Evidence for functional and non-functional classes of peptides translated from 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]Computer Sciences Department, Universitat Politècnica de Catalunya, Barcelona, Spain; [3]Catalan Institution for Research and Advanced Studies, Barcelona, Spain.

*To whom correspondence should be addressed.

Running title: translation of lncRNAs

Keywords: long non-coding RNA, ribosome profiling, translation, RNA-Seq, peptide, polymorphism, natural selection

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 show that a significant proportion of lncRNAs are translated, challenging the view that they are non-coding. These results are based on 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 patterns of purifying selection in proteins translated from lncRNAs. Using the three-nucleotide read periodicity that characterizes actively translated regions, we identified 832 mouse translated lncRNAs. Overall, they produced 1,489 different proteins, most of them smaller than 100 amino acids. Nearly half of the ORFs then showed sequence conservation in rat and/or human transcripts, and many of them are likely to encode functional micropeptides, including the recently discovered Myoregulin. For lncRNAs not conserved in rats or humans, the ORF codon usage bias distinguished between two classes, one with particularly high coding scores and evidence of purifying selection, consistent with the presence of lineage-specific functional proteins, and a second, larger, class of ORFs producing peptides with no significant purifying selection signatures. We obtained evidence that the translation of these lncRNAs depends on 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 lncRNAs remain functionally uncharacterized. Some have nuclear roles, but most are polyadenylated and accumulate 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 (Ribo-Seq) [11]. This technique generates millions of ribosome footprints that can be mapped to a species genome or transcriptome to assess the translation of thousands of open reading frames (ORFs) [12], including low-abundant small peptides that may be difficult to detect by standard proteomics approaches [13–15]. In ribosome profiling experiments, the three-nucleotide periodicity of the reads, resulting from the movement of the ribosome along the coding sequence, can be used to differentiate translated sequences from other possible RNA protein complexes [13,16–19]. A growing number of studies based on this technique have found that a significant proportion of lncRNAs are translated [16,18,20–25], but 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 length, have not been correctly predicted by bioinformatics algorithms [13,15,26,27]. This is likely to include some recently evolved proteins that lack homologues in other species and are even harder to detect than conserved short peptides [24].

One striking feature of the ORFs reported to be translated from lncRNAs is that, in general, they appear to have fewer selective constraints than standard proteins [18,24], raising the possibility that a large fraction of them encode proteins that are not functional, despite being translated in a stable manner. However, evidence for this hypothesis is presently lacking.

Non-synonymous and synonymous single nucleotide polymorphisms in coding sequences provide useful information to distinguish between neutrally evolving proteins and proteins undergoing 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 [28]. The detection of selection signatures provides strong evidence of functionality, whereas non-functional proteins evolve neutrally. The present study takes advantage of existing nucleotide variation data for the house mouse to investigate the selective patterns of peptides translated by lncRNAs. Our findings provide evidence that lncRNAs are pervasively translated and that a large fraction of them produce neutrally evolving peptides. We discuss the importance of these peptides as raw material for the evolution of *de novo* protein-coding genes.

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 (codRNAs) from mouse, using ribosome-profiling RNA-sequencing (Ribo-Seq) data from eight different tissues and cell types (Table 1 and references therein). The samples corresponded to healthy individuals and comprised hippocampus, neural embryonic stem cells, brain, testis, neutrophils, splenic B cells, heart and skeletal muscle. 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 comprised 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 (novel lncRNAs, see Methods). For the assembly of this transcriptome, we used more than 1.5 billion strand-specific RNA sequencing reads from mouse [29].

We selected all expressed transcripts (FPKM > 0.2, see Methods) and predicted all possible canonical ORFs encoding putative proteins with a length of at least 9 amino acids. For each mapped Ribo-Seq experiment, we selected the ORFs covered by at least 10 Ribo-Seq reads and examined the distribution of the predicted ribosome P-sites along the ORF using RibORF software [16] (Fig 1a). ORFs classified as translated by the program showed clear three-nucleotide periodicity and uniformity when

compared to the reads for the rest of the ORFs (Fig 1b and 1c). These two biases are characteristic of regions that are being actively translated [11,20,16,13,17], and are absent from other types of protein-RNA interactions [30]).

This method translated ORFs in ~20% of the loci annotated as lncRNAs and ~90% of the coding genes (Table 1 and Fig 1d). We also identified 286 novel genes that did not overlap with annotated protein-coding genes but contained translated ORFs. A substantial fraction of the codRNAs (29.54%) showed translation of more than one non-overlapping ORF; 2,954 ORFs were located upstream or downstream of the main protein-coding ORF in the same transcript (uORFs and dORFs, respectively), while 3,951 ORFs corresponded to putative alternatively translated products. Moreover, we found that 325 lncRNAs (~39%, including annotated and novel lncRNAs) showed evidence of polycistronic translation, producing two or more peptides.

Sample	GEO (ref)	Annotated codRNA			Annotated lncRNA			Novel lncRNA		
		# genes transcribed	# genes translated	# ORFS translated	# genes transcribed	# genes translated	# ORFS translated	# genes transcribed	# genes translated	# ORFS translated
Brain	GSE51424 (1)	14,092	10,850	11,961	1,416	65	77	3,391	39	44
Testis	GSE50983 (2)	13,369	11,754	13,150	1,370	120	148	3,184	78	84
Neutrophils	GSE22001 (3)	11,074	8,632	9,228	1,080	32	40	2,594	9	9
Heart	GSE41246	11,785	7,245	7,500	582	4	4	2,283	2	3
Skeletal muscle	GSE41246	11,020	4,987	5,137	499	3	3	2,182	0	0
Splenic B cells	GSE62134 (4)	10,741	9,405	10,531	1,200	67	84	2,781	38	40
Neural ES cells	GSE72064 (5)	11,606	10,393	11,668	1,150	77	105	2,837	28	37
Hippocampus	GSE72064 (5)	12,671	12,366	17,930	1,179	425	840	3,077	245	403
Integrated	**-**	**16,681**	**15,013**	**21,918**	**2,580**	**546**	**1,027**	**3,913**	**286**	**462**

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. References (ref): (1) [31], (2) [32], (3) [33], (4) [34], (5) [35].

A significant fraction of the ORFs in codRNAs were transcribed and translated in several samples, whereas lncRNAs, uORFs, and dORFs tended to be sample-specific (S1 Fig). About 75% of the translated lncRNAs encoded putative proteins shorter than 100 amino acids (small ORFs or smORFs). Overall, ORFs in lncRNAs were longer than uORFs and dORFs (median 48-52 *vs.* 26 amino acids, Wilcoxon test, p-value < 10^{-5}), but shorter than the main ORF in protein-coding genes (median 381 amino acids, Wilcoxon test, p-value < 10^{5}). The characteristics of translated transcripts and the size of the translated products were very similar for annotated lncRNAs and for novel expressed loci (Fig 1b, 1c and 1e). Therefore, these two types of transcripts were merged into a single class (lncRNA) for most analyses.

