Published online 5 July 2019

NAR Genomics and Bioinformatics, 2019, Vol. 1, No. 1 ***e2***
doi: 10.1093/nargab/lqz002

Conserved regions in long non-coding RNAs contain abundant translation and protein–RNA interaction signatures

Jorge Ruiz-Orera[1,*] and M. Mar Albà[1,2,*]

[1]Evolutionary Genomics Group, Research Programme on Biomedical Informatics, Hospital del Mar Research Institute, Universitat Pompeu Fabra, Dr Aiguader 88, Barcelona 08003, Spain and [2]Catalan Institution for Research and Advanced Studies, Passeig Lluís Companys 23, Barcelona 08010, Spain

Received May 24, 2019; Revised June 14, 2019; Editorial Decision June 24, 2019; Accepted July 04, 2019

ABSTRACT

The mammalian transcriptome includes thousands of transcripts that do not correspond to annotated protein-coding genes and that are known as long non-coding RNAs (lncRNAs). A handful of lncRNAs have well-characterized regulatory functions but the biological significance of the majority of them is not well understood. LncRNAs that are conserved between mice and humans are likely to be enriched in functional sequences. Here, we investigate the presence of different types of ribosome profiling signatures in lncRNAs and how they relate to sequence conservation. We find that lncRNA-conserved regions contain three times more ORFs with translation evidence than non-conserved ones, and identify nine cases that display significant sequence constraints at the amino acid sequence level. The study also reveals that conserved regions in intergenic lncRNAs are significantly enriched in protein–RNA interaction signatures when compared to non-conserved ones; this includes sites in well-characterized lncRNAs, such as *Cyrano*, *Malat1*, *Neat1* and *Meg3*, as well as in tens of lncRNAs of unknown function. This work illustrates how the analysis of ribosome profiling data coupled with evolutionary analysis provides new opportunities to explore the lncRNA functional landscape.

INTRODUCTION

The advent of high-throughput genomic technologies has revealed that mammalian transcriptomes are more complex than initially thought (1–4). One of the most intriguing findings has been the discovery of thousands of expressed loci that that do not correspond to protein-coding genes; these are known as long non-coding RNAs (lncRNAs) (3,5–9). Similar to coding RNAs, lncRNAs are usually polyadenylated and can have a multi-exonic structure; however, unlike coding RNAs, they lack long conserved open reading frames (ORFs) (10).

Recently, it has been found that many lncRNAs are likely to translate small proteins or peptides (11,12). The strongest evidence comes from the analysis of ribosome profiling experiments (Ribo-Seq), a high-throughput RNA sequencing technique developed to identify ribosome-protected RNA fragments (13). Ribo-Seq read coverage can provide an unambiguous signal of protein translation, arising from the codon-by-codon movement of the ribosome; the three-nucleotide periodicity of the reads is only observed in actively translated regions (14–16). Frequent translation of small ORFs in lncRNAs has been detected in species as diverse as humans (17–20), mice (11,12,21), zebrafish (22) and yeast (13,23). It is not yet clear how many of these peptides are functional. This set of translated small ORFs is highly heterogeneous, ranging from conserved functional micropeptides that have been missed in gene annotation pipelines, proteins involved in recent lineage-specific adaptations, and pervasively translated sequences that may not have a function *per se* but can act as raw material for *de novo* gene birth (16,21,24).

At present, the majority of lncRNAs lack known functions. The subset of well-characterized lncRNAs includes the X chromosome inactivation RNA, *Xist*, or the metastasis-associated transcript *Malat1* (25,26). Functional non-coding RNAs often associate with proteins forming ribonucleoprotein complexes (RNPs). For example, the telomerase RNA component (TERC) contains binding sites for the telomerase reverse transcriptase and other associated factors to form the telomerase complex (27). *Cyrano*, an lncRNA with a role in neural development, interacts with multiple proteins (28). Recent work has shown that protein interactions sites in RNAs can be identified using

*To whom correspondence should be addressed. Email: malba@imim.es
Correspondence may also be addressed to Jorge Ruiz-Orera. Email: jorge.ruizorera@mdc-berlin.de
Present address: Jorge Ruiz-Orera, Max Delbrück Center for Molecular Medicine, Berlin, Germany.

ribosome profiling data, by selecting for reads that are of a specific size and that do not display three-nucleotide periodicity (17,29). Thus, it is possible to use the same Ribo-Seq experiment to obtain information on both putative translated ORFs and protein–RNA interaction signatures.

One of the most intense debates in biology centers around whether the thousands of currently annotated lncRNAs are functional or not (30,31). Some researchers consider that many of the uncharacterized lncRNAs could have a regulatory role, modulating chromatin structure or gene expression (32). Others are more skeptical and reason that many lncRNAs are probably the result of transcriptional noise (33,34), as they are poorly conserved across species and show little sequence constraints. Despite this controversy, most researchers would agree that the set of lncRNAs that are conserved across relatively distant species is likely to be enriched in functional lncRNAs (35). Hundreds of lncRNAs show sequence conservation between mice and humans (36,37); unlike lineage-specific lncRNAs, these conserved lncRNAs display purifying selection signatures (38,39), and many of the well-studied lncRNAs contain short conserved sequence segments that are required for their function (40–42).

Studying the sequence and biochemical properties of conserved versus non-conserved lncRNAs is key to further understanding the evolution and function of these molecules. Here, we perform a thorough investigation of the types of ribosome profiling signatures in evolutionary conserved and non-conserved lncRNAs. We find that conserved lncRNAs are significantly enriched in both translated ORFs and RNPs, indicating that the acquisition of coding capacity and/or protein–RNA interactions facilitates the functionalization of young novel transcripts.

MATERIALS AND METHODS

Identification of conserved sequences in the mouse transcriptome

We retrieved mouse sequences and regulatory regions ('promoters') from Ensembl v.89 (43). We excluded pseudogenes and sense intronic lncRNAs, as the latter could represent unannotated regions in protein-coding genes. In order to avoid spurious conservation matches due to the presence of repeats and transposable elements, we masked repetitive sequences with RepeatMasker (44). The masked regions comprised 11.30% of the total coding gene (codRNA) sequence and 11.56% of the lncRNA sequence. We retained those sequences that had a minimum length of 200 nucleotides and a non-masked sequence length of at least 100 nucleotides or 25% of the total transcript length.

We searched for significant sequence similarity hits between mouse annotated transcripts and human transcripts using BLASTN (45). To ensure completeness, we used a human transcriptome that contains annotated genes as well as *de novo* assembled transcripts obtained from high-coverage RNA sequencing data (RNA-Seq) from different tissues (46), which is available from http://dx.doi.org/10.6084/m9.figshare.4702375. The BLASTN parameters we used were: -evalue 10^{-5}, -strand plus, -max_target_seqs 15, -window_size 12. Next, we defined 'conserved regions' in mouse transcripts as the ones showing significant sequence similarity (*E*-value $< 10^{-5}$) to human transcripts, with a minimum length of 30 nucleotides. Results were largely consistent for different *E*-value cut-offs: the number of conserved lncRNAs only increased by 4.65% when relaxing the *E*-value (*E*-value $< 10^{-4}$) and only decreased by 3.36% when making this parameter more stringent (*E*-value $< 10^{-6}$).

Overlapping BLASTN hits from different transcripts of the same gene were merged, so each gene had a unique set of conserved non-redundant regions. We defined the gene as codRNA if at least one of the isoforms was annotated as protein-coding; otherwise, it was defined as lncRNA. We discarded 368 mouse lncRNAs that had homology to sequences annotated as coding in human, as they might represent unannotated proteins or pseudogenized lncRNAs. Additionally, if two conserved regions were separated by <100 nucleotides (nt) we merged them; this value was chosen because only <5% of the annotated coding sequences had internal gaps longer than 100 nt. After merging, the median length of conserved regions in codRNAs increased from 124 to 343 nt (Additional File 1: Supplementary Figure S1) and we recovered more than 95% of the total coding sequence. A more moderate increase in length was observed in the case of lncRNAs, from a median length of 136 to 163 nt (Additional File 1: Supplementary Figure S1). Our method identified conserved regions in 19,779 out of 21,416 codRNAs and 1,547 out of 9,734 lncRNAs.

Analysis of mouse–human genomic synteny alignments from UCSC (47) indicated that about 80% of the mouse lncRNA conserved regions could be aligned to human syntenic regions, whereas this fraction decreased to about 50% for non-conserved regions, including many tandem repeats that were masked by BLAST and that are often over-represented in whole-genome alignments (48). We also quantified the overlap of conserved and non-conserved regions in codRNAs and lncRNAs with regions annotated as promoters in Ensembl, which covered about 1.62% of the genome. These regions are defined in Ensembl for several cell lines and tissues by a combination of data related to open chromatin regions, histone modifications and transcription factor binding assays (49).

Null model for sequences evolving under no constraints

We simulated the evolution of sequences along the mouse and human lineages with Rose (50). The objective was to create a set of alignments for sequences evolving under no constraints and then compare the substitution rates from these alignments and the substitution rates estimated from the alignment of real sequences. We used the annotated mouse lncRNA sequences as starting sequences for the simulation, as this allowed us to automatically control for GC content and k-mer composition. We used the following parameters to simulate sequence evolution in the absence of selection: HKY model with a TT ratio of 4.26, mouse branch mean substitution 0.34 and indel rate 0.018 × 2, human branch mean substitution 0.17 and indel rate 0.009 × 2, indel function = [.50,.18,.10,.08,.06,.04,.04]. These values were based on previous genomic estimates (51–53), and we implemented a 2-fold higher substitution rate in mouse than in human. After the simulations, we ran BLASTN using

the same conditions as for real sequences and recovered the alignments. Up to 59.6% of the mouse simulated sequences had at least one match in the set of human simulated sequences. This corresponded to 20.8% of the total sequence length.

