Functional and non-functional classes of peptides produced by long non-coding RNAs

Jorge Ruiz-Orera[1,*], Pol Verdaguer-Grau[2], José Luis Villanueva-Cañas[1], Xavier Messeguer[2], M.Mar Albà[1,3,*]

[1]Evolutionary Genomics Group, Research Programme on Biomedical Informatics, Hospital del Mar Research Institute, Universitat Pompeu Fabra, Barcelona, Spain; [2]Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona, Spain; [3]Catalan Institution for Research and Advanced Studies, Barcelona, Spain.

*To whom correspondence should be addressed.

Running title: lncRNA translation

keywords: long non-coding RNA, ribosome profiling, translation, RNA-Seq, peptide

Abstract

Cells express thousands of transcripts that show weak coding potential. Known as long non-coding RNAs (lncRNAs), they typically contain short open reading frames (ORFs) having no homology with known proteins. Recent studies have reported that a significant proportion of lncRNAs are translated, challenging the view that they are non-coding. These results are based on the selective sequencing of ribosome-protected fragments, or ribosome profiling. The present study uses ribosome profiling data from eight mouse tissues and cell types, combined with ~330,000 synonymous and non-synonymous single nucleotide variants, to dissect the patterns of purifying selection in proteins translated from lncRNAs. Using the three-nucleotide read periodicity that characterizes actively translated regions, we identify about 1,365 translated lncRNAs. About one fourth of them (350 lncRNAs) show conservation in humans; this is likely to produce functional micropeptides, including the recently discovered myoregulin. For other lncRNAs, the ORF codon usage bias distinguishes between two classes. The first has significant coding scores and evidence of purifying selection, consistent with the presence of lineage-specific functional proteins. The second large class, comprising >500 lncRNAs, produces proteins that show no significant purifying selection signatures. We show that the translation of these lncRNAs depends on the transcript expression level and the chance occurrence of ORFs with a favorable codon composition. Some of these lncRNAs may be precursors of novel protein-coding genes, filling a gap in our current understanding of *de novo* gene birth.

Introduction

In recent years, the sequencing of transcriptomes has revealed that, in addition to classical protein-coding transcripts, the cell expresses thousands of long transcripts with weak coding potential [1–5]. Some of these transcripts, known as long non-coding RNAs (lncRNAs), have well-established roles in gene regulation; for example, Air is an Igf2r antisense lncRNA involved in silencing the paternal Igf2r allele *in cis* [6,7]. However, the vast majority of them remain functionally uncharacterized. While some lncRNAs have nuclear roles, the largest part is polyadenylated and accumulates in the cytoplasm [8]. In addition, many lncRNAs are expressed at low levels and have a limited phylogenetic distribution [9,10].

In 2009, Nicholas Ingolia and co-workers published the results of a new technique to measure translation of mRNAs by deep sequencing of ribosome-protected RNA fragments, called ribosome profiling [11]. This method permits the detection of lowly abundant small proteins, which may be difficult to detect by standard proteomics approaches. In addition, the three-nucleotide periodicity of the reads, resulting from the movement of the ribosome along the coding sequence, differentiates translated sequences from other possible RNA protein complexes. A growing number of studies based on this technique have reported that a significant proportion of lncRNAs are translated [12–19]. However, the functional significance of this finding is not yet clear. Some of the translated lncRNAs may be mis-annotated protein coding genes that encode micropeptides (< 100 amino acids) which, due to their short size, have not been correctly predicted by bioinformatics algorithms [20–23]. This is likely to include some recently evolved proteins that lack homologues in other species and which are even harder to detect than conserved short peptides [16].

One striking feature of the proteins produced by lncRNAs is that, in general, they appear to be under lower selective constraints than standard proteins [16,17]. This raises the possibility that a large fraction of them encode proteins that, despite being translated in a stable manner, are not functional.

Non-synonymous and synonymous single nucleotide polymorphisms in coding sequences provide useful information to distinguish between neutrally evolving proteins and proteins under purifying or negative selection. Under no selection, both kinds of variants accumulate at the same rate, whereas under purifying selection there is a deficit of non-synonymous variants [24]. The detection of selection signatures provides strong evidence of functionality, whereas non-functional proteins evolve neutrally. The present study takes advantage of the existing nucleotide variation data for the domestic mouse (*Mus musculus*) to investigate the selective patterns of proteins translated by lncRNAs. This data provides evidence that lncRNAs are pervasively translated, providing abundant raw material for *de novo* protein-coding gene evolution.

Results

Identification of translated sequences

We sought to identify translated open reading frames (ORFs) in a comprehensive set of long non-coding RNAs (lncRNAs) and protein-coding genes from mouse, using ribosome profiling RNA sequencing (Ribo-Seq) data from eight different tissues and cell types (Table 1 and references therein). In contrast to RNA sequencing (RNA-Seq) reads, which are expected to cover the complete transcript, Ribo-Seq reads are specific to regions bound by ribosomes. We mapped the RNA-Seq and Ribo-Seq reads of each experiment to a mouse transcriptome that included all Ensembl mouse gene annotations, including both coding genes and lncRNAs, as well as thousands of additional *de novo* assembled polyadenylated transcripts derived from non-annotated expressed loci (see Methods). For the assembly of this transcriptome, we used more than 1.5 billion strand-specific RNA sequencing reads from mouse [25].

We selected all expressed transcripts containing at least one open reading frame (ORF) encoding a putative protein of 24 amino acids or longer for further analyses. This size cut-off was chosen on the basis of the smallest known protein in humans, humanin [26]. To facilitate the comparison between translated and non-translated transcripts we focused on the longest translated ORF when several ORFs existed. For most genes this corresponded to the ORF with the largest number of Ribo-Seq reads (see Methods). For each experiment we selected the ORFs covered by at least 10 Ribo-Seq reads and examined the distribution of the reads along the ORF (Figure 1A). We observed a strong three-nucleotide periodicity of the Ribo-Seq reads both for coding genes, annotated lncRNAs and novel transcripts (Figure 1B). This bias towards the correct frame is characteristic of regions which are being actively translated [11,12,18,21,27] and is not expected in other types of protein-RNA interactions [28]). We defined translated ORFs as those in which at least 60%

of the reads mapped to the expected frame. This corresponded to a p-value < 0.05 according to a null model in which the reads were randomly distributed in each ORF (Supplemental file 1 Figure S1). We also observed that the number of false negatives was very small for coding genes and annotated lncRNAs (1-3%) and slightly higher for novel transcripts (about 11%, blue dots above the 60% threshold in Figure 1B). In ORFs classified as translated ("in-frame", red dots above the 60% threshold in Figure 1B), the Ribo-Seq reads typically covered the complete ORF (Figure 1C), providing additional support for our method. We defined non-translated transcripts as those transcribed at significant levels but showing a ribosome profiling signal that was either very weak or nonexistent (<10 Ribo-Seq reads).

This method identified translated ORFs in ~23% of lncRNAs (1,365 lncRNAs, including novel assembled ones) and ~92% of the coding genes (15,588 codRNAs) among genes expressed at significant levels in at least one sample (Table 1 and Figure 1D, Methods). Most coding genes were transcribed and translated in several samples, whereas lncRNAs tended to be sample-specific (Supplemental file 1 Figure S2). About 70% of the translated lncRNAs encoded proteins shorter than 100 amino acids (small ORFs or smORFs). The number of translated transcripts, and the size of the translated products, was very similar for annotated lncRNAs and for novel expressed loci (Figure 1D and 1E). Therefore, these two types of transcripts were merged into a single class (lncRNA) for most analyses.

