Human Long Noncoding RNAs Are Substantially Less Folded than Messenger RNAs

Jian-Rong Yang¹ and Jianzhi Zhang^{*1}

¹Department of Ecology and Evolutionary Biology, University of Michigan

***Corresponding author:** E-mail: jianzhi@umich.edu..

Associate editor: Jeffrey Thorne

Abstract

Long noncoding RNAs (lncRNAs) do not code for proteins but function as RNAs. Because the functions of an RNA rely on either its sequence or secondary structure, lncRNAs should be folded at least as strongly as messenger RNAs (mRNAs), which serve as messengers for translation and are generally thought to lack secondary structure-dependent RNA-level functions. Contrary to this prediction, analysis of genome-wide experimental data of human RNA folding reveals that lncRNAs are substantially less folded than mRNAs even after the control of expression level and GC% (percentage of guanines and cytosines), although both lncRNAs and mRNAs are more strongly folded than expected by chance. In contrast to mRNAs, lncRNAs show neither the positive correlation between folding strength and expression level nor the negative correlation between folding strength and evolutionary rate. These and other results support that although RNA folding undoubtedly plays a role in RNA biology it is also important in translation and/or protein biology.

Key words: evolutionary rate, expression level, lncRNA, mRNA, RNA folding, translation.

Under physiological conditions, RNA molecules typically fold into secondary structures composed of segments of paired nucleotides known as stems and segments of unpaired nucleotides commonly referred to as loops. Proper folding is critical to many RNA-level functions (i.e., biochemical activities unrelated to being a template for translation). For instance, disruption of nucleotide pairing in the stems of transfer RNAs often causes disease and is selectively disfavored (Wittenhagen and Kelley 2003; McFarland et al. 2004; Kondrashov 2005). Another well-known example is in the biogenesis of microRNAs (miRNAs), small RNAs that play widespread roles in posttranscriptional regulation (Flynt and Lai 2008). A stem-loop hairpin structure is required in the processing of pri-miRNAs and pre-miRNAs that gives rise to mature and functional miRNAs (Bartel 2004). Determining RNA secondary structures is thus essential for understanding the structural and mechanistic basis of various RNA functions.

Computational prediction of RNA folding has a long history of research (Zuker and Sankoff 1984; Mathews 2006; Seetin and Mathews 2012), but its accuracy remains moderate (Hajiaghayi et al. 2012; Park et al. 2013). Experimental determination of RNA folding had been time-consuming until the recent development of next-generation sequencing-based methods (Kertesz et al. 2010; Ding et al. 2014; Rouskin et al. 2014; Wan et al. 2014), which can simultaneously probe the folding of all RNAs in a sample, opening the door to reliable genome-wide analysis of the relationships among RNA folding, expression, function, and evolution. For instance, based on the folding data for greater than 3,000 yeast messenger RNAs (mRNAs) (Kertesz et al. 2010), a strong positive correlation was observed between mRNA concentration and folding strength (Zur and Tuller 2012).

We and others found that mRNA folding is on average stronger than the random expectation given the encoded protein sequence and codon usage in the mRNA, indicating that strong mRNA folding has been selectively favored (Katz and Burge 2003; Park et al. 2013). Furthermore, the excess in mRNA folding strength relative to the random expectation increases with mRNA concentration, suggesting that natural selection for mRNA folding intensifies with the rise of expression level (Park et al. 2013). These patterns are surprising because mRNAs possess no general physiological functions mediated by their structures beyond that of transferring genetic information.

The unexpected findings about mRNA folding prompt us to examine noncoding RNA folding. Because noncoding RNAs are expected to function exclusively at the RNA level and RNA folding is important for many (but not all) RNA functions, we hypothesize that noncoding RNAs have similar or stronger folding than mRNAs. We test this hypothesis by comparing human mRNAs with human long noncoding RNAs (lncRNAs), because of the large number of lncRNAs and the availability of their experimentally determined folding data (Wan et al. 2014). By definition, lncRNAs are noncoding RNAs with greater than 200 nucleotides (Ulitsky and Bartel 2013). By June 2014, over 17,000 lncRNA genes have been annotated in the human genome according to the LNCipedia database (Volders et al. 2013), compared with approximately 20,000 protein-coding genes. Although only a tiny fraction of lncRNAs have been functionally characterized, they are known to participate in a large number of biological processes including epigenetic regulation (Lee 2012), transcriptional regulation (Kornienko et al. 2013), and posttranscriptional regulation (Geisler and Collier 2013).