Estimation of substitution rates

We estimated the number of substitutions per site (k *or* k_o) in the BLASTN alignments using the maximum likelihood method 'baseml' from the PAML package (54) with model 7 (REV). If a position was covered by several BLAST hits, we chose the one with the lowest *E*-value. We discarded k values higher than 5 as they were deemed unreliable. We observed that k values tended to be abnormally low for short alignments (<300 nt) of the simulated sequences. This is because in the case of short regions of homology, significance can only be achieved if the sequences are very well conserved. We employed linear regression to calculate the expected k (k_e) given the length of the alignment (L), using the simulated data. For real alignments of a given length, we could then calculate a normalized substitution rate k_o/k_e, which was informative on the deviation from neutrality while correcting for the effect of length. The log-linear regression model was calculated for short (<300 nt) and long (≥300 nt) neutrally evolved sequences separately:

$$\log(k_{e;L\geq 300}) = -0.468900 - L \times 7.865 \times 10^{-5}$$

$$\log(k_{e;L<300}) = -1.562833 + L \times 0.003879$$

where L represents alignment length

The two models were statistically significant according to a T-test (P-value < 0.05).

Classification of genes based on genomic location or small RNA content

Up to 20% of the total sequence length in lncRNAs had exonic overlap with other genes on the opposite strand. We divided conserved and non-conserved regions into 'overlapping' and 'non-overlapping', depending on whether the region was overlapping an antisense conserved feature (detected by BLAST or annotated as conserved in humans in Ensembl Compara). Regions shorter than 30 nt were not considered.

We classified genes into three different categories: ncRNA host, in the case of genes containing annotated small RNAs or being annotated as microRNA or small RNA hosts; antisense, in the case of genes having at least one overlapping region, being expressed from bidirectional promoters (distance between transcription start sites <2 Kb) and/or being annotated as antisense in Ensembl; or intergenic otherwise.

Analysis of ribosome profiling data

We used RNA-Seq and ribosome profiling data (Ribo-Seq) for the mouse hippocampus obtained from the Gene Expression Omnibus accession number GSE72064 (55). We merged sequencing replicates to increase the power to detect translated ORFs. We removed reads mapping to annotated rRNAs and tRNAs. Next, we mapped Ribo-Seq (361 million mapped reads) and RNA-Seq reads (435 million mapped reads) to the mouse genome (mm10) using Bowtie (v. 0.12.7, parameters -k 1 -m 20 -n 1 –best –strata) (56) and we extracted P-sites corresponding to Ribo-Seq reads as done in a previous study (21). For comparison, we analyzed Ribo-Seq data from the rat brain (rn6, 373 million mapped reads) and the human brain (hg19, 50 million mapped reads) obtained from Gene Expression Omnibus accession numbers, GSE66715 and GSE51424 (57), respectively. We mapped the Ribo-Seq reads to the corresponding syntenic genomic regions in rats and humans.

Next, we assigned strand-specific mouse reads to the different predefined transcript regions if at least 1 bp of the computed P-site (Ribo-Seq) spanned the corresponding region. We defined two metrics: a per-base coverage metric based on the number of reads spanning the region per kilobase and a total coverage based on the percentage of sequences covered by reads.

For genes expressed at very low levels, the Ribo-Seq signal may be undetectable. In order to account for this, we imposed a RNA-Seq coverage threshold in which the number of false negatives (annotated coding sequences not covered by Ribo-Seq reads) was lower than 5% (Additional File 1: Supplementary Figure S2, RNA-Seq coverage in region ≥ 56.38 reads/kb); we only considered regions with expression values above this cut-off. 'Conserved genes' were those that contained at least one conserved region above the cut-off. Finally, we eliminated 192 lncRNAs located within 4 kb from a sense protein-coding gene and/or with evidence of being part of the same gene using RNA-Seq data, as these lncRNAs may have been unannotated UTRs.

ORF translation in conserved and non-conserved regions

We predicted all possible canonical ORFs (ATG to STOP) with a minimum length of nine amino acids in the transcripts. This covered 55% of the conserved regions and 57% of the non-conserved ones. Next, we eliminated redundant ORFs by selecting only the longest ORF when several ORFs overlapped in the same frame. We used RibORF (v.0.1) (58) to predict translated ORFs with a minimum threshold of 10 mapped Ribo-Seq reads per ORF. We used the same score cut-off as in a previous study (≥0.7), which had a reported false positive rate of 3.30–4.16% and a false negative rate of 2.54% (21). We assigned translated ORFs to the different defined regions if at least 10% of the translated sequence spanned the region. For the prediction of substitution rates and proteomics analysis, we used the longest ORF per region.

Mass spectrometry data

We used an available mass-spectrometry dataset from the mouse hippocampus (PXD007150) (59) to search for peptide spectra produced by translated ORFs. Mass-spectrometry data were analyzed using Comet (r2018.01 rev. 2) (60) and setting modifications as described in the original study. Peptide tolerance was 20 ppm and the maximum number of missed cleavages was set at 2. We built three custom databases by concatenating 11,345 annotated proteins in SwissProt that were translated in the hippocampus

according to the RibORF analysis, a file with the most common contaminants, and any of the following problem sets: (i) 157 conserved codRNA genes encoding small proteins (small CDSs, <100 amino acids); (ii) 492 translated ORFs in lncRNAs; (iii) 492 SwissProt proteins sampled with the same distribution of RiboSeq reads as translated ORFs in lncRNAs, and subsequently trimmed from 3′ so they had the same length as translated ORFs in lncRNAs.

The Percolator algorithm (v3.02.1) (61) implemented in the Proteome Discoverer software was run to estimate the *q*-value by using reverse decoys as a negative control. We selected peptides with *q*-value <0.01. Finally, any ORF containing at least two peptides not found in any other gene was considered validated by proteomics. In the positive set of SwissProt, we found 36.65% of proteins containing at least two or more peptides.

d*N*/d*S* analysis in translated ORFs

We used the UCSC tool liftOver (-minMatch = 0.75) (62) to extract the corresponding ORF genomic coordinates in humans. For the cases in which we found a matching region, we aligned the ORFs with PRANK (63) and we selected sequences with matching start and stop codons. Additionally, we considered that each alignment should not contain more than 33% of gaps, and that the alignment length should be longer than 10 amino acids. In all cases, the human ORF was at least 50% the length of the mouse ORF.

Next, we used CODEML of the PAML package (54) to compute a d*N*/d*S* ratio (non-synonymous to synonymous substitution rate, or omega) for each aligned ORF. We discarded cases in which the computed d*N* or d*S* was higher than 10, as they were deemed unreliable. We tested whether this ratio was significantly different from 1 by comparing the likelihood of the model to that obtained with a fixed omega of 1. We found 9 mouse/human conserved ORFs in lncRNAs with d*N*/d*S* significantly lower than 1, with an adjusted *P*-value < 0.05.

PhyloP codon analysis in translated ORFs

We used the GenomicScores package (v. 1.2.2) available at Bioconductor (64) to compute the average PhyloP score per codon position (+1, +2, +3) in different sets of translated ORFs. PhyloP is a set of phylogenetic *P*-values for multiple alignments of 59 vertebrate genomes to the mouse genome. GenomicScores rounds PhyloP scores using a lossy compression algorithm. We compared the conservation of the third codon position to the conservation of the first and second codon positions. In functional proteins, the third position is expected to be more variable because of the degeneracy of the genetic code (65).

Analysis of protein–RNA complexes

We used Rfoot (v.0.1) and FLOSS-based measurements to predict regions in lncRNAs involved in protein–RNA interactions or ribonucleoprotein particles (RNPs). Rfoot is based on the lack of three-nucleotide periodicity as well as low Ribo-Seq coverage uniformity for the predictions (66). FLOSS is a metric based on the distribution of Ribo-Seq read lengths that measures the extent of disagreement between the observed distribution and the distribution for ribosomal associations (67). First, we subtracted predicted ORF sequences with a RibORF score ≥ 0.5 and/or read periodicity ≥ 0.66 and then we selected 60 nt sequence windows showing uniformity < 0.6 and a minimum of 10 reads. Next, we analyzed the FLOSS score of the regions, selecting those with a score ≥ 0.35, as only 5% of annotated coding sequences showed a score above 0.35. Any overlapping predicted RNPs were merged into a single RNP. We predicted RNPs in 95% of the RNAs from a positive control set composed of snRNAs and snoRNAs (10 or more Ribo-Seq reads per RNA), and in only about 5% of the translated ORFs in lncRNAs (Additional File 1: Supplementary Figure S3).

We also downloaded peak annotation files from 39 CLIP-seq datasets in POSTAR (68). The computed peaks were already mapped to the mouse genome and we directly checked the overlap with the predicted RNPs. We found positive FLOSS scores (≥0.35) for ∼90% of CLIP-seq peaks in lncRNAs with 10 or more Ribo-Seq reads (Additional File 1: Supplementary Figure S3), supporting the validity of the selected FLOSS threshold for identifying RNPs.

Definition of a set of functional lncRNAs

We obtained a list of 30 functional mouse lncRNAs expressed in the hippocampus by selecting all cases present in lncRNAdb (69) and adding four additional known lncRNAs: *Pantr1* (70), *Firre* (71), *TERC* (72) and *Norad (2900097C17Rik)* (73).

Statistical tests and plots

Plots and statistics were performed with R (74).

RESULTS

Ribo-Seq reads frequently map to lncRNAs

We searched for matches of the complete set of Ensembl mouse annotated transcripts against the human transcriptome using BLASTN (*E*-value < 10^{-5}) (45). We detected 19,779 protein-coding genes (codRNAs) and 1,547 lncRNAs that contained at least one conserved region in humans. The conserved regions were generally shorter in lncRNAs than in codRNAs with a median length of 163 and 343 nucleotides, respectively (Additional File 1: Supplementary Figure S1).

Next, we mapped ribosome profiling and RNA-Seq data to the set of annotated mouse transcripts using previously published ribosome profiling data from the mouse hippocampus (55). This dataset was chosen because of its high coverage, allowing for the unambiguous recovery of many novel translation events using three-nucleotide periodicity (21). Of the annotated transcripts, 13,345 codRNAs and 707 lncRNAs were expressed in the hippocampus dataset. We focused on these transcripts for the rest of the study. Around 98% of the codRNAs (13,084) had at least one conserved region in humans; this fraction decreased to 41% (289) for lncRNAs (Figure 1A; regions described in Additional File 2: Supplementary Tables S1, S2 and S5).