Table 1

Tissue/cell	GEO (reference)	Annotated codRNA		Annotated lncRNA		Novel lncRNA	
		# transcribed	# translated	# transcribed	# translated	#transcribed	#translated
Brain	GSE51424(1)	12,689	11,127	1,141	83	1,614	139
Testis	GSE50983(2)	13,094	10,477	1,251	67	2,176	98
Neutrophils	GSE22001(3)	8,917	7,736	414	23	961	60
Heart	GSE41426	11,009	8,868	652	4	1,062	47
Skeletal Muscle	GSE41426	10,352	8,392	548	3	1,000	37
Splenic B cells	GSE62134(4)	9,504	7,694	871	38	1,129	46
Neural ES cells	GSE72064(5)	13,289	11,879	1,508	201	2,644	231
Hippocampus	GSE72064(5)	13,963	13,258	1,724	469	2,819	638
Integrated	-	**17,319**	**15,588**	**2,598**	**616**	**3,792**	**749**

Table 1. Number of transcribed and translated loci. Integrated refers to the number transcribed/translated in at least one sample. GEO: Gene Expression Omnibus. codRNA: coding gene. ES cells: embryonic stem cells. (1) [29], (2) [30], (3) [31], (4) [32], (5) [33]

Figure 1

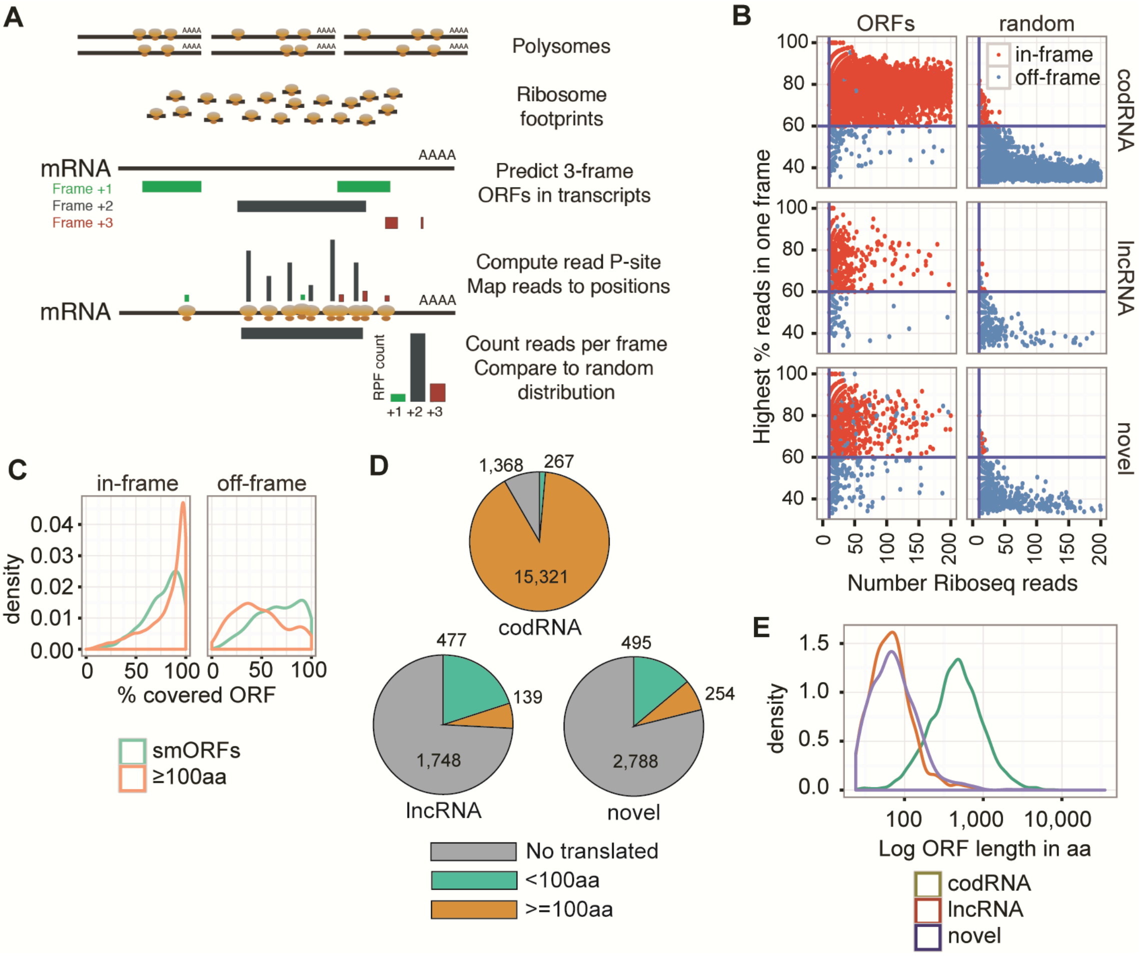

Figure 1. Detection of translated ORFs. A Workflow to identify translated ORFs. Ribosome profiling (Ribo-Seq) reads, corresponding to ribosome-protected fragments, are mapped to all predicted ORFs in transcripts. This is performed with single-nucleotide resolution after computing the read P-site per each read length. In each ORF, reads per frame are counted and compared to the random expectation. **B.** Relationship between the number of reads in a given frame and the number of Ribo-Seq reads that map to the ORF for codRNA. annotated lncRNA, and novel genes. Data shown are for the hippocampus sample; similar results were obtained in other samples. Only ORFs of 24 amino acids or longer were interrogated. ORFs: real data; random: the position of the reads in each ORF was randomized. The ORFs were classified as in-frame when ≥60% reads mapped to the predefined frame (red) or off-frame when <60% reads mapped to that frame or when they mapped to another frame (blue). The in-frame ORFs in the random control indicate the false positive rate (<5%). **C.** Number of translated and not translated expressed transcripts belonging to different classes. When a transcript contained several translated ORFs we selected the longest one. For non-translated transcripts, we took the longest ORF (Met to STOP). The translated ORFs have been divided into small ORFs (< 100 aa) and long ORFs (≥ 100 aa). Off-frame genes were not considered. **D.** Density plot showing the fraction of nucleotide positions in the ORF covered by Ribo-Seq reads, for in-frame and off-frame cases. **E.** Length of translated ORFs for different gene types in logarithmic scale: coding (codRNA), annotated long non-coding RNA (lncRNA) and non-annotated assembled transcripts (novel). The ORFs in the latter two classes were significantly shorter than in codRNAs (Wilcoxon test, p-value < 10^{-5}).

Detection of translated transcripts in different experiments

The number of transcribed and translated genes varied substantially depending on the sample; coding genes and lncRNAs showed parallel trends (Table 1, Figure 2A, Supplemental file 1 Figure S3). We detected the highest number of translated genes in hippocampus, followed by embryonic stem cells. In order to test whether this was due to genuine differences between the biological samples or to differences in the Ribo-Seq sequencing coverage, we randomly selected 10 million reads from each sample and recalculated the number of translated transcripts. This resulted in a much more similar number of translated transcripts in different samples, indicating that sequencing depth was the main cause of the original differences across samples (Supplemental file 1 Figure S3, smORFs). This suggested that the experimental translation signal was not saturated and that the true number of translated lncRNAs may be higher than was estimated here.

Figure 2

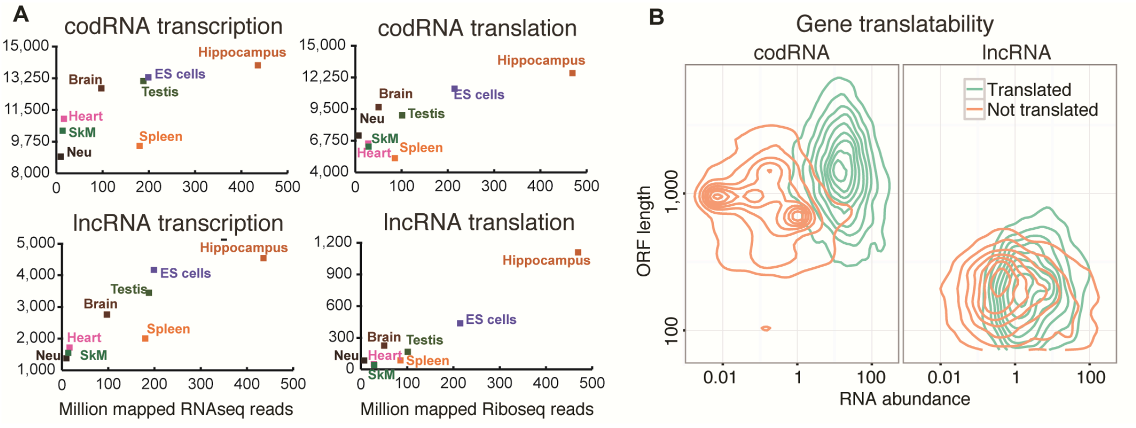

Figure 2. Features of translated transcripts. A. Number of transcribed or translated genes (Y-axis) in relation to the number of sequencing reads (RNAseq or Riboseq) mapped to the transcripts in the different experiments (X-axis). **B.** Relationship between ORF length and RNA abundance in codRNA and lncRNA for translated and non-translated genes. RNA abundance is defined as the maximum FPKM value across the 8 samples.

