Negative Correlation between Expression Level and Evolutionary Rate of Long Intergenic Noncoding RNAs

David Managadze, Igor B. Rogozin, Diana Chernikova, Svetlana A. Shabalina, and Eugene V. Koonin*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

***Corresponding author:** E-mail: koonin@ncbi.nlm.nih.gov.

Accepted: 2 November 2011

Abstract

Mammalian genomes contain numerous genes for long noncoding RNAs (lncRNAs). The functions of the lncRNAs remain largely unknown but their evolution appears to be constrained by purifying selection, albeit relatively weakly. To gain insights into the mode of evolution and the functional range of the lncRNA, they can be compared with much better characterized protein-coding genes. The evolutionary rate of the protein-coding genes shows a universal negative correlation with expression: highly expressed genes are on average more conserved during evolution than the genes with lower expression levels. This correlation was conceptualized in the misfolding-driven protein evolution hypothesis according to which misfolding is the principal cost incurred by protein expression. We sought to determine whether long intergenic ncRNAs (lincRNAs) follow the same evolutionary trend and indeed detected a moderate but statistically significant negative correlation between the evolutionary rate and expression level of human and mouse lincRNA genes. The magnitude of the correlation for the lincRNAs is similar to that for equal-sized sets of protein-coding genes with similar levels of sequence conservation. Additionally, the expression level of the lincRNAs is significantly and positively correlated with the predicted extent of lincRNA molecule folding (base-pairing), however, the contributions of evolutionary rates and folding to the expression level are independent. Thus, the anticorrelation between evolutionary rate and expression level appears to be a general feature of gene evolution that might be caused by similar deleterious effects of protein and RNA misfolding and/or other factors, for example, the number of interacting partners of the gene product.

Key words: long noncoding RNA, ncRNA, RNA expression, genomic alignments, introns, RNA folding.

Introduction

Traditionally, genomes have been perceived mostly as repositories of protein-coding genes. Although this might be largely true in the case of prokaryotes and unicellular eukaryotes, numerous recent studies on the genomes of multicellular eukaryotes, particularly animals, have revealed a vast RNome, that is, a collection of genes for noncoding RNAs (ncRNAs) (Carninci et al. 2005; Mattick and Makunin 2006; Ponting et al. 2009). Strikingly, the total number of genes for ncRNAs that are expressed from a mammalian genome seems to exceed the number of protein-coding genes severalfold (Mattick and Makunin 2006). The classification of ncRNAs, whether these loci should be considered genes or not, and the validation of their functionality remain matters of intensive investigation and debate (van Bakel and Hughes 2009; Ponting and Belgard 2010).

Among many distinct classes of ncRNAs, the long ncRNA (lncRNA) are probably the most enigmatic group. The definition of lncRNA is based solely on the transcript size: lncRNAs are defined as non-coding RNAs longer than 200 nucleotides (Mattick and Makunin 2006; Ponting et al. 2009). Many lncRNAs are spliced, 5′ capped, and polyadenylated (Okazaki et al. 2002; Carninci et al. 2005; Kapranov et al. 2007; Ponjavic et al. 2007). Based on the localization in the genome, lncRNAs can be divided into two distinct classes: 1) transcripts that overlap protein-coding genes and 2) long intergenic noncoding (linc) RNAs that are transcribed from genome regions separating protein-coding genes (Ponting et al. 2009). Many lncRNAs that overlap protein-coding genes are likely to be involved in a sense–antisense regulation (Chen et al. 2005). The current knowledge on the functions of lincRNAs is scarce because very few of the

lincRNAs have been assigned a function experimentally, but their functional range is believed to be broad on the basis of indirect evidence (Bertone et al. 2004; Ponjavic et al. 2007; Mercer et al. 2009; Ponting and Belgard 2010). It has been suggested that lincRNAs could be involved in the regulation of many cellular processes (Mattick and Makunin 2006). For example, they can affect transcription locally at the gene level (Martens et al. 2004; Martianov et al. 2007; Osato et al. 2007; Hirota et al. 2008) as well as target transcription regulators and thus affect transcription of many genes (Feng et al. 2006; Goodrich and Kugel 2006; Huarte et al. 2010). They can also target RNA polymerase II in human and mouse (Espinoza et al. 2007; Mariner et al. 2008) and thus act on an even broader range of genes. Furthermore, lincRNAs participate in the regulation of splicing (Munroe and Lazar 1991; Beltran et al. 2008) and translation (Wang et al. 2005; Centonze et al. 2007). Well-characterized examples of lincRNAs involved in epigenetic processes are Xist (Clemson et al. 1996), Kcnq1ot1 (Umlauf et al. 2004; Pandey et al. 2008) and Air (Nagano et al. 2008) (also see the review by Ponting et al. 2009).

Compared with protein-coding sequences and small RNAs (e.g., miRNA and snoRNA), lncRNAs are weakly conserved: only approximately 5% of the bases have been estimated to be evolutionarily constrained (Marques and Ponting 2009). Accordingly, early studies called these RNAs "transcriptional dark matter" that was considered to be largely nonfunctional although a low level or even lack of appreciable conservation do not necessarily imply lack of function (Pang et al. 2006). A well-studied example is the Xist RNA, which is weakly conserved but is essential for the X chromosome dosage compensation in mammals (Nesterova et al. 2001). Moreover, the limited overall conservation notwithstanding, many of the lncRNAs still contain strongly conserved regions (Siepel et al. 2005). In general, lncRNAs show reduced substitution and insertion/deletion rates compared with random expectation, which has been interpreted as a signature of purifying selection (Ponjavic et al. 2007; Guttman et al. 2009). The recent study by Chodroff et al. (2010) reports the first attempt to characterize lncRNA orthologs present in eutheria, marsupials, and birds. Several lncRNAs analyzed in this work have been shown to possess conserved transcript structures and expression patterns.

In general, the expression levels of lncRNAs tend to be lower than those of protein-coding genes (Mattick and Makunin 2006). A comparative analysis of the expression patterns of intergenic transcripts in brain, heart, testis, and lymphoblastoid cell lines of human and chimpanzee has revealed a tissue-specific conservation pattern, which is similar to that of protein-coding genes. Altogether, approximately half of the transcripts that showed differences in expression between the two species come from the intergenic regions. Thus, lincRNAs might have played an important role in the phenotypic differentiation between these two primates (Khaitovich et al. 2006).

Some lncRNA genes might have evolved from protein-coding genes as exemplified by the thoroughly characterized Xist RNA (Duret et al. 2006; Elisaphenko et al. 2008), suggesting the possibility that some properties of lncRNAs and their genes might be similar to those of protein-coding genes. Protein-coding genes that are expressed highly and broadly across tissues on average are more evolutionarily conserved than genes that have lower expression level and breadth; a significant negative correlation between the expression and evolutionary rate of protein-coding genes has been revealed for all model organisms for which extensive expression data are available (Duret and Mouchiroud 2000; Pal et al. 2001; Krylov et al. 2003; Zhang and Li 2004). This negative correlation extends also to 3′UTRs of protein-coding genes although not to 5′UTRs (Jordan et al. 2004). The universal anticorrelation between the evolutionary rate and the expression level observed for the protein-coding genes has been explained within the framework of the misfolding-driven concept of protein evolution according to which the selective pressure to minimize the misfolding is the strongest for highly expressed proteins (Drummond and Wilke 2008, 2009; Wolf et al. 2010; Yang et al. 2010). Here, we show that the universal dependency between the evolution and the expression holds also for the lincRNA genes and is comparable in magnitude to the anticorrelation detected for protein-coding genes.

Materials and Methods

The lincRNA Sets

Complete mouse and human probe sets were downloaded from NRED database (Dinger et al. 2009) in the tab-delimited and BED (Browser Extensible Data, containing genomic coordinates) formats. The probe sets from platform GNF Atlas 2 (Mouse and Human), with target classification "Noncoding Only" were used for further analysis. This resulted in 5444 mouse and 917 human probe sets. Only the probe sets that mapped to intergenic regions of the mouse and human genomes (i.e., between two adjacent protein-coding genes) were used and analyzed. This was achieved by selecting only the probe sets with the value "Intergenic" in both ProbeGenomicContextSense and ProbeGenomicContextAntisense fields of the tab-delimited file.

As a next step, one-to-many list of probe sets and their corresponding Target Accession IDs (NCBI GenBank IDs of RNAs; TargetAccessionID column of NRED file) was transformed into a one-to-one list, where accession ID corresponded to the single probe set, preferentially the one not with _s_at or _x_at suffix which, according to Affymetrix, non-uniquely map to the genomes in several locations. This list of lincRNAs was further filtered: genes shorter than 200 nucleotides were removed. This procedure yielded the final set of NCBI GenBank Accession IDs of 2390 mouse and

589 human RNAs and their corresponding microarray expression probe sets. There were alternatively spliced isoforms in these data sets, the fraction of such isoforms was <10%.