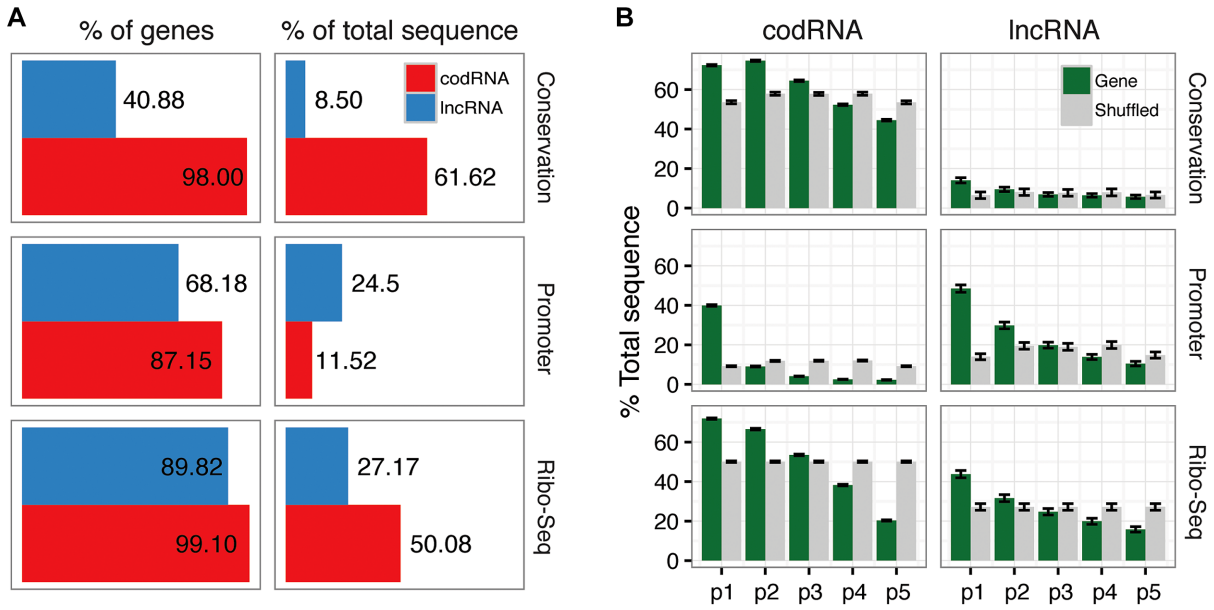

Figure 1. Transcriptome-wide identification of conserved sequences, promoters and Ribo-Seq associations. (**A**) Fraction of mouse genes that showed conservation in humans using BLASTN (Conservation), that overlapped with annotated promoter regions (Promoter), or that were covered by Ribo-Seq reads (Ribo-Seq). The percentage of genes with at least one feature, and the total sequence covered, is indicated. Data are for expressed codRNAs and lncRNAs in the hippocampus (sequences with a minimum RNA-Seq coverage of 56.38 reads/kb). (**B**) Analysis of feature coverage in equally-sized fractions of the genes, from 5′ (p1) to 3′ (p5). Grey bars represent the mean proportion of a shuffled control where the different features per gene were randomly shuffled along the sequence 1000 times. Error bars represent the standard error of the proportion.

Most functionally characterized lncRNAs (27 out of 30 cases) had at least one conserved region (see Additional File 2: Table S3); only *Firre, Adapt33* and *Snhg6* were not found in humans. In terms of the total length, 61.62% of the codRNA sequence and 8.50% of the total lncRNA sequence (25.39% in the case of functionally characterized lncRNAs) were conserved (Figure 1A). Conservation was highest in the 5′ end of the transcripts for both codRNAs and lncRNAs (Figure 1B).

Many transcripts partially overlapped sequences annotated as 'promoter' by Ensembl (49). This affected 87.15% of codRNAs and 68.18% of lncRNAs. The total fraction of the sequence covered by 'promoter' regions was 24.50% for lncRNAs and 11.52% for codRNAs (Figure 1A). As expected, regions annotated as 'promoter' were biased toward the 5′ end of the transcript (Figure 1B and Additional File 1: Supplementary Figure S4). The relatively high overlap of these regions with lncRNAs could be explained by their short size compared to codRNAs; when we focused on the 5′-most 200 nucleotides of the transcript, the percentage of sequence overlapping 'promoter' regions was actually higher for codRNAs than for lncRNAs (80.46% versus 63.38%).

Next, we investigated the presence of ribosome profiling (Ribo-Seq) footprints on the transcripts. Up to 99.10% of codRNAs and 89.82% of lncRNAs had mapped Ribo-Seq sequencing reads. Overall, 50.08% of the codRNA sequence and 27.17% of the lncRNA sequence were covered by at least one Ribo-Seq read (Figure 1A). These results are in line with recent reports of a relatively high coverage of lncRNAs by Ribo-Seq reads (12,17,18,20). We also observed that the Ribo-Seq reads showed a clear 5′ bias, for both conserved and non-conserved transcripts (Figure 1B, Additional File 1: Supplementary Figure S4).

Conserved lncRNA sequences are enriched in promoter regions and Ribo-Seq signatures

Promoter regions in lncRNAs have been described to be more conserved than the rest of the lncRNA sequences (1). Consistent with this, we found that 53.9% of the conserved regions in lncRNAs overlapped promoters, whereas this value was only 21.8% for non-conserved lncRNA regions (test of equal proportions, P-value $< 10^{-5}$).

We next asked whether conserved lncRNA regions were more enriched in Ribo-Seq reads than non-conserved ones. We found that 51.7% of the lncRNA conserved regions, but only 24.9% of the non-conserved ones, were covered by at least one Ribo-Seq read (Test of equal proportions, P-value $< 10^{-5}$). Although conserved regions were significantly more expressed and covered by Ribo-Seq reads (Additional File 1: Supplementary Figure S5), the trend could not be explained by differences in the expression level of the transcript or the amount of overlap with exons from other genes (Additional File 1: Supplementary Figures S6 and S7). Consistent results were observed when analyzing Ribo-Seq data from human and rat brain tissues for the corresponding genomic syntenic sequences (Additional File 1: Supplementary Figure S8).

To analyze the relationship between evolutionary sequence conservation and Ribo-Seq signal in more detail, we divided lncRNAs into intergenic, antisense and host ncRNA types (I, A and H, Figure 2A). Antisense lncRNAs included those lncRNAs annotated as antisense in Ensembl as well as any other lncRNA whose transcription start site

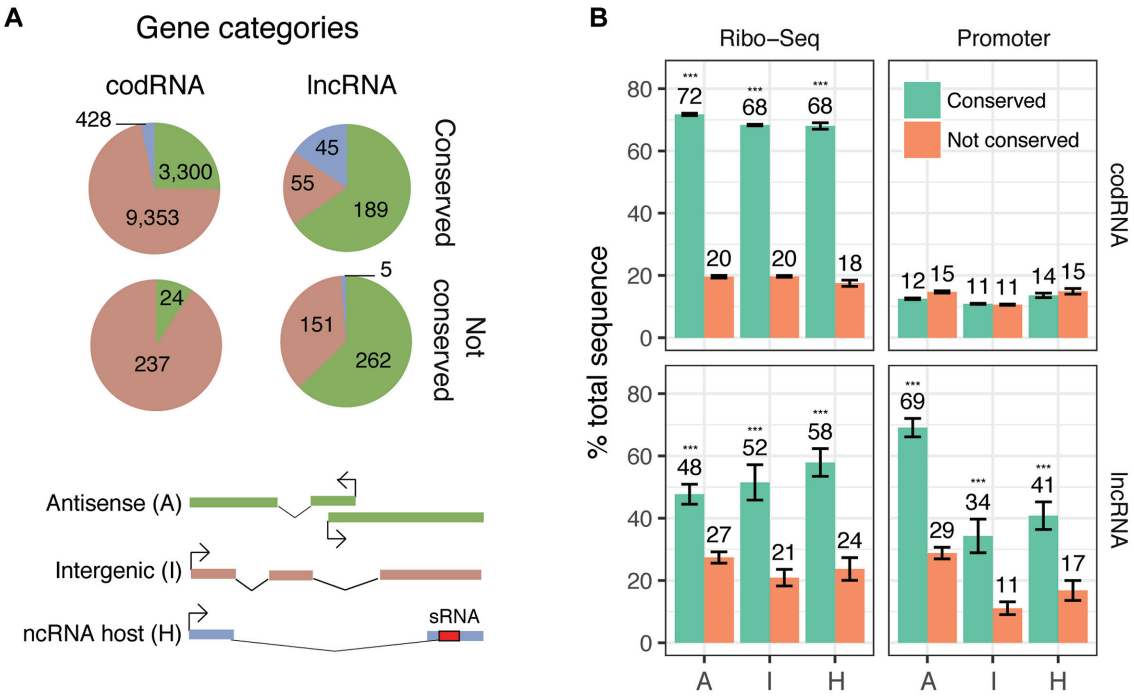

Figure 2. Effect of conservation across lncRNA types. (**A**) Number and fraction of different categories based on position and sequence features. Antisense: Exonic overlap, expression on a bidirectional promoter, and/or annotated as antisense; ncRNA host: Genes with at least one small RNA sequence found in the exonic region; intergenic: rest of the genes. Conserved genes are enriched in antisense and ncRNA host genes. (**B**) Percentage of total sequence that is covered by Ribo-Seq reads (1 or more reads), and annotated promoter cores, for conserved and non-conserved regions in codRNAs and lncRNAs. Conserved lncRNA regions showed a significantly higher proportion of all features compared to not conserved regions or expected randomly (test of equal proportions; *** P-value $< 10^{-5}$). Error bars represent the standard error of the proportion. Categories: A: Antisense; I: Intergenic; H: ncRNA host.

was located less than 2 Kb from the TSS of another gene in an antisense orientation and/or had antisense exonic overlap with another gene. Host ncRNAs corresponded to lncRNAs with embedded short non-coding RNAs in the exons (they contained 33 miRNAs, 41 snoRNAs and 32 miscRNAs). Intergenic lncRNAs were completely independent transcriptional units. Host ncRNAs showed a much higher degree of conservation between mice and humans than the other two lncRNA classes (Figure 2A). Conserved regions from all three classes of transcripts showed a strong enrichment in Ribo-Seq signatures and promoter elements when compared to non-conserved ones (Figure 2B), indicating that the observed trends were largely independent of the lncRNA type. For comparative purposes, we classified the genes annotated as coding in the same three categories as the lncRNAs (Figure 2A). In this case, conserved regions from codRNAs showed enrichment of Ribo-Seq signatures but not of promoter elements (Figure 2B).

