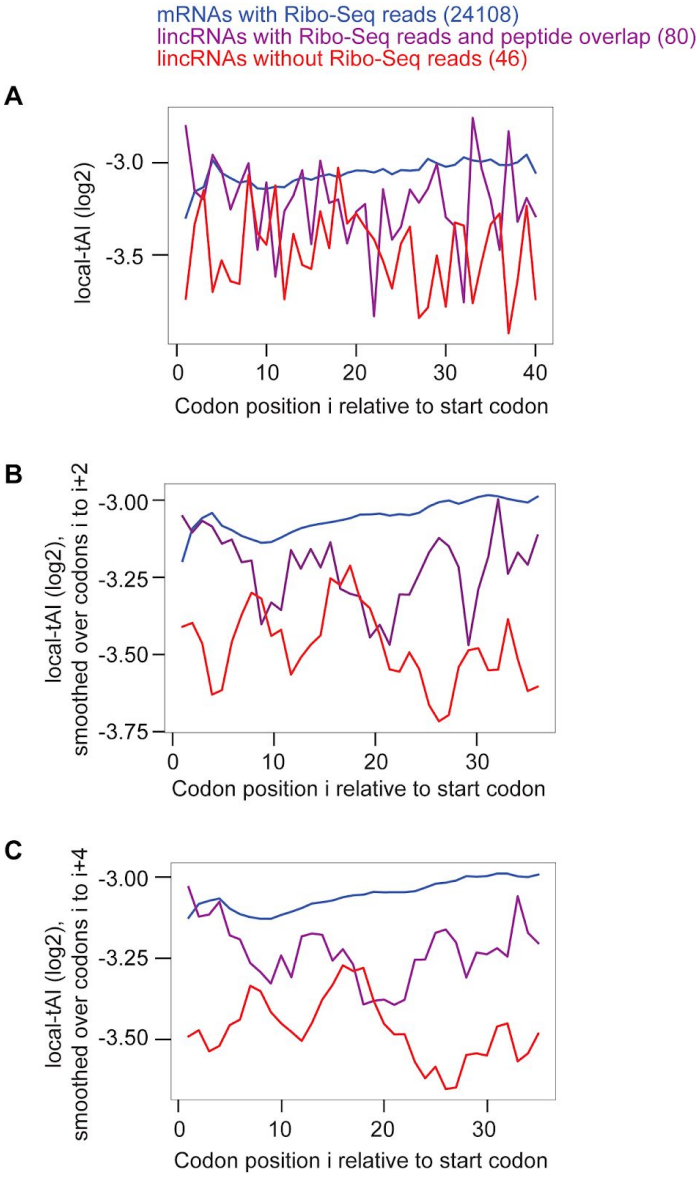


Figure S6: Codon position-specific tAIs for the first 40 codons

Profile of local-tAIs for the first 40 codons downstream of start codons for three groups of transcripts: cytosolic lincRNA transcripts without ribosome-binding (Ribo-seq reads=0; red line), cytosolic lincRNA transcripts with bound ribosomes (Ribo-seq reads>0) that overlapped experimentally supported peptide-coding regions (purple line), and mRNA transcripts (blue line). The total numbers of transcripts in each group are indicated at the top of the figure. The actual local-tAI values at each codon position are shown in (A), while local-tAI values are smoothed by taking the geometric mean over 3 consecutive codons in (B), and over 5 codons in (C).

References

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

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

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M.V., and Komar, A.A. (2016). Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. Mol. Cell *61*, 341–351.

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

Cenik, C., Cenik, E.S., Byeon, G.W., Grubert, F., Candille, S.I., Spacek, D., Alsallakh, B., Tilgner, H., Araya, C.L., Tang, H., et al. (2015). Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. Genome Res. *25*, 1610–1621.

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

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

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

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

Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S., and Lander, E.S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell *154*, 240–251.

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

Hao, Y., Zhang, L., Niu, Y., Cai, T., Luo, J., He, S., Zhang, B., Zhang, D., Qin, Y., Yang, F., et al. (2018). SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. Brief. Bioinform. *19*, 636–643.

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

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

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J.S., Jackson, S.E., Wills, M.R., and Weissman, J.S. (2014). Ribosome profiling reveals pervasive translation outside of

annotated protein-coding genes. Cell Rep. *8*, 1365–1379.

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

Kishore, S., Gruber, A.R., Jedlinski, D.J., Syed, A.P., Jorjani, H., and Zavolan, M. (2013). Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. Genome Biol. *14*, R45.

Komar, A.A., Lesnik, T., and Reiss, C. (1999). Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Letters *462*, 387–391.

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

Love, M.I., Huber, W., and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.

Ma, J., Ward, C.C., Jungreis, I., Slavoff, S.A., Schwaid, A.G., Neveu, J., Budnik, B.A., Kellis, M., and Saghatelian, A. (2014). Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. J. Proteome Res. *13*, 1757–1765.

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

Nakahigashi, K., Takai, Y., Shiwa, Y., Wada, M., Honma, M., Yoshikawa, H., Tomita, M., Kanai, A., and Mori, H. (2014). Effect of codon adaptation on codon-level and gene-level translation efficiency in vivo. BMC Genomics *15*, 1115.

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

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

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

Ruiz-Orera, J., Messeguer, X., Subirana, J.A., and Alba, M.M. (2014). Long non-coding RNAs as a source of new peptides. Elife *3*, e03523.

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

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

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

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

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

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

Acids Res. *38*, 2964–2974.

Walsh, I.M., Bowman, M.A., and Clark, P.L. Synonymous codon substitutions perturb co-translational protein folding and significantly impair cell fitness.

Yeasmin, F., Yada, T., and Akimitsu, N. (2018). Micropeptides Encoded in Transcripts Previously Identified as Long Noncoding RNAs: A New Chapter in Transcriptomics and Proteomics. Front. Genet. *9*, 144.

Yu, C.-H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M.S., and Liu, Y. (2015). Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. Mol. Cell *59*, 744–754.

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

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