bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 essentially non-coding. These results are based on the selective sequencing of ribosome-protected fragments, or ribosome profiling. The present study used ribosome profiling data from eight mouse tissues and cell types, combined with ~330,000 synonymous and non-synonymous single nucleotide variants, to dissect the biological implications of lncRNA translation. Using the three-nucleotide read periodicity that characterizes actively translated regions, we found that about 23% of the transcribed lncRNAs was translated (1,365 out of 6,390). About one fourth of the translated sequences (350 lncRNAs) showed 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 contains functional proteins which are not conserved in humans. The second large class, comprising >500 lncRNAs, produces proteins that show no significant purifying selection signatures. We showed that the neutral translation of these lncRNAs depends on the transcript expression level and the chance occurrence of ORFs with a favorable codon composition. This provides the first evidence to data that many lncRNAs produce non-functional proteins.

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

In recent years, the use of transcriptomics has revealed that, in addition to classical protein-coding transcripts, the cell expresses thousands of long transcripts with weak coding potential (Okazaki et al. 2002; Carninci et al. 2005; Kapranov et al. 2007; Ponjavic et al. 2007; Djebali et al. 2012). 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. (Rinn and Chang 2012; Ulitsky and Bartel 2013). However, the vast majority of them remain functionally uncharacterized. While some lncRNAs have nuclear roles, the majority are polyadenylated and accumulate in the cytoplasm (van Heesch et al. 2014). In addition, many lncRNAs are expressed at low levels and have a limited phylogenetic distribution (Derrien et al. 2012; Necsulea et al. 2014).

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 (Ingolia et al. 2009). 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 (Ingolia et al. 2011, 2014; Juntawong et al. 2014; Ruiz-Orera et al. 2014; Raj et al. 2016; Ji et al. 2015; Chew et al. 2013). 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 (Mackowiak et al. 2015; Bazzini et al. 2014; Crappé et al. 2013). This is likely to include some recently evolved proteins that lack homologues in other species and which are even harder to detect than conserved peptides (Ruiz-Orera et al. 2014).

One striking feature of the proteins produced by lncRNAs is that, in general, they appear to be under lower selective constraints than standard proteins (Ruiz-Orera et al., 2014). This raises the possibility that a large fraction of them encode proteins that, despite being translated in a stable manner, are not functional. The

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

present study is aimed at testing this hypothesis.

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 selection there is a deficit of non-synonymous variants (Nei and Gojobori 1986). 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. We use this information to distinguish between function and non-functional classes of lncRNA translated products.

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. For the assembly, we used more than 1.5 billion strand-specific RNA sequencing reads from mouse (Ruiz-Orera et al. 2015).

We developed a method based on the well-known three-nucleotide periodicity of Ribo-Seq reads in actively translated regions (Ingolia et al. 2009; Bazzini et al. 2014; Ji et al. 2015) in order to very precisely identify translated sequences. First, we selected all expressed transcripts containing at least one ORF encoding a putative protein of 24 amino acids or longer. Then, for each experiment, we defined translated ORFs as those covered by 10 or more Ribo-Seq reads, of which at least 60% matched the expected frame (in-frame, Figure 1A). To avoid redundancy in

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

subsequent gene-focused analyses, we selected the longest translated ORF when several existed. The vast majority of ORFs detected by ribosome profiling had clear three-nucleotide periodicity (Figure 1B), regardless of whether they were in coding genes (codRNA) or in lncRNAs (Supplemental file 1 Figure S1). To determine how often this bias occurs by chance alone, we randomized the position of each read in the ORFs. This random model estimated a false positive rate <5% (Figure 1B). In ORFs classified as translated, 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 1C, Methods). Most coding genes were transcribed and translated in several samples, whereas lncRNAs tended to be sample-specific (Supplemental file 1 Fig 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) (Gonzalez et al. 2014), (2) (Castañeda et al.