Conserved lncRNA sequences are under purifying selection

Although mice and humans are relatively distant species (~90 Million years) (75), some sequence segments may still be sufficiently similar for homology to be detected even if no purifying selection is operating. In order to estimate the conservation expected in the absence of selection, we ran sequence evolution simulations using Rose (50). First, we simulated the evolution of lncRNAs along the mouse and human branches under no evolutionary constraints. Second, we performed BLASTN searches of the evolved mouse sequences against the set of evolved human sequences (see 'Materials and Methods' section for more details). We could find BLASTN homology hits for about 56.2% of the evolved sequences. This fraction happened to be larger than the fraction of real lncRNAs conserved between the two species (40.9%), which is consistent with the idea that a substantial number of mouse lncRNAs have originated after the split with the human lineage (40).

Next, we used the sequence alignments obtained with BLASTN to estimate a normalized observed to expected substitution rate k_o/k_e in different sequence sets (see 'Materials and Methods' section for more details). The normalized substitution rate, k_o/k_e, was significantly lower in real lncRNAs than in neutrally evolving sequences for all three lncRNA types (Additional File 1: Supplementary Figure S9, Wilcoxon test, P-value $< 10^{-5}$). The number of observed substitutions in conserved lncRNA segments was about half the expected. This supports the conclusion that the class of lncRNAs conserved between mice and humans is under significant purifying selection, in accordance with previous findings based on the density of single-nucleotide polymorphisms in this type of lncRNAs (38).

Remarkably, the strength of purifying selection in lncRNA conserved regions was similar for functionally characterized and uncharacterized lncRNAs (median of k_o/k_e 0.49 and 0.50, respectively). This supports the idea that there are hundreds of true bioactive lncRNAs whose functions remain to be uncovered. Coding sequences and ncRNA host transcripts showed somewhat lower k_o/k_e val-

ues (median of 0.37 and 0.46, respectively). Conserved regions overlapping promoters had lower k_o/k_e than the rest of the conserved lncRNA sequences (median 0.47 versus 0.58). However, the difference was relatively small, and when we eliminated the regions overlapping promoters the differences in the substitution rates estimated for neutrally evolving sequences and lncRNAs remained highly significant (Wilcoxon test, P-value $< 10^{-5}$).

Conserved lncRNAs regions are enriched in translated ORFs

Actively translated sequences show a characteristic three-nucleotide read periodicity in ribosome profiling experiments that allows for the identification of novel translation events (14,16,19). We used the program RibORF to score read periodicity and uniformity in all ORFs of size 30 nucleotides or longer (18). Translated ORFs were defined as those with a RibORF score equal to or higher than 0.7, as previously described (21) (Figure 3A). As expected, nearly all codRNAs conserved between mice and humans showed evidence of translation (Figure 3B). The percentage of mouse lncRNAs that contained at least one translated ORF was 52.05%, similar to previous estimates for human lncRNAs (18). Overall, we identified 165 lncRNAs with one or more translated ORFs, including several experimentally characterized lncRNAs (Additional File 2: Supplementary Table S3).

We found that conserved regions in lncRNAs were about nearly three times more covered by translated ORFs than non-conserved regions (14.1% versus 5.65%). The enrichment was consistently observed across the different lncRNA subtypes (test of equal proportions, P-value $< 10^{-5}$), with translation occurring more actively in antisense genes than in other lncRNA classes. A similar result was observed after discarding regions overlapping other genes on the opposite strand (Additional File 1: Supplementary Figure S6). We also observed that the translated ORFs were more abundant in the 5′ end than in the 3′ end of genes, independently of mouse–human sequence conservation (Additional File 1: Supplementary Figure S4). This may be related to the ribosome scanning dynamics (starting at the 5′ end of transcripts) and perhaps also due to the higher GC content in this first part of the gene (Additional File 1: Supplementary Figure S10), which may result in an enrichment of ORFs (76).

We investigated if the putative translated ORFs in lncRNA conserved regions showed signatures of selection at the protein level (Figure 3A), which would indicate that they might be functional. We recovered and aligned the corresponding human sequences using genomic alignments for 93 cases. We then estimated the rate of non-synonymous and synonymous substitutions (dN/dS) in the ORFs, and tested for significant deviation from a dN/dS of 1 (no selection) using a maximum-likelihood-based approach (see 'Materials and Methods' section). Despite the short size of these ORFs (median 56 amino acids) that may limit the identification of significant selection signatures, we identified nine cases in which dN/dS was significantly lower than 1 (chi-square test, P-value < 0.05). All of them were located in uncharacterized lncRNAs, and the size of the proteins ranged from 19 to 128 amino acids (Additional File 2: Supplementary Table S4). There were seven cases that completely overlapped (in antisense orientation) annotated coding sequences; this indicates that this configuration may be more common than previously suspected.

For comparison, we also analyzed the signatures of selection in 157 conserved codRNA genes encoding small proteins (small CDSs, < 100 amino acids). In this case, a much higher proportion of the cases showed significant negative selection signatures (76 out of 124). These cases included a number of known functional peptides that were found 'hidden' in transcripts previously annotated as lncRNAs such as Myoregulin (77,78), NoBody (79) or CASIMO1 (80), and other small functional peptides such as Stannin (81,82) or Sarcolipin (83,84).

As an alternative method we inspected conservation at the three different codon positions in the mouse genome using PhyloP scores (65). The results showed that, as with the dN/dS analysis, small CDSs had stronger purifying selection signatures than conserved ORFs in lncRNAs (Additional File 1: Supplementary Figure S11). We concluded that only a subset of the translated ORFs in lncRNA conserved regions is likely to encode proteins that are under selective constraints.

Whereas ribosome profiling is used to identify translation events, protein-derived mass-spectrometry (MS) data provide direct information on the abundance of the protein in the sample. The drawback is that this technique is less sensitive than high-throughput sequencing methods and many small proteins remain unseen (85). Using an available mouse hippocampus MS dataset, we could find proteomics evidence for only 36.65% of the Swissprot proteins with translation evidence (FDR < 1%, at least two unique peptide matches). In the case of lncRNA ORFs with translation evidence we found no significant hits in the proteomics set. This is not surprising given the short size of the lncRNA predicted peptides coupled with the low expression of the transcripts; when we subsampled the SwissProt proteins by read coverage and length so that they resembled the set of lncRNA ORFs (see 'Materials and Methods' section for more details), only 2.44% of them showed proteomics evidence. A similar negligible proportion of positive cases was observed for the set of small annotated proteins (smCDS)—in this case only 0.83% of the translated ORFs had hits to MS peptides. Given this low success rate, no information about the half-life of the peptides translated from lncRNA ORFs when compared to other proteins could be gathered.

Identification of protein–RNA interactions

Ribosome profiling experiments allow capturing protein–RNA interactions other than ribosomal associations—the two types of signals can be distinguished by their read length distribution (17). When analyzing the regions covered by Ribo-Seq reads, we found that most codRNAs were covered by reads with lengths of 30–32 nucleotides, which correspond to ribosome associations. In lncRNAs the length of the Ribo-Seq reads was more variable, consistent with the presence of non-ribosomal ribonucleoprotein particles (RNPs) in addition to ribosomes. The excess of short (<30 nt) and long (>32 nt) reads could be