Protein-coding genes for which we detected translation were expressed at higher levels and contained longer ORFs than those for which we did not detect translation (Figure 2B). In general, lncRNAs were expressed at much lower levels than coding genes (Figure 2B, lncRNA versus coding), which is consistent with previous reports [9,34]. As for coding genes, translated lncRNAs were expressed at significantly higher levels, and contained longer ORFs, than non-translated lncRNAs (Wilcoxon test, p-value < 10^{-5}).

Phylogenetic conservation and codon usage bias

We examined which fraction of the mouse translated lncRNAs were conserved in humans. For this we generated a *de novo* human transcriptome assembly of a quality similar to that used for mouse (Methods). We performed sequence similarity searches with TBLASTN to identify similar putative proteins in human transcripts (e-value < 10^{-4}). About 25% of the mouse translated lncRNAs showed homology to human transcripts, compared to 98% for protein-coding genes (Figure 3A, Conserved). This is in agreement with previous observations that lncRNAs tend to be much less conserved than coding genes [10,35,36].

Codon usage bias is usually employed to predict coding sequences in conjunction with other variables such as ORF length and sequence conservation [37,38]. In the case of non-conserved short ORFs, such as those translated from many of the lncRNAs, only measures based on codon usage bias can be applied. We previously implemented a metric based on the differences in dicodon (hexamer) frequencies between coding and non-coding sequences, which we used to calculate length-independent coding scores for translated and non-translated ORFs [16]. Based on this metric, we developed a computational tool to identify ORFs with significant coding scores in any set of sequences (evolutionarygenomics.imim.es/CIPHER).

When the CIPHER program was applied to our dataset, translated codRNAs had higher coding scores than non-translated ones (Supplemental file 1 Figure S4). A similar result was observed in the lncRNA set, both for smORFs and for ORFs ≥ 100 amino acids. Using this method, we also found that conserved ORFs had significantly higher coding scores than non-conserved ORFs, both for coding genes and lncRNAs (Figure 3B). We also used CIPHER to divide the translated genes for which we had detected no homologues in human into a group with high coding scores (NC-H, coding score >0.049, significant at p-value <0.05) and another group with lower coding scores (NC-L, ≤0.049).

Testing for signatures of natural selection in translated ORFs

A key question was whether or not the proteins produced by lncRNAs were functional. We addressed it by using a very large number of mouse single nucleotide polymorphism (SNP) variants from dbSNP [39] –157,029 non-synonymous SNPs (PN) and 179,825 synonymous SNPs (PS)– that mapped to the ORFs in our dataset. We used the ratio between non-synonymous and synonymous SNPs to evaluate whether proteins translated from different sets of transcripts were subject to purifying selection. This method has the advantage over using non-synonymous to synonymous substitutions that it can be applied to sequences which do not show phylogenetic conservation. This allowed us to investigate the signatures of selection in hundreds of translated ORFs from lncRNAs which were not conserved in human.

In the absence of selection, and considering that all codons have the same frequency and all mutations between pairs of nucleotides are equally probable, we expect the PN/PS ratio of a sequence or set of sequences to be 2.89 [24]. This neutral model can be refined by using the codon frequencies observed in real coding sequences and providing different weights to diverse types of mutations

[40,41]. For example, transitions (mutations to another nucleotide of the same type, purine or pyrimidine) are known to occur more frequently than transversions (purine to pyrimidine or viceversa) in mammals [42]. Then we can test for purifying selecion by comparing the PN/PS ratio of the set of sequences of interest to that of the null or neutral model. If the PN/PS ratio is significantly lower than the neutral model it means that non-synonymous mutatioins are less well-tolerated than synonymous ones, consistent with negative or purifying selection operating at the protein level.

Here we generated null models of the PN/Ps ratio for the different sets of translated ORFs in coding genes and lncRNAs (C, NC-H and NC-L in Figure 3A). We considered the codon frequencies in each specific sequence set and different transition to transversion ratios; the neutral PN/PS estimates ranged between 2.84 and 3.71 depending on the dataset (area labeled Neutrality in Figure 3C). We used the Fisher test to determine if the real sequences showed lower PN/PS than the corresponding neutral models (Supplemental file 1 Table S1 for the results of all tests).

We found that conserved translated transcripts, both codRNAs and lncRNA, had PN/PS values significantly lower than the neutral expectation (Figure 3C, Supplemental File 1 Table S1, Fisher test p-value $< 10^{-5}$). This strongly suggests that a large fraction of the 350 lncRNAs in this group are protein-coding genes that produce functional small proteins or micropeptides (smORFs). The computational identification of smORFs is especially challenging because they can randomly occur in any part of the genome [43]. Therefore, it is not surprising that some remain hidden in the vast ocean of transcripts annotated as non-coding. For instance, the recently discovered peptide Myoregulin, which is only 46 amino acids long, regulates muscle performance [44] (Figure 3D). Myoregulin was annotated as non-coding when we initiated the study although it has now been re-classified as protein-coding. Another example of conserved micropeptide in our set was Stannin, a mediator of neuronal cell apoptosis conserved across metazoans [45,46] (Figure 3D). Our findings point to the existence of many other similar cases.

The group of non-conserved genes with significant CIPHER scores (NC-H) showed weaker purifying selection than conserved genes; however, PN/PS was still significantly lower than the neutral expectation (Figure 3C, Fisher test p-value <0.005). This indicates that some genes in this class are functional despite not being conserved in humans. In contrast, the PN/PS value in the group of non-conserved genes with low coding scores (NC-L) was consistent with neutral evolution (Figure 3C). In addition, the PN/PS ratio in this group was not significantly different than the PN/PS in the control group of non-translated ORFs with otherwise similar characteristics. These observations strongly argue against protein functionality. The lncRNAs encoding non-functional proteins comprised about 40% of the lncRNAs with evidence of translation; a very small percentage of codRNAs had the same characteristics (~1%).

The group of lncRNAs producing proteins with no selection signatures included several with known non-coding functions, such as *Malat1*, *Neat1*, *Jpx*, and *Cyrano*. These genes are involved in several cellular processes: *Cyrano* is involved in the regulation of embryogenesis [47], *Jpx* functions in X chromosome inactivation [48], *Neat1* has a role in the maintenance and assembly of paraspeckles [49], and *Malat1* regulates the expression of other genes [50]. The translation of these transcripts may be due to promiscuous activity of the ribosome rather than any important role of these proteins in the cell.