In addition, they have been found to play architectural roles (Hirose et al. 2014). However, lncRNA folding has not been extensively studied (Novikova et al. 2012). Here, we report the surprising finding that human lncRNAs are much less folded than mRNAs and discuss its biological implications.

We analyzed the in vitro folding of polyadenylated RNAs (including mRNAs and lncRNAs) in human lymphoblastoid cells quantified by parallel analysis of RNA structure (PARS) (Wan et al. 2014). Following previous studies (Zur and Tuller 2012; Park et al. 2013), we measured the folding strength of an RNA by F , which is the ratio between the relative probability that a site is folded and the relative probability that it is unfolded averaged across all sites in the RNA (see Materials and Methods). Note that F is not the fraction of paired sites divided by the fraction of unpaired sites in the RNA, because F also contains the information of the folding strengths of individual sites. The folding strength is relevant here, because the same RNA can have an ensemble of secondary structures rather than a single conformation and the folding strength reflects how likely one particular site is paired in the ensemble of secondary structures. Furthermore, F is not the mean ratio between absolute probabilities, because the absolute probabilities are unknown due to different enzyme efficiencies and sequencing depths in PARS experiments. In other words, one cannot compute the fraction of paired sites from F , although the higher the F , the stronger the average folding strength per site for the RNA concerned (Kertesz et al. 2010; Zur and Tuller 2012; Park et al. 2013; Wan et al. 2014). Because lncRNAs generally have much lower cellular concentrations than mRNAs, the folding data included a much smaller fraction of lncRNAs than mRNAs. After removing lncRNAs that are unexpressed in any of the eight human organs recently surveyed (Necsulea et al. 2014) and those that are translated (Wilhelm et al. 2014), we obtained a set of 701 lncRNAs with expression and folding data (supplementary table S1, Supplementary Material online). We similarly obtained a set of 16,131 mRNAs with expression and folding data (supplementary table S1, Supplementary Material online). A comparison showed that mRNAs are clearly folded more strongly than are lncRNAs ($P < 10^{-121}$, Mann–Whitney U test; fig. 1A), with the median F value of mRNAs about 3.9 times that of lncRNAs.

In yeast mRNAs, F increases with expression level (Zur and Tuller 2012; Park et al. 2013). Because lncRNAs have lower expression levels than mRNAs, we compared their F values after controlling their expression levels. That is, for each lncRNA, we randomly selected an mRNA with a similar expression level as the lncRNA (see Materials and Methods). We then compared the mean F value of all lncRNAs and that of all selected mRNAs. This process was repeated 1,000 times. In each of the 1,000 times, the mean F was found to be lower for lncRNAs than for the expression level-matched mRNAs (fig. 1B), suggesting that their difference in folding strength is not entirely explainable by their expression level difference.

Because the percentage of guanines and cytosines (GC%) in an RNA sequence can also influence its folding strength and because GC% is on average higher in mRNAs (49.2) than

in lncRNAs (45.3), we also controlled GC%. Comparing lncRNAs with randomly selected mRNAs with both similar expression levels and similar GC% showed that F is still lower in lncRNAs than in mRNAs ($P < 0.001$, randomization test; fig. 1C).

It is possible that the relatively high F values of mRNAs, compared with those of lncRNAs, are byproducts of the encoded protein sequences and/or biased synonymous codon usage of the mRNAs, neither of which would affect lncRNAs. To test this hypothesis, for each mRNA, we generated 100 pseudo-mRNAs by randomly shuffling synonymous codons within the real mRNA; the pseudo-mRNAs all have the same protein sequence, nucleotide frequencies, and synonymous codon frequencies as the real mRNA. Because of the lack of experimental folding data for pseudo-mRNAs, we computationally predicted the folding strengths of these 100 pseudo-mRNAs and estimated their mean fraction of paired sites (F'_p , where the subscript “p” stands for predicted). We also computationally predicted the fraction of paired sites in the real mRNA (F_p). We found that F_p exceeds F'_p in 70.4% of all mRNAs examined, indicating that the folding strengths of human mRNAs tend to be greater than the random expectations given the protein sequences and synonymous codon frequencies ($P < 10^{-300}$, binomial test; fig. 2A). Similar to what was observed in yeast and ten bacteria (Katz and Burge 2003; Park et al. 2013), this result suggests that, even in humans, whose effective population size is at least 3 orders of magnitude smaller than those of yeast and bacteria, strong mRNA folding has been selected for. For comparison, for each lncRNA, we generated 100 pseudo-lncRNAs by randomly shuffling the nucleotides of the lncRNA. We similarly estimated F_p and F'_p by computational prediction of the folding strength of the real lncRNA and those of pseudo-lncRNAs, respectively. We found F_p to exceed F'_p in 57.6% of the lncRNAs examined ($P < 10^{-4}$, binomial test; fig. 2B), suggesting that strong folding has also been selected for in human lncRNAs and supporting the premise of a requirement of strong folding for RNA-level functions. Given that both mRNAs and lncRNAs have higher-than-expected folding strengths, we asked whether the excess in folding strength relative to the random expectation is greater for mRNAs than lncRNAs with comparable expression levels and GC%. This is indeed the case ($P < 0.001$, randomization test; fig. 2C), suggesting stronger selection for folding of mRNAs than comparable lncRNAs. We also confirmed that the fractional excess in folding strength relative to the random expectation (i.e., $F_p/F'_p - 1$) is significantly greater for mRNAs than lncRNAs with comparable expression levels and GC% ($P < 0.001$, randomization test).