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2014), (3) (Guo et al. 2010), (4) (Diaz-Munoz et al. 2015), (5) (Cho et al. 2015)

Figure 1

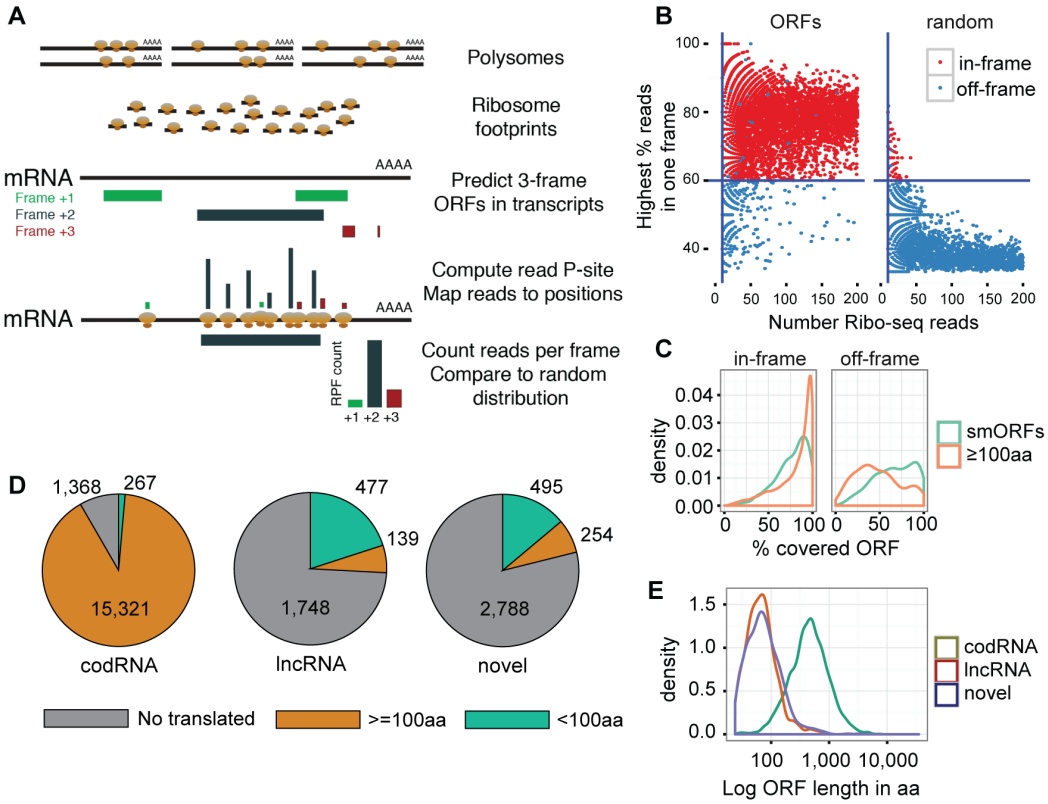


Figure 1. Detection of translated ORFs. A Workflow to identify translated ORFs. Ribosome profiling (Ribo-Seq) reads, corresponding to ribosome-protected fragments, are mapped to all predicted ORFs in transcripts. This is performed with single-nucleotide resolution after computing the read P-site per 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. 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-