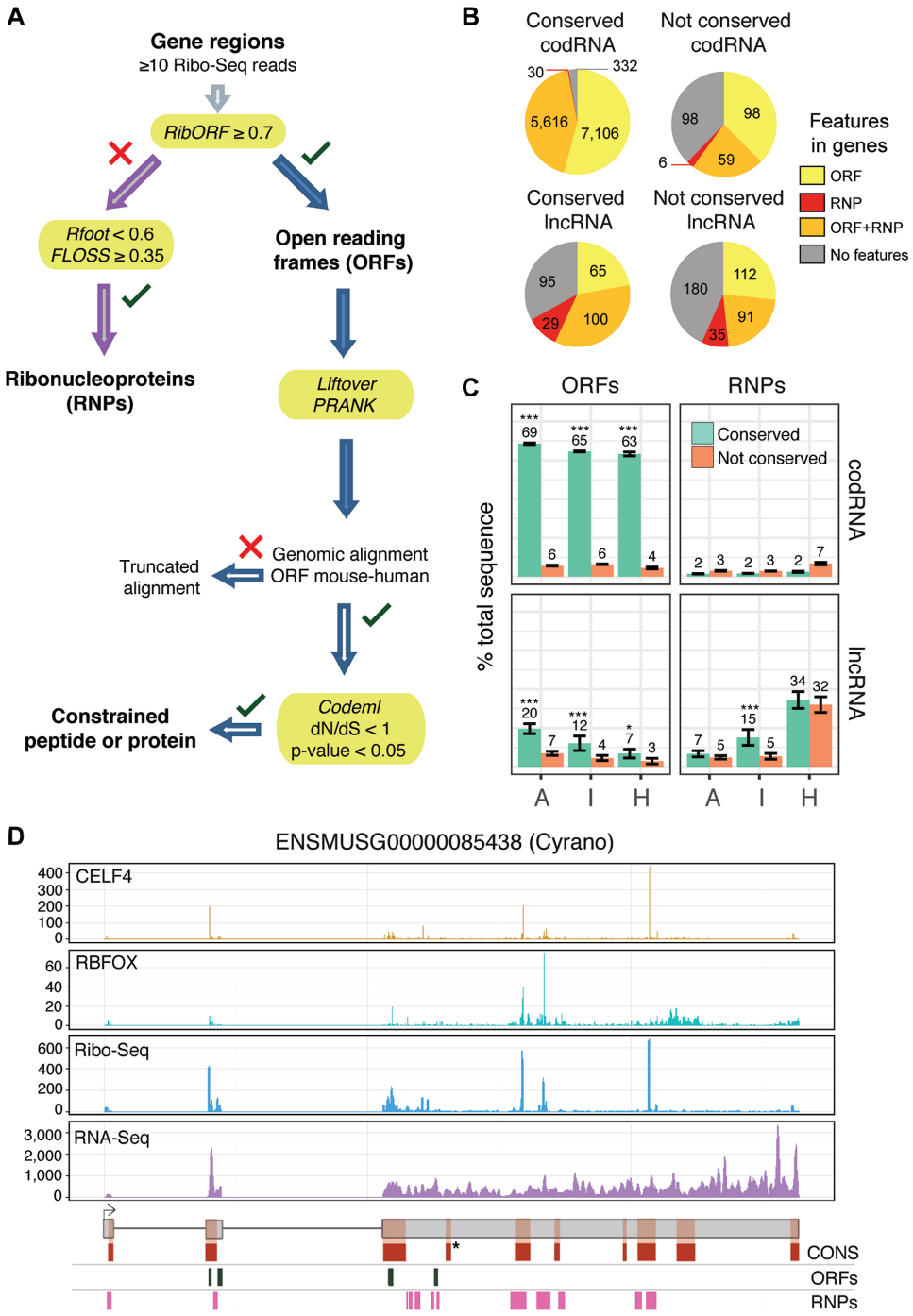

Figure 3. Identification of translated open reading frames and ribonucleoproteins. (**A**) Workflow to identify translated open reading frames (ORFs), putative functional proteins and ribonucleoproteins (RNPs). Ribosome profiling (Ribo-Seq) reads are mapped to candidate gene regions and ORFs with a RibORF score ≥ 0.7 are defined as translated. Rest of regions with Rfoot uniformity score < 0.6 and FLOSS score ≥ 0.35 are defined as RNPs. Next, human ORF syntenic regions are extracted with LiftOver and aligned with PRANK, when possible. Truncated alignments are those for which $>50\%$ of the ORF was aligned, or the gap limit is exceeded (33% or 10-nt). Finally, non-truncated alignments are checked for purifying selection signatures with Codeml to identify putative constrained peptides or proteins (dN/dS ratio < 1; Chi-square test of dN/dS ratio, P-value < 0.05). (**B**) Fraction and number of conserved and not conserved codRNAs and lncRNAs that contain at least one translated open reading frame (ORF), ribonucleprotein (RNP), both features (ORF+RNP), or neither of the two features. (**C**) Percentage of total sequence that is covered by translated ORFs and RNPs, for conserved and non-conserved regions. Overall, about 14.1% of the total conserved region in lncRNAs contained ORFs predicted to be translated (122 ORFs), compared to 5.65% for non-conserved regions (370 ORFs). Test of equal proportions: * P-value < 0.05; *** P-value $< 10^{-5}$. Error bars represent the standard error of the proportion. Categories: A: Antisense; I: Intergenic; H: ncRNA host. (**D**) Example of a functionally characterized lncRNA, *Cyrano*, with RNA-Seq, Ribo-Seq and annotated CLIP-Seq peaks (RBFOX and CELF4). Predicted conserved regions (CONS), ORFs and RNPs are also displayed. There is a high agreement between CLIP-Seq peaks and Ribo-Seq RNPs. * location of a previously described miRNA-binding site.

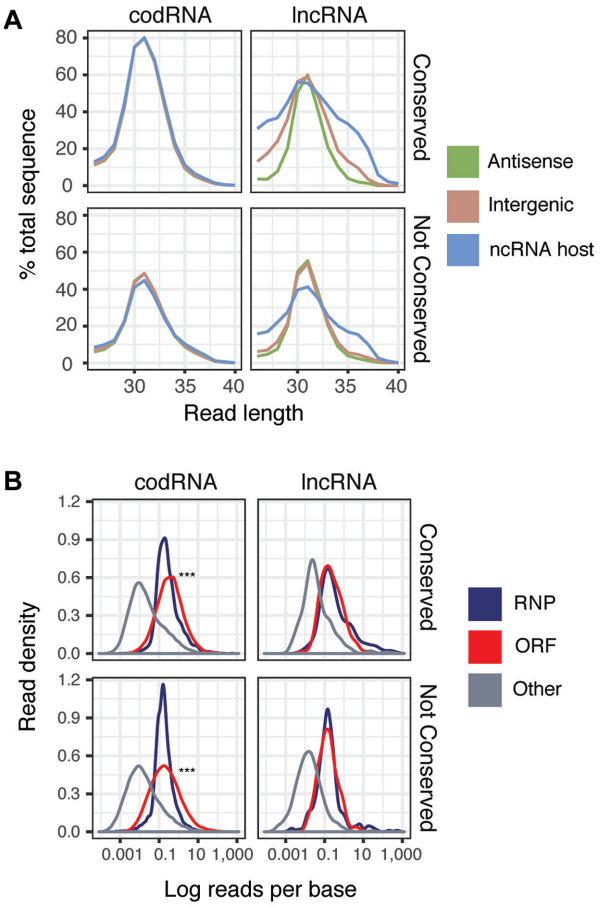

Figure 4. LncRNAs have more heterogenous Ribo-Seq read lengths. (**A**) Fraction of sequence covered by Ribo-Seq that contains reads from a specific length for conserved and not conserved regions in different categories of lncRNAs. While antisense lncRNAs resemble codRNAs in the read distribution, intergenic and ncRNA host regions contain a higher proportion of short and long reads corresponding to non-ribosome associates. (**B**) Ribo-Seq read density for regions predicted as ribonucleoproteins (RNP), translated sequences (ORF) and other regions covered by Ribo-Seq. ORFs in codRNAs have a higher read density than the rest of the sequences (***. Wilcoxon test, *P*-value $< 10^{-5}$).

clearly observed in intergenic lncRNA and the ncRNA host (Figure 4A). In contrast, the vast majority of Ribo-Seq reads mapping to antisense lncRNAs had a size compatible with ribosome-protected fragments (Figure 4A and Additional File 1: Supplementary Figure S12). Consistent patterns were found using independent rat ribosome profiling data for the corresponding syntenic regions (Additional File 1: Supplementary Figure S13).

We predicted the RNP positions by first identifying regions with low Ribo-Seq read uniformity (<0.6) with the program Rfoot (29), and then checking if the Ribo-Seq reads spanning these regions had lengths that were not compatible with ribosome associations using the FLOSS method (Figure 3A). This resulted in 255 out of 707 lncRNAs with RNP signatures (36%) (Figure 3B; Additional File 2: Supplementary Table S2). We analyzed the overlap between the RNPs and the annotated peaks from 39 CLIP-seq datasets, which corresponded to previously identified protein–RNA interactions. We found that RNPs were significantly enriched in CLIP-Seq peaks (35.11% of the RNP sequence was covered by peaks) when compared to all lncRNA sequences (10.38% covered by CLIP-Seq peaks), codRNAs (8.11% covered by CLIP-Seq peaks) or translated ORFs (6.91% covered by CLIP-Seq peaks; test of equal proportions *P*-value $< 10^{-5}$ in all three comparisons). These data confirmed that our pipeline was useful to identify putative protein–RNA interactions.

Furthermore, we found that conserved regions contained a larger number of RNP signatures than non-conserved ones (Figure 3C, conserved versus non-conserved). In terms of the transcripts, 129 conserved lncRNAs (44.6%) contained RNP signatures, compared to 126 non-conserved ones (30.1%) (Figure 3B). Consistently, the conserved RNPs were more extensively covered by CLIP-Seq signatures than the non-conserved ones (53.54% versus 31.58%; test of equal proportions, *P*-value $< 10^{-5}$) or the UTR regions from codRNAs (23%; Test of equal proportions, *P*-value $< 10^{-5}$). In the case of functionally characterized lncRNAs, this percentage increased to 72.65% (in comparison to 29.33% for the rest of the conserved lncRNA RNP regions).

The majority of functionally characterized lcRNAs contained RNP signatures in conserved sequences (21 out of 30) (Additional File 2: Supplementary Table S3). One example was *Cyrano*; in this lncRNA, we identified a previously described highly conserved sequence that is nearly identical to mir-7 microRNA (42), as well as several putative protein interaction sites that were scattered along the sequence (Figure 3D). Other cases were *Malat, Neat1, Meg3, Miat and Lncpint*, known to interact with different protein and splicing factors, and *TERC* that acts as a scaffold for the telomerase complex. We also found RNP signatures in the non-conserved mouse transcript *Firre*, a functional lncRNA that interacts with nuclear factors through a repetitive sequence (71). In this lncRNA, the predicted RNPs matched the repetitive sequences. The study yielded novel predictions for lncRNAs that remain uncharacterized. In particular, we found RNPs in conserved sequences from 32 antisense lncRNAs, 12 intergenic lncRNAs and 19 ncRNA host genes.

Conserved regions in intergenic lncRNAs were significantly enriched in RNP signatures (Figure 3C). According to CLIP-seq data, two proteins were significantly over-represented in these RNPs when compared to non-conserved ones: CELF4 (12.05% versus 2.85%; test of equal proportions, *P*-value $< 10^{-5}$) and RBFOX (10.51% versus 2.98%; test of equal proportions, *P*-value $< 10^{-5}$). These protein–RNA interactions are known to be important for neurosynaptic transmission and cortical development (86,87). Both proteins were found interacting with the functionally characterized lncRNA *Cyrano* (Figure 3D) and two additional intergenic ncRNAs: *6330403K07Rik (ENSMUSG00000018451)* and *Gm7292 (ENSMUSG00000104222)*. In contrast, we could not find any enriched CLIP-seq interactions in the small fraction of antisense lncRNAs covered by RNPs.