Figure 3

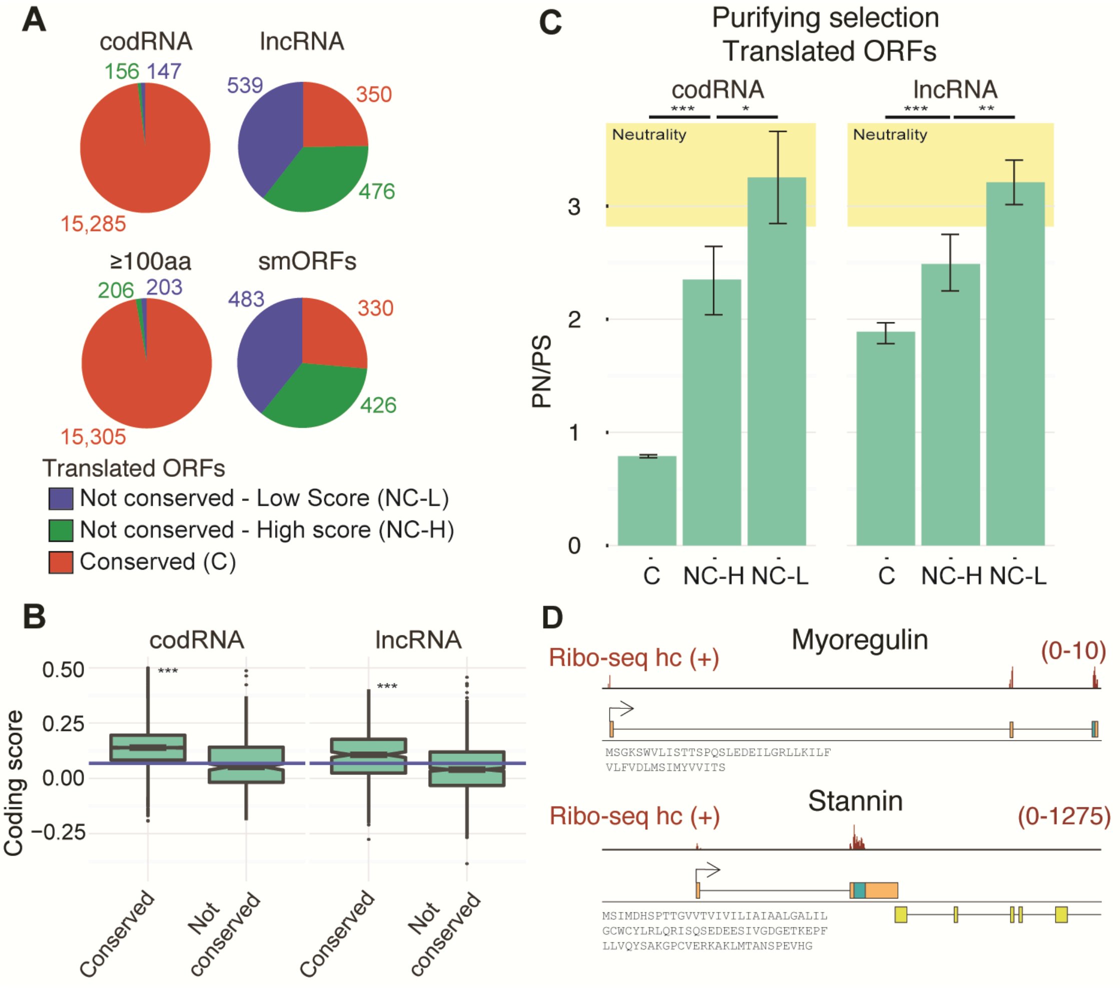

Figure 3. Different classes of translated ORFs. A. Number of translated ORFs that are conserved in human (C), not conserved but showing a high coding score (NC-H, coding score > 0.049, significant at p-value < 0.05) and not conserved with a low coding scorel (NC-L, coding score ≤ 0.049). First, ORFs are divided into coding genes (codRNA) and long non-coding RNAs (lncRNA) and lncRNA, and second, into long (length ≥ 100 amino acids) and small ORFs (smORFs, length < 100 amino acids). **B .** Differences in coding score for conserved (C) and non-conserved ORFs (NC). Conserved ORFs showed significantly higher coding score values than non-conserved ones; Wilcoxon test; ***, p-value < 10^{-5}. Blue line indicates the coding score value used to separate non-conserved ORFs with high coding scores (NC-H) to the rest of non-conserved ORFs. **C**. Analysis of selective constraints in translated ORFs. PN/PS refers to the ratio between non-synonymous (PN) and synonymous (PS) single nucleotide variants. The region labeled "Neutrality" contains all the values for the different neutral models considered (Supplemental file 1 Table S1). Conserved and high-score ORFs show significant purifying selection signatures independently of transcript type (codRNA or lncRNA). In contrast, non-conserved ORFs with low coding scores do not show evidence of purifying selection at the protein level, indicating lack of functionality. Significant differences between PN/PS ratios between the groups are indicated. Fisher test *p- value < 0.05, **p- value < 0.005, ***, p-value < 10^{-5}. Error bars represent the 95% confidence interval. **D .** Distribution of Ribo-seq reads in *Myoregulin*, which encodes a recently discovered micropeptide. Another well-known micropeptide-containing gene, *Stannin*, is shown for comparison. The protein sequences are shown. The data is from hippocampus (hc) ribosome profiling experiments.

What drives the neutral translation of lncRNAs?

Translated ORF not conserved in human and with no evidence of selection (NC-L) comprised 686 genes. They showed the characteristic three nucleotide periodicity of actively translated regions (Figure 4A), which was highly consistent across different tissues (Figure 4B). Why was translation detected in these transcripts but not in others? Selective forces at the level of the protein were not involved, as the ORFs lacked signatures of selection. Neither could we detect differences in the translation initiation sequence context between translated and non-translated ORFs. One factor that was likely to have an influence was the gene expression level. This was supported by the observation that translated lncRNAs were expressed at higher levels than non-translated lncRNAs (Figure 2B) and that experimental samples with more sequencing coverage yielded a larger number of translated products than other samples (Figure 2A).

We hypothesized that the ORF coding score could also affect the "translatability" of the transcript, because codons that are abundant in coding sequences are expected to be more efficiently translated than other, more rare, codons. We indeed found that the translated ORFs exhibited higher coding scores than non-translated ORFs in the group of transcripts with presumably neutral translation (Figure 4C, Wilcoxon test p-value $<10^{-5}$). Importantly, we obtained a similar result after controlling for transcript abundance (Figure 4D for hippocampus, Wilcoxon test p-value $<10^{-5}$; Supplemental file 1 Figure S5 for embryonic stem cells). This is consistent with codon composition having an effect *per se* in ORF translation. Controlling by coding score confirmed that transcript abundance is positively related to the capacity to detect translation (Figure 4E for hippocampus and Supplemental file 1 Figure S5 for embryonic stem cells). In contrast, although translated ORFs tend to be longer than non-translated ORFs (Figure 2B), ORF length had no effect other than that already explained by the coding score (Figure 4E).