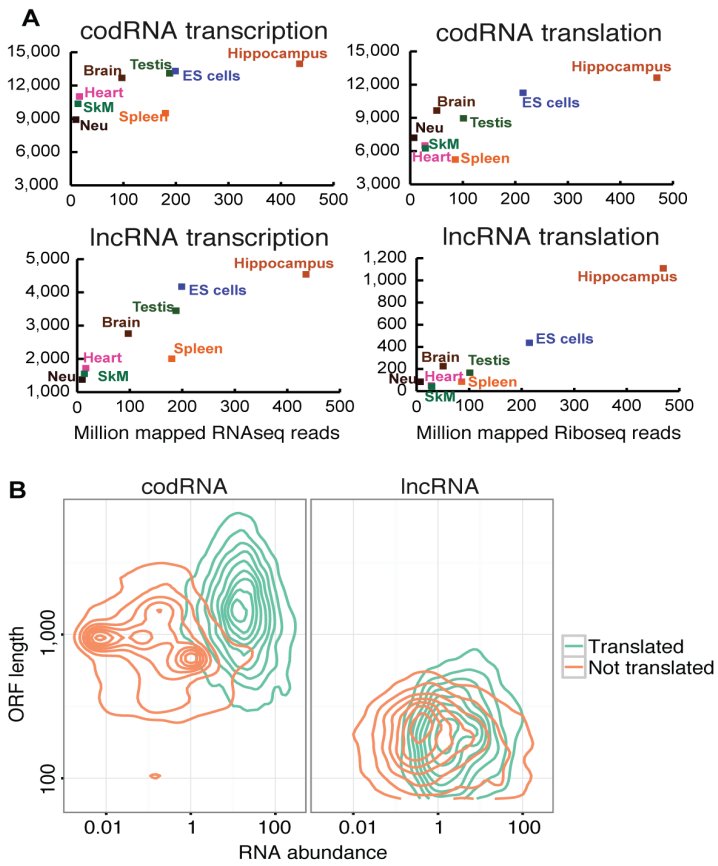
bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 across experiments

The number of transcribed and translated genes varied substantially depending on the sample, especially for lncRNAs and smORFs (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



bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Cabili et al. 2011; Derrien et al. 2012). Translated lncRNAs were also expressed at significantly higher levels, and contained longer ORFs, than non-translated lncRNAs (Wilcoxon test, p-value < 10^{-5}).

A subset of lncRNAs encodes functional micropeptides

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 (Kutter et al. 2012; Wiberg et al. 2015). Here we examined which fraction of the mouse lncRNAs were conserved in humans. First we generated a *de novo* human transcriptome assembly of a quality similar to that used for mouse (Methods). As we were interested in protein translation, conservation was assessed using ORF-based sequence similarity searches (see Methods).

Whereas the vast majority (~98%) of mouse protein coding genes were conserved in the human model, we detected conservation for only about 25% of the translated lncRNAs (Figure 3A). This is not surprising given previous observations that lncRNAs tend to have a limited phylogenetic distribution (Hezroni et al. 2015; Kutter et al. 2012; Necsulea et al. 2014).

A key question was whether or not the proteins produced by conserved lncRNAs were functional. We addressed it by using a very large number of mouse single nucleotide polymorphism (SNP) variants from dbSNP (Sherry et al. 2001) –157,029 non-synonymous SNPs (PN) and 179,825 synonymous SNPs (PS)– that mapped to the ORFs in our dataset. It is well known that there are approximately three times more non-synonymous than synonymous positions in coding sequences. For this reason, we expect PN/PS ~3 in neutrally evolving sequences, which accept all

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mutations (Nei and Gojobori 1986). Values significantly lower than this indicate less tolerance for mutations at non-synonymous positions than at synonymous ones, consistent with negative or purifying selection at the protein level.