In addition to lncRNAs, as many as 5,646 conserved codRNAs had at least one RNP in the UTR regions (Figure 3B). Compared to the main translated ORF, this represented a very small fraction of the reads and had a negli-

gible effect on the overall distribution of read lengths (Figure 4A and Additional File 1: Supplementary Figure S12). In conserved lncRNA regions, RNPs and ORFs occupied a similar percentage of the sequence (17.2% and 14.1%, respectively; Additional File 1: Supplementary Figure S14) and the density of reads in both elements was similar (Figure 4B). In the case of conserved codRNA regions RNPs only occupied 1.7% of the sequence, whereas ORFs occupied 65.5% (Additional File 1: Supplementary Figure S14) and exhibited higher read density values than RNPs (Figure 4B).

DISCUSSION

Here, we have shown that lncRNA regions that are conserved between mice and humans are enriched in translated ORFs and protein–RNA interactions (RNPs). The patterns are quite distinct in different lncRNA classes: antisense lncRNA regions contain more translated ORFs than intergenic or host ncRNAs. RNP signatures are very frequent in host ncRNAs but also in conserved intergenic lncRNA regions. Most antisense lncRNAs overlap coding sequences in the other strand, which may favor the formation of translatable ORFs. Some of these ORFs showed signatures of purifying selection, encouraging future studies to search for functional antisense ORFs. This study also identified many conserved regions in intergenic lncRNAs that may mediate protein–RNA interactions, both in lncRNAs known to be functional and in uncharacterized lncRNAs, providing tens of novel functional candidates.

A number of studies have attempted to identify homologous lncRNAs between species employing blocks of predefined genomic synteny (48,88–94). However, lncRNAs have a high expression turnover (39,46,95), and syntenic conservation does not necessarily imply that the sequence is expressed in the two species. In order to circumvent these limitations, we focused on sequences that showed significant sequence similarity (denoting common ancestry) but that were also expressed in both mice and humans. We identified 289 mouse lncRNAs expressed in the hippocampus that showed homology to human transcripts. Conserved regions in these lncRNAs were usually small; they occupied 8.50% of the total mouse lncRNA sequence length. We observed that, despite their small size, these regions carried signatures of purifying selection, indicating that they are likely to be functionally relevant. This would be in line with previous observations that short regions in lncRNAs are often sufficient to emulate the complete RNA function (48,96).

There are only a limited number of studies on the evolutionary patterns of mammalian lncRNAs (39,94,97–99). These works have shown that transcripts that are conserved across different species evolve more slowly than those that are species-specific, consistent with the existence of selective constraints. However, none of these studies have estimated how much conservation we expect in the absence of selection. In order to investigate this, we simulated the evolution of lncRNAs starting from sequences that would have been present in a common mouse–human ancestor and that would have evolved along the two lineages with no selection. The simulations indicated that over half of the mouse evolved sequences should still show detectable sequence similarity to human sequences. This fraction is higher than that found for real sequences, indicating that many mouse lncRNAs have actually originated after the split from the common mouse–human ancestor and that the very large number of observed lineage-specific lncRNAs cannot be solely explained by rapid sequence evolution. A second interesting observation was that, for aligned lncRNAs, the number of substitutions per site was about half the amount expected under neutral evolution. This indicates that roughly half of the substitutions in conserved lncRNA sequences might be deleterious, providing strong evidence that these regions are functionally relevant. Although we observed a significant association between conservation and the presence of translation and RNP signatures, we cannot exclude the possibility that some may be important at the DNA level, e.g. hosting promoter or enhancer regions.

Our results support the idea that there are a substantial number of small proteins that remain to be characterized (100–103). We detected several ORFs likely to encode micropeptides in transcripts that have only recently been annotated as coding, such as Nbdy (79), and nine new putative cases, which remain to be investigated experimentally. The micropeptide ORFs were detected by a combination of ribosome profiling and the analysis of non-synonymous to synonymous substitutions. The analysis of ribosome profiling data from additional tissues is likely to result in the expansion of this list.

We could not recover significant proteomics support for the small ORFs identified by ribosome profiling. This low success rate was true for small ORFs in lncRNAs but also for small annotated coding sequences. This large gap between ribosome profiling and proteomics-based results is the object of current debate (103). A recent study argues that biases in the lncRNA ORFs, including short ORF size or low expression level, cannot explain the lack of proteomics evidence for lncRNA ORFs, concluding that the transcripts are essentially non-coding (104). However, the three-nucleotide periodicity of the reads observed in many lncRNA ORFs seems difficult to explain if there is no translation activity. Here, we investigated what the effect is of ORF small size and low expression level taken together, not separately as in the previously mentioned study, for obtaining mass-spectrometry hits for a given protein. This is more realistic than considering the biases separately, as the ORFs in lncRNAs are both small and expressed at low levels. By doing this, we concluded that virtually no peptide hits in mass-spectrometry should be expected even if the proteins encoded in the small ORFs were produced. So, lack of proteomics evidence does not preclude the possibility that the proteins actually exist; the experiments are simply not sensitive enough to provide an answer to this question. For this, novel targeted proteomics strategies will need to be developed.

Although we found evidence of purifying selection for some of the conserved translated ORFs in lncRNAs, in many other cases there was not a clear pattern of selection. There are different possible explanations: one possibility is that translation of some ORFs has a regulatory function that is independent of the protein sequence. For example, it has been suggested that the association of lncRNAs with

polysomes may favor their degradation (105). Another possibility is that these ORFs overlap other DNA or RNA elements that are under selection and that their translation is just a consequence of the presence of the RNA in the cytoplasm. In this direction, we found that conserved regions often overlapped gene expression regulatory sequences or 'promoters', which could be important for the expression of the same or other transcripts. Similarly, lncRNA upstream promoter regions were previously noted to have low sequence divergence (106). However, we have to be careful when interpreting these data, as homology detection is not independent of promoter conservation. This is because, in the absence of selection, some transcripts may still retain their ancestral regulatory sequence by chance, and these transcripts will be more easily classified as conserved than the rest.

In some transcript regions there were peaks of Ribo-Seq reads with no three-frame periodicity, suggesting RNA protection by complexes other than ribosomes. Two different methods have been proposed for the identification of ribonucleoprotein particles (RNP) signatures: FLOSS that is based on deviations from the expected RNA length covered by ribosomes (17) and Rfoot that selects regions on the basis of low read uniformity and absence of periodicity (29). We reasoned that protein–RNA interactions should display the two types of signatures to be sufficiently reliable, and designed a specific pipeline that integrated the two approaches. We found a significant association between our method and the regions identified by CLIP-Seq, further validating our approach. We identified putative protein interactions in already characterized lncRNAs, such as *Cyrano*, *Malat1*, *Neat1* and *TERC*, as well as RNPs in conserved regions of uncharacterized lncRNAs, which should encourage future studies. Although at a relatively lower frequency, many non-conserved regions also contained RNP signatures; these cases may be due to promiscuous protein–RNA interactions (107), the existence of young functional lncRNAs that interact with specific proteins (108–110), or lncRNAs that only contain repetitive, very small or poorly conserved sequences. Examples of the latter include the functional repeats described in *Firre* (71) or specific secondary structure elements detected in *Neat1* (111).

The detection of RNP signatures was not incompatible with the existence of translated ORFs in other parts of the transcript. In annotated coding transcripts (codRNAs), there were plenty of RNP signatures in the UTRs, and about 75% of the lncRNAs with RNPs also contained putatively translated small ORFs (Figure 3B). Some of these could be truly bifunctional transcripts; recently described cases are *Lncpint* (112) and *TERC* (113), which have been reported to translate small functional peptides in addition to having a non-coding function.

In conclusion, our study indicates that lncRNAs that have been retained at least since the common ancestor of mice and humans are more likely to encode proteins, and interact with proteins or protein complexes, than lineage-species lncRNAs. The study illustrates the power of combining evolutionary inferences and large-scale experimental measurements to advance our understanding of the transcriptome.

DATA AVAILABILITY

All the generated Python scripts for the identification of conserved and non-conserved regions, RNP prediction, dN/dS calculation and region coverage with specific features, as well as tables with raw data used for the main figures, are available at this GitHub repository: https://github.com/jorruior/TransCons-tools.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank William R. Blevins for useful comments on the manuscript.

FUNDING

Ministerio de Economía e Innovación (Spanish Government) co-funded by FEDER (EU) [BFU2015-65235-P]; Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya [2014SGR1121, 2017SGR01020]. Funding for open access charge: AGAUR 2017SGR1020.

Conflict of interest statement. None declared.

REFERENCES

1. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
2. Kapranov,P., Cheng,J., Dike,S., Nix,D.A., Duttagupta,R., Willingham,A.T., Stadler,P.F., Hertel,J., Hackermüller,J., Hofacker,I.L. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.
3. Okazaki,Y., Furuno,M., Kasukawa,T., Adachi,J., Bono,H., Kondo,S., Nikaido,I., Osato,N., Saito,R., Suzuki,H. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.
4. Ponjavic,J., Ponting,C.P. and Lunter,G. (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.*, **17**, 556–565.
5. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.
6. Liu,J., Jung,C., Xu,J., Wang,H., Deng,S., Bernad,L., Arenas-Huertero,C. and Chua,N.-H. (2012) Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *Plant Cell*, **24**, 4333–4345.
7. Pauli,A., Valen,E., Lin,M.F., Garber,M., Vastenhouw,N.L., Levin,J.Z., Fan,L., Sandelin,A., Rinn,J.L., Regev,A. *et al.* (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.*, **22**, 577–591.
8. Ponting,C.P., Oliver,P.L. and Reik,W. (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.
9. Ulitsky,I. and Bartel,D.P. (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell*, **154**, 26–46.
10. Consortium,T.E.P., Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R. and Gingeras,T.R. (2007) Identification and analysis of functional elements in 1{%} of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
11. Ingolia,N.T., Lareau,L.F. and Weissman,J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, **147**, 789–802.