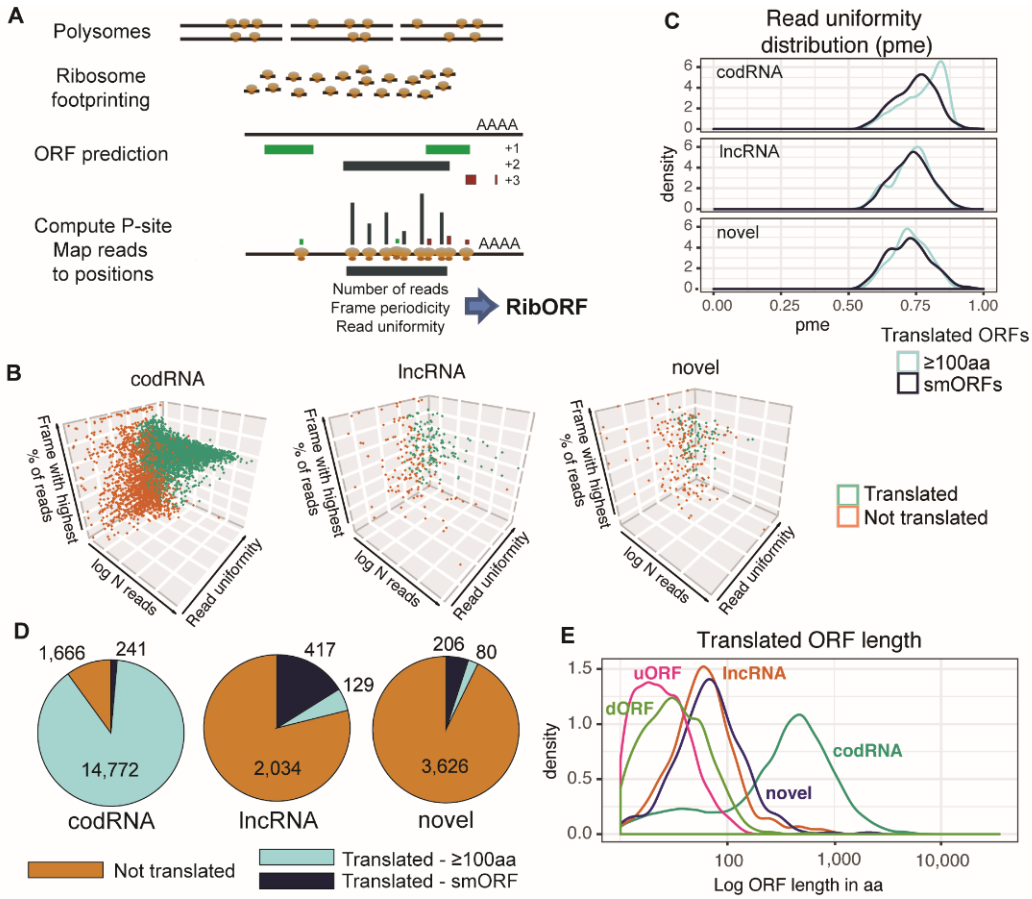

Figure 1. Detection of translated ORFs. A. Workflow to identify translated ORFs. Ribosome profiling (Ribo-Seq) reads, corresponding to ribosome-protected fragments, are mapped to all predicted canonical ORFs in transcripts with a minimum length of 30 nucleotides. This is performed with single-nucleotide resolution after computing the read P-site per each read length. In each ORF, reads per frame and read uniformity are evaluated by RiboORF. **B.** Relationship between the number of reads, frame periodicity and read uniformity in a given ORF for codRNAs, annotated lncRNA and novel genes. Data shown are for the hippocampus sample; similar results were obtained in other samples. **C.** Density plot showing the uniformity of Ribo-Seq reads across codons in the ORF by measuring the Percentage of Maximum Entropy (pme), where 1 represents uniform read distribution across all codons. **D.** Number of translated and non-translated expressed genes belonging to different classes. The translated genes have been divided into small ORFs (< 100 aa) and long ORFs (≥ 100 aa), depending on the length of the longest translated ORF. **E.** Length of translated ORFs for different gene types in logarithmic scale: coding (codRNA), annotated long non-coding RNA (lncRNA), non-annotated assembled transcripts (novel), upstream coding ORFs (uORF) and downstream coding ORFs (dORFs). LncRNA and novel ORFs were significantly shorter than codRNA ORFs (median size 48 and 52, respectively, *versus* 381 for annotated protein-coding genes, Wilcoxon test and p-value < 10^{-5}), but longer than uORFs and dORFs (median size 26, Wilcoxon test, p-value <

10^{-5}).

155 **Properties of translated lncRNAs compared to coding genes**

The number of transcribed and translated ORFs varied substantially depending on the sample (Table 1, Fig 2a). The largest number of translated genes were detected in hippocampus tissue, followed by testis, embryonic stem cells and brain, both for
160 codRNAs and lncRNAs. Similar results were obtained when we focused on ORFs translated in a single tissue (Fig 2b) or separately considered long ORFs and smORFs (S2 Fig). There were two reasons for these differences. The first reason was the number of available Ribo-Seq sequencing reads in each experiment, about three times greater in hippocampus than in other tissues, which provided increased
165 resolution to detect translation of lowly expressed transcripts. As expected, subsampling the number of reads in the hippocampus resulted in a decrease in the number of translated ORFs detected (S2 Fig). The second reason was that, in some tissues, the pool of translated ORFs was highly skewed towards a few very abundant proteins (S3 Fig). For example, in skeletal muscle and heart the five most highly
170 translated genes, which included myosin and titin, gathered 22.5-31.2% of the sequencing reads; this substantially reduced the number of reads available to detect other products of translation. Overall, the data suggested that the experimental translation signal was not saturated and that the true number of translated lncRNAs may be higher than was estimated here.

175

When we compared the translated and non-translated protein-coding genes, the former had higher expression levels and were longer than the latter (Fig 2c, Wilcoxon test, p-value $< 10^{-5}$). In the case of lncRNAs, translation was positively associated with ORF length (Wilcoxon test, p-value $< 10^{-5}$), but we did not detect any
180 relationship between translation status and expression level. In general, lncRNAs were expressed at much lower levels than coding genes (Fig 2C); this is a well-known global difference between the two types of genes [9,36].