Figure 3

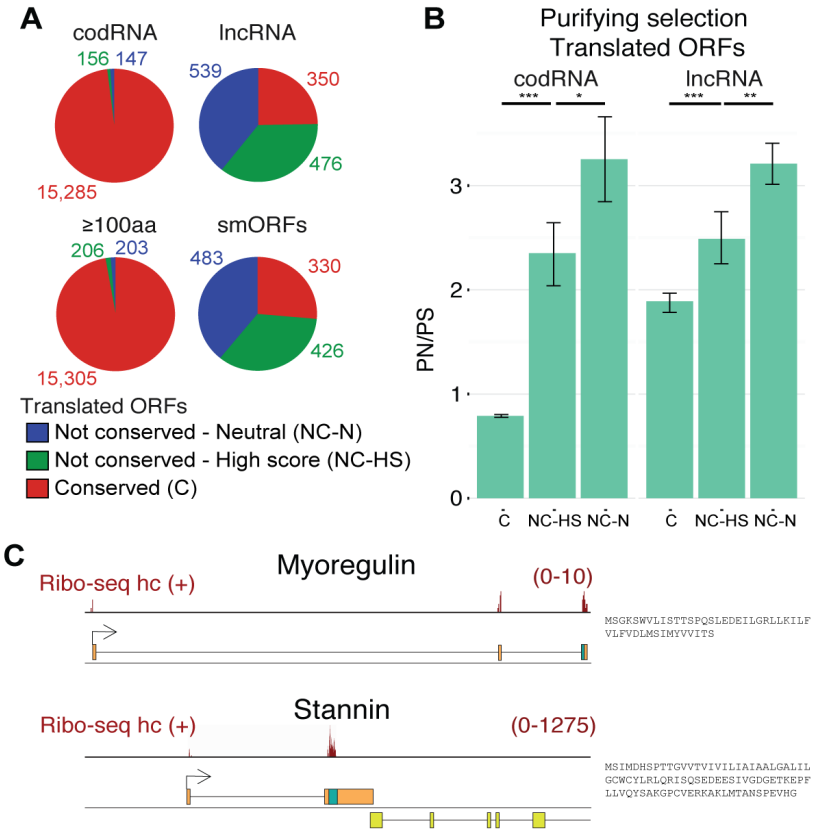


Figure 3. Different classes of translated ORFs. A. Number of translated ORFs that are conserved in human (C), not conserved but showing a high coding score (NC-HS, coding score > 0.049, significant at p-value < 0.05) and neutral (NC-N, coding score ≤ 0.049). First, ORFs are divided into codRNA and lncRNA, and second, into long (length ≥ 100 amino acids) and small ORFs (smORFs, length < 100 amino acids). **B**. Analysis of selective constraints in translated ORFs. PN/PS refers to the ratio between non-synonymous (PN) and synonymous (PS) single nucleotide variants. Conserved and high-score ORFs show significant purifying selection signatures independently of transcript type. 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 are indicated. Fisher test *p-value < 0.05, **p- value < 0.005, ***, p-value < 10^{-5}. Error bars represent the 95% confidence interval. **C.** 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 data is from hippocampus (hc) ribosome profiling experiments.

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We found that conserved translated transcripts, both codRNAs and lncRNA, had PN/PS values significantly lower than the neutral expectation (Fisher test p-value < 10^{-5}, Figure 3B, Supplemental file 1 Table S1). This strongly suggests that most of the 350 lncRNAs in this group are in fact 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 (Dinger et al. 2008). 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 (Anderson et al. 2015) (Figure 3C). Myoregulin was annotated as non-coding when we initiated the study although it has now been re-classified as protein-coding. Many similar cases are expected to emerge in the next years.

Many lncRNAs produce non-functional proteins

The prediction of coding sequences usually takes into account features such as codon frequencies, ORF length, and sequence conservation (Kong et al. 2007; Wang et al. 2013). In the case of non-conserved short ORFs we can only apply measures based on codon composition. 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 (Ruiz-Orera et al. 2014). 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 4A). We then used CIPHER to divide the non-conserved genes into a group with high coding scores (NC-HS, coding score >0.049, significant at p-value <0.05) and another group with lower coding scores (≤0.049). The first group showed weaker purifying selection that the conserved genes discussed in the previous section. However, PN/PS was still significantly lower than the neutral expectation (Fisher test, p-value <0.005), indicating that this class of genes includes a number of mouse genes coding for proteins that, despite not being

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

conserved in humans, are functional. In contrast, the PN/PS value in the group of non-conserved genes with low coding scores (NC-N) was consistent with neutral evolution (Figure 3B, Supplemental file 1 Table S1). 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%). Altogether, this class comprised 686 genes. The Ribo-Seq reads mapping to the corresponding ORFs had clear frame bias (Figure 4B), which was highly consistent across different tissues (Figure 4C); this provided additional evidence that they were indeed translated.