Figure 4

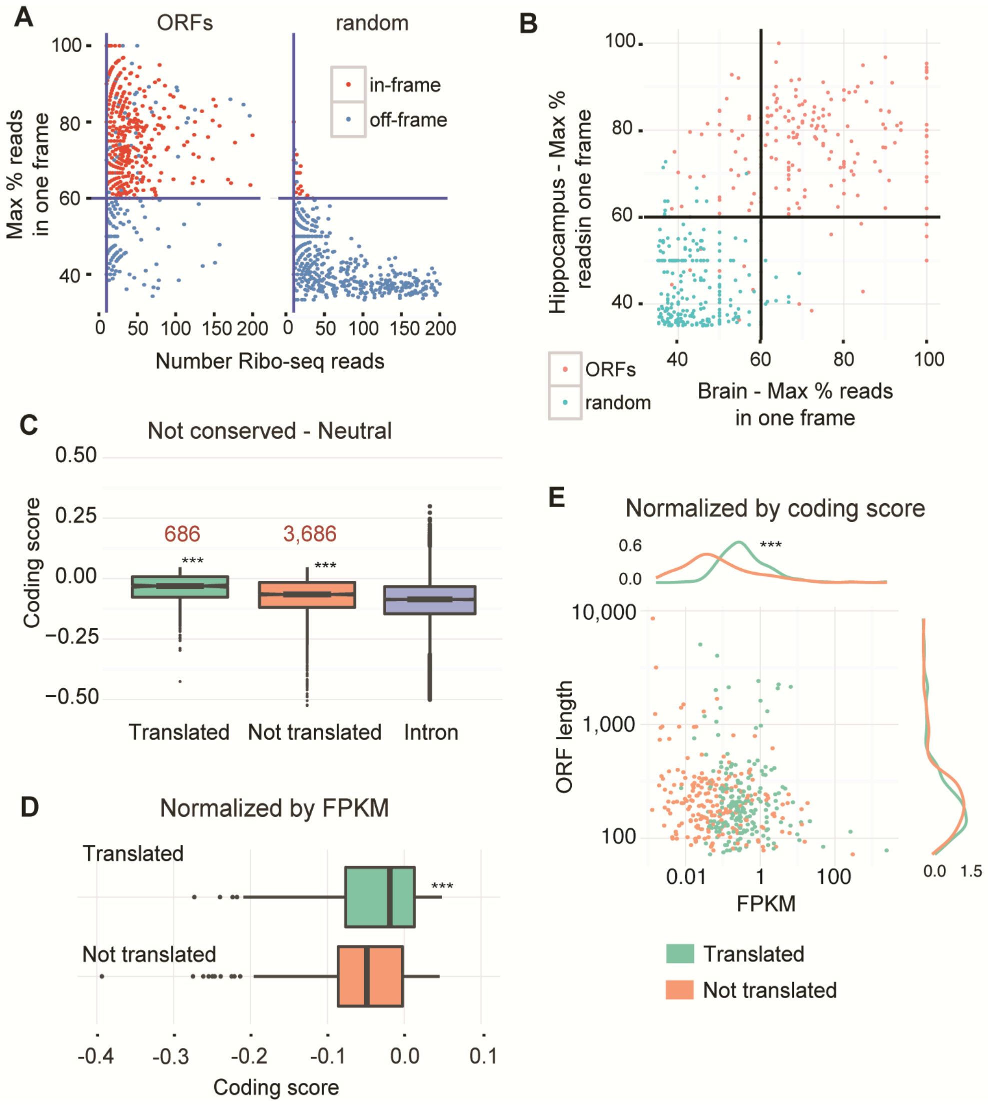

Figure 4. Factors influencing the neutral translation of lncRNAs. A. Relationship between the number of reads in a given frame and the number of Ribo-Seq reads that map to the ORF for non-conserved neutral ORFs (NC-L) for the hippocampus sample. Data is for real transcripts and for controls in which the position of the reads was randomized (random). B**.** Relationship between the percentage of reads falling in a frame in brain and hippocampus samples, for NC-L ORFs. Data is for real transcripts and for controls in which the position of the reads was randomized (random). **C.** Influence of coding score in the translatability of non-conserved neutral ORFs (NC-L). Intronic ORFs are shown for comparison. Translated ORFs showed significantly higher coding score than non-translated ORFs; Wilcoxon test; ***, p-value < 10^{-5}. **D.** Influence of coding score in the translatability of non-conserved neutral ORFs normalized by FPKM expression in hippocampus (median FPKM value = 0.225). Translated ORFs showed significantly higher coding score values than non-translated ORFs; Wilcoxon test; ***, p-value < $10\text{-}^{5.}$ **E.** Influence of FPKM expression and ORF length in the translatability of non-conserved neutral ORFs normalized by coding score in hippocampus (median coding score value = -0.022). Translated ORFs showed significantly higher FPKM values; Wilcoxon test, ***, p-value < 10^{-5}.

DISCUSSION

Several studies have reported that many lncRNAs translate small proteins (Bazzini et al., 2014; Calviello et al., 2016; Ingolia et al., 2014, 2011; Ji et al., 2015; Raj et al., 2016; Ruiz-Orera et al., 2014; this study). This is supported by three-nucleotide periodicity of the Ribo-Seq reads, high translational efficiency values (number of Ribo-Seq reads with respect to transcript abundance), and signatures of ribosome release after the STOP codon. Hundreds or even thousands of lncRNAs with patterns consistent with translation were detected in each of those studies. As lncRNAs are in general expressed at low levels, the stringency of the method, as well as the sequencing depth, are expected to strongly impact the number of translated lncRNAs found in different studies.

The recent discovery that a large number of lncRNAs show ribosome profiling patterns consistent with translation has puzzled the scientific community [51]. Most lncRNAs are not conserved across mammals or vertebrates, which limits the use of substitution-based methods to infer selection. Methods based on the number of non-synonymous and synonymous nucleotide polymorphisms (PN and PS, respectively) detect selection at the population level and can be applied to both conserved and non-conserved ORFs. This analysis is well-suited for pre-defined sets of ORFs; individual coding sequences do not usually contain enough polymorphims to test for selection [52]. In a previous study using ribosome profiling experiments from several species, we found that, in general, ORFs with evidence of translation in lncRNAs have weak but significant purifying selection signatures [16]. Together with previous observations that lncRNAs tend to be lineage-specific [10] and that young proteins evolve under relaxed purifying selection [53], this finding led us to suggest that lncRNAs are enriched in young protein-coding genes.

In the present study we have employed recently generated mouse and human deep transcriptome sequencing, together with extensive mouse variation data and codon usage bias to investigate the patterns of selection in translated ORFs from lncRNAs. LncRNAs conserved across species are more likely to be functional than those which are not conserved. This is supported by studies measuring the sequence constraints of lncRNAs with different degrees of phylogenetic conservation [36,54]. Here we estimate that about 5% of the lncRNAs encode functional micropeptides (smORFs). Standard proteomics techniques have important limitations for the detection of micropeptides and it is likely that the smORFs currently annotated in databases are only a small part of the complete set [55–58]. As shown here, and in other recent studies [20,21], computational prediction of ORFs coupled with ribosome profiling is a promising new avenue to unveil many of these peptides. In our study, the majority of transcripts encoding micropeptides were not annotated as coding, emphasizing the power of using whole transcriptome analysis instead of only annotated genes to characterize the so-called smORFome. Analysis of other tissues, and case-by-case experimental validation, will no doubt lead to a sustained increase in the number of micropeptides with characterized functions.

Remarkably, the largest class of lncRNAs appears to translate non-functional proteins. These ORFs can be distinguished from the rest because they are not conserved across species and have low coding scores. Although the existence of non-functional proteins may seem counterintuitive at first, we have to consider that most lncRNAs tend to be expressed at low levels and so the associated energy costs may be negligible. It has also been estimated that the cost of transcription and translation in multicellular organisms is probably too small to overcome genetic drift [59]. In other words, provided the peptides are not toxic, the negative selection coefficient associated with the cost of producing them may be too low for natural selection to effectively remove them. We observed that the translation patterns of many of these peptides were similar across tissues, indicating that their translation is relatively stable and reproducible. The neutral translation of lncRNAs provides an answer for the conundrum of why transcripts that have been considered to be non-coding appear to be coding when viewed through the lens of ribosome profiling.

According to our results, the neutral translation of certain lncRNAs, but not others, may be due to the chance existence of ORFs with a favorable codon composition. This is consistent with the observation that abundant codons enhance translation elongation [60], whereas rare codons might affect the stability of the mRNA and activate decay pathways [61]. Other researchers have hypothesized that the distinction between translated and non-translated lncRNAs may be related to the relative amount of the lncRNA in the nucleus and the cytoplasm [18]. However, we found evidence that some lncRNAs with nuclear functions, such as *Malat1* and *Neat1*, are translated, suggesting that the cytosolic fraction of any lncRNA may be translated independently of the role or preferred location of the transcript. In the absence of experimental evidence, the codon composition of an ORF can provide a first indication of whether the ORF will be translated or not. Differences in codon frequencies between genes reflect the specific amino acid abundance as well as the codon usage bias, which is the differential use of synonymous codons. These differences can arise from a combination of selection, mutation, and drift [62,63]. Algorithms to predict coding sequences often use dicodon instead of codon frequencies, as the former also capture dependencies between adjacent amino acids or nucleotide triplets. We found that ORFs with very low coding scores are in general not translated. One example of this sort was the previously described *de novo* gene *Poldi* [64]. The group of ORFs that had high coding scores, but lacked conservation in humans had weak but significant purifying selection signatures. The patterns were very similar for coding and non-coding genes, indicating that for lineage-specific genes there are essentially no differences between genes annotated as coding and genes annotated as non-coding.

There is accumulating evidence that some protein-coding genes have originated *de novo* from previously non-functional genomic regions [65–73]. These *de novo* genes encode proteins with unique sequences that are likely to mediate lineage-specific adaptations. Young proteins are usually small and disordered, and have been hypothesized to become longer and more complex over time [69,74,75]. It has been hypothesized that many of these genes originated from lncRNAs [25,76,77], which would be consistent with the large number of species-specific transcripts with lncRNA features identified in comparative transcriptomics studies [25,78–80]. The discovery that some non-coding RNAs are translated makes the transition from non-coding/non-functional to coding/functional more plausible, as deleterious polypeptides can be purged by selection [81]. The translation of lncRNAs provides a large number of brand new proteins that can be tested for new functions. However, the observation that lncRNAs are translated is by itself inconclusive, as one could also argue that translated lncRNAs are simply mis-annotated functional protein-coding genes. Here we have shown that, for the bulk of translated lncRNAs, this is not the case, because many of the peptides do not show signatures of purifying selection. We propose that the evolutionary neutral translation of lncRNAs represents the missing link between transcribed genomic regions with no coding function and the eventual birth of proteins with new functions.