To control for the possibility of contamination of the lincRNA data set with protein-coding genes, two tests were performed: 1) mouse lincRNA sequences were searched against protein sequence databases using the BlastX program, 2) the coding potential of lincRNAs was predicted using the SYNCODE program (Rogozin et al. 1999).

Gene Expression Data

To analyze the transcriptome of normal tissues without bias from cancerous tissues, only data for normal (non-cancerous) tissues (73 human and 61 mouse tissues) were used. Log_2-normalized expression levels (A-values) <7.0 (threshold representing a conservative expression level above background, according to the NRED database) were ignored, and median values of expression for each probe set across the tissues were calculated and designated as probe set's Median Expression Level. Expression Breadth, the number of tissues where the probe set has been expressed above the A-value threshold (7.0), was also calculated. The probe sets with _s_at and _x_at suffixes were discarded as promiscuous because they map to several regions in the genome, and it would be incorrect to associate their expression levels to one particular lincRNA ID.

As the final step, median expression level and breadth values of probe sets were associated with their corresponding lincRNA IDs and used throughout the entire work. For mouse, this data set contained 2013 lincRNAs; human data set contained 519 lincRNAs.

As an alternative method of measuring expression levels, counting of the Expressed Sequence Tags (ESTs) was employed. The sequences of lincRNAs were extracted from the UCSC Table Browser (see Sequences, Alignments, and Evolutionary Distances in the Materials and Methods section). These sequences were searched against the human and mouse subsets of EST database (archives est_human.tar.gz and est_mouse.tar.gz, available from the NCBI FTP site at ftp://ftp.ncbi.nih.gov/blast/db/) using the program BlastN 2.2.24+ (Camacho et al. 2009) from the Blast package. The number of ESTs with >97% identity and alignment length of at least 200 nucleotides or longer was counted. The procedure was repeated twice, with and without masking the human/mouse repeat sequences; median value was calculated and assigned to lincRNAs as an expression level based on EST count.

The mouse RNA-seq data for eight tissues (the ENCODE project; modENCODE Consortium, 2009) was downloaded from the UCSC genome browser Web site (ftp://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeLicrRnaSeq/) and pooled together. The number of RNA-seq hits (M) was calculated for each mouse lincRNA. The overall expression of each lincRNA was estimated using a $\log_2$ normalization: Exp-RNA-seq $= \log_2[(M + 1)/L] + 10$, where L is the length of the lincRNA. This normalization has been suggested as a robust estimator of expression using RNA-seq data (Lee et al. 2011).

Sequences, Alignments, and Evolutionary Distances

The genomic coordinates and sequences of exons and introns of Target RNA Accession IDs (NCBI GenBank IDs of RNAs; TargetAccessionID column of NRED file) were downloaded from the UCSC Table Browser (Karolchik et al. 2004), from all_mrna tables of mouse mm8 and human hg18 assemblies. Multiple alignments of these regions were fetched from Galaxy (Blankenberg et al. 2010; Goecks et al. 2010). Two different 17-way multiZ alignments—with human (hg18) and mouse (mm8) reference genomes—were used; only mouse (mm8), human (hg18), chimp (panTro1), macaque (rheMac2), rat (rn4), and dog (canFam2) alignments were downloaded. Alignments of exons and introns <100 nt were discarded. The exon and intron alignments for each lincRNA were concatenated, and two alignments were produced per lincRNA: "stitched" exons and "stitched" introns.

Calculation of percentage of insertions/deletions (indels) in the alignments was performed using an in-house tool written in C++. The program employed the following algorithm: the number of indel positions of pairwise alignments and the alignment length were computed. Finally, the ratio/percentage of indels and alignment length were calculated. In order to eliminate unreliable alignments containing an excess of indels, only alignments with the total length of indels below a threshold were used for subsequent analysis; three indel thresholds (15%, 30%, and 45%) were applied. Pairwise evolutionary distance matrices for concatenated alignments of exons or introns from human and mouse linc RNA genes were calculated using the DNADIST program from the PHYLIP package (Felsenstein 1996), with the Kimura nucleotide substitution model.

The C.A.MAM program (Bohning et al. 1992, 1998) was used to reveal outliers in the distributions of evolutionary distances and expression. This program attempts to decompose each distribution into two or more normal or log-normal distributions. If the decomposition procedure produced one distribution, no outliers were removed. If the decomposition produced several distributions, only one distribution with the largest number of data points was used for further correlation analysis. The Pearson (r), Spearman (rho), and Kendall (tau) correlation coefficients and their corresponding P values were calculated using an ad hoc R-language script. To eliminate the potential effect of contamination by the protein-coding genes on the observed correlations, all the lincRNAs with a similarity to the protein-coding genes were removed from the lincRNA set and the correlation

coefficients were recalculated. To obtain the list of lincRNAs similar to proteins, exons (separate as well as concatenated) were compared with the mouse RefSeq proteins using the BlastX program. Hits with the E-value $<10^{-4}$ and alignment length >20 amino acids were considered suspect. The two lists of suspect lincRNAs originated from the analyses of concatenated and separate exons (853 and 860 lincRNAs, respectively) were merged into the final set of 907 lincRNAs potentially containing protein-coding regions, which were removed from the mouse lincRNA set, and the correlation coefficients were recalculated for the remaining lincRNAs.

Gene Sequences and Alignments for Orthologous Protein-Coding Genes in Human–Mouse

To compare the results obtained for lincRNA genes with the trends observed for protein-coding genes, a control gene set was compiled consisting of 7,711 well-annotated orthologous human and mouse protein-coding genes that yielded high-quality genome alignments. For each human and mouse pair from the UCSC list of the human-mouse orthologs (http://genome.ucsc.edu/cgi-bin/hgTables), we identified the best Blast hit and estimated the overlap of the protein-coding sequences. Only full-length protein-coding transcripts with links to the RefSeq database and with up to 75% of protein-coding region aligned were included in the control group and used for the subsequent analysis. Mouse genome sequences were downloaded from ftp://hgdownload.cse.ucsc.edu/goldenPath/mm8/chromosomes/. Genomic coordinates of extended human gene loci were transferred to the mouse genome sequence using the UCSC Lift Genome Annotations tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Mammalian genomic repeats were masked, and extended genomic loci of orthologous human–mouse genes were aligned using the OWEN program (Ogurtsov et al. 2002) and annotated. In case of alternatively spliced forms, the longest CDSs and UTRs were considered. For the protein-coding regions, the alignment of nucleotide sequences was guided by the amino acid sequence alignment. Core hits with E-values $<10^{-3}$ produced by OWEN program were extracted for analysis as described previously (Ogurtsov et al. 2008). Synonymous and nonsynonymous divergence (K_s and K_a, respectively) were calculated using the PAML program (ftp://abacus.gene.ucl.ac.uk/pub/paml) with default parameters and the yn00 estimation method (Yang 1997).

RNA Secondary Structure Prediction

RNA secondary structures were predicted using two methods, which are based on the global and local free energy estimations, respectively. The lincRNAs were computationally "folded" and the predicted minimum free energy of the secondary structure was calculated, using our implementation of the algorithm that employs nearest neighbor parameters for evaluation of free energy (Zuker 2003). Energy minimization was performed by the dynamic programming method that finds the secondary structure with the minimum free energy with sums contributing from stacking loop length using an improved algorithm for evaluation of internal loops; this program "folds" sequences up to the 28,000 nucleotide long (Ogurtsov et al. 2006). Local free energy was estimated for the pairs of highly similar slow evolvind sequences, extracted from the human–mouse alignments (Kondrashov and Shabalina 2002). The secondary structure of the expressed lincRNAs was inferred by intersecting their chromosomal positions with the positions of the RNAz structural predictions made across the entire mouse genome, as previously described in the NRED database (Washietl et al. 2005; Mercer et al. 2008; Dinger et al. 2009). Conserved RNA secondary structures were considered significant at the confidence threshold level of $P > 0.5$, where P is the significance of the classification, which is quantified as "RNA-class probability" (Gruber et al. 2007).

Results

The Mouse and Human lincRNA Sets

To avoid potential complications caused by the coordinated expression of protein-coding genes and lncRNAs, we chose to analyze only the sets of mammalian lincRNAs. The data of 5,444 "Noncoding Only" mouse probe sets were downloaded from NRED database (Dinger et al. 2009). After discarding the probe sets that did not map to intergenic space and establishing one-to-one relationship between RNA IDs and their corresponding probe set IDs (see Materials and Methods), we obtained the final set of 2,390 mouse lincRNAs (NCBI GenBank Accession IDs of RNAs) of which 977 contained introns. After discarding the probe sets with very low median expression levels and those with equivocal genome mapping, the final set of 2,013 mouse lincRNAs, including 918 intron-containing ones, was obtained (for details, see Materials and Methods). For humans, the data for 917 probe sets were downloaded, and the same procedure of removing low-expressed or equivocally mapped lincRNAs yielded the final set of 519 lincRNAs, including 211 intron-containing genes.