The observation that the excess in folding strength is greater for mRNAs than for lncRNAs is unexpected under the presumption that lncRNA functions often require certain RNA secondary structures whereas the mRNA function (as a messenger) does not. There are several potential causes of our observations. First, a potential trivial explanation is that most of the lncRNAs studied here are functionless and hence need not have stronger-than-expected RNA folding. Although the fraction of greater than 17,000 human lncRNAs that are

FIG. 1. Human lncRNAs are less folded than mRNAs. (A) A comparison in experimentally determined in vitro RNA folding strength (F) between mRNAs and lncRNAs. The bottom and top of each box are the first and third quartiles of all data points in respective categories, and the band inside the box shows the median. The whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range from the box edges. The gray dots show outliers, which lie outside the range shown by the whiskers. The P value is based on Mann–Whitney U test. (B) Folding strength is lower for lncRNAs than for expression-matched mRNAs. The histogram shows the frequency distribution (from 1,000 sets) of mean F of the same number of randomly picked mRNAs as lncRNAs that have similar expression levels as lncRNAs. The mean F of lncRNAs is indicated by the arrow. (C) Folding strength is lower for lncRNAs than for expression- and GC%-matched mRNAs. The histogram shows the frequency distribution (from 1,000 sets) of mean F of the same number of randomly picked mRNAs as lncRNAs that have similar expression levels and GC% as lncRNAs. The mean F of lncRNAs is indicated by the arrow.

functionless is debated (Ponjavic et al. 2007; Haerty and Ponting 2013; Pennisi 2014), the 701 lncRNAs in our data set are less likely than average lncRNAs to be functionless, because the RNA folding data enrich relatively highly expressed lncRNAs, which have a higher expectation than average lncRNAs to be functional. We also analyzed the subset of lncRNAs that are expressed in both human and macaque (supplementary table S2, Supplementary Material online) and thus are more likely than the average of the 701 lncRNAs to be functional. We again found mRNAs to fold more strongly than these lncRNAs, with ($P < 0.001$) or without ($P < 10^{-64}$) the control of expression level and GC% (supplementary fig. S1, Supplementary Material online). These considerations suggest that the potential inclusion of functionless lncRNAs in our data does not fully explain why lncRNAs are less folded than mRNAs. Second, it is possible that many mRNAs also have RNA-level functions such as binding to protein or RNA molecules and providing signals for localization. But most of these functions depend on RNA sequence motifs rather than secondary structures. In fact, strong secondary structures may even reduce the accessibility of certain sequence-dependent regulatory elements (Kertesz et al. 2007). Thus, the potential RNA-level functions of mRNAs are unlikely to explain their strong folding. Third, strong RNA folding may be primarily used to extend the RNA half-life. It is possible that on average mRNAs require longer half-lives than lncRNAs because mRNAs but not lncRNAs need to be translated to produce functional molecules. Experimental data indeed showed that the average half-life is longer for mRNAs than lncRNAs in a mouse cell line (Clark et al. 2012). Nevertheless, folding strength and half-life are uncorrelated among mRNAs (Zur and Tuller 2012). Thus,