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 (Ulitsky et al. 2011), *Jpx* functions in X chromosome inactivation (Tian et al. 2010), *Neat1* has a role in the maintenance and assembly of paraspeckles (Clemson et al. 2009), and *Malat1* regulates the expression of other genes (Tripathi et al. 2010). Our results support the targeting of lncRNAs by the translational machinery, probably as a result of promiscuous activity of the ribosome rather than any important role of these proteins in the cell.

A key question was why we detected translation of some lncRNAs but not of others. This is especially relevant for lncRNAs that translate non-functional proteins, as presumably no selective forces are involved. One obvious likely factor is the gene expression level. This is 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, for non-functional proteins, the translated ORFs exhibited higher coding scores than non-translated ORFs (Figure 4D, Wilcoxon test p-value $<10^{-5}$). Importantly, we obtained a similar result after controlling for transcript abundance (Figure 4E for hippocampus, Wilcoxon test p-value $<10^{-5}$; Supplemental file 1 Figure S5 for

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 4F 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 4F).

Figure 4

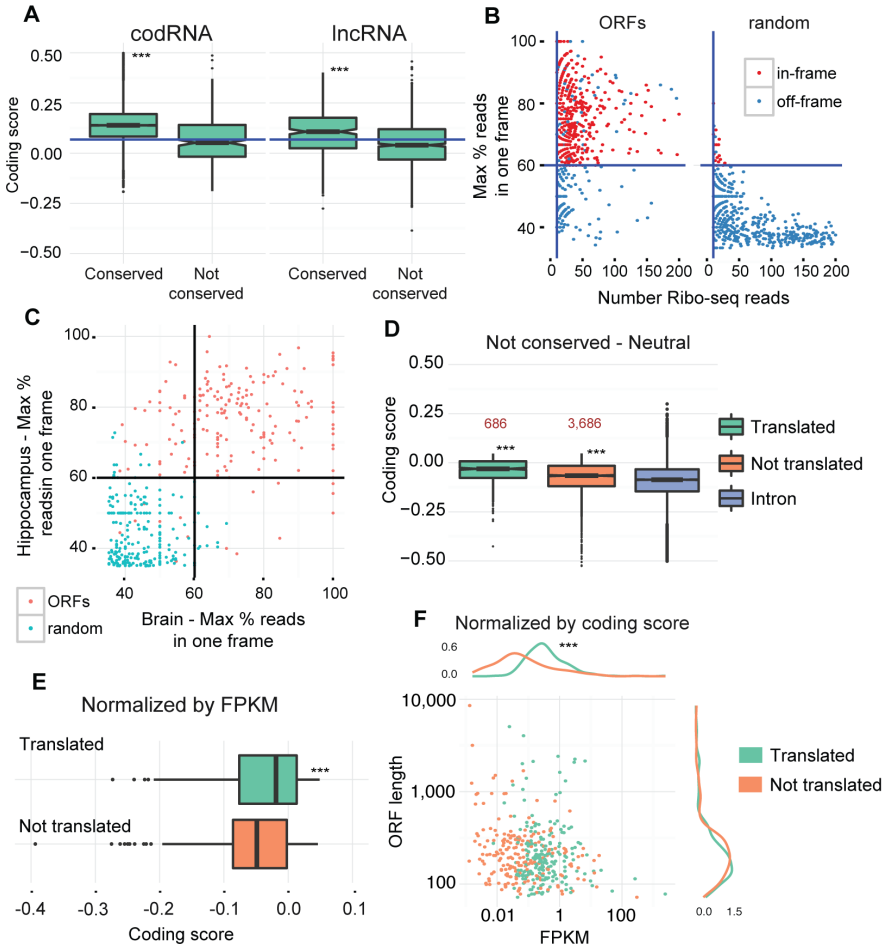


Figure 4. Factors influencing the translatability of lncRNAs. A. Differences in coding score for conserved (C) and non-conserved ORFs (NC). Conserved ORFs showed significantly higher coding score values than non-conserved ones; Wilcoxon test; ***, p-value < 10^{-5}. Blue line indicates the coding score value used to separate non-conserved ORFs with high coding scores (NC-HS) to the rest of non-conserved ORFs. **B .** 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-N) for the hippocampus sample. Data is for real transcripts and for controls in which the position of the reads was randomized (random). **C.** Relationship