In conclusion, our data support that the analysis of lncRNA translation patterns by ribosome profiling can lead to the discovery of many new functional micropeptides. We also observed that many lncRNAs produce small proteins that lack a function; these peptides can serve as raw material for the evolution of new protein-coding genes. We have found that the translated ORFs in these lncRNAs are enriched in coding-like hexamers when compared to non-translated or intronic ORFs, which implies that the sequences available for the formation of new proteins are not random but have coding-like features from the start.

METHODS

Transcriptome assembly

The polyA+ RNA-Seq from mouse comprised 18 strand-specific paired end data publicly available in the Gene Expression Omnibus under accession numbers GSE69241 [25], GSE43721 [82], and GSE43520 [10]. Data corresponded to 5 brain, 2 liver, 1 heart, 3 testis, 3 ovary and 4 placenta samples.

The polyA+ RNA-Seq from human comprised 8 strand-specific paired end data publicly available in the Gene Expression Omnibus under accession number GSE69241 [25]. Data corresponded to 2 brain, 2 liver, 2 heart and 2 testis samples.

RNA-seq reads were filtered by length and quality. We retrieved genome sequences and gene annotations from Ensembl v. 75. We aligned the reads to the correspondent reference species genome with Tophat (v. 2.0.8, –N 3, -a 5 and –m 1) [83]. Multiple mapping to several locations in the genome was allowed unless otherwise stated.

We assembled the transcriptome with Stringtie [84] merging the reads from all the samples, with parameters -f 0.01, and -M 0.2. We used the species transcriptome as guide (Ensembl v.75) but permitting the assembly of annotated and novel isoforms and genes as well. We selected genes with a minimum size of 300 nucleotides. To eliminate potential pseudogenes we discarded genes that showed exonic overlap with annotated pseudogenes or which contained small ORFs that had significant sequence similarity to proteins. We selected genes with a per-nucleotide read coverage ≥ 5 in at least one sample. This ensures a high degree of transcript completeness, as shown in Ruiz-Orera et al. (2015).

Ribosome profiling data

We used 8 different data sets that included both strand-specific ribosome profiling (Ribo-seq) and RNA-seq experiments that we obtained from Gene Expression Omnibus under accession numbers GSE51424 [29], GSE50983 [30], GSE22001 [31], GSE62134 [32], GSE72064 [33], and GSE41426. Data corresponded to brain, testis, neutrophils, splenic B cells, ES cells, hippocampus, heart and skeletal muscle (Table 1).

Ribo-seq data sets were depleted of anomalous reads (length < 26 or > 33 nt) and small RNAs after discarding reads that mapped to annotated rRNAs and tRNAs. Next, reads were mapped to the assembled mouse transcriptome with Bowtie (v. 0.12.7, -k 1 -m 20 -n 1 --best --strata –norc).

We used the mapping of the Ribo-seq reads to the complete set of annotated coding sequences in mouse again to define the exact read point that corresponds to the ribosome and compute the offset position (P-site) for each read length (≥45% total reads), as in other studies [11,18,21,27]. If no offset was clear for a specific length, reads with that length were not considered for subsequent analysis.

Detection of translated ORFs

For each ribosome profiling experiment, we calculated the minimum gene expression level that was required to detect translation using the distribution of fragments per Kilobase per Million reads (FPKM) of coding sequence genes together with information on the samples in which we detected translation. We built a null model in which failure to detect translation of a protein-coding gene in a sample was attributed to poor sequencing coverage, provided that its translation was detected in at least three other samples. Then we chose the RNA-seq expression

FPKM cutoff that corresponded to a p-value of less than 5% using the previously defined model. This minimum gene expression level was determined in a sample-based manner to accommodate differences in the sequencing depth of the Ribo-Seq experiments.

We predicted all possible ORFs in every transcript (ATG to TGA/TAA/TAG) with a minimum length of 24 amino acids. For every gene, we selected the longest ORF across all transcripts and the longest in-frame ORF, if any translated. In most cases the longest ORF was also the ORF with the largest number of mapped Ribo-Seq reads (75.7% for codRNAs and 84% for lncRNAs). We did not consider annotated pseudogenes and excluded ORFs in lncRNAs that showed significant sequence similarity to know protein-coding sequences. We used the selected ORFs to perform all gene-based analyses. We differentiated between genes with small ORFs (smORFs) and long ORFs. In the fist class the longest ORF in the gene encoded a protein of less than 100 amino acids. The criteria employed excluded non-canonical ORFs, secondary translated ORFs, or translated short isoforms.

ORFs with fewer than 10 mapped reads were classified as non-translated. Otherwise, we analyzed whether ≥60% of the Ribo-seq reads were classified in the correct frame, with a minimum of 10 mapped reads (in-frame). Ambiguous cases with alternative or ill-defined frames were not considered in subsequent analyses (off-frame). This approach correctly classified 97.73% of translated protein-coding genes with more than 10 mapped reads as in-frame. As a control, the position of the reads in each ORF was randomized and the false positive rate of our pipeline was estimated in the different experiments; the rate was <5% in all cases.

Sequence conservation

We searched for homologues of the mouse ORFs in the human transcript assembly using TBLASTN (limited to one strand, e-value < 10^{-4}) [85]. In the case of non-translated ORFs the longest ORF per gene was taken. The longest translated ORF was used for translated genes. In some instances we detected homology even if the ORF was not translated (i.e. conserved non-translated). In these cases we may have indirectly captured sequence similarity at the DNA level or, alternatively, similarity between proteins that were not translated in the tissues analyzed.

Single nucleotide polymorphism data

The SNPs were extracted from dbSNP Build 138 [39], which includes data from 56 different sources. We classified SNPs in ORFs as non-synonymous (PN, amino acid altering) and synonymous (PS, not amino-acid altering). We calculated the PN/PS ratio in each ORF group by using the sum of PN and PS in all the sequences. In general, estimation of PN/PS ratios of individual sequences was not reliable due to lack of sufficient SNP data per ORF. We obtained confidence intervals (95%) using the proportion test. We calculated a neutrally expected (NE) PN/PS for each ORF set by counting the number of nonsynonymous and synonymous positions in the sequences. We also estimated a normalized NE considering a transition to transversion ratio of 1.5 (k=1.5). We used the k value to give different weights to the non-synonymous and synonymous positions depending on whether they were transitions or transversions.

Computation of coding scores with CIPHER

For each hexanucleotide (hexamer), we calculated the relative frequency of the hexamer in the complete set of mouse annotated coding sequences encoding experimentally validated proteins and in the ORFs of a large set of randomly selected intronic sequences [16]. Hexamer frequencies were calculated in frame, using a sliding window and 3 nucleotide steps. Subsequently, we obtained the logarithm of each hexamer frequency in coding sequences divided by the frequency in non-coding sequences. This log likelihood ratio was calculated for each possible hexamer *i* and termed CS$_{hexamer(i)}$. The coding score of an ORF (CS$_{ORF}$) was defined as the average of the hexamer coding scores in the ORF.

The following equations were employed:

$$CS_{hexamer(i)} = \log\left(freq_{coding}\left(hexamer(i)\right) \Big/ freq_{non-coding}\left(hexamer(i)\right) \right)$$

$$CS_{ORF} = \frac{\sum_{i=1}^{i=n} CS_{hexamer(i)}}{n}$$

We have developed a computational tool, CIPHER, that uses this metric to calculate the coding score of the ORFs in any set of sequences. It also predicts the subset of ORFs that are likely to be translated by performing an empirical calculation of p-values derived from the distribution of coding scores in ORFs from introns. Specific parameters have been derived for several eukaryotic species. The code and executable file is freely available at https://github.com/jorruior/CIPHER. The program can also be accessed at http://evolutionarygenomics.imim.es/cipher/.

Using this metric, we divided the set of non-conserved genes into a group of genes with high coding score (NC-H) and a group of genes with low coding score (NC-L). The coding score was measured on the longest ORF with evidence of translation. Genes in the NC-H group were defined as those with a coding score > 0.049 (significant at p-value < 0.05).

Statistical data analyses

The generation of plots and statistical tests was performed with the R package [86].