it is unlikely that the folding strength difference between mRNAs and lncRNAs is related to their different half-lives. Fourth, strong mRNA folding does not help translational initiation. In fact, low mRNA folding at the beginning of an mRNA (5'-untranslated region and the first ~30 coding nucleotides) increases the translational initiation rate (Kudla et al. 2009; Gu et al. 2010; Tuller et al. 2010). Thus, requirement for translational initiation cannot explain why mRNAs fold more strongly than lncRNAs. Fifth, it was previously hypothesized that strong mRNA folding may decrease mRNA aggregation (i.e., pairing between two RNA molecules) and hence improve translational efficiency (Zur and Tuller 2012). However, although mRNA aggregation may decrease translational efficiency, mRNA folding has a qualitatively similar effect (Qu et al. 2011). Furthermore, biophysical modeling does not support this hypothesis, because mRNA folding is almost always energetically favored, compared with aggregation (Park et al. 2013). It was reported that even among the most weakly expressed genes, which should not have experienced selection for reduced mRNA aggregation, mRNA folding is energetically favored over aggregation for all but one gene (Park et al. 2013), suggesting virtually no risk for mRNA aggregation. Sixth, very recently, a tradeoff between translational elongation speed and translational accuracy was discovered (Yang et al. 2014). It was proposed that strong mRNA folding decreases elongation speed (Tuller et al. 2011; Yang et al. 2014) and hence increases translational accuracy (Yang et al. 2014). Indeed, a yeast transgene experiment demonstrated that increasing mRNA folding strength position-specifically enhances translational accuracy (Yang et al. 2014). If strong mRNA folding is selected primarily for enhanced translational fidelity, lncRNAs, which are not

Fig. 2. Computationally predicted folding strengths of mRNAs exceed expectations more than lncRNAs do. (A) Computationally predicted folding strength of an mRNA (F_p) tends to exceed its corresponding random expectation (F'_p), which is measured by the average folding strength of 100 pseudo-mRNAs generated by randomly shuffling synonymous codons in the mRNA. Each dot represents one real mRNA (y axis) and its pseudo-mRNAs (x axis). The numbers of dots above (red) and below (blue) the black diagonal line of $F_p = F'_p$ are indicated. The P value is from a binomial test of equal numbers of dots above and below the diagonal. (B) Computationally predicted folding strength of an lncRNA (F_p) tends to exceed its corresponding random expectation (F'_p), which is measured by the average folding strength of 100 pseudo-lncRNAs generated by randomly shuffling all nucleotides in the lncRNA. Each dot represents one real lncRNA (y axis) and its pseudo-lncRNAs (x axis). The numbers of dots above (red) and below (blue) the black diagonal line of $F_p = F'_p$ are indicated. The P value is from a binomial test of equal numbers of dots above and below the diagonal. (C) Comparison between the mean ($F_p - F'_p$) of mRNAs and lncRNAs. The histogram shows the frequency distribution of mean ($F_p - F'_p$) from 1,000 random sets of mRNAs with expression levels and GC% matched with lncRNAs. The arrow indicates the mean ($F_p - F'_p$) of lncRNAs.

translated, would not be subject to the same selection. Thus, this difference can in principle explain, at least qualitatively, why lncRNAs are less folded than mRNAs. However, it is unknown whether enhancing translational accuracy is the sole reason why mRNAs are folded more strongly than lncRNAs. Seventh, if strong mRNA folding slows translational elongation (Tuller et al. 2011; Yang et al. 2014), mRNA folding may also be deployed to regulate cotranslational protein folding, because slower elongation provides more time for generally better cotranslational folding (O'Brien et al. 2012). Indeed, the characterization of the secondary structure of the genomic RNA of human immunodeficiency virus (HIV) revealed a correlation between RNA structure and sequences that encode interdomain loops in HIV proteins, consistent with the idea that RNA folding modulates ribosome elongation to promote native protein folding (Watts et al. 2009). Whether this explanation is true to a large number of human mRNAs should be investigated in the future.

As mentioned, yeast mRNAs show a strong positive correlation between folding strength and expression level (Zur and Tuller 2012; Park et al. 2013). We discovered that the same correlation exists in human mRNAs, although its magnitude is small (Pearson's $R = 0.091$, $P < 10^{-300}$; table 1). This correlation for mRNAs can be explained by an increased demand for translational accuracy of more highly expressed genes (Yang et al. 2014), because mistranslation is wasteful in material and energy and may even be cytotoxic (Drummond and Wilke 2008; Yang et al. 2012). Interestingly, human lncRNAs do not show significant correlation between expression level and folding strength ($R = 0.044$, $P = 0.24$; table 1). Because RNAs are relatively inexpensive to synthesize, compared with proteins, the fitness cost due to material and energy waste in synthesizing

nonfunctional lncRNA molecules is negligible. Therefore, if misfolded lncRNAs are not toxic, the selection for lncRNA folding is not expected to increase with expression level. Our result contrasts a previous study (Managadze et al. 2011) that reported a significant positive correlation between mammalian lncRNA expression level and folding strength, based on computationally predicted folding data of a much larger set of mouse lncRNAs. This disparity may be caused by 1) RNA folding prediction errors; 2) difference in data size, because the previously reported correlation also diminishes when a smaller human lncRNA data set was used (Managadze et al. 2011); and/or 3) different criteria for lncRNA expression used by the two studies.