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between the percentage of reads falling in a frame in brain and hippocampus samples, for NC-N ORFs. Data is for real transcripts and for controls in which the position of the reads was randomized (random). **D.** Influence of coding score in the translatability of non-conserved neutral ORFs (NC-N). Intronic ORFs are shown for comparison. Translated ORFs showed significantly higher coding score than non-translated ORFs; Wilcoxon test; ***, p-value < 10^{-5}. **E .** 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^{-5}. **F .** 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

There is mounting evidence 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. In comparing data from different mouse experiments, we observed that the number of translated lncRNAs detected depends not only on the stringency of the method but also on the sequencing depth.

The recent discovery that a large number of lncRNAs show ribosome profiling patterns consistent with translation has puzzled many scientists (Housman G and Ulitsky I 2016). Most 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. The analysis is undertaken for predefined groups of genes, as individual coding sequences do not usually contain enough polymorphims for sound statistical analysis (Gayà-Vidal and Albà 2014). 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 (Ruiz-Orera et al. 2014). Together with previous observations that lncRNAs tend to be lineage-specific (Necsulea et al.

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2014) and that young proteins evolve under relaxed purifying selection (Cai and Petrov 2010), this finding led us to suggest that lncRNAs are enriched in young protein-coding genes. Taking one step further, in the present study we employed recently generated mouse and human deep transcriptomes data, together with extensive mouse variation data, to delineate the exact nature of the relationship between species conservation and function of the translated ORFs. This resulted in the identification of three broad classes of translated lncRNAs: conserved and functional, non-conserved and functional (but with low constraints), and non-conserved and non-functional.

We estimate that about 5% of the analyzed lncRNAs encode functional micropeptides (smORFs) that are conserved in humans. 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 (Crappé et al. 2015; Slavoff et al. 2013; Saghatelian and Couso 2015; Pauli et al. 2015). As shown here, and in other recent studies (Bazzini et al. 2014; Mackowiak et al. 2015), 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 important 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 (Lynch and Marinov 2015). 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.

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Yu et al. 2015). 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 (Ji et al. 2015). 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 (dos Reis and Wernisch 2009; Doherty and McInerney 2013). 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* (Heinen et al. 2009). The group of ORFs that had high coding scores, but lacked conservation in humans had significant purifying selection signatures. This was independent of the annotated coding status of the transcript, reinforcing the idea that the differences between coding and non-coding genes in this group are very tenuous (Ruiz-Orera et al., 2014).

There is accumulating evidence that some protein-coding genes have originated *de novo* from previously non-functional genomic regions (Reinhardt et al. 2013; Toll-Riera et al. 2009; McLysaght and Guerzoni 2015; Carvunis et al. 2012; Knowles and McLysaght 2009). These *de novo* genes encode proteins with unique sequences that may have played a role in lineage-specific adaptations. It has been hypothesized that many of these genes originated from lncRNAs (Xie et al. 2012; Chen et al. 2015; Ruiz-Orera et al. 2015), which would be consistent with the large number of species-specific transcripts with lncRNA features identified in comparative transcriptomics studies (Palmieri et al. 2014; Zhao et al. 2014; Neme and Tautz 2016; Ruiz-Orera et al. 2015). The discovery that some non-coding RNAs are translated makes the transition from non-coding/non-functional to coding/functional more probable than previously anticipated. (Wilson and Masel

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2011). This is because the translation products, even if generated by pure accident, can be tested for useful 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. Therefore, the unproductive translation of lncRNAs can be regarded as 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 the notion that the analysis of lncRNA translation patterns is expected to lead to many new discoveries related to the world of 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 (Ruiz-Orera et al. 2015), GSE43721 (Soumillon et al. 2013), and GSE43520 (Necsulea et al. 2014). 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 (Ruiz-Orera et al. 2015). 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) (Kim et al. 2013). Multiple mapping to several locations in the genome was allowed unless otherwise stated.