Thus, the current set of experimentally verified human lincRNAs was several times smaller than the corresponding mouse set, in agreement with previous observations (e.g., Rearick et al. 2010). Most likely, this difference reflects the different states of lncRNA annotations rather than a genuine excess of lincRNAs in rodents.

Table 1 summarizes the characteristics of the sets of mouse and human lincRNAs analyzed here.

Evolutionary Rates and Expression of lincRNA

The traditional gauge of selection in protein-coding genes is the ratio of nonsynonymous (K_a) over synonymous (K_S)

Table 1
Statistics of lincRNA Data Sets

	Mouse	Human
Probe sets	5444	917
All lincRNAs	2390	589
Median length, nt	2,535	2,626
Average length, full lincRNA	11,775	16,855
Fused exons	1,843	1,998
Fused introns	24,246	36,686
GC% (aggregate[a]/median), full length	0.42/0.44	0.42/0.44
Fused exons	0.45/0.45	0.45/0.45
Fused introns	0.42/0.44	0.41/0.43
Intron-containing (introns with length<40 nt discarded)	979	245
Exons in intron-containing lincRNAs	3,439	1,194
Introns in intron-containing lincRNAs	2,462	949
Introns shorter than 40 nt	424	94
Exons shorter than 15 nt	41	7
Introns per lincRNA	2.52	3.86
Exons per lincRNA	3.52	4.85
Average length, nt	25,816	38,264
Average exon length	478	383
Average intron length	9,574	9,435
Intronless genes (one exon only)	1411	344
lincRNAs with A $\geq$ 7.0, uniquely mapping to genomes	2,013	519
Intron-containing	918	211

[a] Aggregate GC% is calculated from the sequences of all samples concatenated together.

substitutions. $K_a/K_S < 1$ is thought to indicate purifying selection, whereas $K_a/K_S > 1$ is construed as the signature of positive selection (Hurst 2002). In the case of lncRNA genes, the substitution rate of exons (Ke) may be considered analogous to K_a, whereas the substitution rate in intronic sequences (Ki) is a logical choice of the proxy for neutral evolution, analogously to the traditional use of K_S. Indeed, apart from pseudogenes, introns are among the best candidates for neutrally evolving sequences (Louie et al. 2003; Hoffman and Birney 2007; Resch et al. 2007). Purifying selection in lncRNA exons potentially can be defined by Ke/Ki < 1.

Substitution rates for exons and introns in intron-containing lincRNA genes (~41% human and mouse lincRNAs contain introns; see table 1) are shown in table 2 and supplementary table S1, Supplementary Material online. The indel cutoff (we used alignments with the total length of indels below a threshold of 15%, 30%, or 45%) employed to vary alignment stringency does not qualitatively influence the results although higher statistical significance was observed under the more stringent criteria (15% and 30% indels) (table 2 and supplementary table S1, Supplementary Material online). The substitution rate of mouse lincRNA exons was found to be significantly lower compared with mouse introns (Ke/Ki $<$ 1; table 2 and fig. 1). These results suggest that purifying selection acts on exons of lincRNA genes and are consistent with earlier observations (Ponjavic et al. 2007; Guttman et al. 2009). Additionally, the distribution of substitution rates was notably wider for concatenated exons of mouse lincRNAs than it was for concatenated introns (fig. 1 and supplementary figure S1, Supplementary Material online), indicative of the variance in the intensity of the purifying selection on mouse lincRNA genes. The lower substitution rates of the exons compared with the introns were observed for human lincRNA genes as well (supplementary table S1, Supplementary Material online) although the statistical support was weaker due to the smaller size of the human lincRNA set. The significantly reduced substitution rate in the mouse and human lincRNA exons was reproduced when all lincRNAs (intron containing and intronless) were used for the calculation of exon substitution rates (supplementary tables S2 and S3, Supplementary Material online). The reduced substitution rate of the lincRNA exons compared with the adopted neutral baseline (in this case, the substitution rate for introns) and the broad distribution of Ke values qualitatively resemble the case of protein-coding exons, which are almost universally subject to purifying selection of widely varying strengths ($K_a/K_S < 1$) (Koonin and Wolf 2010). However, the purifying selection on the exons in the lincRNAs is much weaker than on nonsynonymous positions in the protein-coding genes (supplementary figure S1, Supplementary Material online). Both the strength and the shape of the distribution of the substitution rates in lincRNA exons more closely resemble synonymous than nonsynonymous substitutions in protein-coding genes (supplementary figure S1, Supplementary Material online) although the differences in methods used to estimate substitution rates in noncoding and coding sequences preclude a direct quantitative comparison (Resch et al. 2007). As is the case with K_S, neutral

Table 2
Evolutionary Rates of Mouse Intron-Containing lincRNA Genes

		Exons			Introns			Student *t*-test	
Species Pair	**Threshold (Indel %)[a]**	**Data Points**	**Mean Rate**	**Variance**	**Data Points**	**Mean Rate**	**Variance**	***H***	***P* Value**
Mouse–Human	15	290	0.375	0.012	141	0.425	0.012	1	1.4E-05
	30	468	0.394	0.011	259	0.430	0.009	1	6.0E-06
	45	599	0.404	0.010	458	0.439	0.006	1	1.9E-05
	100	871	0.418	0.011	863	0.449	0.006	1	1.9E-12
Mouse–Chimp	15	270	0.375	0.013	117	0.431	0.010	1	6.5E-15
	30	444	0.398	0.012	230	0.431	0.008	1	2.5E-07
	45	582	0.405	0.011	433	0.438	0.006	1	4.5E-06
	100	863	0.417	0.011	840	0.448	0.006	1	4.0E-12
Mouse–Macaque	15	251	0.374	0.013	115	0.428	0.012	1	7.2E-05
	30	408	0.392	0.011	221	0.434	0.009	1	9.8E-07
	45	540	0.403	0.011	375	0.442	0.007	1	2.0E-15
	100	847	0.419	0.012	829	0.451	0.006	1	1.2E-11
Mouse–Rat	15	840	0.149	0.002	815	0.168	0.003	1	3.1E-10
	30	878	0.150	0.002	887	0.169	0.003	1	8.9E-10
	45	890	0.151	0.002	897	0.169	0.003	1	3.6E-08
	100	910	0.152	0.003	913	0.171	0.003	1	1.2E-13
Mouse–Dog	15	191	0.387	0.015	91	0.475	0.021	1	1.1E-09
	30	355	0.429	0.017	195	0.502	0.016	1	1.0E-14
	45	483	0.445	0.015	321	0.503	0.013	1	4.0E-11
	100	816	0.467	0.018	823	0.509	0.010	1	7.9E-13

[a] We used alignments with the total length of indels below a threshold; three indel thresholds (15%, 30%, 45% and 100%, that is, no threshold) were applied.

evolution of intron sequences is only an approximation because some introns contain embedded genes (e.g., coding for microRNAs) and can be constrained for other reasons as well (Haddrill et al. 2005; Gazave et al. 2007). Therefore, the estimates obtained here represent the low bound of the selective pressure affecting exons of lincRNAs.

Negative Correlation between Evolutionary Rates and Expression Levels of lincRNAs

The range of expression levels of lincRNAs measured using microarrays is narrow, with the vast majority of lincRNAs having median expression <9 ($\log_2$-normalized A-values, see Materials and Methods) (supplementary figure S2, Supplementary Material online). This distribution is quite different from the distribution of the expression levels of protein-coding genes, which includes a much greater fraction of highly expressed genes (supplementary figure S3, Supplementary Material online). In addition to microarrays, we also used the raw number of ESTs hits as an alternative measure of expression level (supplementary figure S4 and tables S4 and S5, Supplementary Material online). There was a strong highly significant correlation between the expression levels extracted from the microarray data and those obtained using EST (Pearson CC $= 0.39$, $P < 10^{-62}$) (supplementary figure S5, Supplementary Material online).