In yeast, the evolutionary rate of a protein is negatively correlated with the mRNA folding strength, with or without the control of the mRNA expression level, suggesting that strong mRNA folding is selectively favored and mutations weakening RNA folding tend not to be fixed (Park et al. 2013). We found these patterns to hold for human mRNAs as well. Specifically, we estimated the protein evolutionary rate by percent sequence difference between human and macaque orthologs in aligned regions and found it to correlate negatively with the mRNA folding strength, with ($R = -0.058$ by partial correlation, $P < 10^{-10}$) or without ($R = -0.052$, $P < 10^{-8}$) the control of mRNA expression level (table 1). This trend is expected, because a higher demand for mRNA folding reduces the probability that a random mutation is neutral and thus constrains evolution (Park et al. 2013).

We similarly analyzed lncRNAs. Because lncRNAs have rapid evolutionary turnovers (Kutter et al. 2012; Necsulea et al. 2014), to enrich functional lncRNAs in our data, we

Table 1. Correlations between Expression Level, Evolutionary Rate, and RNA Folding Strength for Human mRNAs and lncRNAs.

Variables Correlated	Variable Controlled	Pearson's <i>R</i>	<i>P</i> value	No. of genes
mRNAs				
Folding strength and expression level	Expression level	0.0912	<10 ^{−300}	16,131
Folding strength and protein evolutionary rate		−0.0518	<10 ^{−8}	13,442
Folding strength and protein evolutionary rate		−0.0582	<10 ^{−10}	13,442
lncRNAs				
Folding strength and expression level	Expression level	0.0441	0.24	701
Folding strength and evolutionary rate		−0.0372	0.50	338
Folding strength and evolutionary rate		−0.0367	0.50	338

required an lncRNA to be expressed in both human and macaque to be included. Applying this criterion lowered our sample size to 338 lncRNAs. We chose to compare human with macaque because their divergence is low enough to allow a relatively large sample of lncRNAs but not too low to hamper reliable estimation of the rate of lncRNA sequence evolution. We then measured the evolutionary rate of an lncRNA by the fraction of sites that are different between human and macaque in aligned regions. We found no significant correlation between the evolutionary rate of an lncRNA and its folding strength, with ($R = -0.037$ by partial correlation, $P = 0.5$) or without ($R = -0.037$, $P = 0.5$) the control of its expression level (table 1), consistent with the finding of a recent study (Schuler et al. 2014) that used computationally predicted RNA folding strengths. These results suggest that the variation in lncRNA folding strength is not functionally important. This is possible if there is a functional requirement for lncRNA folding to exceed a certain threshold but any folding strength above this threshold is equally good. Nonetheless, it remains possible that lncRNA folding strength does predict evolutionary conservation but the current data set is too small to detect this signal. This interpretation is consistent with the observed negative but statistically insignificant correlation between lncRNA folding strength and evolutionary rate (table 1).

In summary, we found comparable human mRNAs to fold more strongly than lncRNAs. This result is inconsistent with the assumption that RNA folding is involved only in RNA biology, but is consistent with the recent finding of the role of mRNA folding in regulating translational elongation speed and accuracy (Yang et al. 2014). Whether RNA folding plays other roles in translation and/or protein biology (e.g., cotranslational protein folding) in addition to RNA biology is an interesting subject worth exploration.

Materials and Methods

Human mRNA and lncRNA expression levels measured in RPKM (reads per kilobase of exon per million mapped reads) from eight organs were retrieved from a recent study (Necsulea et al. 2014) and then averaged across organs. The lncRNA list excluded miRNAs, RNase P, and other RNA genes annotated in Ensembl. The RNA secondary structures were experimentally determined by PARS using RNAs prepared from lymphoblastoid cells of a family trio with HapMap samples GM12878, GM12891, and GM12892 (Wan et al. 2014).