185

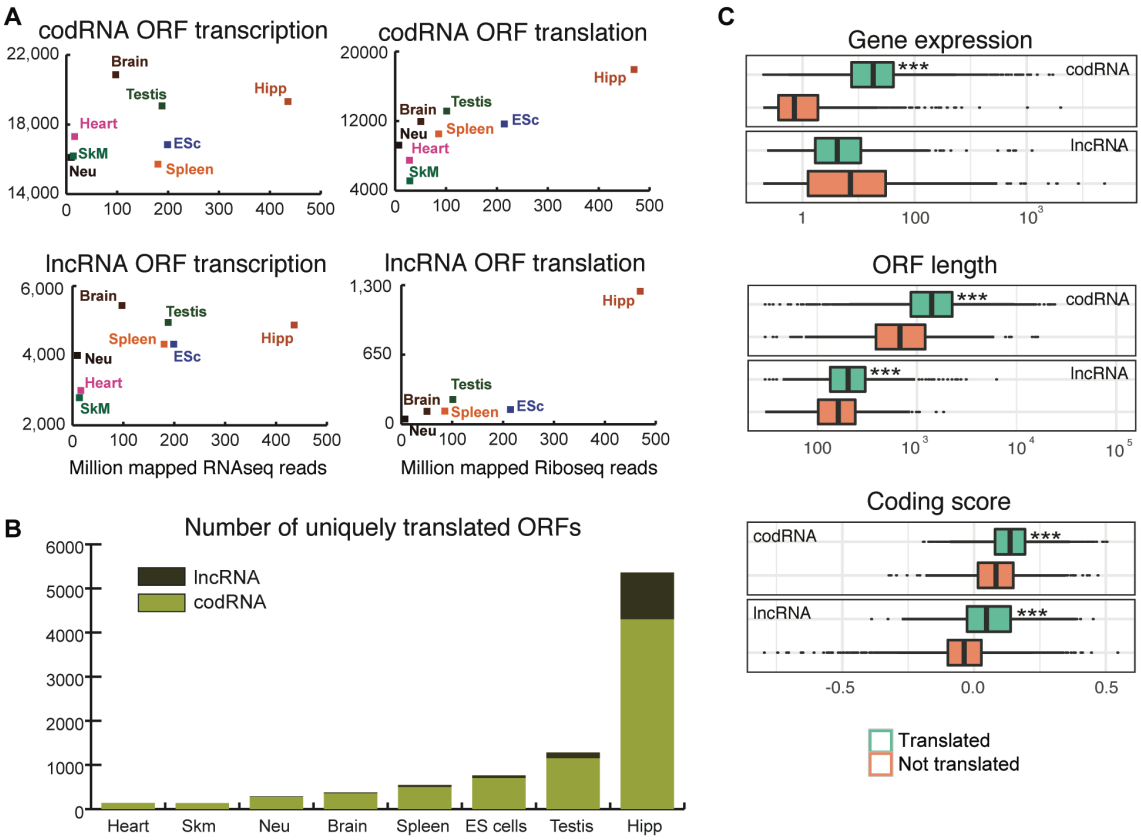

Figure 2. Features of translated sequences. A. Number of transcribed or translated ORFs (Y-axis) in relation to the number of sequencing reads (RNA-Seq or Ribo-Seq) mapped to the genome in the different experiments (X-axis) for codRNA and lncRNA. Hipp: hippocampus; ES cells (ESc): Neural embryonic stem cells; Neu: neutrophils; SkM: skeletal muscle. Translation of lncRNAs in heart and skeletal muscle was very low (< 5) and therefore the points are not represented. **B.** Number of ORFs translated in a single tissue, for codRNAs and lncRNAs. **C.** Distribution of ORF length, gene expression level, and coding score values in codRNA and lncRNA for the longest translated and non-translated ORF per gene. Expression level is defined as the maximum FPKM value across the 8 samples. *** Wilcoxon test, p-value < 10^{-3}.

Phylogenetic conservation and codon usage bias

We next examined which fraction of the mouse lncRNAs with evidence of translation were conserved in rat and/or human transcripts. For this we employed *de novo* human and rat transcript assemblies of a quality similar to that used for mouse (see Methods). We searched for homologues of the putatively translated mouse ORFs in the human and rat transcripts using TBLASTN (e-value < 10^{-4}). We found hits in human and/or rat for 41% of the mouse-translated ORFs in lncRNAs, compared to 92% for protein-coding genes (Fig 3a). This is in line with previous studies showing that lncRNAs tend to be much less conserved than protein-coding genes [10,37,38].

Codon usage bias is usually employed to predict coding sequences in conjunction with other variables such as ORF length and sequence conservation [39,40]. In the case of non-conserved smORFs, such as those translated from many of the lncRNAs, only measures based on codon usage bias can be applied. We have previously implemented [24] a metric based on the differences in dicodon (hexamer) frequencies between coding and non-coding sequences, which we have used to calculate length-independent coding scores for translated and non-translated ORFs in different species. Based on this metric, we developed a computational tool to identify ORFs with significant coding scores in any set of sequences of interest, which is available online (evolutionarygenomics.imim.es/CIPHER).

We observed a positive relationship between the RibORF translation score and the coding score produced by CIPHER, both for codRNAs and in lncRNAs (Fig 2c, S4 Fig). We also found that conserved ORFs (group C) had significantly higher coding scores than non-conserved ORFs, both for codRNAs and lncRNAs (Fig 3b). We reasoned that ORFs with a very biased codon usage may correspond to functional proteins even if not conserved across species. We used CIPHER to divide the non-conserved genes into a group with high coding scores (NC-H, coding score > 0.079, above the median value for conserved coding genes) and another group with lower coding scores (NC-L, coding score ≤ 0.079). We also searched for proteomics evidence in PRIDE [41]. Using stringent criteria, we found proteomics evidence for 37 of the ORFs in the lncRNAs, with similar numbers in the different groups (11 in C, 12 in NC-H and 14 in NC-L).

Testing for signatures of natural selection in translated ORFs

We had detected ~1,500 putatively translated ORFs in lncRNAs, but it was unclear if they were likely to encode functional proteins. To address this, we investigated the signatures of natural selection in the ORFs using a large collection of mouse single nucleotide polymorphisms (SNP) for the house mouse subspecies *Mus musculus castaneus* [42,43]. 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 an advantage over non-synonymous to synonymous substitutions in 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 mouse lncRNAs that were not conserved in human or rat transcripts.

250

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 would expect the PN/PS ratio of a sequence or set of sequences to be 2.89 [44]. However, not all codons are equally frequent in coding sequences, and the probability of mutation
255 differs between pairs of nucleotides [45–47]. These parameters can be estimated from real data and subsequently used to compute an expected PN/PS under neutrality. The difference between the observed and expected PN/PS ratios informs us on the strength of purifying selection. If the observed PN/PS normalized by the expected PN/PS is not significantly different from 1, the observed proportion of non-
260 synonymous and synonymous SNPs is consistent with neutral evolution. If it is significantly lower than 1, there is a depletion of non-synonymous SNPs. Such a depletion is consistent with purifying selection acting at the amino acid sequence level and provides a strong argument for functionality.

265 We mapped a total of 324,729 SNPs from *Mus musculus castaneus* to the previously defined ORFs from codRNAs and lncRNAs. For each sequence, and sequence dataset (C, NC-H and NC-L), we calculated the ratio between observed non-synonymous and synonymous SNPs (PN/PS(obs)) and divided it by the ratio expected under neutrality (PN/PS(exp)), obtaining a normalized PN/PS. The expected PN/PS was obtained using
270 a table of nucleotide mutation frequencies in *Mus musculus castaneus*, which we derived from SNPs in intronic sequences (S1 and S2 Tables), and the observed codon frequencies in the sequences of interest. The values ranged from 2.31 to 2.47 for different sequence datasets (S3 Table). We used the chi-square test to determine if the sequences under analysis showed a PN/PS that deviated significantly from that
275 expected under neutrality (Fig 3c, S4 Table).

We found that conserved translated ORFs, both in codRNAs and lncRNAs, had PN/PS values significantly lower than the neutral expectation (Fig 3c, chi-square test p-value $< 10^{-5}$). In lncRNAs, there was an approximately 40% depletion of non-
280 synonymous SNPs over the expected value, strongly suggesting that a sizable fraction of the 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 [48]. Therefore, it is not surprising that some remain
285 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 [49]. Another example is NoBody, which encodes a protein just

68 amino acids long and has recently been shown to interact with the mRNA decaying complex [50]. NoBody was annotated as non-coding when we initiated the study, and Myoregulin was annotated with a different non-canonical ORF, although their annotations are now fully consistent with our findings. Other examples of conserved smORFs in our set were Stannin, a mediator of neuronal cell apoptosis conserved across metazoans [51,52], and Apela, a peptide ligand that acts as an embryonic regulator and increases cardiac contractility in mouse [53,54]. The distribution of the Ribo-Seq reads in these examples is shown in Figure 4.

The group of ORFs which were not conserved across species but had high coding scores (NC-H) showed weaker purifying selection than conserved genes; however, PN/PS was significantly lower than the neutral expectation (Fig 3c, chi-square p-value = 1.6×10^{-5} for codRNAs and p-value = 0.0026 for lncRNAs). Despite the lack of detectable homologues in rat and human, this finding indicates some of the proteins in this group are probably functional.

In contrast, the normalized PN/PS in the rest of the non-conserved ORFs (NC-L) was not different from 1 and therefore consistent with neutral evolution (Fig 3c). This result was equivalent to that found for randomly selected ORFs from introns (Table S4). Despite the lack of evidence of selection, these ORFs showed strong three-nucleotide read periodicity and uniformity (Fig 4, Fig 5a), indicating *bona fide* translation. The lack of selection signatures was evident both in transcripts annotated as coding RNAs and as lncRNAs. Although these cases represented a very small minority of the protein-coding genes (~1%), they were a much larger fraction of the lncRNAs (~40%).

The above analyses grouped the sequences into classes before computing the PN/PS ratio. In general, ORF-by-ORF analysis was not possible because the ORFs were small and contained too few SNPs. Nevertheless, a small fraction of the ORFs in lncRNAs contained 10 or more SNPs, and we computed a normalized PN/PS ratio for these cases. The results were very much in line with those obtained with the complete sequence sets and supported our previous conclusions (Fig 3d).