Table 3 summarizes the Pearson, Spearman, and Kendall correlation coefficients between evolutionary distances and the median expression levels of lincRNAs estimated using microarrays. For the exons of mouse lincRNAs, statistically significant negative correlation was consistently observed between the sequence evolution rate and the expression level, with correlation coefficients mostly in the range of 0.1–0.16 (table 3, figs. 2*A*–*C*, and supplementary figure S6, Supplementary Material online). In contrast, for introns, the correlation coefficients were very low and statistically not significant, that is, there was negligible or no connection between evolutionary rate and expression (table 3). The results for the human lincRNAs corroborated the mouse data and also revealed negative correlations for exons but not for introns, although the statistical significance of the results was inevitably lower due to the smaller size of the data set (fig. 2*D*, supplementary figure S7 and table S6, Supplementary Material online). As an alternative measure of expression level, we employed EST counts and showed that the rate-expression correlation pattern obtained using this approach was similar to the microarray results, that is, there was a significant negative correlation for human and mouse lincRNA exons but not for introns (supplementary tables S7 and S8, Supplementary Material online). Using maximum expression levels or the 75% quantile of the distribution of the expression levels instead of the median produced similar results, confirming that the observed negative correlation between expression and evolutionary rate is a robust property of lincRNAs (supplementary tables S9 and S10, Supplementary Material online). Analysis of the expression breadth across the tissues also showed a significant negative correlation between the breadth and the lincRNA

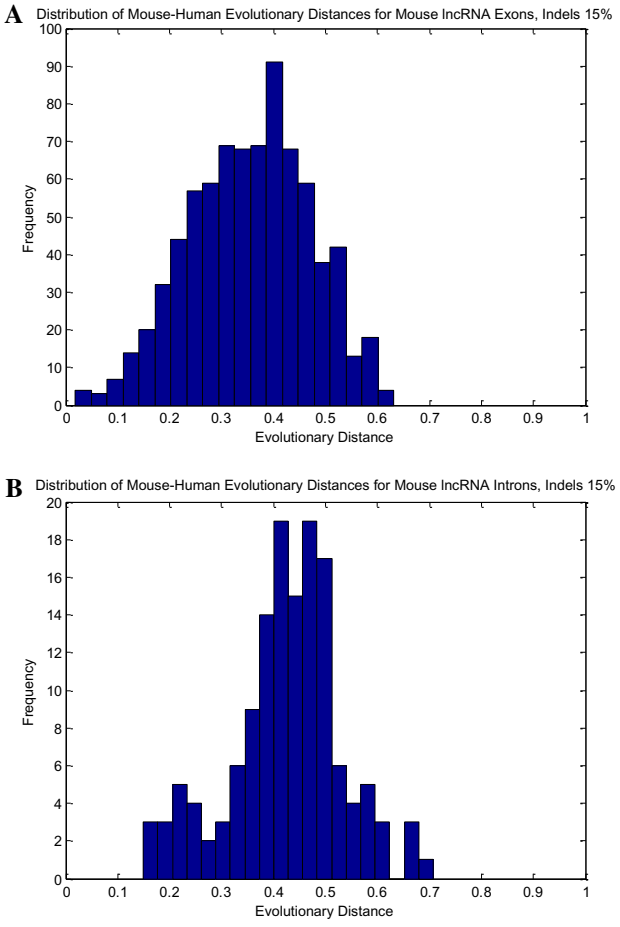

Fig. 1.—Distribution of evolutionary distances for exons (*A*) and introns (*B*) of the orthologous lincRNAs from human and mouse. All exon and intron sequences from each gene from the respective data sets were concatenated prior to the analysis.

evolutionary rates although the typical correlation coefficients (0.04–0.12) were smaller compared with the median of expression and EST counts (results not shown).

Further analysis using the mouse RNA-seq data strongly supported the consistent negative correlation between the expression of lincRNAs and the rate of evolution of lincRNA exons (fig. 3 and table 4). In this case, the observed correlations were a uniformly highly statistically significant support: all *P* values were <0.000001. The correlation coefficients obtained when the RNA-seq data were used as the measure of expression (table 4) were larger than the corresponding correlation coefficients obtained with the microarray data (table 3). This difference could be due to the truncation of microarray data (the threshold 7.0 was applied, see Materials and Methods) because of which the distributions of expression data were asymmetrical (fig. 2), causing problems for correlation analysis. The RNA-seq data showed no such asymmetry (fig. 3).

The authors of the NRED database have employed different techniques to remove contaminating protein-coding genes from the lincRNA data set (Dinger et al. 2009). Nevertheless, to control for the possibility that the observed correlations were caused by a contamination of the set of lincRNAs with protein-coding genes, we removed all sequences with significant similarity to protein-coding genes from the mouse lincRNA data set and recalculated the

Table 3

Correlations between Evolutionary Rates and Expression Levels (Microarrays) of Mouse lincRNAs

Species	Human	Chimp	Macaque	Rat	Dog
Exons, Indels: 15%					
Pearson	−0.105	−0.157	−0.139	−0.113	−0.143
P value	0.0040	<0.0001	0.0003	<0.0001	0.0009
Spearman	−0.112	−0.132	−0.121	−0.107	−0.142
P value	0.0017	0.0004	0.0016	<0.0001	0.0008
Kendall	−0.074	−0.087	−0.080	−0.070	−0.093
P value	0.0019	0.0005	0.0018	<0.0001	0.0011
Datapoints	779	720	684	1735	558
Exons, Indels: 30%					
Pearson	−0.103	−0.128	−0.108	−0.099	−0.114
P value	0.0006	<0.0001	0.0007	<0.0001	0.0009
Spearman	−0.102	−0.112	−0.095	−0.099	−0.123
P value	0.0005	0.0002	0.0022	<0.0001	0.0003
Kendall	−0.067	−0.074	−0.064	−0.065	−0.082
P value	0.0006	0.0003	0.0023	<0.0001	0.0003
Datapoints	1148	1096	1027	1930	877
Exons, Indels: 45%					
Pearson	−0.113	−0.117	−0.103	−0.098	−0.105
P value	<0.0001	<0.0001	0.0003	<0.0001	0.0005
Spearman	−0.098	−0.097	−0.091	−0.097	−0.100
P value	0.0002	0.0003	0.0010	<0.0001	0.0007
Kendall	−0.065	−0.064	−0.060	−0.064	−0.066
P value	0.0003	0.0003	0.0012	<0.0001	0.0008
Datapoints	1411	1381	1286	1950	1138
Introns, Indels: 15%					
Pearson	−0.014	0.004	−0.004	−0.011	−0.009
P value	0.8696	0.9646	0.9653	0.7617	0.9322
Spearman	−0.018	−0.001	−0.026	−0.029	−0.043
P value	0.8345	0.9922	0.7823	0.4053	0.6907
Kendall	−0.010	0.005	−0.015	−0.019	−0.028
P value	0.8655	0.9324	0.8071	0.4202	0.7019
Datapoints	141	117	115	814	89
Introns, Indels: 30%					
Pearson	−0.014	−0.017	−0.038	−0.015	−0.038
P value	0.81701	0.7919	0.5721	0.6525	0.5952
Spearman	−0.021	−0.063	−0.047	−0.031	−0.067
P value	0.7424	0.3421	0.4890	0.3592	0.3511
Kendall	−0.013	−0.041	−0.031	−0.020	−0.045
P value	0.7571	0.3532	0.4869	0.3782	0.3569
Datapoints	259	230	221	885	194
Introns, Indels: 45%					
Pearson	−0.009	−0.013	−0.025	−0.010	−0.043
P value	0.8450	0.7926	0.6252	0.7540	0.4421
Spearman	−0.003	−0.021	−0.056	−0.024	−0.046
P value	0.9529	0.6700	0.2814	0.4777	0.4086
Kendall	−0.001	−0.012	−0.038	−0.015	−0.029
P value	0.9625	0.6993	0.2735	0.5041	0.4341
Datapoints	458	433	375	895	320

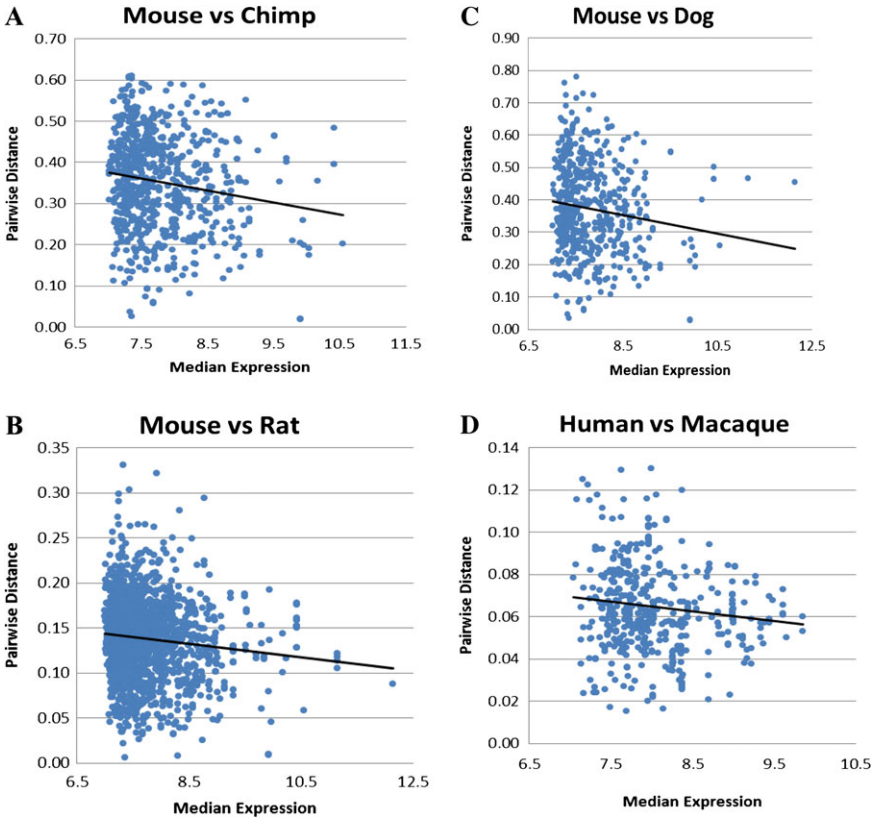

Fig. 2.—Correlation between the expression level and evolutionary rate for mouse (*A*–*C*) and human (*D*) lincRNAs based on microarray data. The data are for the indel threshold = 15%.