The PARS experiment separately treats a sample by RNase V1 and RNase S1, which specifically cleaves the 3' of double-stranded and single-stranded RNA, respectively, followed by next generation sequencing. The short reads generated in PARS were previously mapped to the human transcriptome (Wan et al. 2014) and were downloaded from NCBI GEO with the accession number of GSE50676. To increase the overall coverage, we summed the number of mapped reads from the three samples for each nucleotide. To estimate the RNA folding strength, for each nucleotide in each gene, we calculated a PARS score by $(v + 1)/(s + 1)$, where v and s are the numbers of reads mapped to the site in the RNase V1 and S1 treated RNA samples, respectively. Only sites with $v + s > 0$ are used in our analysis. The folding strength (F) of an RNA was calculated as the average PARS score across all sites in the RNA. We used EnsEMBL v75 (Flicek et al. 2014) to unify the annotations of protein-coding genes in the expression and PARS data. For lncRNAs, due to the lack of a consensual annotation, we extracted all noncoding transcripts from the PARS study (Wan et al. 2014) and BLASTed (Camacho et al. 2009) them against the annotated lncRNA genes in the expression data (Necsulea et al. 2014). We considered a PARS transcript to be derived from an lncRNA gene only when the BLAST alignment covers greater than 95% of the transcript and has a greater than 95% sequence identity, allowing minor mismatches in annotations caused by alternative splicing or less well-defined 5'- and 3'-ends of lncRNA genes/transcripts. When a single lncRNA gene is hit by multiple PARS transcripts, only the longest transcript is used. Finally, we removed any lncRNA that is annotated as being translated in a recently completed mass-spectrometry-based draft of the human proteome (Wilhelm et al. 2014), but note that protection of an lncRNA by ribosome as reflected in ribosome-profiling data is not a proof of translation (Guttman et al. 2013). As a result, we obtained 16,131 mRNAs and 701 lncRNAs as our core data set (supplementary table S1, Supplementary Material online). We avoided using experimentally measured in vivo RNA folding data (Rouskin et al. 2014), because the data contained only read counts for unpaired A's and C's without the corresponding counts for paired A's and C's and because there were no data for G's and U's. These deficiencies, coupled with the confounding factor of different sequencing depths for different RNAs (due to their different concentrations), make it difficult to infer the mean nucleotide pairing probability of an RNA from the in vivo RNA folding data.

We conducted our evolutionary analysis by comparing human genes with macaque genes. An initial list of human–macaque orthologous lncRNA pairs was downloaded from a recent study (Necsulea et al. 2014). To ensure that a human lncRNA was already an lncRNA in the common ancestor of human and macaque, we required that the orthologous lncRNAs have nonzero transcriptional activity in at least one of the eight human organs (cortex or whole brain, cerebellum, heart, kidney, liver, placenta, ovary, and testis) and at least one of the six macaque organs (cortex or whole brain, cerebellum, heart, kidney, liver, and testis) probed by RNA sequencing (Necsulea et al. 2014). Each orthologous pair was then aligned by CLUSTALW (Thompson et al. 2002). The lncRNAs whose alignments contained greater than 40% gaps were removed, resulting in a final data set of 338 lncRNAs (supplementary table S2, Supplementary Material online). The alignments are available online at http://www.umich.edu/~zhanglab/download/Jianrong_MBE2014/index.htm (last accessed December 26, 2014). The fractional difference between human and macaque orthologous lncRNAs was calculated by the number of nucleotide differences in aligned (gapless) regions divided by the total length of the aligned regions. For protein-coding genes, we used the one-to-one orthologous protein-coding genes in human and macaque defined in Ensembl v75 (Flicek et al. 2014). We also required human–macaque orthologous genes to have nonzero transcription in at least one of the eight human tissues and at least one of the six macaque tissues (Necsulea et al. 2014). The protein sequences of human and macaque orthologous genes were downloaded from Ensembl and similarly processed as lncRNAs. Briefly, they were first aligned by CLUSTALW (Thompson et al. 2002), followed by the calculation of the fractional difference in the protein sequence using the aligned (gapless) regions.