DATA ACCESS

Transcript assemblies and ribosome profiling-based translation predictions have been deposited at figshare (http://dx.doi.org/10.6084/m9.figshare.3486503). Supplemental file 1 contains supplementary Figures and Tables. Supplemental file 2 in an Excel file that contains detailed information on the translated and non-translated ORFs identified in this study.

SUPPORTING INFORMATION

Table S1. Number of non-synonymous and synonymous single nucleotide polymorphisms (PN and PS, respectively) in diferent ORF classes. PN and PS data was obtained from dbSNP. Conserved: ORFs with homology to human expressed sequences. NC-H: ORFs with no homology to human expressed sequences, coding score > 0.049. NC-L: rest of ORFs with no homology to human expressed sequences. Translated ORF: longest translated ORF in genes with evidence of translation. Non-translated ORF: longest ORF in genes with no evidence of translation. NE refers to the neutral expectation based on the number of non-synonymous and synonymous positions in the ORFs, NE (k=1.5) refers to the same calculation but the frequencies of non-synonymous and synonymous positions were normalized using a transition/transversion ratio of 1.5. Fisher exact test, observed *versus* NE *p-value<0.05, ** p-value < 10^{-3}.

Figure S1. Relationship between the number of reads in a given frame and the number of Ribo-Seq reads that map to the ORF for the hipoccampus sample. Data

is for real ORFs and for a control in which the position of the reads was randomized (random). All transcripts with at least one ORF (Met-STOP) for a protein of 24 amino acids or longer covered by at least 10 Ribo-Seq reads were considered. The ORFs were classified as in-frame if ≥60% reads mapped to the correct frame starting with the Met, or off-frame when less than 60% reads mapped to that frame or when they mapped to another frame. The in-frame ORFs in the random control correspond to the estimated false positive rate (<5% of the ORFs).

Figure S2. Number of tissues in which genes are transcribed, according to the computed FPKM threshold per sample, or translated. Whilst codRNAs are often expressed and translated in most of the samples, annotated and novel lncRNAs often exhibit tissue-specific patterns of expression and translation.

Figure S3. Number of translated small ORFs (smORFs, length < 100 aa) in the eight different samples. Total smORFs: Total in-frame translated smORFs; Unique smORFs: Total in-frame tissue-specific translated smORFs; Sampled smORFs: Total in-frame translated smORFs after randomly subsampling 10 million Ribo-seq reads.

Figure S4. Differences in coding score for conserved and not conserved ORFs divided by length, biotype and presence or absence of translation. For non-translated genes, the longest ORF was considered. Translated ORFs showed significantly higher coding score values than non-translated ones; Wilcoxon test; **, p- value < 0.005, ***, p-value < 10^{-5}. Number of conserved non-translated codRNA smORFs was too low (12) to perform a statistic test.

Figure S5. Left: Influence of coding score in the translatability of non-conserved neutral ORFs normalized by FPKM expression in ES cells (median FPKM value = 0.195). Translated ORFs showed significantly higher coding score values than non-translated ORFs; Wilcoxon rank-sum test; ***, p-value < 10-5. Right: Influence of FPKM expression and ORF length in the translatability of non-conserved neutral ORFs normalized by coding score in ES cells (median score value = -0.029). Translated ORFs showed significantly higher FPKM values; Wilcoxon rank-sum test, ***, p-value < 10-5.

FUNDING

We acknowledge funding from research grants BFU2012-36820,BFU2015-65235 and TIN2015-69175-C4-3-R from the Spanish Government (MINECO), co-funded by the European Regional Development Fund (FEDER, EU), and. We also received funds from Agència de Gestió d'Ajuts Universitaris i de Recerca Generalitat de Catalunya (AGAUR), grant number 2014SGR1121.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: JRO MA. Performed the experiments: JRO, PVG, JVC, XM, MA. Analyzed the data: JR, MA. Contributed reagents/materials/analsys tools: JRO, PVG. Wrote the paper: JRO, MA.

REFERENCES

1. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420: 563–573. doi:10.1038/nature01266.

2. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559–1563. doi:10.1126/science.1112014.

3. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316: 1484–1488. doi:10.1126/science.1138341.

4. Ponjavic J, Ponting CP, Lunter G (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res 17: 556–565. doi:10.1101/gr.6036807.

5. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, et al. (2012) Landscape of transcription in human cells. Nature 489: 101–108. doi:10.1038/nature11233.

6. Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. Annu Rev Biochem 81: 145–166. doi:10.1146/annurev-biochem-051410-092902.

7. Ulitsky I, Bartel DP (2013) lincRNAs: genomics, evolution, and mechanisms. Cell 154: 26–46. doi:10.1016/j.cell.2013.06.020.

8. Van Heesch S, van Iterson M, Jacobi J, Boymans S, Essers PB, et al. (2014) Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. Genome Biol 15: R6. doi:10.1186/gb-2014-15-1-r6.

9. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, 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. doi:10.1101/gr.132159.111.

10. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, et al. (2014) The

evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505: 635–640. doi:10.1038/nature12943.

11. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324: 218–223. doi:10.1126/science.1168978.

12. Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF, et al. (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res 22: 2219–2229. doi:10.1101/gr.133249.111.

13. Ingolia NT, Lareau LF, Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147: 789–802. doi:10.1016/j.cell.2011.10.002.

14. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJS, et al. (2014) Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. Cell Rep 8: 1365–1379. doi:10.1016/j.celrep.2014.07.045.

15. Juntawong P, Girke T, Bazin J, Bailey-Serres J (2014) Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. Proc Natl Acad Sci U S A 111: E203–E212. doi:10.1073/pnas.1317811111.

16. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM (2014) Long non-coding RNAs as a source of new peptides. Elife 3: e03523. doi:10.7554/eLife.03523.

17. Raj A, Wang SH, Shim H, Harpak A, Li YI, et al. (2016) Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. Elife 5. doi:10.7554/eLife.13328.

18. Ji Z, Song R, Regev A, Struhl K (2015) Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife 4: e08890. doi:10.7554/eLife.08890.

19. Chew G-L, Pauli A, Rinn JL, Regev A, Schier AF, et al. (2013) Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. Development 140: 2828–2834. doi:10.1242/dev.098343.

20. Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, et al. (2015) Extensive identification and analysis of conserved small ORFs in animals. Genome Biol 16: 1–21. doi:10.1186/s13059-015-0742-x.

21. Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, et al. (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J 33: 981–993. doi:10.1002/embj.201488411.

22. Crappé J, Van Criekinge W, Trooskens G, Hayakawa E, Luyten W, et al. (2013) Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. BMC Genomics 14: 648. doi:10.1186/1471-2164-14-648.

23. Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, et al. (2014)

Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. Elife 3: e03528. doi:10.7554/eLife.03528.

24. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–426.

25. Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, et al. (2015) Origins of De Novo Genes in Human and Chimpanzee. PLOS Genet 11: e1005721. doi:10.1371/journal.pgen.1005721.

26. Lee C, Yen K, Cohen P (2013) Humanin: a harbinger of mitochondrial-derived peptides? Trends Endocrinol Metab 24: 222–228. doi:10.1016/j.tem.2013.01.005.

27. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, et al. (2016) Detecting actively translated open reading frames in ribosome profiling data. Nat Meth 13: 165–170.

28. Ji Z, Song R, Huang H, Regev A, Struhl K (2016) Transcriptome-scale RNase-footprinting of RNA-protein complexes. Nat Biotechnol 34: 410–413. doi:10.1038/nbt.3441.

29. Gonzalez C, Sims JS, Hornstein N, Mela A, Garcia F, et al. (2014) Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. J Neurosci 34: 10924–10936. doi:10.1523/JNEUROSCI.0084-14.2014.

30. Castañeda J, Genzor P, van der Heijden GW, Sarkeshik A, Yates JR, et al. (2014) Reduced pachytene piRNAs and translation underlie spermiogenic arrest in Maelstrom mutant mice. EMBO J 33: 1999–2019. doi:10.15252/embj.201386855.

31. Guo H, Ingolia NT, Weissman JS, Bartel DP (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 466: 835–840. doi:10.1038/nature09267.

32. Diaz-Munoz MD, Bell SE, Fairfax K, Monzon-Casanova E, Cunningham AF, et al. (2015) The RNA-binding protein HuR is essential for the B cell antibody response. Nat Immunol 16: 415–425.