correlation coefficients (see Materials and Methods). The correlation between the evolutionary rate and expression level remained negative and statistically significant in all cases (supplementary table S11, Supplementary Material online). We then performed an additional experiment to control for a possible admixture of protein-coding genes in the analyzed lincRNA data sets: the coding potential of lincRNAs was predicted using the SYNCOD program (Rogozin et al. 1999). This method has been shown to produce a relatively low rate of overpredicted protein-coding regions (Rogozin et al. 1999). The SYNCOD analysis identified 94 potential protein-coding regions in 2390 lincRNAs (exons only) and 527 potential protein-coding regions in 2462 introns of lincRNA genes. The mean density of protein-coding regions in lincRNAs (one potential protein-coding region per 47 Kb) was close to that in introns (one per 45 Kb). The frequency of potential protein-coding regions was similar in the direct and complementary strands for both sets (~50%, all differences are not statistically significant). The false positive rate for the SYNCOD method has been estimated at ~0.06–0.07 (Rogozin et al. 1999), which is similar to the fraction of potential protein-coding regions in the exons (94/2390 = 0.04). Thus, taken together, the results of these analyses appear to effectively rule out a significant contamination of the analyzed lincRNA data set with protein-coding genes.

To control for the possibility that the observed negative correlation could be (at least, partially) due to regional substitution biases across the genome (Resch et al. 2007), we analyzed the substitution rate of exons divided by the substitution rate of introns within the same gene (Ke/Ki, supplementary table S12, Supplementary Material online). This ratio is analogous to K_a/K_S and is expected to reflect the strength of purifying selection that affects the exons of lincRNAs. Negative correlation between the Ke/Ki ratio and the expression level was consistently observed although some values were not statistically significant due to small sample sizes (supplementary table S12, Supplementary Material online). Thus, a moderate but highly significant negative correlation between the evolutionary rates (or selection strengths) and the expression levels of human and mouse lincRNA exons is a consistent feature of the evolution of lincRNAs.

We further sought to compare the magnitude of the negative correlation between the evolutionary rate and the expression level for lincRNAs and for protein-coding genes. Rates of nonsynonymous substitutions and synonymous substitutions were calculated using human–mouse pairwise alignments of protein-coding genes. For the purpose of this comparison, we used a sampling procedure that was repeated 1,000 times (for details, see table 5). Each sample

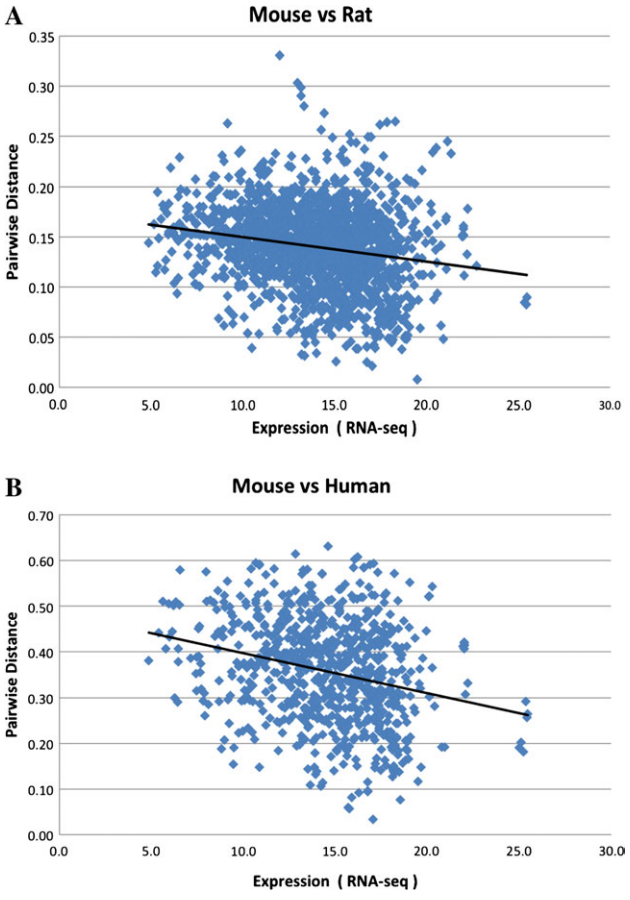

Fig. 3.—Correlation between expression level and evolutionary rate for mouse lincRNAs based on RNA-Seq data. The data are for the indel threshold = 15%.

of the orthologous protein-coding genes from the human and mouse had the size and the mean evolutionary distance approximately equal (according to the Student's *t*-test, table 5) to those for lincRNA pairwise comparisons (table 2). For all genes used to draw these control samples, the K_a/K_S values were below unity (supplementary figure S8, Supplementary Material online) indicating that these were bona fide protein-coding genes. Analysis of correlations between the evolutionary rate and the expression level for these protein-coding genes showed that Pearson correlation coefficient values for lincRNAs were well within the distributions of correlation coefficients for nonsynonymous substitution rates (K_a) across the samples of protein-coding genes, and in some cases, on the negative tail of these distributions (table 5 and fig. 4). Thus, the negative correlation between the evolutionary rate and expression level of lincRNAs is at least as strong as that for nonsynonymous substitution rates in the mammalian protein-coding genes at the same level of divergence. When compared with the synonymous rates in the same samples of protein-coding genes, lincRNAs showed a stronger correlation (table 5 and fig. 4).

Connections between Secondary Structure and Expression of lincRNAs

The demonstration that the magnitude of the correlation between the evolutionary divergence and the expression level is similar for lincRNAs and for protein-coding genes raises the key question about the biological factors underpinning such correlations. It has been shown that RNA folding is crucial for mRNA stability and functionality and is correlated with the expression level and breadth (Nackley et al. 2006; Shabalina et al. 2006; Parmley and Hurst 2007; Zhang et al. 2010). Here we analyzed folding characteristics of lincRNAs and the connections between predicted lincRNA secondary structure stability, expression level, and the rate of evolution. We found an abundance of predicted stable folding (calculated as the fraction of paired nucleotides in the optimal folding of the full-length transcript) in the lincRNA data set. The distributions of the fraction of base-paired nucleotides were similar for lincRNAs and the mRNA control set (supplementary figure S9, Supplementary Material online), which was compiled taking into account the nucleotide content, length, and gene structure of the lincRNAs (Shabalina et al. 2006). A significant positive correlation was detected between the fraction of paired nucleotides in the predicted optimal folding of mouse lincRNAs and their expression level, which was calculated from the EST counts (fig. 5*A*) or from GenAtlas 2 database of the microarray data (fig. 5*B*). A similar connection between the folding and the expression level was observed for the human lincRNAs (supplementary table S13, Supplementary Material online). Free energy (ΔG) normalized against the transcript length in the optimal folding showed the same trend (data not shown).