To ensure a fair comparison in folding strength between lncRNAs and mRNAs, for each lncRNA, we randomly sampled an mRNA with a similar expression level as the lncRNA for comparison, using the following criteria: i) If the expression level of the lncRNA is $x = 0$, a randomly picked mRNA with expression level $y < 0.1$ was chosen; ii) if $x > 0$, a randomly picked mRNA satisfying $|y - x|/x < 0.1$ was chosen; iii) if no mRNA satisfies criteria (i) or (ii) for the lncRNA, an mRNA randomly picked from the 100 mRNAs with the smallest $|y - x|$ was chosen; iv) after an mRNA was chosen for each lncRNA, a Mann–Whitney U test was used to compare the expression levels between all lncRNAs and all chosen mRNAs. The set of mRNAs was accepted only when the P value exceeds 0.05. Note that no mRNA could appear twice in the set. Higher P -value cutoffs (e.g., 0.1) were avoided to allow sufficient stochastic variation among different replicates of this procedure. After the set of expression-matched mRNAs was accepted, we compared the folding strengths between the lncRNAs and chosen mRNAs. This procedure was repeated 1,000 times.

When the GC% of an RNA is also controlled, in addition to the aforementioned criteria for expression level, we also required that the GC% of a chosen mRNA v) differs from the given lncRNA by less than 5, or vi) is among the 100 mRNAs that is most similar to the given lncRNA in GC%.

We further required that vii) a Mann–Whitney U -test between lncRNAs and the chosen mRNAs in GC% should have a P value exceeding 0.05. Again, 1,000 sets of mRNAs with matching expressions and GC% were chosen to compare with lncRNAs.

We used RNAfold in ViennaRNA Package 2 (Lorenz et al. 2011) to computationally predict the minimum free energy RNA secondary structure using a sliding window approach, with a window size of 150 bases (Lange et al. 2012) and a step size of 10 bases. The folding strength of a nucleotide is calculated by the number of windows in which the nucleotide is predicted to be paired, divided by the total number of windows covering this nucleotide. The folding strength of an RNA molecule is the average folding strength of all of its nucleotides. Note that the folding strength defined here is the fraction of nucleotides paired rather than folding energy, because folding energy is less comparable to the experimentally determined PARS scores (Wan et al. 2014) due to the different energies associated with different nucleotide pairs.

To assess the random expectation of the folding strength of an mRNA, we randomly shuffled synonymous codons within the coding region of the mRNA to create 100 pseudo-mRNAs. We did not control dinucleotide frequencies when generating these pseudo-mRNAs (Katz and Burge 2003), because dinucleotide frequencies may have been altered by natural selection for RNA folding (Chamary and Hurst 2005). The average folding strength of these 100 pseudo-mRNAs was regarded as the expected folding strength (F'_p) of the mRNA. We also computationally predicted the folding strength (F_p) of the real mRNA. For each lncRNA, its entire nucleotide sequence was randomly shuffled to create 100 pseudo-lncRNAs. The average folding strength of these 100 pseudo-lncRNAs was regarded as the expected folding strength of the lncRNA. We also computationally predicted the folding strength of the real lncRNA. Our finding that $F_p - F'_p$ is smaller for lncRNAs than for mRNAs is likely to be conservative, because the pseudo-lncRNA sequences are probably more different from their corresponding real lncRNA sequences than they should be, due to the lack of knowledge and hence lack of consideration of potential functional constraints in lncRNAs. The Perl scripts for calculating F_p and F'_p are available online at http://www.umich.edu/~zhanglab/download/Jianrong_MBE2014/index.htm (last accessed December 26, 2014).

Supplementary Material

Supplementary tables S1 and S2 and figure S1 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgment

The authors thank Xiaoshu Chen for valuable comments. This work was supported in part by research grant R01GM103232 from the US National Institutes of Health to J.Z.

References

- Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116:281–297.

- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.
- Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* 6:R75.
- Clark MB, Johnston RL, Inostroza-Ponta M, Fox AH, Fortini E, Moscato P, Dinger ME, Mattick JS. 2012. Genome-wide analysis of long noncoding RNA stability. *Genome Res.* 22:885–898.
- Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. 2014. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* 505:696–700.
- Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.
- Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res.* 42:D749–D755.
- Flynt AS, Lai EC. 2008. Biological principles of microRNA-mediated regulation: shared themes amid diversity. *Nat Rev Genet.* 9:831–842.
- Geisler S, Collier J. 2013. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat Rev Mol Cell Biol.* 14: 699–712.
- Gu W, Zhou T, Wilke CO. 2010. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol.* 6:e1000664.
- Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. 2013. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154:240–251.
- Haerty W, Ponting CP. 2013. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol.* 14:R49.
- Hajjaghaiy M, Condon A, Hoos HH. 2012. Analysis of energy-based algorithms for RNA secondary structure prediction. *BMC Bioinformatics* 13:22.
- Hirose T, Mishima Y, Tomari Y. 2014. Elements and machinery of non-coding RNAs: toward their taxonomy. *EMBO Rep.* 15:489–507.
- Katz L, Burge CB. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.* 13: 2042–2051.
- Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. 2007. The role of site accessibility in microRNA target recognition. *Nat Genet.* 39: 1278–1284.
- Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Nature* 467:103–107.
- Kondrashov FA. 2005. Prediction of pathogenic mutations in mitochondrially encoded human tRNAs. *Hum Mol Genet.* 14: 2415–2419.
- Kornienko AE, Guenzl PM, Barlow DP, Pauler FM. 2013. Gene regulation by the act of long non-coding RNA transcription. *BMC Biol.* 11:59.
- Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324: 255–258.
- Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC. 2012. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.* 8:e1002841.
- Lange SJ, Maticzka D, Mohl M, Gagnon JN, Brown CM, Backofen R. 2012. Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res.* 40:5215–5226.
- Lee JT. 2012. Epigenetic regulation by long noncoding RNAs. *Science* 338: 1435–1439.
- Lorenz R, Bernhart SH, Honer Zu Siederdisen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol.* 6:26.
- Managadze D, Rogozin IB, Chernikova D, Shabalina SA, Koonin EV. 2011. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol Evol.* 3: 1390–1404.
- Mathews DH. 2006. Revolutions in RNA secondary structure prediction. *J Mol Biol.* 359:526–532.
- McFarland R, Elson JL, Taylor RW, Howell N, Turnbull DM. 2004. Assigning pathogenicity to mitochondrial tRNA mutations: when “definitely maybe” is not good enough. *Trends Genet.* 20: 591–596.
- Necsulea A, Soumilion M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505: 635–640.
- Novikova IV, Hennelly SP, Sanbonmatsu KY. 2012. Sizing up long non-coding RNAs: do lncRNAs have secondary and tertiary structure? *Bioarchitecture* 2:189–199.
- O'Brien EP, Vendruscolo M, Dobson CM. 2012. Prediction of variable translation rate effects on cotranslational protein folding. *Nat Commun.* 3:868.
- Park C, Chen X, Yang JR, Zhang J. 2013. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A.* 110:E678–E686.
- Pennisi E. 2014. Cell biology. Lengthy RNAs earn respect as cellular players. *Science* 344:1072.
- Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* 17:556–565.
- Qu X, Wen JD, Lancaster L, Noller HF, Bustamante C, Tinoco I Jr. 2011. The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature* 475:118–121.
- Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. 2014. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* 505:701–705.
- Schuler A, Ghanbarian AT, Hurst LD. 2014. Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Mol Biol Evol.* 31: 3164–3183.
- Seetin MG, Mathews DH. 2012. RNA structure prediction: an overview of methods. *Methods Mol Biol.* 905:99–122.
- Thompson JD, Gibson TJ, Higgins DG. 2002. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics. Chapter 2:Unit 2.3.*
- Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppén E, Ziv-Ukelson M. 2011. Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol.* 12:R110.
- Tuller T, Waldman YY, Kupiec M, Ruppén E. 2010. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci U S A.* 107:3645–3650.
- Ulitsky I, Bartel DP. 2013. lincRNAs: genomics, evolution, and mechanisms. *Cell* 154:26–46.
- Volders PJ, Helsen K, Wang X, Menten B, Martens L, Gevaert K, Vandesompele J, Mestdagh P. 2013. LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.* 41:D246–D251.
- Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z, Zhang J, Spitale RC, Snyder MP, Segal E, et al. 2014. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 505: 706–709.
- Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW Jr, Swanstrom R, Burch CL, Weeks KM. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* 460:711–716.
- Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* 509:582–587.
- Wittenhagen LM, Kelley SO. 2003. Impact of disease-related mitochondrial mutations on tRNA structure and function. *Trends Biochem Sci.* 28:605–611.

- Yang JR, Chen X, Zhang J. 2014. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol.* 12:e1001910.
- Yang JR, Liao BY, Zhuang SM, Zhang J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc Natl Acad Sci U S A.* 109:E831–E840.
- Zuker M, Sankoff D. 1984. RNA secondary structures and their prediction. *Bull Math Biol.* 46:591–621.
- Zur H, Tuller T. 2012. Strong association between mRNA folding strength and protein abundance in *Scerevisiae*. *EMBO Rep.* 13: 272–277.