12. Ruiz-Orera,J., Messeguer,X., Subirana,J.A. and Alba,M.M. (2014) Long non-coding RNAs as a source of new peptides. *Elife*, **3**, e03523.
13. Ingolia,N.T., Ghaemmaghami,S., Newman,J.R.S. and Weissman,J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.
14. Bazzini,A.A., Johnstone,T.G., Christiano,R., Mackowiak,S.D., Obermayer,B., Fleming,E.S., Vejnar,C.E., Lee,M.T., Rajewsky,N., Walther,T.C. *et al.* (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.*, **33**, 981–993.
15. Calviello,L. and Ohler,U. (2017) Beyond Read-Counts: Ribo-seq data analysis to understand the functions of the transcriptome. *Trends Genet.*, **33**, 728–744.
16. Ruiz-Orera,J. and Albà,M.M. (2019) Translation of small open reading frames: Roles in regulation and evolutionary innovation. *Trends Genet.*, **35**, 186–198.
17. Ingolia,N.T., Brar,G.A., Stern-Ginossar,N., Harris,M.S., Talhouarne,G.J.S., Jackson,S.E., Wills,M.R. and Weissman,J.S. (2014) Ribosome profiling reveals pervasive translation outside of annotated Protein-Coding genes. *Cell Rep.*, **8**, 1365–1379.
18. Ji,Z., Song,R., Regev,A. and Struhl,K. (2015) Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife*, **4**, e08890.
19. Calviello,L., Mukherjee,N., Wyler,E., Zauber,H., Hirsekorn,A., Selbach,M., Landthaler,M., Obermayer,B. and Ohler,U. (2016) Detecting actively translated open reading frames in ribosome profiling data. *Nat. Meth.*, **13**, 165–170.
20. Raj,A., Wang,S.H., Shim,H., Harpak,A., Li,Y.I., Engelmann,B., Stephens,M., Gilad,Y. and Pritchard,J.K. (2016) Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife*, **5**, e13328.
21. Ruiz-Orera,J., Verdaguer-Grau,P., Villanueva-Cañas,J.L., Messeguer,X. and Albà,M.M. (2018) Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat. Ecol. Evol.*, **35**, 186–198.
22. Chew,G.-L., Pauli,A., Rinn,J.L., Regev,A., Schier,A.F. and Valen,E. (2013) Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. *Development*, **140**, 2828–2834.
23. Aspden,J.L., Eyre-Walker,Y.C., Phillips,R.J., Amin,U., Mumtaz,M.A.S., Brocard,M., Couso,J.-P., Andrews,S., Rothnagel,J., Arava,Y. *et al.* (2014) Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *Elife*, **3**, e03528.
24. Wilson,B.A. and Masel,J. (2011) Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol. Evol.*, **3**, 1245–1252.
25. Ponting,C.P., Oliver,P.L. and Reik,W. (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.
26. Ulitsky,I. and Bartel,D.P. (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell*, **154**, 26–46.
27. Moriarty,T.J., Huard,S., Dupuis,S. and Autexier,C. (2002) Functional multimerization of human telomerase requires an RNA interaction domain in the N terminus of the catalytic subunit. *Mol. Cell. Biol.*, **22**, 1253–1265.
28. Smith,K.N., Starmer,J. and Magnuson,T. (2018) Interactome determination of a Long Noncoding RNA implicated in Embryonic Stem Cell Self-Renewal. *Sci. Rep.*, **8**, 17568.
29. Ji,Z., Song,R., Huang,H., Regev,A. and Struhl,K. (2016) Transcriptome-scale RNase-footprinting of RNA-protein complexes. *Nat. Biotechnol.*, **34**, 410–413.
30. Ponjavic,J., Ponting,C.P. and Lunter,G. (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.*, **17**, 556–565.
31. Pertea,M., Shumate,A., Pertea,G., Varabyou,A., Breitwieser,F.P., Chang,Y.-C., Madugundu,A.K., Pandey,A. and Salzberg,S.L. (2018) CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.*, **19**, 208.
32. Morris,K. V. and Mattick,J.S. (2014) The rise of regulatory RNA. *Nat. Rev. Genet.*, **15**, 423–437.
33. Wang,J., Zhang,J., Zheng,H., Li,J., Liu,D., Li,H., Samudrala,R., Yu,J. and Wong,G.K.-S. (2004) Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. *Nature*, **431**, 1–2.
34. Struhl,K. (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.*, **14**, 103–105.
35. Ulitsky,I. (2016) Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.*, **17**, 601–614.
36. Necsulea,A., Soumillon,M., Warnefors,M., Liechti,A., Daish,T., Zeller,U., Baker,J.C., Grützner,F. and Kaessmann,H. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, **505**, 635–640.
37. Hezroni,H, Koppstein,D, Schwartz,MG, Avrutin,A, Bartel,DP and Ulitsky,I (2016) Priniciples of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.*, **11**, 1110–1122.
38. Wiberg,R.A.W., Halligan,D.L., Ness,R.W., Necsulea,A., Kaessmann,H. and Keightley,P.D. (2015) Assessing recent selection and functionality at long noncoding RNA loci in the mouse genome. *Genome Biol. Evol.*, **7**, 2432–2444.
39. Kutter,C., Watt,S., Stefflova,K., Wilson,M.D., Goncalves,A., Ponting,C.P., Odom,D.T. and Marques,A.C. (2012) Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.*, **8**, e1002841.
40. Kapusta,A. and Feschotte,C. (2014) Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet.*, **30**, 439–452.
41. Ulitsky,I. (2016) Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.*, **17**, 601–614.
42. Ulitsky,I., Shkumatava,A., Jan,C.H., Sive,H. and Bartel,D.P. (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, **147**, 1537–1550.
43. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–55.
44. Smit,A.F.A., Hubley,R. and Green,P. (2018) *RepeatMasker Open-4.0*. http://www.repeatmasker.org (16 June 2018,date last accessed).
45. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
46. Ruiz-Orera,J., Hernandez-Rodriguez,J., Chiva,C., Sabidó,E., Kondova,I., Bontrop,R., Marqués-Bonet,T. and Albà,M.M. (2015) Origins of de novo genes in human and chimpanzee. *PLOS Genet.*, **11**, e1005721.
47. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
48. Hezroni,H., Koppstein,D., Schwartz,M.G., Avrutin,A., Bartel,D.P. and Ulitsky,I. (2015) Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.*, **11**, 1110–1122.
49. Zerbino,D.R., Wilder,S.P., Johnson,N., Juettemann,T. and Flicek,P.R. (2015) The ensembl regulatory build. *Genome Biol.*, **16**, 56.
50. Stoye,J., Evers,D. and Meyer,F. (1998) Rose: generating sequence families. *Bioinformatics*, **14**, 157–163.
51. Consortium, M.G.S. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
52. Lunter,G. (2007) Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics*, **23**, i289–i296.
53. Ogurtsov,A.Y., Sunyaev,S. and Kondrashov,A.S. (2004) Indel-Based evolutionary distance and mouse–human Divergence. *Genome Res.*, **14**, 1610–1616.
54. Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.
55. Cho,J., Yu,N.-K., Choi,J.-H., Sim,S.-E., Kang,S.J., Kwak,C., Lee,S.-W., Kim,J., Choi,D. Il, Kim,V.N. *et al.* (2015) Multiple repressive mechanisms in the hippocampus during memory formation. *Science*, **350**, 82–87.
56. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