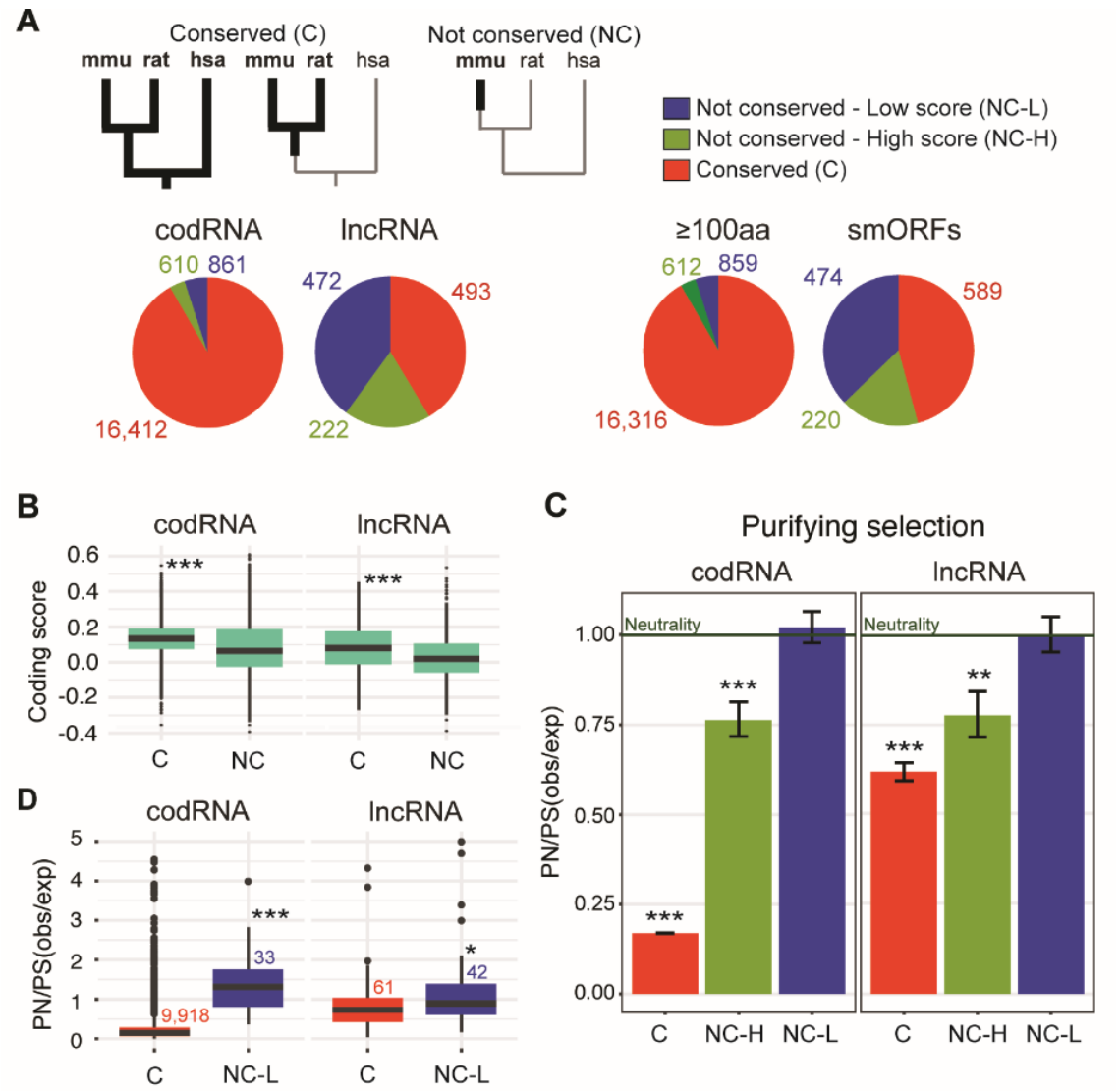

Figure 3. Different classes of translated ORFs. A. Number of translated ORFs that were conserved in human (C), not conserved but showing a high coding score (NC-H, coding score > 0.079, significant at p-value < 0.05) and not conserved with a low coding scorel (NC-L, coding score ≤ 0.079). First, ORFs were 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}. **C**. Analysis of selective constraints in translated ORFs. PN/PS (obs/exp) refers to the normalized ratio between non-synonymous (PN) and synonymous (PS) single nucleotide polymorphisms, using data from a population of 20 individuals of *Mus musculus castaneus*; a value of 1 is expected in the absence of selection at the protein level. Conserved and high-score ORFs showed significant purifying selection signatures independently of transcript type (codRNA or lncRNA). In contrast, non-conserved ORFs with low coding scores (NC-L) did not show evidence of purifying selection at the protein level in either codRNAs or lncRNAs. Significant differences between observed and expected PN/PS are indicated: chi-square test with one degree of freedom, * p- value < 0.05,

** p- value < 0.005, *** p-value < 10^{-5}. Error bars indicate the standard error of the sample proportion. **D.** Distribution of normalized PN/PS values for individual ORFs in different gene classes. Only ORFs with at least 10 SNPs were considered; the NC-H group contained too few cases to be analyzed. The differences between C and NC-L were significant in both codRNAs (Wilcoxon test, p-value <10^{-5}) and lncRNAS (Wilcoxon test, p-value < 0.05).

The group of lncRNAs producing proteins with no selection signatures included several RNAs with known non-coding functions, such as *Malat1*, *Neat1*, *Jpx*, and *Cyrano*. These genes are involved in several cellular processes: *Cyrano* in the regulation of embryogenesis [55], *Jpx* in X chromosome inactivation [56], *Neat1* in the maintenance and assembly of paraspeckles [57], and *Malat1* in regulating the expression of other genes [58]. Many other translated ORFs were located in transcripts with no known function. Two examples are shown in Figure 4. Due to the absence of selection signatures, one must conclude that the translation of these transcripts is probably due to promiscuous activity of the ribosome machinery. This may lead to the production of thousands of novel non-functional small proteins in different cell types and tissues.