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We assembled the transcriptome with Stringtie (Pertea et al. 2015) 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 (Gonzalez et al. 2014), GSE50983 (Castañeda et al. 2014), GSE22001 (Guo et al. 2010), GSE62134 (Diaz-Munoz et al. 2015), GSE72064 (Cho et al. 2015), and GSE41426. Data corresponded to brain, testis, neutrophils, splenic B cells, ES cells, hippocampus, heart and skeletal muscle (Table 1).

Reads were filtered as in the previous datasets and 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 (Calviello et al. 2016; Ji et al. 2015; Bazzini et al. 2014; Ingolia et al. 2009). 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

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 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. We used these selected ORFs to perform all gene-based analyses. Genes with smORFs were defined as having <100 amino acids in the longest ORF. These criteria 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 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}) (Altschul et al. 1997). 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 (Sherry et al. 2001), which

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Ruiz-Orera et al. 2014). 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

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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-HS) and a group of genes with low coding score (NC-N). The coding score was measured on the longest ORF with evidence of translation. Genes in the NC-HS group were defined as those with a coding score > 0.049. This group has a lower PN/PS ratio than the NC-N group.

Statistical data analyses

The generation of plots and statistical tests was performed with the R package (R Development Core Team 2016).

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 2 contains detailed information on the translated and non-translated ORFs.

ACKNOWLEDGEMENTS

We are grateful to Elaine Lilly, Ph.D., for text revision. 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.

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–402.

Anderson DM, Anderson KM, Chang C-L, Makarewich CA, Nelson BR, McAnally JR, Kasaragod P, Shelton JM, Liou J, Bassel-Duby R, et al. 2015. A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* **160**: 595–606.

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, et al. 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**: 981–93.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**: 1915–27.

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.

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

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309**: 1559–63.

Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature* **487**: 370–4.

Castañeda J, Genzor P, van der Heijden GW, Sarkeshik A, Yates JR, Ingolia NT, Bortvin A. 2014. Reduced pachytene piRNAs and translation underlie spermiogenic arrest in Maelstrom mutant mice. *EMBO J* **33**: 1999–2019.

Chen J-Y, Shen QS, Zhou W-Z, Peng J, He BZ, Li Y, Liu C-J, Luan X, Ding W, Li S, 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.

Chew G-L, Pauli A, Rinn JL, Regev A, Schier AF, Valen E. 2013. Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. *Development* **140**: 2828–34.

Cho J, Yu N-K, Choi J-H, Sim S-E, Kang SJ, Kwak C, Lee S-W, Kim J, Choi D Il, Kim VN, et al. 2015. Multiple repressive mechanisms in the hippocampus during memory formation. *Science* **350**: 82–87.

Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, Lawrence JB. 2009. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol Cell* **33**: 717–726.

Crappé J, Van Criekinge W, Trooskens G, Hayakawa E, Luyten W, Baggerman G, Menschaert G, Pdf TP, Genomics BMC, Crapp J, et al. 2013. Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics* **14**: 648.

Crappé J, Ndah E, Koch A, Steyaert S, Gawron D, De Keulenaer S, De Meester E, De Meyer T, Van Criekinge W, Van Damme P, et al. 2015. PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration.

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Nucleic Acids Res **43**: e29.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**: 1775–89.

Diaz-Munoz MD, Bell SE, Fairfax K, Monzon-Casanova E, Cunningham AF, Gonzalez-Porta M, Andrews SR, Bunik VI, Zarnack K, Curk T, et al. 2015. The RNA-binding protein HuR is essential for the B cell antibody response. *Nat Immunol* **16**: 415–425.

Dinger ME, Pang KC, Mercer TR, Mattick JS. 2008. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol* **4**: e1000176.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489**: 101–108.

Doherty A, McInerney JO. 2013. Translational selection frequently overcomes genetic drift in shaping synonymous codon usage patterns in vertebrates. *Mol Biol Evol* **30**: 2263–7.