To disentangle the relationships between the structural features, evolution and expression of lincRNAs, we constructed a linear regression model to predict gene expression patterns based on RNA folding and/or evolutionary rates. Taking into account the connection between the expression level and the evolutionary rate of lincRNA genes, linear regression analysis showed that the evolutionary variable (Ke for the mouse–rat comparison) was predictive with respect to the expression level of the mouse lincRNA genes (EST abundance), independent of RNA structural features ($R = 0.122$ on the validation set and $R = 0.105$ on the training set). Conversely, a model that used the RNA structural parameter (fraction of paired nucleotides in the mouse lincRNA folding, PRF) alone yielded $R = 0.167$ on the validation set and $R = 0.14$ on the training set. The two variables had orthogonal predictive power, that is, R^2 values for cumulative structural and evolutionary predictions ($R^2 = 0.033 = 0.182^2$) were close to the sum of R^2 (0.033–0.036 = 0.023 + 0.013) values for independent structural ($R^2 = 0.023 = 0.152^2$) and evolutionary ($R^2 = 0.013 = 0.112^2$) predictions. Thus, the multiple regression models indicate that the two variables, Ke and PRF, independently correlate with lincRNA gene expression (*F*-test, $P < 0.01$). These findings are in agreement with the

Table 4
Correlation between the Evolutionary Rates and Expression Levels (RNA-seq) for Mouse

Species	Human	Chimp	Macaque	Rat	Dog
Exons, Indels: 15%					
Pearson	−0.224	−0.256	−0.223	−0.178	−0.252
P value	<0.000001	<0.000001	<0.000001	<0.000001	<0.000001
Spearman	−0.247	−0.279	−0.236	−0.203	−0.280
P value	<0.000001	<0.000001	<0.000001	<0.000001	<0.000001
Kendall	−0.168	−0.190	−0.162	−0.138	−0.190
P value	<0.000001	<0.000001	<0.000001	<0.000001	<0.000001
Data points	772	712	678	1766	547
Exons, Indels: 30%					
Pearson	−0.248	−0.242	−0.249	−0.171	−0.278
P value	<0.000001	<0.000001	<0.000001	<0.000001	<0.000001
Spearman	−0.271	−0.268	−0.268	−0.196	−0.305
P value	<0.000001	<0.000001	<0.000001	<0.000001	<0.000001
Kendall	−0.185	−0.181	−0.182	−0.134	−0.207
P value	<0.000001	<0.000001	<0.000001	<0.000001	<0.000001
Data points	1136	1086	1016	1851	863
Exons, Indels: 45%					
Pearson	−0.234	−0.224	−0.227	−0.181	−0.269
P value	<0.000001	<0.000001	<0.000001	<0.000001	<0.000001
Spearman	−0.257	−0.250	−0.247	−0.199	−0.288
P value	<0.000001	<0.000001	<0.000001	<0.000001	<0.000001
Kendall	−0.174	−0.168	−0.167	−0.135	−0.195
P value	<0.000001	<0.000001	<0.000001	<0.000001	<0.000001
Data points	1402	1366	1273	1873	1124

corresponding observations for mRNAs (SAS, unpublished data). Consistent with these observations, we did not find significant correlations between the evolutionary rates of lincRNAs and the predicted folding.

Discussion

The lincRNAs comprise a substantial part of the mammalian RNome but very little is currently known about their functions and evolution. Together with previous observations, the results described here suggest (even if indirectly) that many lincRNAs are indeed functional molecules that are subject to relatively weak but significant purifying selection as determined from the Ke/Ki ratio. As such, lincRNAs genes provide evolutionary biologists with a unique data set to investigate the general and more idiosyncratic features of evolution by comparing their evolutionary patterns with those of protein-coding genes. Unlike the highly conserved

Table 5
Correlations between Evolutionary Rates and Expression Levels (Microarrays) for Samples of Alignments of Orthologous Protein-Coding Genes from Human and Mouse Simulating lincRNA Sets

Comparison	Mean Correlation Coefficient CC_{PC}	Fraction of Samples with the $CC_{PC} \leq CC_{lincRNA}$	95% Confidence Intervals for CC_{PC}
Nonsynonymous sites			
Human–Chimp	−0.16	0.08	−0.08: −0.22
Human–Macaque	−0.14	0.62	−0.06: −0.20
Human–Dog	−0.06	0.98	+0.07: −0.17
Mouse–Rat	−0.10	0.44	−0.05: −0.15
Synonymous sites			
Human–Chimp	−0.04	0.84	+0.04: −0.12
Human–Macaque	−0.04	0.98	+0.04: −0.13
Human–Dog	−0.05	0.99	+0.09: −0.19
Mouse–Rat	−0.04	0.94	+0.02: −0.11

NOTE.—To compare protein-coding genes (PC) and lincRNAs, we used a sampling procedure repeated 1,000 times. Each sample has the size and the mean value of evolutionary distance approximately equal to those for the subsets of the lincRNAs (Table 1 and supplementary table 1, Supplementary Material online), the difference between the mean evolutionary distance for the PC genes and the mean distance for the lincRNAs is not significant according to the Student *t*-test. Pearson correlation coefficient was used to measure correlation between the expression and the divergence for protein-coding genes (CC_{PC}). The median value of Pearson correlation coefficients for 15%, 30%, and 45% thresholds was used as $CC_{lincRNA}$. For pairwise comparisons other than those listed in the table, the sampling procedure did not converge due to insufficient number of protein-coding genes with large evolutionary distance.

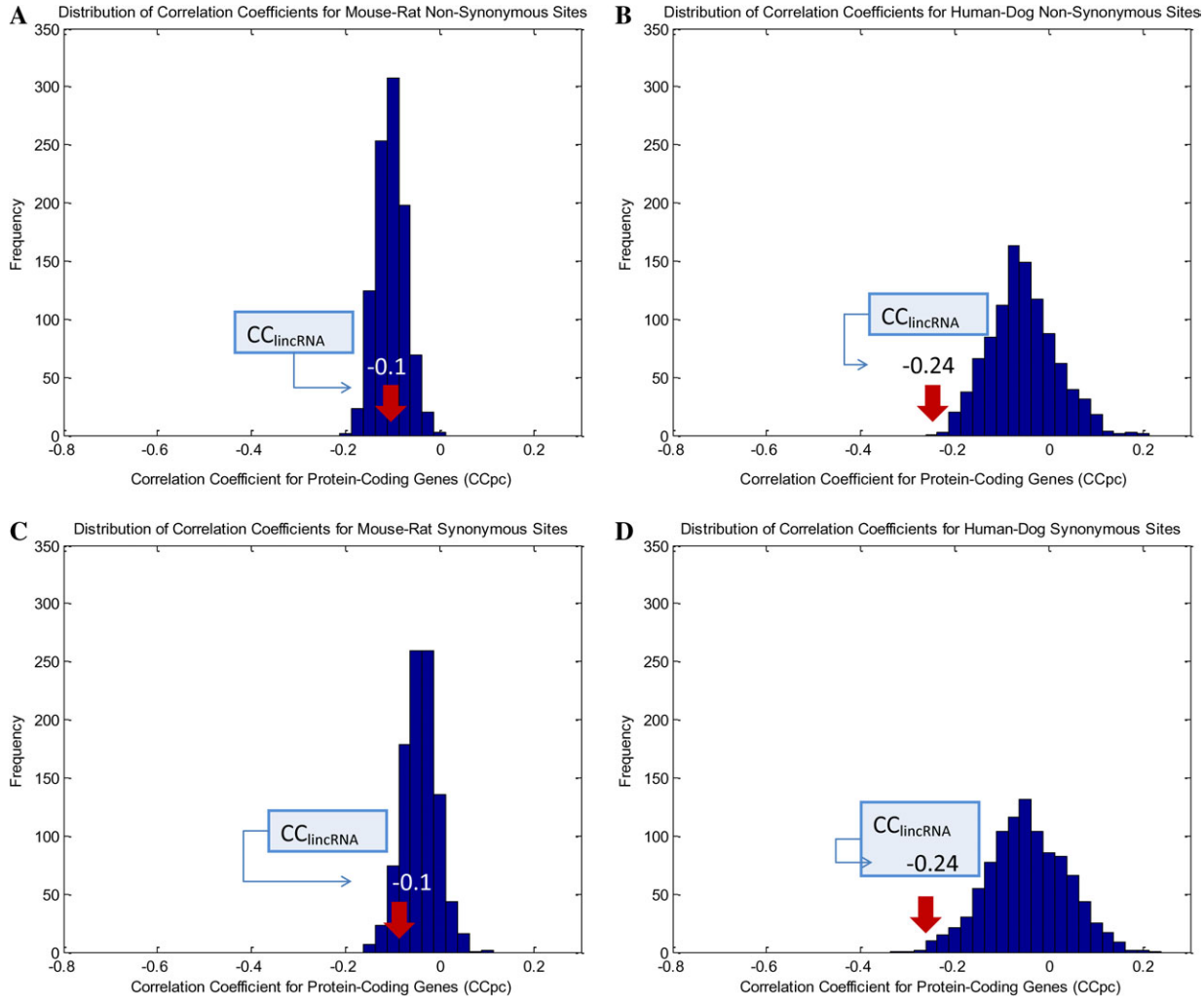

Fig. 4.—Distributions of correlation coefficients between the evolutionary rates and the expression levels for samples of alignments of the human–mouse protein-coding genes and lincRNAs. (*A*) Mouse–rat, nonsynonymous sites. (*B*) Human–dog, nonsynonymous sites. (*C*) Mouse–rat, synonymous sites. (*D*) Human–dog, synonymous sites. The distributions are for 1,000 samples of protein-coding genes simulating the lincRNA sets (for details see text and Table 5). The lincRNA correlation coefficients are shown by red arrows. The expression values were from the human and mouse microarray data sets.

structural RNAs (rRNAs and tRNAs) or small microRNAs, the lincRNA genes closely resemble protein-coding genes in terms of diversity, size, and gene architecture. The fundamental difference is that the transcripts of these genes are not translated into proteins but rather function directly as RNA molecules. Evolution of protein-coding genes shows correlations of varying strengths with several molecular phenomic variables (Koonin and Wolf 2006; Wolf et al. 2006). The most consistent and typically strongest is the negative correlation between the rate of sequence evolution and expression level of protein-coding genes or protein abundance (Drummond and Wilke 2008, 2009; Wolf et al. 2010). This relationship between evolution and expression of protein-coding genes inspired the hypothesis that evolution of proteins is driven primarily by selection for robustness to misfolding, which is partly caused by the errors of translation (Drummond and Wilke 2008, 2009). Evolutionary models built on the assumption that the deleterious effect of misfolding is the primary fitness cost associated with mutations in the protein-coding genes have been shown to be compatible both with the dependency between the evolutionary rate and expression and with the universal distribution of the evolutionary rates of protein evolution (Drummond and Wilke 2009; Lobkovsky et al. 2010). In view of this unifying hypothesis of protein evolution, we were interested to determine whether the evolution of lincRNAs is similarly connected with expression.