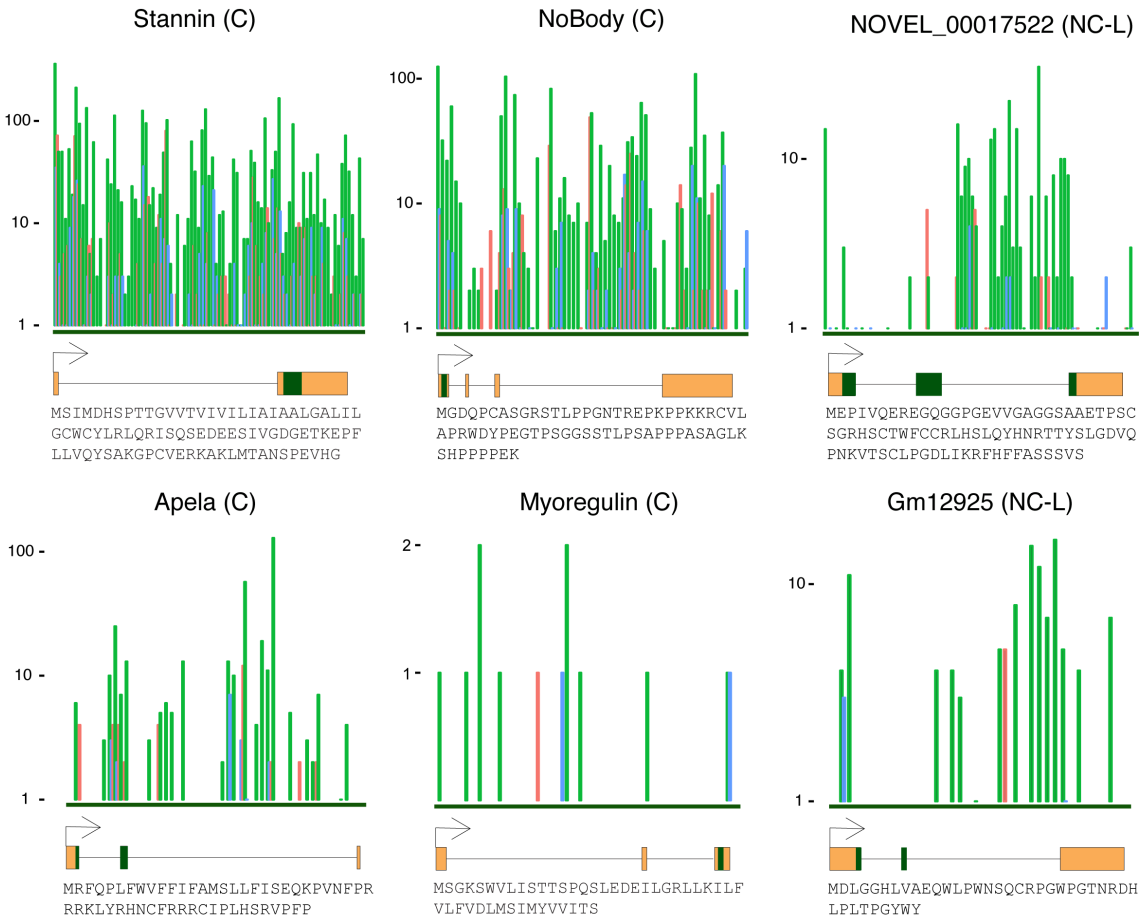

Figure 4. Examples of translated ORFs. The mapping of Ribo-Seq reads on ORFs classified as translated by RibORF is shown. The Y axis represents the number of reads, the X axis the positions in the coding sequence. The reads show strong frame bias, with a preponderance of in-frame reads (green) versus off-frame reads (red and blue). The exon/intron structure and the amino acid sequence is also shown. C: Conserved; NC-L: Not conserved – Low score. The genes NOVEL_00017522 (ORF_id: CUFF_mmu_all_st_00017522-1, gene_id: XLOC_007614) and Gm12925 are lncRNAs with no known function.

What drives the translation of lncRNAs?

Translated ORFs in lncRNAs lacking conservation in rat and/or human and with no evidence of selection (NC-L) comprised 369 genes translating 472 different ORFs. These genes were not coding in the usual sense, as we observed no signatures of selection but yet they produced proteins. The ORFs in the NC-L group of genes showed the characteristic three-nucleotide periodicity of actively translated regions (Fig 5a). In addition, the ORF frame bias was highly reproducible across tissues and the correlation coefficient similar to that computed for conserved, well-established, codRNAS (Fig 5b and S5-S7 Fig).

Why was translation detected in these lncRNAs but not in others? The lack of selection signatures at the protein level precluded their being mis-annotated classical protein-coding genes. We inspected the translation initiation sequence context but did not detect any significant differences between translated and non-translated ORFs. We then hypothesized that the ORF coding score could 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. Consistent with this hypothesis, we found that the translated ORFs in this group exhibited higher coding scores than the ORFs in non-translated genes (Fig 5c, Wilcoxon test, p-value $< 10^{-5}$). Importantly, we obtained a similar result after controlling for gene expression level (Fig 5d, Wilcoxon test, p-value $< 10^{-5}$). This is consistent with codon composition having an effect *per se* in ORF translation. We also detected significant differences in the expression level of translated and non-translated ORFs when controlling by coding score (Fig 5e). This may reflect better capacity to detect translation in the case of highly expressed transcripts. In contrast, although translated ORFs tend to be longer than non-translated ORFs (Fig 2c), ORF length appeared to have no effect *per se* in translatability (Fig 5e).

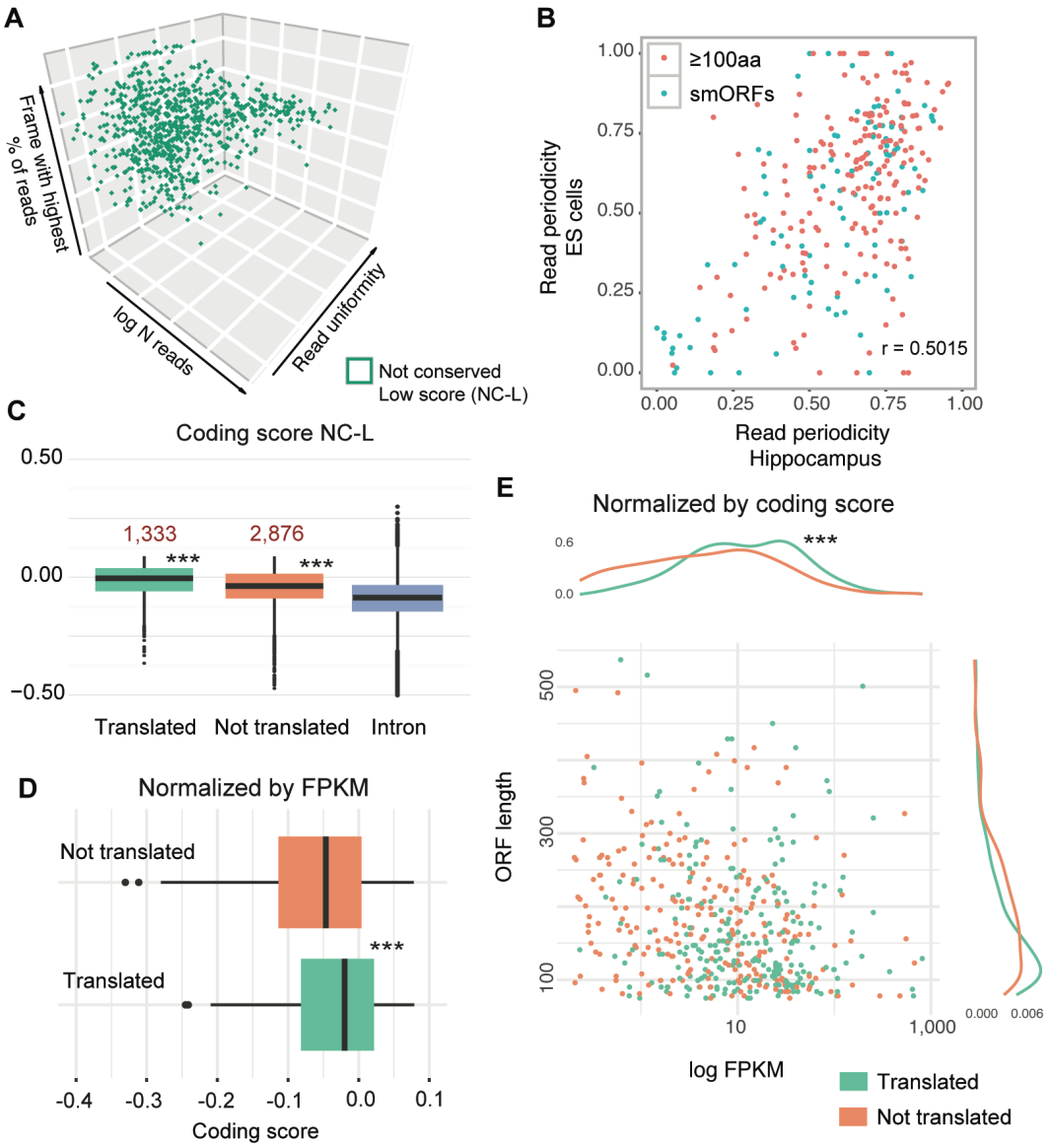

Figure 5. Factors influencing the translation of lncRNAs. A. Relationship between the number of reads, frame periodicity, and read uniformity in a given translated ORF for conserved and non-conserved neutral (NC-L) ORFs. Data shown are for the hippocampus sample; similar results were obtained in other samples. **B.** Relationship between the percentage of reads falling in the correct frame in ES cells and hippocampus samples, for small and long NC-L ORFs. Pearson correlation coefficient (R=0.5015) was highly significant (p-value < 10^{-5}) and similar to conserved protein-coding ORFs (R= 0.5037, S5 Fig). **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 maximum FPKM gene expression (median FPKM value = 13.19). Translated ORFs showed significantly higher coding score values than non-translated ORFs; *** Wilcoxon test, p-value < $10^{-5.}$ **E.** Influence of maximum FPKM gene expression and ORF length in the translatability of neutral ORFs normalized by

coding score (median coding score value = -0.013). Translated ORFs showed significantly higher FPKM values; *** Wilcoxon test, p-value < 10^{-5}.