33. Cho J, Yu N-K, Choi J-H, Sim S-E, Kang SJ, et al. (2015) Multiple repressive mechanisms in the hippocampus during memory formation. Science 350: 82–87. doi:10.1126/science.aac7368.

34. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25: 1915–1927. doi:10.1101/gad.17446611.

35. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, et al. (2015) Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. Cell Rep 11: 1110–1122. doi:10.1016/j.celrep.2015.04.023.

36. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, et al. (2012) Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet 8: e1002841. doi:10.1371/journal.pgen.1002841.

37. Kong L, Zhang Y, Ye Z-Q, Liu X-Q, Zhao S-Q, et al. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35: W345–W349. doi:10.1093/nar/gkm391.

38. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, et al. (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res 41: e74–e74. doi:10.1093/nar/gkt006.

39. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308–311. doi:10.1093/nar/29.1.308.

40. Zhao Z, Fu Y-X, Hewett-Emmett D, Boerwinkle E (2003) Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene 312: 207–213.

41. Gorlov IP, Kimmel M, Amos CI (2006) Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome. Hum Mol Genet 15: 1143–1150. doi:10.1093/hmg/ddl029.

42. Li W-H (1997) Molecular Evolution. Sinauer Associates.

43. Dinger ME, Pang KC, Mercer TR, Mattick JS (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol 4: e1000176. doi:10.1371/journal.pcbi.1000176.

44. Anderson DM, Anderson KM, Chang C-L, Makarewich CA, Nelson BR, et al. (2015) A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. Cell 160: 595–606. doi:10.1016/j.cell.2015.01.009.

45. Buck-Koehntop BA, Mascioni A, Buffy JJ, 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. doi:10.1016/j.jmb.2005.09.038.

46. Pueyo JI, Magny EG, Sampson CJ, Amin U, Evans IR, et al. (2016) Hemotin, a Regulator of Phagocytosis Encoded by a Small ORF and Conserved across Metazoans. PLoS Biol 14: e1002395. doi:10.1371/journal.pbio.1002395.

47. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell 147: 1537–1550. doi:10.1016/j.cell.2011.11.055.

48. Tian D, Sun S, Lee JT (2010) The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. Cell 143: 390–403. doi:10.1016/j.cell.2010.09.049.

49. Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, et al. (2009) An

architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol Cell 33: 717–726. doi:10.1016/j.molcel.2009.01.026.

50. Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, et al. (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell 39: 925–938. doi:10.1016/j.molcel.2010.08.011.

51. Housman G, 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.

52. Gayà-Vidal M, Albà MM (2014) Uncovering adaptive evolution in the human lineage. BMC Genomics 15: 599. doi:10.1186/1471-2164-15-599.

53. Cai JJ, Petrov DA (2010) Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. Genome Biol Evol 2: 393–409. doi:10.1093/gbe/evq019.

54. Wiberg RAW, Halligan DL, Ness RW, Necsulea A, Kaessmann H, et al. (2015) Assessing Recent Selection and Functionality at Long Noncoding RNA Loci in the Mouse Genome. Genome Biol Evol 7: 2432–2444. doi:10.1093/gbe/evv155.

55. Crappé J, Ndah E, Koch A, Steyaert S, Gawron D, et al. (2015) PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. Nucleic Acids Res 43: e29. doi:10.1093/nar/gku1283.

56. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, et al. (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol 9: 59–64. doi:10.1038/nchembio.1120.

57. Saghatelian A, Couso JP (2015) Discovery and characterization of smORF-encoded bioactive polypeptides. Nat Chem Biol 11: 909–916. doi:10.1038/nchembio.1964.

58. Pauli A, Valen E, Schier AF (2015) Identifying (non-)coding RNAs and small peptides: challenges and opportunities. Bioessays 37: 103–112. doi:10.1002/bies.201400103.

59. Lynch M, Marinov GK (2015) The bioenergetic costs of a gene. Proc Natl Acad Sci U S A 112: 15690–15695. doi:10.1073/pnas.1514974112.

60. Yu C, Dang Y, Zhou Z, Wu C, Zhao F, et al. (2015) Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. Mol Cell 59: 744–754.

61. Presnyak V, Alhusaini N, Chen Y-H, Martin S, Morris N, et al. (2015) Codon optimality is a major determinant of mRNA stability. Cell 160: 1111–1124. doi:10.1016/j.cell.2015.02.029.

62. Dos Reis M, Wernisch L (2009) Estimating translational selection in eukaryotic genomes. Mol Biol Evol 26: 451–461. doi:10.1093/molbev/msn272.

63. Doherty A, McInerney JO (2013) Translational selection frequently overcomes genetic drift in shaping synonymous codon usage patterns in vertebrates. Mol Biol Evol 30: 2263–2267. doi:10.1093/molbev/mst128.

64. Heinen TJAJ, Staubach F, Häming D, Tautz D (2009) Emergence of a new gene from an intergenic region. Curr Biol 19: 1527–1531. doi:10.1016/j.cub.2009.07.049.

65. Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ (2006) Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci U S A 103: 9935–9939. doi:10.1073/pnas.0509809103.

66. Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, et al. (2013) De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS Genet 9: e1003860. doi:10.1371/journal.pgen.1003860.

67. Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, et al. (2009) Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol 26: 603–612. doi:10.1093/molbev/msn281.

68. McLysaght A, Guerzoni D (2015) New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philos Trans R Soc Lond B Biol Sci 370. doi:10.1098/rstb.2014.0332.

69. Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, et al. (2012) Proto-genes and de novo gene birth. Nature 487: 370–374. doi:10.1038/nature11184.

70. Knowles DG, McLysaght A (2009) Recent de novo origin of human protein-coding genes. Genome Res 19: 1752–1759. doi:10.1101/gr.095026.109.

71. Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E (2013) Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol Evol 5: 439–455. doi:10.1093/gbe/evt009.

72. Bornberg-Bauer E, Schmitz J, Heberlein M (2015) Emergence of de novo proteins from "dark genomic matter" by "grow slow and moult". Biochem Soc Trans 43: 867–873. doi:10.1042/BST20150089.

73. Schlötterer C (2015) Genes from scratch – the evolutionary fate of de novo genes. Trends Genet. doi:10.1016/j.tig.2015.02.007.

74. Albà MM, Castresana J (2005) Inverse relationship between evolutionary rate and age of mammalian genes. Mol Biol Evol 22: 598–606. doi:10.1093/molbev/msi045.

75. Toll-Riera M, Bostick D, Albà MM, Plotkin JB (2012) Structure and age jointly influence rates of protein evolution. PLoS Comput Biol 8: e1002542. doi:10.1371/journal.pcbi.1002542.

76. Xie C, Zhang YE, Chen J-Y, Liu C-J, Zhou W-Z, et al. (2012) Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. PLoS

Genet 8: e1002942. doi:10.1371/journal.pgen.1002942.

77. Chen J-Y, Shen QS, Zhou W-Z, Peng J, He BZ, et al. (2015) Emergence, Retention and Selection: A Trilogy of Origination for Functional De Novo Proteins from Ancestral LncRNAs in Primates. PLoS Genet 11: e1005391. doi:10.1371/journal.pgen.1005391.

78. Palmieri N, Kosiol C, Schlötterer C (2014) The life cycle of Drosophila orphan genes. Elife 3: e01311.

79. Zhao L, Saelao P, Jones CD, Begun DJ (2014) Origin and spread of de novo genes in Drosophila melanogaster populations. Science 343: 769–772. doi:10.1126/science.1248286.

80. Neme R, Tautz D (2016) Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. Elife 5. doi:10.7554/eLife.09977.

81. Wilson BA, Masel J (2011) Putatively noncoding transcripts show extensive association with ribosomes. Genome Biol Evol 3: 1245–1252. doi:10.1093/gbe/evr099.

82. Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, et al. (2013) Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Rep 3: 2179–2190. doi:10.1016/j.celrep.2013.05.031.

83. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, et al. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14: R36. doi:10.1186/gb-2013-14-4-r36.

84. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, et al. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotech 33: 290–295.

85. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.

86. R Development Core Team (2016) R: a language and environment for statistical computing.