The results presented here reveal the existence of a relatively weak but consistent and highly significant negative correlation between the evolutionary rate and expression

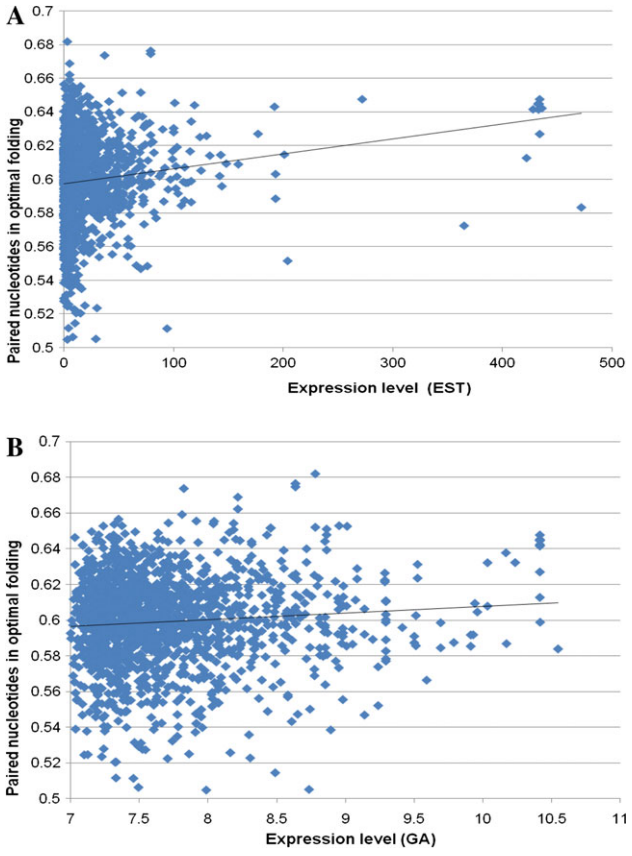

Fig. 5.—Correlation between the predicted level of nucleotide pairing in optimal folding and expression level for mouse lincRNAs measured by EST abundance (*A*) and estimated from GenAtlas database (*B*).

level of lincRNAs. Introns of lincRNA genes provide an internal control: the absence of correlation for the intronic sequences indicates that the observed connection between evolution and expression has to do with structure and function (or robustness to malfunction) of the mature lincRNA molecules. We further showed that the level of correlation between evolutionary distances and expression is similar for lincRNAs and protein-coding genes evolving under comparable constraints. The connection between expression and evolution in mammals is relatively weak for both lincRNA and protein-coding genes, with only 1–2% of the variance in evolutionary rates accounted for by expression. These findings are compatible with the previous observations that the negative correlation between the sequence evolutionary rate and the expression level is the weakest in mammals among all tested model organisms (Drummond and Wilke 2008). It seems most likely that this limited dependency is caused by the general weakness of purifying selection in mammals due to their characteristic low effective population sizes (Lynch and Conery 2003; Lynch 2006). Accordingly, mammals might not be the best choice of the model to study the causes of the dependency between evolution and expression for protein-coding gene. However, by the same token, this seems to be the only model on which a comparison of the evolutionary regimes of protein-coding genes and "protein-like" lincRNAs is possible because large diverse repertoires of long ncRNAs apparently could not evolve in organisms subject to strong selective constraints (Lynch 2007; Koonin and Wolf 2010).

We then examined potential connections between the predicted stability of lincRNA folding, their expression, and the rate of evolution. A limited in magnitude but significant positive correlation was detected between the predicted folding and expression: lincRNA molecules with greater folding potential show a tendency to be highly expressed. A positive correlation between the (predicted) RNA stability and expression level has been described previously for mammalian mRNAs (Shabalina et al. 2006). However, we found no significant link between folding and the rate of evolution of lincRNAs and further observed that RNA folding and sequence evolution rate contributed to the expression level of lincRNAs independently.

The findings reported here show that the link between evolution and expression is a fundamental dependency that is not limited to protein-coding genes. Whether or not the deleterious effects of misfolding, leading to the formation of nonfunctional protein or RNA molecules, represent the principal factor behind this universal link remains to be determined. Certainly, the process of RNA folding is fundamentally different from protein folding as the two processes are based on different types of molecular interactions. Nevertheless, there is also undeniable general similarity between the folding processes of these two classes of biomolecules. Indeed, both proteins and RNAs are heteropolymers that fold to form well-defined secondary structure elements through local interactions followed by the formation of a unique 3D conformation through nonlocal interactions. Moreover, RNA misfolding is common if not thoroughly understood, and the increasingly apparent prevalence of RNA chaperones attests to its biological relevance (Cristofari and Darlix 2002; Bhaskaran and Russell 2007; Rajkowitsch et al. 2007; Russell 2008; Semrad 2011). At face value, the observations reported here on the lack of connection between predicted RNA folding and evolutionary rate and the independence of the contributions of predicted folding and evolutionary rate to lincRNA expression can be taken as argument against a causal connection between lincRNA misfolding and the evolution–expression coupling. However, these observations should be interpreted with much caution. Prediction of the base-pairing potential is a blunt instrument that certainly does not reveal the true complexity of the RNA folding process and might not be able to distinguish well between correctly folded and misfolded RNA molecules. However, according to our estimations, about 60% of nucleotides are paired in lincRNAs and mRNAs (Shabalina et al. 2006), which is comparable

with the base pairing values for some experimentally characterized mRNAs (Kertesz et al. 2010). Also, some of the local predicted structures for lincRNAs are in agreement with the structures predicted by biochemical probing, for example, for the A region of Xist RNA (Maenner et al. 2010).

The distinct possibility remains that misfolded lincRNAs are deleterious similar to misfolded proteins, and this effect might explain the connection between their evolutionary rate and expression. Certainly, alternative explanations for this universal link could be relevant as well, for example, the potentially greater number of both functional and nonfunctional interactions in highly expressed proteins and RNAs constraining their evolution. Furthermore, it is impossible to rule out that, although the correlations between expression and evolution are of the same sign and similar in magnitude for proteins and lincRNAs, the underlying causes are substantially different (even if this possibility is less than parsimonious).

Conclusions

The functions of the numerous lncRNAs remain largely unknown but the results presented here support previous findings that many of these RNAs are subject to purifying selection, albeit relatively weak, and so are predicted to be functional. We found that lincRNAs recapitulate the universal negative correlation between the evolutionary rate and expression that has been reported for protein-coding genes from diverse model organisms. Moreover, the magnitude of the correlation for the lincRNAs was comparable to the magnitude of the correlation that we identified in equal-sized control sets of protein-coding genes with levels of sequence conservation similar to those observed for lincRNAs. The expression level of the lincRNAs also was significantly and positively correlated with the predicted extent of lincRNA molecule folding (base paring). However, there was no significant correlation between lincRNA folding and evolutionary rate, and the contributions of the evolutionary rate and folding to the expression level were found to be independent. The results of this work indicate that the anticorrelation between evolutionary rate and expression level is a general feature of gene evolution. The causative factors behind this fundamental dependency that might include similar fitness effects of protein and RNA misfolding remain to be elucidated.

Supplementary Material

Supplementary figures S1–S9 and tables S1–S13 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Liran Carmel, Joshua Cherry, Jean Thierri-Mieg, Mikhail Galperin, Alexander Lobkovsky, Kira Makarova, and Yuri Wolf for useful discussions. This work was supported by the Intramural Research Program of the National Library of Medicine at National Institutes of Health (US Department Health and Human Services).

Literature Cited

Beltran M, et al. 2008. A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. Genes Dev. 22:756–769.

Bertone P, et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. Science 306:2242–2246.