440 **DISCUSSION**

Several studies have reported that many lncRNAs translate small proteins [13,16–18,22]. Each study detected hundreds or even thousands of lncRNAs with patterns consistent with translation. Varying criteria have been used to differentiate active
445 translation from other signals, including 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. As lncRNAs are, in general, expressed at low levels, the stringency of the method, as well as the sequencing depth, can be expected to strongly impact the number of
450 translated lncRNAs identified.

The recent discovery that a large number of lncRNAs show ribosome profiling patterns consistent with translation has puzzled the scientific community [59]. Most lncRNAs are not conserved across mammals or vertebrates, which limits the use of
455 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 in mammals do not always contain enough
460 polymorphisms to test for selection [60]. 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 [24]. Together with previous observations that lncRNAs tend to be lineage-specific [10] and that young proteins evolve under relaxed purifying selection [61], this finding
465 led us to hypothesize that lncRNAs are enriched in young protein-coding genes.

The present study employed recently generated mouse, rat 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.
470 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 [38,62]. Here we found that about 5% of the lncRNAs in databases may encode conserved functional micropeptides (smORFs). Standard proteomics techniques have important limitations

475 for the detection of micropeptides and there is evidence that the smORFs currently annotated in databases are only a small part of the complete set [63–66]. As shown here, and in other recent studies [13,15], 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
480 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.

485 Aside from lncRNAs which translate functional microproteins, the present study identified another large class of lncRNAs that appears to evolve neutrally and thus to translate proteins that do not perform any useful function. These ORFs can be distinguished from the rest because they were not conserved across species and did not exhibit high coding scores. As the test of neutrality was applied to the complete
490 group, it remains possible that a few of the ORFs were under selection, but this is likely to be a very small number. An interesting observation is that the lack of selection signatures was not only observed in lncRNAs but also in coding RNAs that share the same characteristics. This blurs any differences between the two classes of genes when we focus on genes showing limited phylogenetic conservation. Overall,
495 we detected 1,333 proteins that appeared to be translated but showed no signs of selection. This could be a gross underestimate, considering that many cell types and tissues have not yet been sampled.

Although the existence of non-functional proteins may seem counterintuitive at first,
500 we must consider that most of these transcripts (lncRNAs and non-conserved codRNAs) tend to be expressed at low levels and so the associated energy costs of this activity may be negligible. This is in agreement with recent estimates that the cost of transcription, and even translation, in multicellular organisms is probably too small to overcome genetic drift [67]. In other words, provided the peptides are not
505 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
510 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 more favorable codon composition. This is consistent with the observation that abundant codons enhance translation elongation [68], whereas rare codons might affect the stability of the mRNA and activate decay pathways [69]. 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 [16]. 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, and are influenced both by selection and drift [70,71]. 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* non-coding gene *Poldi* [72], which lacked any evidence of translation in the data we analyzed. The group of ORFs that had high coding scores, but lacked conservation in human or rat transcripts, had weak but significant purifying selection signatures. Although there may be different reasons why we did not detect any homologues, such as rapid evolution linked to very short protein size, or loss of the gene in different lineages, this set is probably enriched in genes that have recently evolved *de novo* [73,74]. For this type of genes, the annotation as coding and non-coding appears to be highly irrelevant, as the two types of genes displayed very similar features in all the analyses performed.

A growing number of protein-coding genes have been reported to have originated *de novo* from previously non-functional genomic regions [75–83]. These genes encode proteins with unique new sequences, which may have important roles in lineage-specific adaptations. The encoded proteins are usually small and disordered, and have been hypothesized to become longer and more complex over time [79,84,85]. Interestingly, there is recent evidence that many of these genes may have originated from lncRNAs [29,86,87]. This is also consistent with the large number of species-specific transcripts with lncRNA features identified in comparative transcriptomics studies [29,88–90]. 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 [91], and the remaining ones 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-
555 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.
560

In conclusion, our data support the use of ribosome profiling and conservation analysis to uncover putative 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 found that the
565 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 may have coding-like features from the start.

570 **METHODS**

Transcriptome assembly

The polyA+ RNA-Seq from mouse was comprised of 18 strand-specific paired-end
575 data publicly available in the Gene Expression Omnibus under accession numbers GSE69241 [29], GSE43721 [92], 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 [29]. Data corresponded to
580 2 brain, 2 liver, 2 heart and 2 testis samples.

RNA-Seq sequencing reads were filtered by length (> 25 nucleotides) and by quality using Condetri (v.2.2) with the following settings: -hq = 30 –lq = 10. We retrieved genome sequences and gene annotations from Ensembl v. 75. We aligned the reads
585 to the corresponding reference species genome with Tophat (v. 2.0.8, –N 3, -a 5 and –m 1) [93]. Multiple mapping to several locations in the genome was allowed unless otherwise stated.

We assembled the transcriptome with Stringtie [94], merging the reads from all the samples, with parameters -f 0.01, and -M 0.2. We used the species transcriptome as a guide (Ensembl v.75), including all annotated isoforms, but permitting the assembly of annotated and novel isoforms and genes (antisense, intergenic and intronic) as well. We complemented our human and mouse transcript assemblies with an additional rat transcript assembly generated in a parallel study [95]. The latter assembly was derived from RNA-seq data from 11 tissues: adrenal gland, brain, heart, kidney, liver, lung, muscle, spleen, testis, thymus, and uterus.

In mouse, we selected genes with a minimum size of 300 nucleotides and with a per-nucleotide read coverage ≥ 5 in at least one sample. This ensures a high degree of transcript completeness, as shown previously [29]. The resulting transcriptome comprised 16,679 protein-coding genes (average of 6.64 isoforms/gene); 2,580 long non-coding RNAs (average of 2.35 isoforms/gene) defined as assembled genes that overlapped annotated genes that were not annotated as protein-coding; 3,912 novel non-annotated genes (average of 1.07 isoforms/gene), and 3,467 genes overlapping pseudogenes.

Ribosome profiling data

We used 8 different data sets that included both strand-specific ribosome profiling (Ribo-Seq) and RNA-seq data and which we obtained from Gene Expression Omnibus under accession numbers GSE51424 [31], GSE50983 [32], GSE22001 [33], GSE62134 [34], GSE72064 [35], and GSE41246. Data corresponded to brain, testis, neutrophils, splenic B cells, neural embryonic stem cells, hippocampus, heart and skeletal muscle (Table 1). Only datasets corresponding to normal samples were considered. Any replicates were merged before the analyses. For all analyses we considered only genes expressed at significant levels in at least one sample (fragments per kilobase per Million mapped reads (FPKM) > 0.2).

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 in mouse. Next, reads were mapped to the assembled mouse genome (mm10) with Bowtie, allowing read multimapping (v. 0.12.7, -k 1 -m 20 -n 1 --best --strata –norc) and controlling for strand information.

We used the mapping of the Ribo-Seq reads to the complete set of annotated coding sequences in mouse to compute the position of the P-site, corresponding to the tRNA

binding-site in the ribosome complex, for reads of different length, as in other studies [11,13,16,17]. If no P-site offset was clear for a specific length, reads with that length were not considered for subsequent analysis. Considering that the ORFs had to be extensively covered by reads to be considered translated (high uniformity), we decided to include multiple mapped reads so as not to compromise the detection of paralogous proteins (S8 Fig).