57. Gonzalez,C., Sims,J.S., Hornstein,N., Mela,A., Garcia,F., Lei,L., Gass,D.A., Amendolara,B., Bruce,J.N., Canoll,P. *et al.* (2014) Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *J. Neurosci.*, **34**, 10924–10936.
58. Ji,Z., Song,R., Regev,A. and Struhl,K. (2015) Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife*, **4**, e08890.
59. Leon,J., Moreno,A.J., Garay,B.I., Chalkley,R.J., Burlingame,A.L., Wang,D. and Dubal,D.B. (2017) Peripheral elevation of a klotho fragment enhances brain function and resilience in young, aging, and α-Synuclein transgenic mice. *Cell Rep.*, **20**, 1360–1371.
60. Eng,J.K., Jahan,T.A. and Hoopmann,M.R. (2012) Comet: an open-source MS/MS sequence database search tool. *Proteomics*, **13**, 22–24.
61. Kall,L., Canterbury,J.D., Weston,J., Noble,W.S. and MacCoss,M.J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Meth.*, **4**, 923–925.
62. Tyner,C., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Eisenhart,C., Fischer,C.M., Gibson,D., Gonzalez,J.N., Guruvadoo,L. *et al.* (2017) The UCSC genome browser database: 2017 update. *Nucleic Acids Res.*, **45**, D626–D634.
63. Loytynoja,A. and Goldman,N. (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 10557–10562.
64. Puigdevall,P. and Castelo,R. (2018) GenomicScores: seamless access to genomewide position-specific scores from R and Bioconductor. *Bioinformatics*, **34**, 3208–3210.
65. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
66. Ji,Z., Song,R., Huang,H., Regev,A. and Struhl,K. (2016) Transcriptome-scale RNase-footprinting of RNA-protein complexes. *Nat. Biotechnol.*, **34**, 410–413.
67. Ingolia,N.T., Brar,G.A., Stern-Ginossar,N., Harris,M.S., Talhouarne,G.J.S., Jackson,S.E., Wills,M.R. and Weissman,J.S. (2014) Ribosome profiling reveals pervasive translation outside of annotated Protein-Coding genes. *Cell Rep.*, **8**, 1365–1379.
68. Hu,B., Yang,Y.-C.T., Huang,Y., Zhu,Y. and Lu,Z.J. (2017) POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res.*, **45**, D104–D114.
69. Quek,X.C., Thomson,D.W., Maag,J.L. V, Bartonicek,N., Signal,B. and Clark,M.B. (2014) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res*, **43**, D168–D173.
70. Goff,L.A., Groff,A.F., Sauvageau,M., Trayes-Gibson,Z., Sanchez-Gomez,D.B., Morse,M., Martin,R.D., Elcavage,L.E., Liapis,S.C., Gonzalez-Celeiro,M. *et al.* (2015) Spatiotemporal expression and transcriptional perturbations by long noncoding RNAs in the mouse brain. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 6855–6862.
71. Hacisuleyman,E., Goff,L.A., Trapnell,C., Williams,A., Henao-Mejia,J., Sun,L., McClanahan,P., Hendrickson,D.G., Sauvageau,M., Kelley,D.R. *et al.* (2014) Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat. Struct. Mol. Biol.*, **21**, 198–206.
72. Feng,J., Funk,W.D., Wang,S.S., Weinrich,S.L., Avilion,A.A., Chiu,C.P., Adams,R.R., Chang,E., Allsopp,R.C. and Yu,J. (1995) The RNA component of human telomerase. *Science*, **269**, 1236–1241.
73. Lee,S., Kopp,F., Chang,T.-C., Sataluri,A., Chen,B., Sivakumar,S., Yu,H., Xie,Y. and Mendell,J.T. (2016) Noncoding RNA NORAD regulates genomic stability by sequestering PUMILIO proteins. *Cell*, **164**, 69–80.
74. R Core Team (2014) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, http://www.R-project.org.
75. Hedges,S.B., Marin,J., Suleski,M., Paymer,M. and Kumar,S. (2015) Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.*, **32**, 835–845.
76. Vakirlis,N., Hebert,A.S., Opulente,D.A., Achaz,G., Hittinger,C.T., Fischer,G., Coon,J.J. and Lafontaine,I. (2018) A molecular portrait of de novo genes in yeasts. *Mol. Biol. Evol.*, **35**, 631–645.
77. Anderson,D.M., Anderson,K.M., Chang,C.-L., Makarewich,C.A., Nelson,B.R., McAnally,J.R., Kasaragod,P., Shelton,J.M., Liou,J., Bassel-Duby,R. *et al.* (2015) A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell*, **160**, 595–606.
78. Yu,X., Zhang,Y., Li,T., Ma,Z., Jia,H., Chen,Q., Zhao,Y., Zhai,L., Zhong,R., Li,C. *et al.* (2017) Long non-coding RNA Linc-RAM enhances myogenic differentiation by interacting with MyoD. *Nat. Commun.*, **8**, 14016.
79. D'Lima,N.G., Ma,J., Winkler,L., Chu,Q., Loh,K.H., Corpuz,E.O., Budnik,B.A., Lykke-Andersen,J., Saghatelian,A. and Slavoff,S.A. (2017) A human microprotein that interacts with the mRNA decapping complex. *Nat. Chem. Biol.*, **13**, 174–180.
80. Polycarpou-Schwarz,M., Groß,M., Mestdagh,P., Schott,J., Grund,S.E., Hildenbrand,C., Rom,J., Aulmann,S., Sinn,H.-P., Vandesompele,J. *et al.* (2018) The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. *Oncogene*, **37**, 4750–4768.
81. Buck-Koehntop,B.A., Mascioni,A., Buffy,J.J. and Veglia,G. (2005) Structure, dynamics, and membrane topology of stannin: A mediator of neuronal cell apoptosis induced by trimethyltin chloride. *J. Mol. Biol.*, **354**, 652–665.
82. Pueyo,J.I., Magny,E.G., Sampson,C.J., Amin,U., Evans,I.R., Bishop,S.A. and Couso,J.P. (2016) Hemotin, a regulator of phagocytosis encoded by a small ORF and conserved across metazoans. *PLoS Biol.*, **14**, e1002395.
83. Wawrzynow,A., Theibert,J.L., Murphy,C., Jona,I., Martonosi,A. and Collins,J.H. (1992) Sarcolipin, the 'proteolipid' of skeletal muscle sarcoplasmic reticulum, is a unique, amphipathic, 31-residue peptide. *Arch. Biochem. Biophys.*, **298**, 620–623.
84. Magny,E.G., Pueyo,J.I., Pearl,F.M.G., Cespedes,M.A., Niven,J.E., Bishop,S.A. and Couso,J.P. (2013) Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science*, **341**, 1116–1120.
85. Wang,D., Eraslan,B., Wieland,T., Hallström,B., Hopf,T., Zolg,D.P., Zecha,J., Asplund,A., Li,L.-H., Meng,C. *et al.* (2019) A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.*, **15**, e8503.
86. Damianov,A., Ying,Y., Lin,C.-H., Lee,J.-A., Tran,D., Vashisht,A.A., Bahrami-Samani,E., Xing,Y., Martin,K.C., Wohlschlegel,J.A. *et al.* (2016) Rbfox proteins regulate splicing as part of a large multiprotein complex LASR. *Cell*, **165**, 606–619.
87. Dasgupta,T. and Ladd,A.N. (2012) The importance of CELF control: molecular and biological roles of the CUG-BP, Elav-like family of RNA-binding proteins. *Wiley Interdiscip. Rev. RNA*, **3**, 104–121.
88. Ulitsky,I., Shkumatava,A., Jan,C.H., Sive,H. and Bartel,D.P. (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, **147**, 1537–1550.
89. He,Y., Ding,Y., Zhan,F., Zhang,H., Han,B., Hu,G., Zhao,K., Yang,N., Yu,Y., Mao,L. *et al.* (2015) The conservation and signatures of lincRNAs in Marek's disease of chicken. *Sci. Rep.*, **5**, 15184.
90. Mohammadin,S., Edger,P.P., Pires,J.C. and Schranz,M.E. (2015) Positionally-conserved but sequence-diverged: identification of long non-coding RNAs in the Brassicaceae and Cleomaceae. *BMC Plant Biol.*, **15**, 217.
91. Li,D. and Yang,M.Q. (2017) Identification and characterization of conserved lncRNAs in human and rat brain. *BMC Bioinform.*, **18**, 489.
92. Necsulea,A., Soumillon,M., Warnefors,M., Liechti,A., Daish,T., Zeller,U., Baker,J.C., Grützner,F. and Kaessmann,H. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, **505**, 635–640.
93. Guttman,M., Amit,I., Garber,M., French,C., Lin,M.F., Feldser,D., Huarte,M., Zuk,O., Carey,B.W., Cassady,J.P. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
94. Marques,A.C. and Ponting,C.P. (2009) Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol.*, **10**, R124–R124.

95. Neme,R. and Tautz,D. (2016) Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *Elife*, **5**, e09977.
96. Quinn,J.J., Ilik,I.A., Qu,K., Georgiev,P., Chu,C., Akhtar,A. and Chang,H.Y. (2014) Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. *Nat. Biotechnol.*, **32**, 933–940.
97. Wiberg,R.A.W., Halligan,D.L., Ness,R.W., Necsulea,A., Kaessmann,H. and Keightley,P.D. (2015) Assessing recent selection and functionality at long noncoding RNA loci in the mouse genome. *Genome Biol. Evol.*, **7**, 2432–2444.
98. Pegueroles,C. and Gabaldón,T. (2016) Secondary structure impacts patterns of selection in human lncRNAs. *BMC Biol.*, **14**, 1–13.
99. Haerty,W. and Ponting,C.P. (2013) Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol.*, **14**, R49.
100. Ladoukakis,E., Pereira,V., Magny,E.G., Eyre-Walker,A. and Couso,J.P. (2011) Hundreds of putatively functional small open reading frames in Drosophila. *Genome Biol.*, **12**, R118.
101. Pauli,A., Norris,M.L., Valen,E., Chew,G.-L., Gagnon,J.A., Zimmerman,S., Mitchell,A., Ma,J., Dubrulle,J., Reyon,D. *et al.* (2014) Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science*, **343**, 1248636.
102. Saghatelian,A. and Couso,J.P. (2015) Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat. Chem. Biol.*, **11**, 909–916.
103. Housman,G and Ulitsky,I (2016) Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive purpose of translation of long noncoding RNAs. *Biochim. Biophys. Acta*, **1859**, 31–40.
104. Verheggen,K., Volders,P.-J., Mestdagh,P., Menschaert,G., Van Damme,P., Gevaert,K., Martens,L. and Vandesompele,J. (2017) Noncoding after All: Biases in proteomics data do not explain observed absence of lncRNA translation products. *J. Proteome Res.*, **16**, 2508–2515.
105. Carlevaro-Fita,J., Rahim,A., Guigo,R., Vardy,L.A. and Johnson,R. (2016) Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. *RNA*, **22**, 867–882.
106. Derrien,T., Johnson,R., Bussotti,G., Tanzer,A., Djebali,S., Tilgner,H., Guernec,G., Martin,D., Merkel,A., Knowles,D.G. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.*, **22**, 1775–1789.
107. Davidovich,C., Zheng,L., Goodrich,K.J. and Cech,T.R. (2013) Promiscuous RNA binding by Polycomb Repressive Complex 2. *Nat. Struct. Mol. Biol.*, **20**, 1250–1257.
108. Heinen,T.J., Staubach,F., Häming,D. and Tautz,D. (2009) Emergence of a new gene from an intergenic region. *Curr. Biol.*, **19**, 1527–1531.
109. Rigoutsos,I., Lee,S.K., Nam,S.Y., Anfossi,S., Pasculli,B., Pichler,M., Jing,Y., Rodriguez-Aguayo,C., Telonis,A.G., Rossi,S. *et al.* (2017) N-BLR, a primate-specific non-coding transcript leads to colorectal cancer invasion and migration. *Genome Biol.*, **18**, 98.
110. Durruthy-Durruthy,J., Sebastiano,V., Wossidlo,M., Cepeda,D., Cui,J., Grow,E.J., Davila,J., Mall,M., Wong,W.H., Wysocka,J. *et al.* (2015) The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. *Nat. Genet.*, **48**, 44–52.
111. Lin,Y., Schmidt,B.F., Bruchez,M.P. and McManus,C.J. (2018) Structural analyses of NEAT1 lncRNAs suggest long-range RNA interactions that may contribute to paraspeckle architecture. *Nucleic Acids Res.*, **46**, 3742–3752.
112. Zhang,M., Zhao,K., Xu,X., Yang,Y., Yan,S., Wei,P., Liu,H., Xu,J., Xiao,F., Zhou,H. *et al.* (2018) A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. *Nat. Commun.*, **9**, 4475.
113. Rubtsova,M., Naraykina,Y., Vasilkova,D., Meerson,M., Zvereva,M., Prassolov,V., Lazarev,V., Manuvera,V., Kovalchuk,S., Anikanov,N. *et al.* (2018) Protein encoded in human telomerase RNA is involved in cell protective pathways. *Nucleic Acids Res.*, **46**, 8966–8977.