Bhaskaran H, Russell R. 2007. Kinetic redistribution of native and misfolded RNAs by a DEAD-box chaperone. Nature 449:1014–1018.

Blankenberg D, et al. 2010. Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. Chapter 19:Unit 19.10.1–19.10.21.

Bohning D, Dietz E, Schlattmann P. 1998. Recent developments in computer-assisted analysis of mixtures. Biometrics 54:525–536.

Bohning D, Schlattmann P, Lindsay B. 1992. Computer-assisted analysis of mixtures (C.A.MAM): statistical algorithms. Biometrics 48:283–303.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Carninci P, et al. 2005. The transcriptional landscape of the mammalian genome. Science 309:1559–1563.

Centonze D, et al. 2007. The brain cytoplasmic RNA BC1 regulates dopamine D2 receptor-mediated transmission in the striatum. J Neurosci. 27:8885–8892.

Chen J, et al. 2005. Human antisense genes have unusually short introns: evidence for selection for rapid transcription. Trends Genet. 21:203–207.

Chodroff RA, et al. 2010. Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biol. 11:R72.

Clemson CM, McNeil JA, Willard HF, Lawrence JB. 1996. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. J Cell Biol. 132:259–275.

Cristofari G, Darlix JL. 2002. The ubiquitous nature of RNA chaperone proteins. Prog Nucleic Acid Res Mol Biol. 72:223–268.

Dinger ME, et al. 2009. NRED: a database of long noncoding RNA expression. Nucleic Acids Res. 37:D122–D126.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352.

Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet. 10:715–724.

Duret L, Chureau C, Samain S, Weissenbach J, Avner P. 2006. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science 312:1653–1655.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 17:68–74.

Elisaphenko EA, et al. 2008. A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. PLoS One 3:e2521.

Espinoza CA, Goodrich JA, Kugel JF. 2007. Characterization of the structure, function, and mechanism of B2 RNA, an ncRNA repressor of RNA polymerase II transcription. RNA 13:583–596.

Felsenstein J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 266:418–427.
Feng J, et al. 2006. The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev. 20:1470–1484.
Gazave E, Marques-Bonet T, Fernando O, Charlesworth B, Navarro A. 2007. Patterns and rates of intron divergence between humans and chimpanzees. Genome Biol. 8:R21.
Goecks J, Nekrutenko A, Taylor J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11:R86.
Goodrich JA, Kugel JF. 2006. Non-coding-RNA regulators of RNA polymerase II transcription. Nat Rev Mol Cell Biol. 7:612–616.
Gruber AR, Neubock R, Hofacker IL, Washietl S. 2007. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. Nucleic Acids Res. 35:W335–W338.
Guttman M, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458:223.
Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content. Genome Biol. 6:R67.
Hirota K, et al. 2008. Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. Nature 456:130–134.
Hoffman MM, Birney E. 2007. Estimating the neutral rate of nucleotide substitution using introns. Mol Biol Evol. 24:522–531.
Huarte M, et al. 2010. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 142:409–419.
Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. 18:486.
Jordan IK, Mariño-Ramírez L, Wolf YI, Koonin EV. 2004. Conservation and coevolution in the scale-free human gene coexpression network. Mol Biol Evol. 21:2058–2070.
Kapranov P, et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316: 1484–1488.
Karolchik D, et al. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32:D493–D496.
Kertesz M, et al. 2010. Genome-wide measurement of RNA secondary structure in yeast. Nature 467:103–107.
Khaitovich P, et al. 2006. Functionality of intergenic transcription: an evolutionary comparison. PLoS Genet. 2:e171.
Kondrashov AS, Shabalina SA. 2002. Classification of common conserved sequences in mammalian intergenic regions. Hum Mol Genet. 11:669–674.
Koonin EV, Wolf YI. 2006. Evolutionary systems biology: links between gene evolution and function. Curr Opin Biotechnol. 17:481–487.
Koonin EV, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. Nat Rev Genet. 11:487–498.
Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13:2229–2235.
Lee S, et al. 2011. Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. Nucleic Acids Res. 39:e9.
Lobkovsky AE, Wolf YI, Koonin EV. 2010. Universal distribution of protein evolution rates as a consequence of protein folding physics. Proc Natl Acad Sci U S A. 107:2983–2988.
Louie E, Ott J, Majewski J. 2003. Nucleotide frequency variation across human genes. Genome Res. 13:2594–2601.
Lynch M. 2006. The origins of eukaryotic gene structure. Mol Biol Evol. 23:450–468.
Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A. 104(Suppl 1): 8597–8604.
Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401–1404.
Maenner S, et al. 2010. 2-D structure of the A region of Xist RNA and its implication for PRC2 association. PLoS Biol. 8:e1000276.
Mariner PD, et al. 2008. Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. Mol Cell. 29:499–509.
Marques AC, Ponting CP. 2009. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol. 10:R124.
Martens JA, Laprade L, Winston F. 2004. Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. Nature 429:571–574.
Martianov I, et al. 2007. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. Nature 445:666–670.
Mattick JS, Makunin IV. 2006. Non-coding RNA. Hum Mol Genet. 15:R17–R29 Spec No 1.
Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. Nat Rev Genet. 10:155–159.
Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. 2008. Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci U S A. 105:716–721.
Munroe SH, Lazar MA. 1991. Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA. J Biol Chem. 266:22083–22086.
Nackley AG, et al. 2006. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science 314:1930–1933.
Nagano T, et al. 2008. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science 322:1717–1720.
Nesterova TB, et al. 2001. Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. Genome Res. 11:833–849.
Ogurtsov AY, et al. 2008. Expression patterns of protein kinases correlate with gene architecture and evolutionary rates. PLoS One 3:e3599.
Ogurtsov AY, Roytberg MA, Shabalina SA, Kondrashov AS. 2002. OWEN: aligning long collinear regions of genomes. Bioinformatics 18:1703–1704.
Ogurtsov AY, Shabalina SA, Kondrashov AS, Roytberg MA. 2006. Analysis of internal loops within the RNA secondary structure in almost quadratic time. Bioinformatics 22:1317–1324.
Okazaki Y, et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420:563–573.
Osato N, Suzuki Y, Ikeo K, Gojobori T. 2007. Transcriptional interferences in cis natural antisense transcripts of humans and mice. Genetics 176:1299–1306.
Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158:927–931.

Pandey RR, et al. 2008. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell. 32:232–246.
Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet. 22:1–5.
Parmley JL, Hurst LD. 2007. How do synonymous mutations affect fitness? Bioessays 29:515–519.
Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. 17:556–565.
Ponting CP, Belgard TG. 2010. Transcribed dark matter: meaning or myth? Hum Mol Genet. 19:R162–R168.
Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long noncoding RNAs. Cell 136:629–641.
Rajkowitsch L, et al. 2007. RNA chaperones, RNA annealers and RNA helicases. RNA Biol. 4:118–130.
Rearick D, et al. 2010. Critical association of ncRNA with introns. Nucleic Acids Res. 39:2357–2366.
Resch AM, et al. 2007. Widespread positive selection in synonymous sites of mammalian genes. Mol Biol Evol. 24:1821–1831.
Rogozin IB, D'Angelo D, Milanesi L. 1999. Protein-coding regions prediction combining similarity searches and conservative evolutionary properties of protein-coding sequences. Gene 226:129–137.
Russell R. 2008. RNA misfolding and the action of chaperones. Front Biosci. 13:1–20.
Semrad K. 2011. Proteins with RNA chaperone activity: a world of diverse proteins with a common task-impediment of RNA misfolding. Biochem Res Int. 2011:532908.
Shabalina SA, Ogurtsov AY, Spiridonov NA. 2006. A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic Acids Res. 34:2428–2437.
Siepel A, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.
Umlauf D, et al. 2004. Imprinting along the Kcnq1 domain on mouse chromosome 7 involves repressive histone methylation and recruitment of Polycomb group complexes. Nat Genet. 36: 1296–1300.
van Bakel H, Hughes TR. 2009. Establishing legitimacy and function in the new transcriptome. Brief Funct Genomic Proteomic. 8:424–436.
Wang H, et al. 2005. Dendritic BC1 RNA in translational control mechanisms. J Cell Biol. 171:811–821.
Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci U S A. 102:2454–2459.
Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. Proc R Soc B Biol Sci. 273:1507–1515.
Wolf YI, Gopich IV, Lipman DJ, Koonin EV. 2010. Relative contributions of intrinsic structural-functional constraints and translation rate to the evolution of protein-coding genes. Genome Biol Evol. 2:190–199.
Yang JR, Zhuang SM, Zhang J. 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. Mol Syst Biol. 6:421.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Zhang F, Saha S, Shabalina SA, Kashina A. 2010. Differential arginylation of actin isoforms is regulated by coding sequence-dependent degradation. Science 329:1534–1537.
Zhang L, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol. 21:236–239.
Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31:3406–3415.

Associate editor: Bill Martin