Detection of translated ORFs

We predicted all possible ORFs (ATG to TGA/TAA/TAG) with a minimum length of 30 nucleotides (9 amino acids) in transcripts expressed at FPKM > 0.2 in any sample. Next, we ran RibORF (v.0.1) [16] to obtain a set of translated ORFs per sample. This program is a support vector machine classifier and we used a score threshold of 0.7 to classify an ORF as translated, as in the original study. This cutoff is considered to be very stringent, with a false positive rate of 0.67% and a false negative rate of 2.5% [16]. With the 0.7 threshold, no translated ORFs were found in annotated small RNAs, providing additional support for our approach. We only considered ORFs with 10 or more mapped reads; the rest were classified as non-translated. The ORFs classified as translated by the program showed high three-nucleotide read periodicity and uniformity when compared to ORFs classified as non-translated (Figure 1b).

For every gene, we selected all the ORFs that were translated in any of the samples and merged overlapping ORFs in clusters represented by the longest ORF in the group, for conservation and coding score analyses. If any of the ORFs were found upstream or downstream of another longer ORF in an annotated protein-coding transcript, we defined them as upstream ORF (uORF) or downstream ORF (dORF). If a gene was not translated, we selected the longest ORF across all transcripts for comparative purposes. In translated ORFs, the ORF with the highest number of mapped Ribo-Seq reads was usually the longest ORF (75.7% for codRNAs and 84% for lncRNAs).

We differentiated between genes with small ORFs (smORFs) and those with longer ORFs. In the first class, the longest ORF in the gene encoded a protein of less than 100 amino acids. We did not consider genes overlapping annotated pseudogenes and excluded smORFs in lncRNAs that showed significant sequence similarity to known protein-coding sequences, since they might be pseudogenized regions.

Sequence conservation

665

We searched for homologues of the mouse ORFs in the human and rat transcript assemblies using TBLASTN (limited to one strand, e-value < 10^{-4}) [96]. The aim was to define a set of proteins which were conserved in human, rat, or both, and a set of non-conserved proteins for which homologues in the transcriptomes of these species
670 could not be identified. An additional requirement to classify a protein as conserved was that the alignment covered at least 50 amino acids or 75% of the total ORF length. The smallest conserved protein was 19 amino acids long. In the non-conserved ORFs we only considered proteins of size 24 amino acids or longer, as homologues of shorter proteins may be difficult to detect even if they exist. For
675 simplicity, in the analysis of the signatures of purifying selection we also discarded a small number of non-conserved ORFs that were in the same transcript than conserved ORFs (uORFs and dORFs were not taken into account here).

Single nucleotide polymorphism data

680

Single nucleotide polymorphism (SNP) data was obtained from Harr et al. [43], and included complete genotyping information from 20 individuals of the house mouse subspecies *Mus musculus castaneus*. We classified SNPs in ORFs as non-synonymous (PN, amino acid altering) and synonymous (PS, not amino-acid altering). We
685 discarded any nucleotide variants that are fixed in the population used. We calculated the PN/PS ratio in each ORF group by using the sum of PN and PS in all the sequences ((PN/PS)obs). In general, estimation of PN/PS ratios of individual sequences was not reliable due to lack of a sufficiently large number of SNPs per ORF; we only performed this calculation in cases with at least 10 SNPs in the ORF.

690

We calculated the expected PN/PS under neutrality ((PN/PS)exp) using the mutation frequencies between pairs of nucleotides in *Mus musculus castaneus* and the codon composition of the different sequences or sets of sequences under study. The mutation frequencies were estimated from SNPs in introns from the same population
695 of mice (supplemental file, Table S1). The transition to transversion ratio was 4.42, very similar to the 4.26 value obtained in early observations based on mouse-rat divergence data [97]. As a test of neutrality on the coding sequence, we used a chi-square test with one degree of freedom that compares the observed and expected PN and PS values in the sequences of interest. In the absence of selection we expect
700 (PN/PN)obs/(PN/PS)exp to be approximately 1. Under purifying selection, this ratio will be lower than one. Positively selected mutations are rapidly fixed in the population and their effect is expected to be negligible when using SNP data.

Analysis of proteomics data

We used the proteomics database PRIDE [41] to search for peptide matches in the proteins encoded by various gene sets. For a protein to have proteomics evidence, we required at least two distinct perfect matches of peptides that did not map to any other protein in the dataset, allowing for up to two mismatches. These are very stringent conditions, for which a false positive rate lower than 0.2% has been previously estimated [95].

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 [24]. 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 ORFs with a high likelihood to be translated by using 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). Genes in the NC-H group were defined as those with a coding score over the median value of conserved coding sequences (> 0.079, with CIPHER significant at p-value < 0.025).

745

Statistical data analyses

The generation of plots and statistical tests was performed with the R package [98].

750 **DATA AVAILABILITY**

Transcript assemblies, open reading frames (ORFs) predicted to be translated, and code to calculated the PN/PS expected under neutrality in mouse sequences have been deposited at figshare (http://dx.doi.org/10.6084/m9.figshare.4702375).
755 Supplemental file contains detailed information on the translated ORFs identified in this study (S1 Dataset).

SUPPORTING INFORMATION

760 **S1 Fig. Number of tissues in which ORFs are translated.** codRNAs are expressed and translated in most of the samples. Annotated and novel lncRNAs, as well as upstream and downstream ORFs (u/dORFs), often exhibit sample-specific patterns of translation.

765 **S2 Fig. Effect of Ribo-Seq number of reads in the detection of translation.** Number of translated long and small ORFs (length < 100 aa) after randomly subsampling mapped Ribo-Seq reads from hippocampus using different read numbers are shown. For comparison, points represent the number of ORFs translated on the eight different samples.

770

S3 Fig. Percentage of Ribo-Seq reads mapping to the top 5 translated genes in every sample. The five most highly translated genes in skeletal muscle and heart gathered a significant proportion of the mapped reads, decreasing the capability to detect translation of other proteins in the sample.

775

S4 Fig. Relationship between coding score and RibORF score. ORFs are groups according to the RibORF score; data shown are for hipoccampus. Only ORFs with at least 10 mapped Ribo-Seq reads were considered. A positive relationship between

coding score and RibORF score was observed in both codRNAs and lncRNAs/novel transcripts.

S5 Fig. Reproducibility of the Ribo-Seq frame bias in codRNAs in brain and hippocampus samples. Relationship between the percentage of reads falling in the correct frame in brain and hippocampus samples. The data is for codRNAs classified as Conserved. Pearson correlation coefficient was R=0.5037, significant at p-value < 10^{-5}.

S6 Fig. Reproducibility of the Ribo-Seq frame bias in lncRNAs in neural embryonic stem cells and hippocampus samples. Relationship between the percentage of reads falling in the correct frame in brain and hippocampus samples. The data are for lncRNAs classified as NC-LS. Pearson correlation coefficient was R=0.3654, significant at p-value < 10^{-5}.

S7 Fig. Reproducibility of the Ribo-Seq frame bias in codRNAs in neural embryonic stem cells and hippocampus samples. Relationship between the percentage of reads falling in the correct frame in brain and hippocampus samples. The data is for codRNAs classified as Conserved. Pearson correlation coefficient was R=0.3512, significant at p-value < 10^{-5}.

S8 Fig. Effect of using multiple mapped reads versus uniquely mapped reads in the detection of translation. Maximum frame bias refers to the fraction of reads that map to the correct frame. pme is the percentage of maximum entropy and measures read uniformity along the ORF. The values of the parameters were measured with RibORF. "All" is when multiple mapping with Bowtie2 was allowed (up to 20 different locations), "unique" is when only reads that map to a single location with Bowtie2 were considered. "Translated" refers to the cases in which RibORF score was equal or higher than 0.7. CodRNA: protein-coding ORFs. LncRNA: annotated and novel long-noncoding ORFs. Use of multimapped reads resulted in a better performance of the parameters classifying protein-coding genes and long non-coding genes as translated.