Gayà-Vidal M, Albà MM. 2014. Uncovering adaptive evolution in the human lineage. *BMC Genomics* **15**: 599.

Gonzalez C, Sims JS, Hornstein N, Mela A, Garcia F, Lei L, Gass DA, Amendolara B, Bruce JN, Canoll P, et al. 2014. Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *J Neurosci* **34**: 10924–36.

Guo H, Ingolia NT, Weissman JS, Bartel DP. 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**: 835–40.

Van Heesch S, van Iterson M, Jacobi J, Boymans S, Essers PB, de Bruijn E, Hao W, Macinnes AW, Cuppen E, Simonis M. 2014. Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol* **15**: R6.

Heinen TJAJ, Staubach F, Häming D, Tautz D. 2009. Emergence of a new gene from an intergenic region. *Curr Biol* **19**: 1527–31.

Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. 2015. Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Rep* **11**: 1110–22.

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.

Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJS, Jackson SE, Wills MR, Weissman JS. 2014. Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Rep* **8**: 1365–79.

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**: 218–23.

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.

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.

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**: E203–12.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**: 1484–8.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.

Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res* **19**: 1752–9.

Kong L, Zhang Y, Ye Z-Q, Liu X-Q, Zhao S-Q, Wei L, Gao G. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35**: W345–W349.

Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC. 2012. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet* **8**: e1002841.

Lynch M, Marinov GK. 2015. The bioenergetic costs of a gene. *Proc Natl Acad Sci U S A* **112**: 15690–5.

Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, Mastrobuoni G, Rajewsky N, Kempa S, Selbach M, et al. 2015. Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol* **16**: 1–21.

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**.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**: 635–40.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**: 418–26.

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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**.

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**: 563–73.

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

Pauli A, Valen E, Schier AF. 2015. Identifying (non-)coding RNAs and small peptides: challenges and opportunities. *Bioessays* **37**: 103–12.

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

Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17**: 556–65.

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

Raj A, Wang SH, Shim H, Harpak A, Li YI, Engelmann B, Stephens M, Gilad Y, Pritchard JK. 2016. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* **5**.

Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. 2013. De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet* **9**: e1003860.

Dos Reis M, Wernisch L. 2009. Estimating translational selection in eukaryotic genomes. *Mol Biol Evol* **26**: 451–61.

Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. *Annu Rev Biochem* **81**: 145–66.

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, Marqués-Bonet T, Albà MM. 2015. Origins of De Novo Genes in Human and Chimpanzee ed. J. Noonan. *PLOS Genet* **11**: e1005721.

Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. 2014. Long non-coding RNAs as a source of new peptides. *Elife* **3**: e03523.

Saghatelian A, Couso JP. 2015. Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat Chem Biol* **11**: 909–16.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308–11.

Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A. 2013. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* **9**: 59–64.

Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthès P,

bioRxiv preprint doi: https://doi.org/10.1101/064915; this version posted July 20, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Kokkinaki M, Nef S, Gnirke A, et al. 2013. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep* **3**: 2179–90.

Tian D, Sun S, Lee JT. 2010. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell* **143**: 390–403.

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Albà MM. 2009. Origin of primate orphan genes: a comparative genomics approach. *Mol Biol Evol* **26**: 603–12.

Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA, et al. 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell* **39**: 925–38.

Ulitsky I, Bartel DP. 2013. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**: 26–46.

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**: 1537–50.

Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* **41**: e74–e74.

Wiberg RAW, Halligan DL, Ness RW, Necsulea A, Kaessmann H, Keightley PD. 2015. Assessing Recent Selection and Functionality at Long Noncoding RNA Loci in the Mouse Genome. *Genome Biol Evol* **7**: 2432–44.

Wilson BA, Masel J. 2011. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol Evol* **3**: 1245–52.

Xie C, Zhang YE, Chen J-Y, Liu C-J, Zhou W-Z, Li Y, Zhang M, Zhang R, Wei L, Li C-Y. 2012. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet* **8**: e1002942.

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

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in Drosophila melanogaster populations. *Science* **343**: 769–72.