S1 Table. Number and type of single nucleotide polymorphisms (SNPs) in introns from *Mus musculus castaneus*. The SNP data were from 20 different individuals and obtained from [43]. We mapped the SNP data to 23,667 randomly taken introns from mouse genes annotated in Ensembl version 75. T: transition; V: transversion.

S2 Table. Relative mutation frequencies in *Mus musculus castaneus*. The frequencies were derived from the intronic SNPs in Table S1. We used them to calculate the expected ratio between non-synonymous and synonymous SNPs in the absence of selection in any sets of sequences ((PN/PS)exp, see Table S3). For practical reasons, the table was normalized so that the sum of each row was 1.

S3 Table. Number of non-synonymous and synonymous single nucleotide polymorphisms (PN and PS, respectively) in translated ORFs from coding genes and lncRNAs. PN and PS was from 20 different individuals of Mus musculus castaneus, as provided by Harr et al. (2016). Conserved: ORFs with homology to human expressed sequences. NC-HS: ORFs with no homology to human expressed sequences and with a coding score > 0.079. NC-N: rest of ORFs with no homology to human expressed sequences. (PN/PS)exp refers to the neutral expectation based on the estimated mutational frequencies (Table S2) and the observed codon composition of each set of sequences. ORFs in introns: 4,835 randomly taken ORFs in introns. Chi-square test with 1 degree of freedom, observed versus expected, ** p-value < 10^{-2}, *** p-value < 10^{-3}, N.S.: non-significant.

S4 Table. Observed versus expected PN/PS ratio. SE: standard error of sample proportion. The p-value obtained with the chi-square test employed to compare the observed and expected data in Table S3 is shown. The datasets are the same as in Table S3.

S1 Dataset. Sequence data and annotations. Information on the open reading frames (ORFs) and genes used in this research is annotated by translation status, sample, conservation class, coding scores, non-synonymous polymorphisms (PN), synonymous polymorphisms (PS), expected PN/PS under neutrality, proteomics evidence.

ACKNOWLEDGEMENTS

We thank colleagues, reviewers, and associated editor, for very useful comments that helped improved this work. We are grateful to Elaine Lilly, Ph.D., for text revision

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.

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.

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.

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

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

6. Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. Annu Rev Biochem 81: 145–166.

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

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.

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.

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.

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.

12. Ingolia NT (2014) Ribosome profiling: new views of translation, from single codons to genome scale. Nat Rev Genet 15: 205–213.

13. 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.

14. Aspden JL, Eyre-Walker YC, Philips RJ, Amin U, Mumtaz MAS, et al. (2014) Extensive translation of small ORFs revealed by Poly-Ribo-Seq. Elife 3: e03528.

15. 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.

16. 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.

17. 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.

18. 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: e13328.

19. Hsu PY, Calviello L, Wu H-YL, Li F-W, Rothfels CJ, et al. (2016) Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. Proc Natl Acad Sci 113: E7126–E7135.

20. 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.

21. 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.

22. 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.

23. 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.

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

25. 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.

26. 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.

27. 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.

28. Sunyaev S, Kondrashov FA, Bork P, Ramensky V (2003) Impact of selection, mutation rate and genetic drift on human genetic variation. Hum Mol Genet 12: 3325–3330.

29. 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.

935 30. Ji Z, Song R, Huang H, Regev A, Struhl K (2016) Transcriptome-scale RNase-footprinting of RNA-protein complexes. Nat Biotechnol 34: 410–413.

31. 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.

940 32. 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.

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

945 34. 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.

35. 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–
950 87.

36. 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.

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

38. 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.

960 39. 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.

40. 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.
965 Nucleic Acids Res 41: e74–e74.

41. Vizcaíno JA, Csordas A, del-Toro N, Dianes JA, Griss J, et al. (2016) 2016 update of the PRIDE database and its related tools. Nucleic Acids Res 44: D447–D456.

42. Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD (2010) Evidence
970 for Pervasive Adaptive Protein Evolution in Wild Mice. PLoS Genet 6: e1000825.

43. Harr B, Karakoc E, Neme R, Teschke M, Pfeifle C, et al. (2016) Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. Sci Data 3: 160075.

975 44. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of

synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–426.

45. 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.

46. 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.

47. Li W-H (1997) Molecular evolution. Sinauer Associates.

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

49. 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.

50. D'Lima NG, Ma J, Winkler L, Chu Q, Loh KH, et al. (2017) A human microprotein that interacts with the mRNA decapping complex. Nat Chem Biol 13: 174–180.

51. 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.

52. 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.

53. Perjes A, Kilpio T, Ulvila J, Magga J, Alakoski T, et al. (2016) Characterization of apela, a novel endogenous ligand of apelin receptor, in the adult heart. Basic Res Cardiol 111: 2.

54. Pauli A, Norris ML, Valen E, Chew G-L, Gagnon JA, et al. (2014) Toddler: an embryonic signal that promotes cell movement via Apelin receptors. Science 343: 1248636.

55. 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.

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

57. 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.

58. 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.

59. 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.

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

61. 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.

62. 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.

63. 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.

64. 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.

65. Saghatelian A, Couso JP (2015) Discovery and characterization of smORF-encoded bioactive polypeptides. Nat Chem Biol 11: 909–916.

66. Pauli A, Valen E, Schier AF (2015) Identifying (non-)coding RNAs and small peptides: challenges and opportunities. Bioessays 37: 103–112.

67. Lynch M, Marinov GK (2015) The bioenergetic costs of a gene. Proc Natl Acad Sci U S A 112: 15690–15695.

68. 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.

69. 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.

70. dos Reis M, Wernisch L (2009) Estimating translational selection in eukaryotic genomes. Mol Biol Evol 26: 451–461.

71. 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.

72. Heinen TJAJ, Staubach F, Häming D, Tautz D (2009) Emergence of a new gene from an intergenic region. Curr Biol 19: 1527–1531.

73. Begun DJ, Lindfors HA, Kern AD, Jones CD (2006) Evidence for de Novo Evolution of Testis-Expressed Genes in the Drosophila yakuba/Drosophila erecta Clade. Genetics 176: 1131–1137.

74. Neme R, Tautz D (2014) Evolution: dynamics of de novo gene emergence. Curr Biol 24: R238-40.

75. 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.

76. 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.

77. 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.

78. 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 370:20140332

79. 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.

80. Knowles DG, McLysaght A (2009) Recent de novo origin of human protein-coding genes. Genome Res 19: 1752–1759.

81. 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.

82. 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.

83. Schlötterer C (2015) Genes from scratch – the evolutionary fate of de novo genes. Trends Genet.

84. Albà MM, Castresana J (2005) Inverse relationship between evolutionary rate and age of mammalian genes. Mol Biol Evol 22: 598–606.

85. Toll-Riera M, Bostick D, Albà MM, Plotkin JB (2012) Structure and age jointly influence rates of protein evolution. PLoS Comput Biol 8: e1002542.

86. 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.

87. 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.

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

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

90. 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.

91. Wilson BA, Masel J (2011) Putatively noncoding transcripts show extensive association with ribosomes. Genome Biol Evol 3: 1245–1252.

92. 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.

93. 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.

94. 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.

95. Villanueva-Cañas J, Ruiz-Orera J, Agea M, Gallo M, Andreu D, et al. (2017) New genes and functional innovation in mammals. bioRxiv. https://doi.org/10.1101/090860.

96. 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.

97. Rosenberg MS, Subramanian S, Kumar S (2003) Patterns of Transitional Mutation Biases Within and Among Mammalian Genomes. Mol Biol Evol 20: 988–993.

98. R Development Core Team (2016) R: a language and environment for statistical computing. http://www.R-project.org/.